

Research article

Open Access

## Evolutionary flexibility of protein complexes

Michael F Seidl<sup>1,2</sup> and Jörg Schultz\*<sup>1</sup>

Address: <sup>1</sup>Department of Bioinformatics, Biozentrum, University Würzburg, Am Hubland, 97074 Würzburg, Germany and <sup>2</sup>Theoretical Biology and Bioinformatics Group, Department of Biology, Faculty of Science, Utrecht University, Padualaan 8, 3584CE Utrecht, The Netherlands

Email: Michael F Seidl - m.f.seidl@uu.nl; Jörg Schultz\* - joerg.schultz@biozentrum.uni-wuerzburg.de

\* Corresponding author

Published: 7 July 2009

Received: 1 August 2008

*BMC Evolutionary Biology* 2009, **9**:155 doi:10.1186/1471-2148-9-155

Accepted: 7 July 2009

This article is available from: <http://www.biomedcentral.com/1471-2148/9/155>

© 2009 Seidl and Schultz; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Proteins play a key role in cellular life. They do not act alone but are organised in complexes. Throughout the life of a cell, complexes are dynamic in their composition due to attachments and shared components. Experimental and computational evidence indicate that consecutive addition and secondary losses of components played a major role in the evolution of some complexes, mostly without affecting the core function. Here, we analysed in a large scale approach whether this flexibility in evolution is only limited to a distinct number of complexes or represents a more general trend.

**Results:** Focussing on human protein complexes, we based our analysis on a manually curated dataset from HPRD. In total, 1,060 complexes with 6,136 proteins from 2,187 unique genes were considered. We computed interologs in 25 different species and predicted the composition of complexes. Over the analysed species, the composition of most complexes was highly flexible and only 25% of all genes were never lost. Even if one component was lost at a particular point in time, the fraction of observed second, independent losses of additional components was high (75% of all complexes affected). Still, loss of whole complexes happened rarely. This biological signal deviated significantly from random models. We exemplified this trend on the anaphase promoting complex (APC) where a core is highly conserved throughout all metazoans, but flexibility in certain components is observable.

**Conclusion:** Consecutive additions and losses of distinct units is a fundamental process in the evolution of protein complexes. These evolutionary events affecting genes coding for units in human protein complexes showed a significantly different phylogenetic pattern compared to randomly selected genes. Determination of taxon specific attachments or losses might be linked to specific cellular or morphological features. Thus, protein complexes contain not only structural and functional, but also evolutionary cores.

### Background

Proteins are, next to RNA, the fundamental unit of biological activity. But, they do not act alone. Many biological and cellular processes require a precise organisation of proteins in time and space [1]. These multi protein com-

plexes, also called molecular- or protein-machines, are among the fundamental entities of molecular organisation [1,2]. Recent high throughput studies identified and analysed the components of protein interaction networks and how they are organised to functional units [1,3-5].

On a higher level, multi-protein complexes are embedded in a network linking cellular processes [6]. Here, the complexes are connected by shared components, e.g. proteins present in more than one complex. Most of these shared components are associated peripherally and are not integral members of the complexes suggesting a role in the regulation of molecular-machines [6]. Complementary to this network view, protein complexes can be partitioned in a core which is modulated by different attachments. By adding different attachments, isoforms of a complex are built, possibly with slightly different functions. Some of these attachments, which can consist of multiple proteins itself, can be connected to different core complexes. These mobile regulatory units are often called modules [1]. The combination of core functional units with variably attached modules increases the number of different complexes and thereby the complexity of the cell. This complexity, comprising both the functional and structural entities of protein complexes, raises the question how the interplay of core complexes with variable attachments evolved. As a first step in this direction, it has been shown that yeast complexes enriched with gene products having an ortholog in human preferentially interact with other gene products that also have a human ortholog [3]. Comparing the constitution of cores and modules in other species revealed that they are unlikely to be present partially [1]. This could be interpreted as an 'ortholog proteome' that resembles the backbone necessary to facilitate fundamental functions of an eukaryotic cell [7].

Complementary to these large scale analyses, an in-depth study of the SMN complex which is involved in splicing revealed a high degree of evolutionary flexibility of its components [8]. The studied complex is responsible for mediating assembling of the UsnRNPs (uridine rich small nuclear ribonucleoproteins). In humans, it consists of eight components, namely SMN and the Gemin 2–8. This complexity arose via addition of distinct entities to

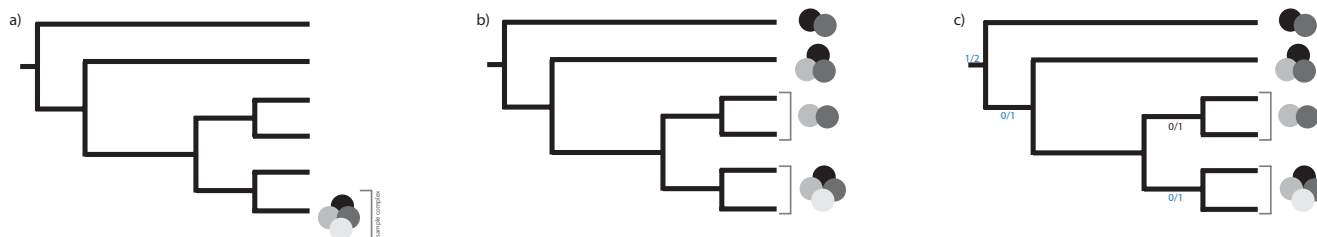
the ancestral core of SMN and Gemin 2 which can already be found in protists. Contrary to this trend, diptera have lost three of the components but still contain a functional SMN complex. Similar losses were found in further organism, indicating evolutionary dynamics of the complex.

Here, we addressed the question whether evolutionary flexibility is limited to a distinct number of machines or represents a general feature of the evolution of protein complexes.

