Published online 29 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D175–D180***
doi:10.1093/nar/gkn815

CentrosomeDB: a human centrosomal proteins database

Rubén Nogales-Cadenas[1], Federico Abascal[2], Javier Díez-Pérez[2], José María Carazo[2] and Alberto Pascual-Montano[1,*]

[1]Computer Architecture Department, Complutense University of Madrid and [2]National Center for Biotechnology, CNB-CSIC, Madrid, Spain

Received August 9, 2008; Revised September 30, 2008; Accepted October 12, 2008

ABSTRACT

Active research on the biology of the centrosome during the past decades has allowed the identification and characterization of many centrosomal proteins. Unfortunately, the accumulated data is still dispersed among heterogeneous sources of information. Here we present centrosome:db, which intends to compile and integrate relevant information related to the human centrosome. We have compiled a set of 383 likely human centrosomal genes and recorded the associated supporting evidences. Centrosome:db offers several perspectives to study the human centrosome including evolution, function and structure. The database contains information on the orthology relationships with other species, including fungi, nematodes, arthropods, urochordates and vertebrates. Predictions of the domain organization of centrosome:db proteins are graphically represented at different sections of the database, including sets of alternative protein isoforms, interacting proteins, groups of orthologs and the homologs identified with blast. Centrosome:db also contains information related to function, gene–disease associations, SNPs and the 3D structure of proteins. Apart from important differences in the coverage of the set of centrosomal genes, our database differentiates from other similar initiatives in the way information is treated and analyzed. Centrosome:db is publicly available at http://centrosome.dacya.ucm.es.

INTRODUCTION

Centrosomes are present in single copy in most animal cells in a location close to the nucleus. The high-order structure of the centrosome consists of a pair of centrioles surrounded by an apparently amorphous matrix, the pericentriolar material (PCM). Roles of the centrosome are diverse and apparently disparate and intriguing (1), including organization of the cytoskeleton and the mitotic spindle, cell division and regulation of cell cycle or protein degradation processes (2).

Recent mass-spectrometry characterization of the human centrosomal proteome allowed the identification of up to 114 proteins (3), many of them being large coiled-coil proteins, which are likely constituents of the structural scaffold of the PCM. In addition to these 114 identified proteins, the bibliography provides evidence of centrosomal localization for many others.

In this contribution we present centrosome:db, a human centrosomal proteins database that aims at storing, organizing and analyzing known centrosomal proteins. As far as we know, there is only one repository of centrosomal proteins, the MiCroKit database (4). MiCroKit, which was last updated in June 2006, is a curated multi-species not-yet published database of proteins related to the centrosome, the midbody and the kinetochore. As an alternative to MiCroKit, we have compiled a list of human centrosomal genes on the basis of different types of evidences (see below) and obtained different information from several repositories and programs. The resulting information has been organized to provide insights into the centrosomal proteome. The comparison of centrosome:db and MiCroKit reveals differences in the degree of coverage as well as in the information associated to each gene.

METHODS

Definition of the set of human centrosomal proteins

A total of 383 human genes were considered as centrosomal on the basis of several types of evidences. A list of 108 genes was obtained from the proteomics analysis of Andersen *et al.* (3). As a complementary resource, human gene annotations in public databases were used

*To whom correspondence should be addressed. Tel: +34 913944420; Fax: +34 913944687; Email: pascual@fis.ucm.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

as another type of evidence. Up to 55 genes were annotated in Ensembl (5) with Gene Ontology (GO) (6) Cellular Component terms related to the centrosome ('Centrosome', 'Spindle pole'). From a total of these 55 genes, 23 were also supported by the results of Andersen *et al*. The remaining 32 genes were incorporated into the set of centrosomal genes. GO terms related to biological processes characteristic of the centrosome (i.e. centrosome cycle, centrosome duplication, centrosome separation, centrosome localization, mitotic centrosome separation, and centrosome organization and biogenesis) were also considered as potential markers of centrosomal localization. This type of evidence supported the inclusion of 16 genes, three of which were not supported by any of the previous types of evidences. The Human Protein Reference Database (HPRD) (7) was found to be a valuable source of information since many genes described there were annotated as centrosomal on the basis of published scientific bibliography. A total of 117 genes were recovered from HPRD, allowing the inclusion of 60 additional genes into the set of centrosomal proteins. In addition, orthology relationships with closely related species were also considered as an alternative source of evidence of centrosomal localization. We obtained a list of 34 mouse genes annotated as centrosomal and identified their human orthologs. Out of these 34 genes, four were not supported by any of the previous evidences and were incorporated into the set. As a result of this compilation process we obtained a set of 207 human genes.