## Results and discussion

### A parsimony based approach for inferring the evolutionary history of protein complexes

We focussed our analysis on human protein complexes annotated in the human protein reference database (HPRD), as this database is manually curated and, accordingly, of high quality [9]. At the time of the analysis, the HPRD dataset contained 2,197 distinct genes which were found in 1,060 protein complexes. As a first step, we identified orthologs of these genes in the genomes of a selected subset of species (see Fig. 1a–c for a hypothetical example of the applied approach). To provide a wide spectrum, we chose 25 annotated eukaryotic species including 17 metazoan, six fungi, one choanoflagellate and one amoebozoan as an outgroup (see Tab. 1). Using literature data, a phylogenetic tree for these species was reconstructed (see Methods). For ortholog detection InParanoid [10] combined with an iterative searching approach was implemented (see Methods for details). Using the concept of interolog mapping [11,12] allowed the prediction of the constitution of 'orthologous' complexes in each species (see Fig. 1b). This prediction will vary from the 'real' complex, as we did not consider gene duplications. A duplication in the other (non human) species should not influence the results, as one of the copies is expected to stay as a member of the protein complex. If the duplication is human specific, two scenarios have to be distin-



**Figure 1**

**Identification of 'ortholog' complexes and their evolutionary history.** Example explaining the identification of 'ortholog' complexes and the maximum parsimony approach to infer the evolutionary history according to a phylogenetic tree. A hypothetical complex consisting of four components is derived from HPRD (a). Computing the ortholog genes using InParanoid and deriving the constitution of the complex in all species of interest (b). Using a maximum parsimony approach to infer the evolutionary history, gene emergence and loss events, of every component of the complex. The numbers in blue indicate complex or gene emergence, the black numbers loss events (c).

**Table 1: Table of the examined species, the source and the version.**

Name	Version	Release date	Source	Reference
<i>Anopheles gambiae</i>	AgamP3	Feb. 2006	Ensembl	[35]
<i>Apis mellifera</i>	v2.0	unknown	Beebase	[36]
<i>Aspergillus niger</i>	v1.0	Nov. 2005	JGI	-
<i>Branchiostoma floridae</i>	v1.0	Mar. 2006	JGI	[37]
<i>Caenorhabditis elegans</i>	WS180	Sep. 2007	Ensembl	[38]
<i>Ciona intestinalis</i>	JGI2	Mar. 2005	Ensembl	[39]
<i>Danio rerio</i>	ZFISH7	Jul. 2006	Ensembl	[40]
<i>Daphnia pulex</i>	v1.0	Sep. 2006	JGI	-
<i>Dictyostelium discoideum</i>	unknown	Jan. 2008	Dictybase	[41]
<i>Drosophila melanogaster</i>	BDGP4-3	Jan. 2006	Ensembl	[42]
<i>Encephalitozoon cuniculi</i>	unknown	Jan. 08	NCBI	[18]
<i>Homo sapiens</i>	NCBI36	Nov. 2006	Ensembl	[14,15]
<i>Laccaria bicolor</i>	v1.0	Mar. 2005	JGI	[43]
<i>Monosiga brevicollis</i>	v1.0	Jul. 2006	JGI	[44]
<i>Mus musculus</i>	NCBIM37	Apr. 2007	Ensembl	[45]
<i>Nematostella vectensis</i>	v1.0	2006	JGI	[46]
<i>Oryzias latipes</i>	MEDAKA1	Oct. 2005	Ensembl	[47]
<i>Phycomyces blakesleeanus</i>	v1.0	Sep. 2006	JGI	-
<i>Rattus norvegicus</i>	RGSC3-4	Nov. 2004	Ensembl	[48]
<i>Saccharomyces cerevisiae</i>	SGD1	Dec. 2006	Ensembl	[49]
<i>Schizosaccharomyces pombe</i>	v19.0	unknown	Sanger	[50]
<i>Takifugu rubripes</i>	FUGU4	Jun. 2005	Ensembl	[51]
<i>Tetraodon nigroviridis</i>	TETRAODON7	Apr. 2003	Ensembl	[52]
<i>Trichoplax adhaerens</i>	v1.0	Jul. 2006	JGI	[53]
<i>Xenopus tropicalis</i>	JGI4-1	Aug. 2005	Ensembl	-

Names of the examined species in alphabetical order, the source (Ensembl, JGI, species related databases), the version, the release date and the reference if available.