Moreover, we incorporated those genes tagged as centrosomal in the MiCroKit database. The whole MiCroKit database contains 473 human genes, of which 301 genes are annotated as centrosomal. The comparison of these 301 genes and the set of 207 that we compiled revealed a relatively small overlap as shown in Figure 1. Up to 176 genes present in MiCroKit were not included in our initial set. On the other hand, we identified 82 candidates that were not described in MiCroKit. Hence, we decided to combine both sets of genes, resulting in a set of 383 likely centrosomal human genes. The remaining 162 human genes present in MiCroKit were not included because they were associated to the midbody or the kinetochore, but not to the centrosome.

The analysis of the evidences supporting each gene showed that most genes (246) are supported by only one evidence (176 are supported by the MiCroKit database, 43 by the Andersen *et al*. results and the remaining by one of the other types of evidences). Up to 137 genes were supported by two or more evidences, 68 by three or more, 31 by four or more, 11 by five or more, and finally, two genes (Cep110 and Ninein) were supported by six evidences. We also found that the most frequent sources of evidence were the MiCroKit (301) and HPRD (117) databases, as well as the work of Andersen *et al*. (108).

Information retrieval

In order to describe and better understand the function of human centrosomal genes, we compiled information from several repositories. Data from the Ensembl database (5,8) was retrieved through the R BiomaRt (8) package. Such data comprised information related to genes, isoforms, genomic location, orthology relationships (from the Compara database), Gene Ontology annotations and single nucleotide polymorphisms. Protein–protein

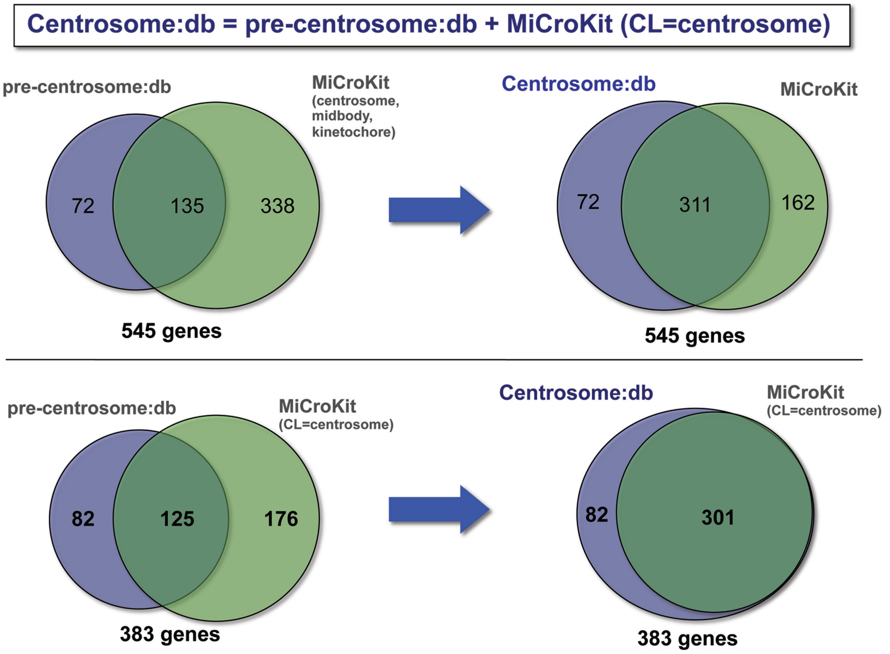


Figure 1. Comparison of the coverage of the centrosome:db and MiCroKit databases. The comparison is displayed both against the whole MiCroKit (upper part) and the centrosomal-section of MiCroKit (lower part). The left and right parts, respectively, show the comparison before and after the inclusion of MiCroKit in centrosome:db. This comparison was done at the level of genes since a few of the proteins contained in MiCroKit mapped to the same gene (e.g. entries MCK-HS-00327 and MCK-HS-00119 both correspond to the gene ENSG00000136861).

interactions among centrosomal proteins were obtained from the HPRD database (7). We obtained information about the association of genes and diseases from the OMIM database (9). Three-dimensional structural information, when available, was retrieved from the MSD database (10) using the BiomaRt package. Scientific references associated to each gene were automatically obtained from NCBI's entrez. The level of association with centrosomes of each of these references was estimated by applying simple text-mining rules. Finally, gene-expression experiments relevantly associated to each centrosomal gene, were retrieved by means of the Array Express web service (11).

Orthology relationships

To facilitate the analysis of the centrosomal proteome in the light of evolution, we identified the orthologs of the human centrosomal genes in 38 other species, including: yeast, the nematode *Caenorhabditis elegans*, three arthropods, two urochordates, and a total of 27 vertebrates (five fishes, one amphibian, one ave and 20 mammals). We used the Ensembl Compara database to identify the orthology relationships. Then, each orthologous gene was incorporated into centrosome:db, and the information related to that gene retrieved using BiomaRt. As a result of this orthology expansion process, the database finally contained 12 172 genes and 17 515 protein isoforms.

Domain assignments

We predicted the domain organization of each of the 17 515 proteins included in centrosome:db. Predictions were made with the sequence-profile comparison program Rps-blast (12) and the Pfam 21.0 (13) and Superfamily 1.69 (14) databases. The results of the Rps-blast execution over these databases were post-processed to remove redundancies and to eliminate the less-scoring domains when two domains of the same type (Pfam or Superfamily) overlapped to a large extent. In addition, the COILS program (15) was used to predict the presence of coiled-coil regions. It is known that coiled-coils are particularly important and frequent in the centrosome (3).

Focusing on the 1115 human proteins, we observed that as a result of the domain assignment process, there were 1486 Pfam domain assignments (*e*-value < 1e-04), corresponding to 806 proteins, which in turn are isoforms of 301 distinct genes. In the case of the Superfamily domains, there were 1274 assignments affecting to 688 proteins and 255 genes. We found 2515 coiled-coil regions, corresponding to 485 proteins and 182 genes. Interestingly, the 71.3% of the proteins identified by Andersen *et al.* contained at least one coiled-coil region. This percentage, however, decreased to 45.5% for the proteins supported by the MiCroKit database (47.5% for centrosome:db). This suggests that the centrosomes isolated and mass spectrometry characterized by Andersen *et al.* (3) are likely enriched in structural proteins whereas the MiCroKit set, and consequently centrosome:db, may contain more regulatory or transient centrosome-visitor proteins.