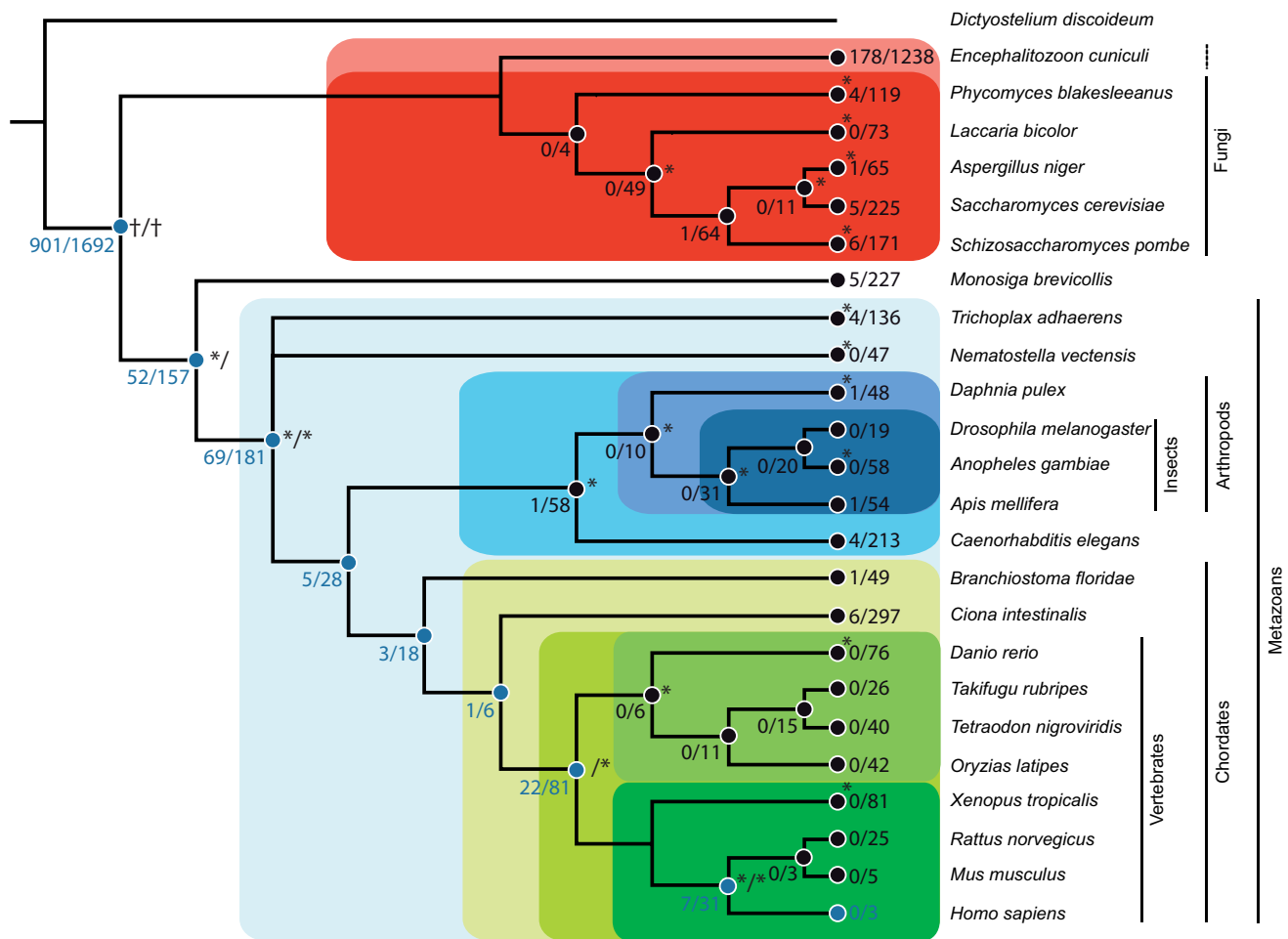
guished. In the first, both human genes are components of different protein complexes. In this case, their ancestor was probably a member of both complexes [13]. In the second scenario, only one of the duplicated proteins is a member of a complex. In cases where this functionality evolved after the speciation, a false positive will be seen. Thus, gene duplications will only slightly influence the prediction of the 'ortholog' complexes. Based on the presence and absence pattern of complexes and the forming components we inferred the evolutionary history using on a parsimony based approach (see Methods and Fig. 1c for more information).

#### **Emergence of protein complexes and their components**

As a first step, the emergence of each gene coding for a component was reconstructed according to the species tree (Fig. 2, blue numbers). For 77% of the genes orthologs were found in at least one fungus, indicating that their origin lay before the split of fungi and metazoans. Branches with a substantial addition of orthologs were the base of choanoflagellates-metazoans (157) and from there to the metazoan lineage (181). Based on the species sampling, these 'inventions' could also represent fungi specific gene losses. It has been suggested that the observable complexity of organisms is not mainly reflected by the gene number [14,15] but, among many

other factors, by the number of protein interactions and the resulting interaction networks [6]. Indeed, the estimated size of different interactomes, in which protein complexes are embedded [6], is correlated with the biological complexity [16]. Thus, the emergence of genes localises with the increase in morphological complexity and the evolution of certain traits, like for the vertebrates (81) and mammals (31).

In a second step, we focused on the more complex centric view and analysed the emergence of whole complexes. We applied three alternative definitions specifying the emergence of a complex. (i) The point where at the first time two or more components of the complex were found (subsequently added or present at once), according to a definition that at least two components are necessary to constitute a complex [17]. (ii) The point of occurrence of the largest set of components at one time. (iii) The point of occurrence of all HPRD annotated components. Obviously, these definitions are oversimplifications as the minimal number of components necessary to constitute a functional complex could be different for every complex. Still, with our definitions we provided an upper and lower boundary to estimate complex emergence. With the most general definition, most of the complexes were already present in the last common ancestor of human and fungi



**Figure 2**  
**Phylogenetic Tree with gene and complex emergence and losses.** The pattern of gene and complex emergence and the secondary losses of components of whole complexes is displayed along the tree according to the absence and presence pattern of the ortholog genes in terminal species or in subsets of species concluding the loss in the last common ancestor of all subsequent species. The numbers of gene and complex emergence is indicated in blue (complex emergence/gene emergence). The number of secondary losses are shown in black per affected node. It was discriminated between whole complex losses and gene losses (complex losses/gene losses). The significance of emergence (discriminated between complex and gene emergence) and loss (only gene) events compared to the random model are indicated with '\*'. As we restricted our analysis to fungi and metazoans, evolutionary events which have been mapped to the base of the tree ('+') could have evolved at any time before the split.

(approximately 85%), with an increase at the base of the choanoflagellates-metazoans lineage, the metazoans, vertebrates and mammalians, respectively (Fig. 2). Comparable results were found with the second definition. Even with the most conservative definition a high number of complexes were observable at the last common ancestor of human and fungi or before (approximately 42%) and huge accretions at the base of the choanoflagellates-metazoans lineage (not considering fungi specific gene losses) and the metazoans. Overall, nearly 82% of all complexes had already emerged at that point. To test whether our results reflect an evolutionary signal and not just random

fluctuations in complex composition we compared them to a random model. We chose a random subset of human genes identical in size to the original dataset and calculated the emergence of genes and complexes. This was repeated 10,000 times and compared to the biological signal. For most of the nodes (highlighted with a '\*' in Fig. 2), the number of gene and complex emergence events differed significantly between the biological signal and the random model (all p-values smaller than an alpha (0.05) corrected for multiple testing, see Methods). In all significant nodes, fewer genes evolved than expected from

the random model. Thus, a gene coding for a protein of a human complex tends to be older than the average gene.

The initial emergence of a complex is followed by a sequential addition of further components which might be linked to cellular or morphological features. Moreover, most components of protein complexes emerged early in the species tree and tend to be older than randomly chosen human genes.

### Secondary loss

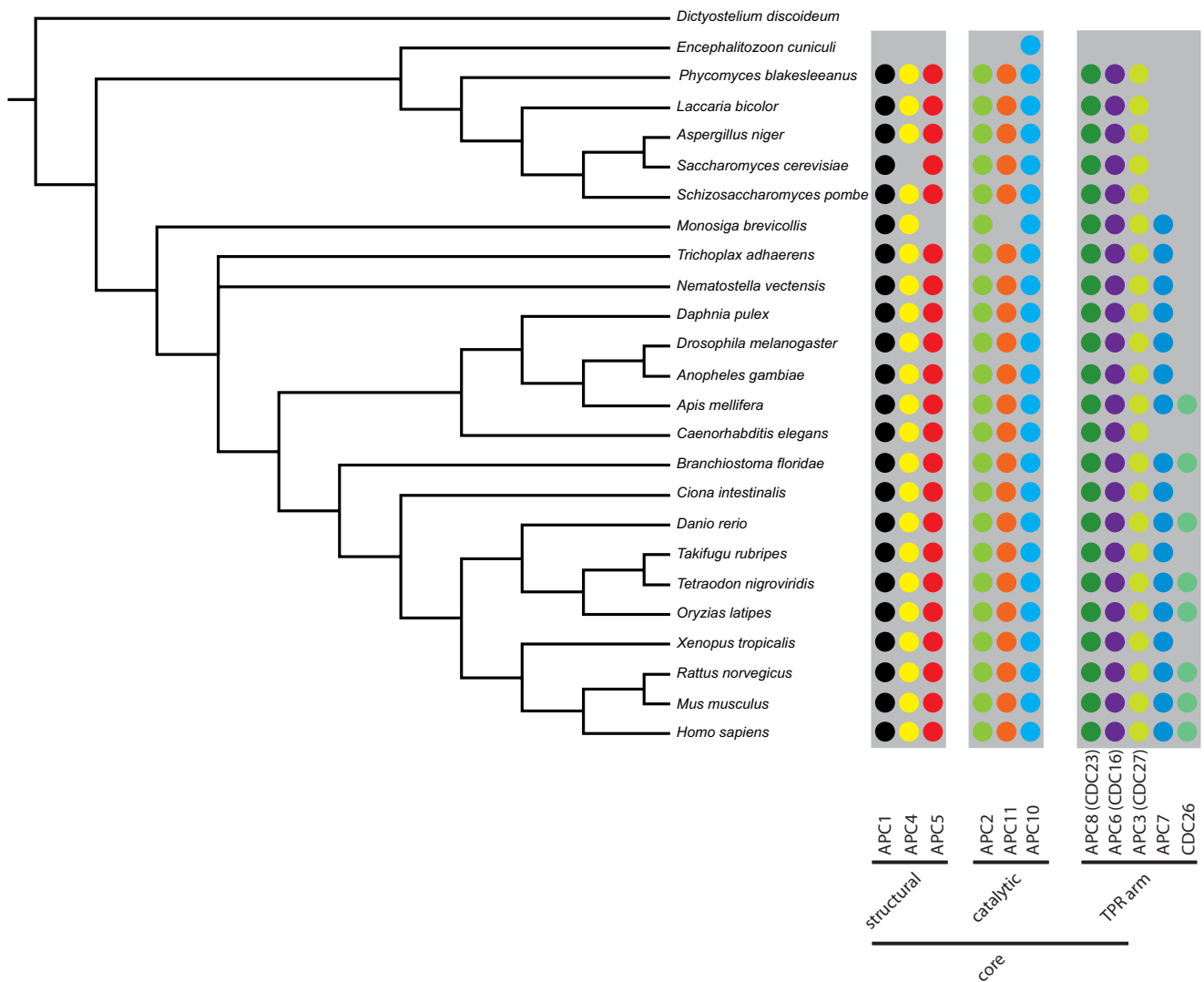
Having calculated the point of emergence for each component of a human protein complex, we were now able to address the question of secondary losses of genes and whole complexes. For each gene present in a human protein complex, we predicted species missing its ortholog and, to identify the likely branch of gene loss, mapped gene losses to the last common ancestor. To test the significance of the observed pattern, we compared our results to a random model which took into account the observed bias of emergence events. In all significant cases (with *Aspergillus niger*, *Phycomyces blakesleeanus* and *Anopheles gambiae* as exceptions) fewer losses were observed than expected from the random model. Nevertheless, a high number of losses occurred along the tree (Fig. 2, black numbers). Interestingly, *Encephalitozoon cuniculi* has lost approximately 73.2% of the genes present in the last common ancestor of fungi and metazoan/choanoflagellates lineage. This might be the result of the intracellular parasitic nature with a reduced gene set, complete losses of biochemical pathways and a reduced protein-protein interaction network [18]. Comparable, but not equally large gene losses were observed in *Saccharomyces cerevisiae*, *Monosiga brevicollis*, *Trichoplax adhaerens*, *Caenorhabditis elegans* and *Ciona intestinalis*. A general trend for the loss of genes was already described for fungi, insects and *C. elegans* [19-21]. When looking only at genes with orthologs in human protein complexes we recall this trend for fungi and *C. elegans*. In contrast, we did not find any outstanding number of losses in insects in general or diptera in particular. The high number of losses found in *C. intestinalis* might be caused by errors in gene prediction. In the analysis of the SMN complex orthologs for *C. intestinalis* were not identified on the proteomic level due to annotation problems, but in a search against the whole genome shotgun sequences [8]. This example highlights the dependency of this analysis on the quality of the available genome data. Here, we focussed on proteins with a function in a protein complex which evolve comparably slow [22]. As most gene annotation pipelines utilize homology prediction, the rate of false positives will be lower than for randomly chosen proteins.