THE HUMAN CENTROSOMAL PROTEOME

Brief overview of the evolutionary origin of the human centrosomal proteome

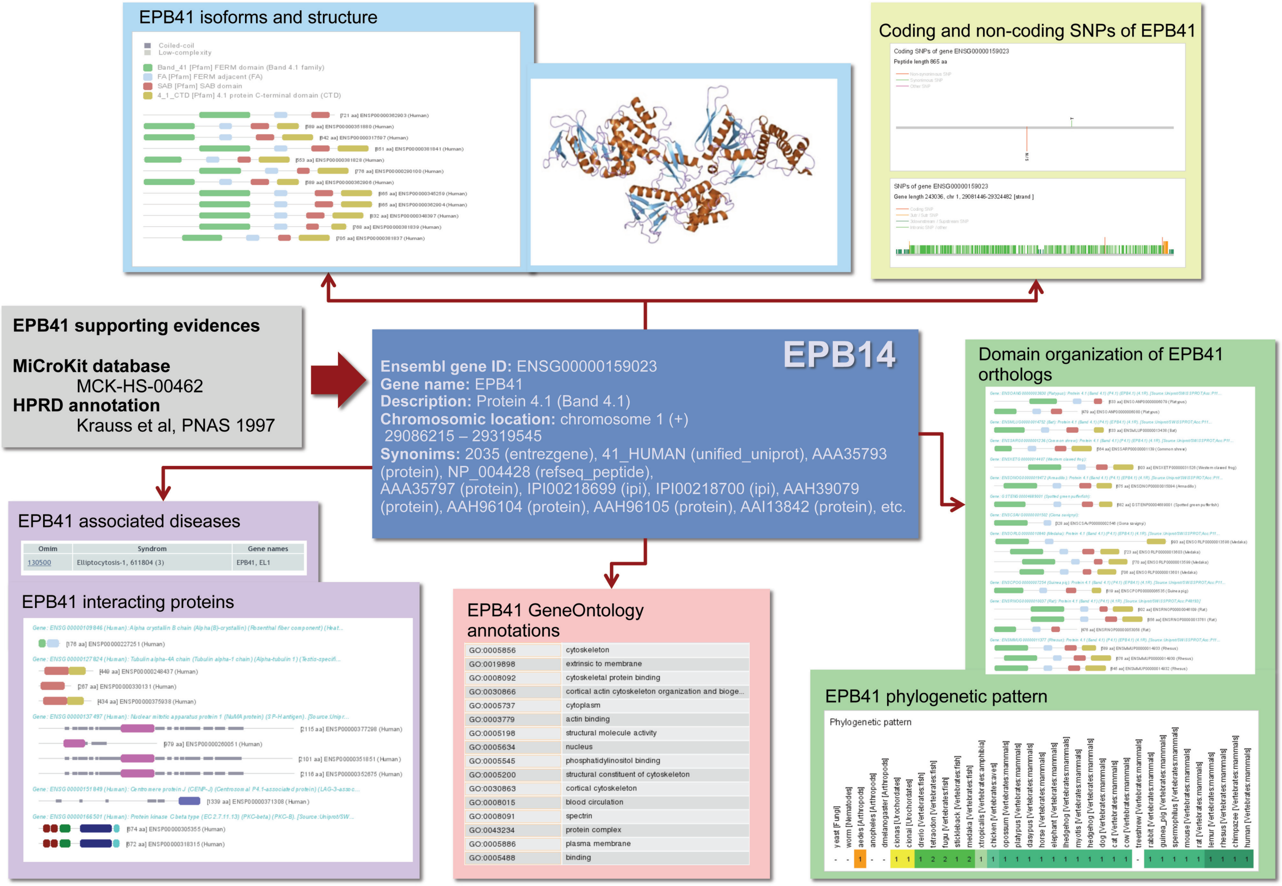
We determined, according to the Compara database and the set of species analyzed, which is the most likely evolutionary origin of each centrosomal gene. We found that 144 genes are neither present in yeast, arthropods nor nematodes. Hence, the origin of those genes could be at the ancestor of chordates. A total of 99 genes are only found in vertebrates, whereas a total of 43, 37 and 30 genes are exclusive of tetrapods, amniotes and mammals, respectively. Finally, four of the 383 centrosomal genes are only found in primates.

The number of human genes having orthologs in each of the compared species is provided in the Supplementary Data. In addition, an analysis of significantly enriched functional annotations for the whole set of genes can also be found in the Supplementary Data.

CENTROSOME:DB

The structure and organization of centrosome:db is summarized in Figure 1 of the Supplementary Data. The database can be accessed at: http://centrosome.dacya.ucm.es and queried in different ways. Gene queries can be conducted either by browsing the list of gene names, by supplying a third-party database identifier, or in full-text mode. Supported identifiers are: Ensembl, Uniprot, Entrez, Refseq, IPI, UniGene and standard gene names (HGNC). In addition, the database can be sequence searched with blast. This can be helpful when the gene identifier is not known, or when we want to compare a sequence from a species that is not included in centrosome:db. As an add-on value, the blast results are accompanied by a domain organization graph of the identified homologs. Such domain-organization representations, in which the user can choose between three alternative representations (Pfam, Superfamily or Coils), are displayed across almost all of the sections of centrosome:db. Remarkably, the analysis of the domain organization of a protein can help to understand or explain its function, although not to deduce it (16). It can also prove useful in understanding the evolution of a group of orthologs or a protein family, in deciphering the differences of alternative protein isoforms, or the molecular basis of protein–protein interactions.