In total, only 25% of the genes found in human protein complexes were present in all species subsequent to the

initial emergence. Of this 522 genes, 302 (approximately 58%) have already emerged before the fungi/metazoan split. The fraction of at least one secondary loss in the HPRD dataset of 2,197 human genes was 76.2%. This highlights the evolutionary flexibility of genes coding for components which are part of protein complexes. 913 genes were affected by more than one loss event, which is approximately 55% of all the genes affected by secondary losses. Thus, genes which are affected by a loss once, are more likely to be affected by additional further losses.

Nearly 44% of all 2,197 analysed genes were present in more than one complex and 36 of them were found in more than 10 different complexes. Of the nine genes that are shared between more than 15 complexes those with the highest occurrence were never lost, especially Integrin beta-1 precursor [Ensembl:ENSG00000150093] which is present in 54 complexes. The mean number of losses in genes that are present in more than 10 complexes was 1.25 (range 0-5), the mean number of losses found in only a single complex was 1.65 (range 0-13). Genes coding for proteins that are present in multiple complexes and therefore form a high number of interactions tend to evolve more slowly and seem to be more conserved than genes coding for proteins with few interactions, however the magnitude of difference was not dramatic [19,23]. Our analysis corroborates these observations.

Contrasting a high variability of the components of protein complexes, we rarely observed a loss of a whole complex. An exception was again *E. cuniculi*, which had lost many complexes completely. Thus, the loss of certain parts of already established complexes seems to be tolerable for the fitness of the organism. Overall, only 32 complexes annotated in HPRD (excluded complexes with the size of one) did not suffer from any secondary loss (3%) and 96.13% had at least one secondary loss of any component present (1,019). 75% of the complexes had at least two losses, indicating that functional modules or single components of different subunits were lost. Still, the core functionality of the complex has to be conserved, either as the result of the remained functionality or by the recruitment of non-ortholog, but functional equivalent, gene products. When predicting the composition of human complexes in other species, our analysis suggest that the composition is evolutionary highly flexible. However, the absence of whole complexes was rarely observed, indicating that either the remaining component are sufficient or additional, species specific, components are recruited to preserve the main function of the complex in the given context. In contrast, the partial loss or presence of ortholog components in different species in either core or modules has not been reported for yeast [1]. This difference might be the result of the heterogeneity of the HPRD datasets, comprising core, modules and attachments or



**Figure 3**  
**The APC complex.** Graphical representation of the presence-absence pattern of single components of the APC complex, grouped by the sub-complexes (the composition of the sub-complexes have been derived from the literature [25] and is not reflected in HPRD). The structural, the catalytic and the TPR arm create the core complex. Presence of a component is indicated by a circle, the spectrum of examined species by the grey underlying bar (*D. discoideum* as outgroup was not considered).

the fact that the protein interaction network of human, compared to yeast, is larger, generating more hypothetical possibilities of flexibility.

**Evolutionary dynamics of the APC Complex**

As a case study, we analysed the anaphase-promoting complex (APC), also called cyclosome, in detail. The APC plays a key role in the degradation of cyclins and other factors of cell cycle regulation mediated by the attachment of multiple ubiquitine chains to a lysine residue in the target protein (for a review on ubiquitination see [24]). The human cyclosome is a large, 1.5 MDa complex consisting of 11 core components (annotated in HPRD as 'COM\_144'; one additional component, Apc13 is not

described in HPRD) and two additional transient attachments (also not found in HPRD) required to bridge the interaction with the substrate [25] and activate the APC [26]. Two components, Apc2 and Apc11, built the catalytic core of the complex [25] and both are conserved throughout most eukaryotes and essential in the examined species [27,28]. The whole complex can be divided in four different sub-complexes, composed of the structural part (Apc1/Apc4/Apc5), the catalytic arm (Apc2/Apc11/Apc10), a tetratricopeptide repeat (TPR) arm (Apc8/Apc6/Apc3/Apc7/Cdc26/Apc13 [Ensembl:ENSG00000129055]) involved in adaptor binding and the attachments bridging the interaction to

substrate (Cdc20/Cdh1; [Ensembl:ENSG00000117399]/[Ensembl:ENSG00000105325]).

We predicted the composition of the APC complex in 24 species using the described InParanoid procedure. For species where a loss was inferred we manually checked the absence of the particular gene product by using a reciprocal best hit approach against the NCBI non redundant database (nrdb).

The structural part of the complex was already present in the last common ancestor of human and fungi (Fig. 3, additional file 1 for the corresponding gene identifier). Apc1 was ubiquitous found in all species except *E. cuniculi*. The ortholog in *Danio rerio* was identified by a manual search against nrdb. Apc4 was lost in *E. cuniculi* and seems to be lost in *S. cerevisiae*. Experiments revealed a protein functionally corresponding to Apc4 in *S. cerevisiae*, but it was highly divergent and showed only a weak similarity to the human and the *Schizosaccharomyces pombe* Apc4 [27,28]. *E. cuniculi* and *M. brevicollis* have furthermore lost Apc5. The ortholog of Apc5 in *C. elegans* was not predicted by InParanoid, however could be inferred by a search against nrdb.