Gene queries can retrieve either general information of the gene (see below), or detailed orthology information, which includes the list of orthologs, a graphical representation of the corresponding phylogenetic pattern and the domain organization of the group of orthologs. Centrosome:db phylogenetic patterns summarizes in a graph the orthology relationships, to make easier the effort of finding in which species a gene is either present or absent, and at which species or phylogenetic range gene duplications have occurred. For instance, according to Compara, the phylogenetic pattern of Ckap5 indicates that this gene is present in single copy in all of the species


EPB41 isoforms and structure
Coding and non-coding SNPs of EPB41
Domain organization of EPB41 orthologs
EPB41 interacting proteins

EPB41 supporting evidences

MiCroKit database
MCK-HS-00462
HPRD annotation
Krauss et al, PNAS 1997

EPB14

Ensembl gene ID: ENSG00000159023
Gene name: EPB41
Description: Protein 4.1 (Band 4.1)
Chromosomic location: chromosome 1 (+)
29086215 – 29319545
Synonims: 2035 (entrezgene), 41_HUMAN (unified_uniprot), AAA35793 (protein), NP_004428 (refseq_peptide),
AAA35797 (protein), IPI00218699 (ipi), IPI00218700 (ipi), AAH39079 (protein), AAH96104 (protein), AAH96105 (protein), AAI13842 (protein), etc.

EPB41 associated diseases

Omim	Syndrom	Gene names
130500	Elliptocytosis-1, 611804 (3)	EPB41, EL1

EPB41 GeneOntology annotations

GO:0005856	cytoskeleton
GO:0019898	extrinsic to membrane
GO:0008092	cytoskeletal protein binding
GO:0030866	cortical actin cytoskeleton organization and bioge...
GO:0005737	cytoplasm
GO:0003779	actin binding
GO:0005198	structural molecule activity
GO:0005634	nucleus
GO:0005545	phosphatidylinositol binding
GO:0005200	structural constituent of cytoskeleton
GO:0030863	cortical cytoskeleton
GO:0008015	blood circulation
GO:0008091	spectrin
GO:0043234	protein complex
GO:0005886	plasma membrane
GO:0005488	binding

EPB41 phylogenetic pattern

Phylogenetic pattern

yeast [Fungi]	-
worm [Nematodes]	-
aedes [Arthropods]	1
anopheles [Arthropods]	-
dmelanogaster [Arthropods]	-
cionas [Urochordates]	1
cionai [Urochordates]	1
drerio [Vertebrates:fish]	1
tetraodon [Vertebrates:fish]	2
fugu [Vertebrates:fish]	2
stickleback [Vertebrates:fish]	1
medaka [Vertebrates:fish]	2
xtropicalis [Vertebrates:amphibia]	1
chicken [Vertebrates:aves]	1
opossum [Vertebrates:mammals]	1
platypus [Vertebrates:mammals]	1
dasypus [Vertebrates:mammals]	1
horse [Vertebrates:mammals]	1
elephant [Vertebrates:mammals]	1
lhedgehog [Vertebrates:mammals]	1
myotis [Vertebrates:mammals]	1
hedgehog [Vertebrates:mammals]	1
dog [Vertebrates:mammals]	1
cat [Vertebrates:mammals]	1
cow [Vertebrates:mammals]	1
treeshrew [Vertebrates:mammals]	-
rabbit [Vertebrates:mammals]	1
guinea_pig [Vertebrates:mammals]	1
spermophilus [Vertebrates:mammals]	1
mouse [Vertebrates:mammals]	1
rat [Vertebrates:mammals]	1
lemur [Vertebrates:mammals]	1
rhesus [Vertebrates:mammals]	1
chimpazee [Vertebrates:mammals]	1
human [Vertebrates:mammals]	1

Figure 2. The particular example of the Epb41 gene is shown. Centrosome:db provides information indicating which evidence(s) support the consideration of this gene as centrosomal. The domain organization is shown for the 12 epb41 alternative isoforms, the centrosomal interactors and the group of orthologs. Relevant functional information (OMIM relevant diseases, SNPs, Gene Ontology-based annotations, etc.) is also provided.

included in centrosome:db. In contrast, the phylogenetic pattern of Bcc3 indicates that Bcc3 is specific of mammals.