The components of the catalytic arm of the multi-protein enzyme were also present in the last common ancestor of fungi and human. Apc10, promoting substrate binding [25], was the most conserved subunit found in every examined species. Apc2 and Apc11, both part of the catalytic core, were identified throughout our species selection, except for *E. cuniculi* and in the case of Apc11 in *M. brevicollis*. The orthologs of Apc11 in *Xenopus tropicalis*, *Drosophila melanogaster* and *C. intestinalis* were identified by a manual search against nrdb.

The TPR arm components were also present in the last common ancestor of fungi and humans. Apc3, Apc6 and Apc8 were found in all analysed metazoan genomes and are even conserved throughout most fungi [25], highlighting the importance of the subunits to associate the attachments to the APC. Apc7, another component of the TPR arm sub-complex, has been described as vertebrate specific. Recent studies [29] indicated a genuine ortholog in *D. melanogaster*. We identified further orthologs in all metazoan and in *M. brevicollis* with the only exception of *C. elegans*. Additional orthologs were identified in plants and *Dictyostelium discoideum*. Thus, fungi seem to have lost this gene. The Cdc26 subunit, a small protein of 86 amino acids, was only identified in chordates and arthropods. Functional equivalents were described in *S. cerevisiae* (also named Cdc26) and *S. pombe* (named Hcn1) [30]. A manual PSI-BLAST [31] search with the *S. cerevisiae* Cdc26 protein and the *S. pombe* Hcn1, respectively, did not report any sequence similarity to other proteins in our dataset.

The APC complex demonstrate that both high evolutionary flexibility and conservation of entities in human complexes could be observed. Moreover, we show examples that the loss of a gene can be compensated by the displacement with a non-homologous gene product to sustain the functionality of the complex.

## Conclusion

How do protein complexes evolve? Do they emerge with all components at a specific branch in the phylogenetic tree or is it a more gradual process over longer time scale? Looking from human complexes back into phylogenetic history, we found that both is true. In most cases the emergence of some members of the complex is followed by the addition of further components. Still most components of protein complexes tend to be older than randomly chosen genes. Although the components show fewer losses than observed in a random model we also revealed frequent secondary losses of genes involved in a specific complex. Are these losses of genes with a possibly important function in the human complex real? A critical point in the analysis is the sequence based ortholog detection. If proteins evolve too fast, homologs might not be identified but still be present leading to false negatives and thereby to increased loss rates. An analysis of the BLAST algorithm underlying InParanoid showed that BLAST consistently identified homologs even over larger phylogenetic distances than used here [32]. We further improved sensitivity by using InParanoid, one of the best programs for ortholog detection [33] and applying iterative pairwise comparisons. Finally, the analysis focussed on proteins with a function in a complex which evolve slower than randomly selected proteins. We therefore expect only a small influence by false negative orthologs. We identified secondary gene losses on the sequence level, without the possibility to infer the function of the resulting complexes in the examined species. The SMN complex demonstrated that even with a reduced set of genes a complex can be still functional. Moreover, as seen in the APC complex, the loss of a gene can be compensated by the displacement with a non-homologous gene product. In many cases these enzymes have evolved by shifting the substrate specificity of a related but distinct enzyme [34].

Despite these limitation, our results indicate that losses can happen even for genes which are tightly bound into an interaction network like a protein complex. Together with the gradual emergence this has several consequences. First, one can identify an evolutionary core of a protein complex complementary to structural or functional cores. Second, taxon specific attachments or losses of complexes might be linked to specific cellular or morphological features. Third, the identification of the 'smallest' version of a complex might enable an easier experimental characterisation.

## Methods

### Genomic Data

Genomes used in this study as well as their source and version are given in Tab. 1 [14,15,18,35-53].

### Species Tree

The phylogenetic tree used to guide the analysis and the ortholog identification was based on literature data. The position of *D. discoideum* as the outgroup to all other sampled species has been shown in [41] where a phylogeny based on ortholog clusters between different species had been calculated. The relationship of the fungi was derived from [54] where a concatenated six gene marker was used to infer the positions of the species. The position of the microsporidia (e.g. *E. cuniculi*) within the fungi is currently under debate, due to accelerated rate of sequence evolution. Early results suggested that microsporidia are among the earliest diverging protist lineages within the eukaryotes [55], however this seems to be an artefact of 'long branch attraction (LBA)' [56,57]. Recent phylogenetic [54,58,59] and molecular results [60-62] have implied that microsporidia are in fact atypical fungi [63] (Fig. 2 - red/light-red box). For the choanoflagellate *M. brevicollis* the position on the basis as the closest known relative to the metazoan clade was extracted from [44]. The basic relationship within the metazoan was found in [64] (Fig. 2 - light-blue box). The nematod *C. elegans* was placed as a sister group to the arthropods, according to the ecdysozoa hypothesis (Fig. 2 - blue box). An analysis based on the coelomata hypothesis did not lead to substantially different results (supplemental material, additional file 2). The precise order in the arthropods was gathered from the honey bee genome publication [36], for the fishes from a phylogenomics approach focusing on the Hox gene cluster [65]. The position of the lancelets and the urochordates to the vertebrates was chosen based on recent molecular data, suggesting that the urochordates, and not the lancelets [66], are the closest relatives to vertebrates [67]. As the exact order of divergence of the placozoan and cnidaria has not been determined beyond doubts [68], it was represented as a trifurcation.