Centrosome:db can also be queried by types of domains. This allows the identification of which proteins/genes from either all or a particular species have a given domain. The domain organization of the matching proteins is graphically represented to reveal those domains to which the query domain co-occurs. A phylogenetic pattern indicating how many genes in each species encode a protein isoform predicted to have the query domain is also provided. Finally, the database can be interrogated by either the type of evidence or the number of supporting evidences. The list of genes supported by the work of Andersen *et al.* (3), or the list of genes supported by three or more evidences are retrieved with this option.

User case: the epb41 gene

In order to provide a more comprehensive understanding of the type of information that can be retrieved using centrosome:db, we present a full example using the epb41 gene (Figure 2). According to Pfam, the 12 alternative protein isoforms encoded by the epb41 gene have four different domains: Band_41, FA, SAB and 4_1_CTD (note that in the current version of Pfam, the Band_41 domain has been divided in three domains: FERM_N, FERM_M and FERM_C). The domain-organization graph indicates that some of these isoforms lack completely or partially the C-terminal domain 4_1_CTD. According to the Superfamily database, these proteins have three domains of known 3D-structure, which belong to the PH domain-like, Ubiquitin-like and Second domain of FERM superfamilies. These three Superfamily domains correspond to the Band_41 and FA Pfam domains. There is a predicted coiled-coil region inside the SAB domain.

The ebp41 gene is described as Protein 4.1 and according to OMIM has been associated with the Ellipocytosis syndrome. The list of Gene Ontology terms associated to Epb41 is large and includes cellular component terms such as plasma membrane, cytoplasm and nucleus, as well as biological process terms such as actin-binding, blood circulation and cortical cytoskeleton organization and biogenesis. The P4.1 protein is known to interact with four other centrosomal proteins: TubA4A, Numa1, CenpJ and PrkCB1.

Up to 81 scientific references are associated to this gene and are sorted according to their likelihood of association with the centrosome.

The phylogenetic pattern of epb41 indicates that there are orthologs in all vertebrates (except for *Tupaia belangeri*, the treeshrew), in urochordates and in the yellow fever mosquito. It also indicates that there are in-paralogs (relatively recent duplications) in some fishes. Consequently, according to our data, this gene is absent in two arthropods, *C. elegans* and yeast. Since the deciphering of orthology relationships represents a highly complex problem, we contrasted the information contained in centrosome:db, which is based on the Compara database, with external sources. The Inparanoid database (17) suggests that orthologs of Epb41 can also be found in *C. elegans*, the malaria mosquito and the fly. In contrast, the roundup database (18) supports a vertebrate origin of Epb41. The answer to these discrepancies can be devised with the TreeFam database (19), in which we can look at the phylogenetic tree of the family. According to the TreeFam tree, the most likely hypothesis is that the Epb41 gene is vertebrate specific. We realized that there are many paralogs of Epb41 (Epb41L1, Epb41L3, Farp1, Farp2, Frmd7) that complicate the proper identification of orthologs in an automatic fashion.

Hence, in the case of Epb41, the Compara results seem to be erroneous. Interestingly, the domain organization of the Compara group of orthologs in centrosome:db indicates that the SAB domain, which is present in vertebrates, is absent in arthropods and urochordates, what is in accordance with the hypothesis that those genes are not orthologs but paralogs.

Since the complete graph of the domain organization of epb41 orthologs is very large, only a part of it is shown in Figure 2.