### Ortholog detection

For the analysis of the ortholog relationships we used InParanoid [10] in version 2.0, with standard parameters and an outgroup. The outgroup was chosen as the closest sister taxon of the compared species. The underlying BLAST search was performed with the usage of the '-F m S' option enabling soft filtering of low complexity regions. This option will result in the highest number of identified orthologs and minimal error rates for BLAST based identification methods [69]. In order to increase the sensitivity of the ortholog identification we applied an iterative, triangular approach searching from a given gene to all identified orthologs in other species and used them as the

starting point for another search until no new ortholog were identified. This should further increase the sensitivity of the InParanoid algorithm, which has been reported to be about 80% [70], with both specificity and sensitivity, and therefore the best performing ortholog detection method [70,71]. Moreover, the test dataset used in [70] comprised six different eukaryotes (*Arabidopsis thaliana*, *C. elegans*, *D. melanogaster*, *Homo sapiens*, *S. cerevisiae* and *S. pombe*) spanning an even broader range of the eukaryotic tree of life. To further increase the sensitivity the BLAST searches were performed on protein sequences, whereas the definition of orthology is based on genes. Therefore, the resulting ortholog clusters had to be matched to genes. Following, overlapping or identical clusters, in the case of isoforms through alternative splicing, had to be resolved. In the clearest scenario a cluster consisted of more than two proteins from one species and, after mapping to the corresponding coding gene, the cluster had two identical genes. For this cluster one of the identical genes was deleted during the collapsing process. If two independent clusters consisted of several proteins and the clusters became identical after mapping, one of this clusters was deleted. In the case of overlapping clusters after mapping the clusters were merged.

As a result of the iterative search and the possibility of false positive assignments, the specificity might decrease. As our focus was on the secondary losses and the resulting evolutionary flexibility, this should only weakly influence our predictions. Moreover, this iterative search procedure should reduce the effect of fast evolving genomes and differences in the evolutionary rate of the examined species because the ortholog prediction is not merely based on direct ortholog identification starting from human, but predicting orthologs from more closely related species.

We defined gene emergence as the point in the lineage leading to the most recent common ancestor of the species in which the ortholog genes were present [72] (see Fig. 1c). This maximum parsimony approach will give a too recent origin of the gene if it was lost in the sister group of the derived last common ancestor. Considering the species sampling, this effect might be prominent for genes lost in fungi, which will be classified as metazoan specific. Similarly, a secondary loss was defined as the point in the phylogenetic tree where no ortholog of a given gene could be identified. This could be in a species or in the last common ancestor of several species if subsequent to the ancestor no ortholog was identified [19]. Thus, no multiple independent losses were counted (see Fig. 1c).

### Interaction data

The protein-complex dataset was based on HPRD [9] version 7 (9. Jan. 2007). We extracted only data derived by



affinity purification techniques leading to 1,060 complexes with 6,136 annotated proteins. The latter were mapped to 4,939 genes in total. These represented 2,197 unique genes due to homo-dimerisation of the gene products within a complex as well as gene products present in more than one complex.

### Comparison of phylogenetic pattern with random sets

To test, whether the observed evolutionary trends reflected a specific feature of protein complexes, we compared our results with a random model. We randomly drew 2,197 human genes out of the human dataset (approximately 23,000 genes). Based on this dataset, we applied the iterative ortholog detection method and retrieved the phylogenetic pattern of emergence. Moreover, based on the random dataset of 2,197 distinct genes we calculated random complexes with the same size distribution observed in the HPRD dataset (1,060 random complexes with 4,939 genes; genes must not be present twice or more in a given, but can be present in multiple complexes). We computed 10,000 repeats and compared this random model to the phylogenetic pattern observed for the HPRD dataset. As secondary losses depend on the point of emergence, we created a subset of randomly chosen 2,197 distinct genes out of the human dataset according to the observed distribution of emergence events along the tree. Furthermore, we created random complexes with the same size distribution observed in the HPRD dataset. For these dataset we computed 1,000 repeats and compared the phylogenetic pattern of secondary losses with the HPRD dataset. To estimate whether the biological signal deviated from the random model, we counted how many times a larger or lower signal, depending on the under- or overrepresentation of evolutionary events, was found in the random set. This count was divided by the number of random experiments to obtain a p-value estimate for every node. We corrected the alpha-value 0.05 for multiple testing according to the rough false discovery rate and marked the nodes with a p-value smaller than the corrected alpha as significant.

### Authors' contributions

JS designed the study. The analysis was performed by MFS. Both drafted and contributed to writing the paper, read the final manuscript and approved it.

### Additional material

#### Additional file 1

*Gene identifier of the ortholog genes predicted for the APC complex. Tabular collection of the obtained ortholog gene identifier of the human APC complex predicted by the iterative orthologs identification procedure and manual curation.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-9-155-S1.pdf>]

#### Additional file 2

*Phylogenetic tree with gene and complex emergence and losses (according to the coelomata hypothesis). The pattern of gene and complex emergence and the secondary losses of components of whole complexes is displayed along the tree according to the absence and presence pattern of the ortholog genes in terminal species or in subsets of species concluding the loss in the last common ancestor of all subsequent species. The numbers of gene and complex emergence is indicated in blue (complex emergence/gene emergence). The number of secondary losses are shown in red per affected node. It was discriminated between whole complex losses and gene losses (complex loss/gene losses).*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-9-155-S2.pdf>]

### Acknowledgements

We would like to thank Frank Förster for his help in programming the recursive tree mapping algorithm, proofreading the manuscript and for fruitful discussions. We would also like to thank Like Fokkens and John van Dam for discussion and comments on the manuscript. Some of the sequences used in this analysis were produced by the US Department of Energy Joint Genome Institute <http://www.jgi.doe.gov/>, the Tetraodon Sequencing Project at the Broad Institute of MIT and Harvard <http://www.broad.mit.edu> and the *Danio rerio* Sequencing Group at the Sanger Institute <http://www.sanger.ac.uk/Projects/D%5Frerio/>.