The analysis of the domain organization of the Epb41 orthologs reveals some other interesting observations. We noticed that in most orthologs there is a coiled-coil region next to the C-terminal region, right in the place were the SAB domain is located. However, in fugu there is no coiled-coil inside the SAB domain but at the N-terminal region of the protein, before the Band_41 domain. Hence, it is possible that some reorganization of the molecular functions have taken place in the fugu ortholog. In one of the two *Tetraodon nigroviridis* orthologs, the coiled-coil region, which is also located at the N-terminal of the protein, is of much larger extent that in any of the other orthologs (Figure 2 of the Supplementary Data). In addition, the same *T. nigroviridis* ortholog apparantely lacks the FA, SAB, and 4_1_CTD domains. The zebra fish ortholog displays a similar pattern of domain organization. In summary, this example illustrates how the analysis of the domain organization of proteins under the light of evolution can provide interesting clues for understanding proteins's function. This example also highlights the possible pitfalls, which are mainly related to the difficulties in deciphering orthology relationships among species.

DISCUSSION AND FUTURE DIRECTIONS

Centrosome:db is currently focused on the human centrosome. However, the centrosomes of other species are indirectly considered by means of orthology relationships with human genes. Comparatively, the main advantage of centrosome:db over existing databases like MiCroKit is at the information content and the visualization of the information. Centrosome:db has benefited from the data deposited in MiCroKit to enlarge our set of centrosomal genes. In order to recognize this point, each gene supported by MiCroKit has been linked out to the MiCroKit database, in which the particular references supporting its centrosomal localization can be looked up.

We plan to maintain centrosome:db updated, including additional genes when new evidences appear. Importantly, we provide a submission form to allow users to submit new genes or modify the information related to the already existing genes. Such a contribution from the scientific community would significantly improve the quality of this repository and will help in converting centrosome:db in a full curated database Finally, we are considering the development of new versions of centrosome:db related to other species commonly used for the study of the centrosome or similar subcellular structures such as the spindle pole body.

SUPPLEMENTARY DATA

Supplementary web site available at: http://centrosome.dacya.ucm.es/centrosome/supmat.

ACKNOWLEDGEMENTS

We would like to acknowledge César Vicente and the National Institute for Bioinformatics (INB) for technical support. A.P.M. acknowledges the support of the Spanish *Ramón y Cajal* program.

FUNDING

This work has been partially funded by the Spanish grants S-Gen-0166/2006, BIO2007-67150-C03-02, CSD2006-00023 and TIN2005-5619. I3P contract of the Spanish MEC (to F.A.). Funding for open access charge: S-Gen-0166/2006.

Conflict of interest statement. None declared.

REFERENCES

1. Rieder,C.L., Faruki,S. and Khodjakov,A. (2001) The centrosome in vertebrates: more than a microtubule-organizing center. *Trends Cell Biol.*, **11**, 413–419.
2. Doxsey,S., Zimmerman,W. and Mikule,K. (2005) Centrosome control of the cell cycle. *Trends Cell Biol.*, **15**, 303–311.
3. Andersen,J.S., Wilkinson,C.J., Mayor,T., Mortensen,P., Nigg,E.A. and Mann,M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, **426**, 570–574.
4. Xue, Y., Zhou, F., Fu, C., Jin, C., Pei, S., Xu, Y. and Yao, X. MiCroKit: Midbody, Centrosome and Kinetochore proteins. *NAR Molecular Biology Database Collection*, entry number 1022.
5. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.

6. Consortium,T.G.O. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
7. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
8. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
9. McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
10. Tagari,M., Tate,J., Swaminathan,G.J., Newman,R., Naim,A., Vranken,W., Kapopoulou,A., Hussain,A., Fillon,J., Henrick,K. *et al.* (2006) E-MSD: improving data deposition and structure quality. *Nucleic Acids Res.*, **34**, D287–D290.
11. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
14. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
15. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
16. Attwood,T.K. (2000) Genomics. The Babel of bioinformatics. *Science*, **290**, 471–473.
17. Alexeyenko,A., Tamas,I., Liu,G. and Sonnhammer,E.L. (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, **22**, e9–e15.
18. Deluca,T.F., Wu,I.H., Pu,J., Monaghan,T., Peshkin,L., Singh,S. and Wall,D.P. (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**, 2044–2046.
19. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.