### References

- Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440(7084)**:631-6.
- Alberts B: **The cell as a collection of protein machines: preparing the next generation of molecular biologists.** *Cell* 1998, **92(3)**:291-4.
- Gavin AC, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415(6868)**:141-7.
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CWV, Figeys D, Tyers M: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415(6868)**:180-3.
- Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, Ong PS, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF: **Global landscape of protein**

- complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, **440(7084)**:637-43.
6. Krause R, von Mering C, Bork P, Dandekar T: **Shared components of protein complexes-versatile building blocks or biochemical artefacts?** *Bioessays* 2004, **26(12)**:1333-43.
  7. Rubin GM, Yandell MD, Wortman JR, Miklos GLG, Nelson CR, Har-iharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vossball LB, Zhang J, Zhao Q, Zheng XH, Lewis S: **Comparative genomics of the eukaryotes**. *Science* 2000, **287(5461)**:2204-15.
  8. Kroiss M, Wiesner J, Chari A, Sickmann A, Fischer U: **Evolution of an RNP assembly system: a minimal SMN complex facilitates formation of UsnRNPs in *Drosophila melanogaster***. 2008, **105(29)**:10045-50.
  9. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schieman WP, Constantinescu SN, Huang L, Khosravifar R, Steen H, Tewari M, Ghaffari S, Blobel GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans**. *Genome Res* 2003, **13(10)**:2363-71.
  10. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons**. *J Mol Biol* 2001, **314(5)**:1041-52.
  11. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M: **Protein interaction mapping in *C. elegans* using proteins involved in vulval development**. *Science* 2000, **287(5450)**:116-22.
  12. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs**. *Genome Res* 2004, **14(6)**:1107-18.
  13. Szklarczyk R, Huynen MA, Snel B: **Complex fate of paralogs**. *BMC Evol Biol* 2008, **8**:337.
  14. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huseon DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts JR, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hosten D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome**. *Science* 2001, **291(5507)**:1304-51.
  15. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrum J, Mesirov JP, Miranda C, Morris W, Naylor J, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Showstee R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendt MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minochima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Böcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korfi I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrino A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowski J, Consortium IHGS: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409(6822)**:860-921.
  16. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C: **Estimating the size of the human interactome**. *Proc Natl Acad Sci USA* 2008, **105(19)**:6959-64.
  17. Devos D, Russell RB: **A more complete, complexed and structured interactome**. *Curr Opin Struct Biol* 2007, **17(3)**:370-7.
  18. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P, Delbac F, Alaoui HE, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarès CP: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi***. *Nature* 2001, **414(6862)**:450-3.
  19. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution**. *Genome Res* 2003, **13(10)**:2229-35.
  20. Kortschak RD, Samuel G, Saint R, Miller DJ: **EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss**

- and rapid sequence divergence in the model invertebrates.** *Curr Biol* 2003, **13(24)**:2190-5.
21. Wyder S, Kriventseva E, Schröder R, Kadowaki T, Zdobnov E: **Quantification of ortholog losses in insects and vertebrates.** *Genome Biol* 2007, **8(11)**:R242.
  22. Wuchty S, Oltvai ZN, Barabási AL: **Evolutionary conservation of motif constituents in the yeast protein interaction network.** *Nat Genet* 2003, **35(2)**:176-9.
  23. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296(5568)**:750-2.
  24. Hershko A, Ciechanover A: **The ubiquitin system.** *Annu Rev Biochem* 1998, **67**:425-79.
  25. Thornton BR, Toczyski DP: **Precise destruction: an emerging picture of the APC.** *Genes & Development* 2006, **20(22)**:3069-78.
  26. Gmachl M, Gieffers C, Podtelejnikov AV, Mann M, Peters JM: **The RING-H2 finger protein APC11 and the E2 enzyme UBC4 are sufficient to ubiquitinate substrates of the anaphase-promoting complex.** *Proc Natl Acad Sci USA* 2000, **97(16)**:8973-8.
  27. Yu H, Peters JM, King RW, Page AM, Hieter P, Kirschner MW: **Identification of a cullin homology region in a subunit of the anaphase-promoting complex.** *Science* 1998, **279(5354)**:1219-22.
  28. Zachariae W, Shevchenko A, Andrews PD, Ciosk R, Galova M, Stark MJ, Mann M, Nasmyth K: **Mass spectrometric analysis of the anaphase-promoting complex from yeast: identification of a subunit related to cullins.** *Science* 1998, **279(5354)**:1216-9.
  29. Pál M, Nagy O, Ménesi D, Udvardy A, Deák P: **Structurally related TPR subunits contribute differently to the function of the anaphase-promoting complex in Drosophila melanogaster.** *J Cell Sci* 2007, **120(Pt 18)**:3238-48.
  30. Harper JW, Burton JL, Solomon MJ: **The anaphase-promoting complex: it's not just for mitosis any more.** *Genes & Development* 2002, **16(17)**:2179-206.
  31. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-402.
  32. Albà MM, Castresana J: **On homology searches by protein Blast and the characterization of the age of genes.** *BMC Evol Biol* 2007, **7**:53.
  33. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: **Benchmarking ortholog identification methods using functional genomics data.** *Genome Biol* 2006, **7**:R31.
  34. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12(9)**:334-6.
  35. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturvedi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigó R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL: **The genome sequence of the malaria mosquito Anopheles gambiae.** *Science* 2002, **298(5591)**:129-49.
  36. Honeybee Genome Sequencing Consortium: **Insights into social insects from the genome of the honeybee Apis mellifera.** *Nature* 2006, **443(7114)**:931-49.
  37. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernández J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-i T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PWH, Satoh N, Rokhsar DS: **The amphioxus genome and the evolution of the chordate karyotype.** *Nature* 2008, **453(7198)**:1064-71.
  38. elegans Sequencing Consortium C: **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282(5396)**:2012-8.
  39. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, Tomaso AD, Davidson B, Gregorio AD, Gelpke M, Goodstein DM, Harafuji N, Hastings KEM, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-Bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS: **The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins.** *Science* 2002, **298(5601)**:2157-67.
  40. **The Danio rerio Sequencing Group at the Sanger Institute** [<http://www.sanger.ac.uk/Projects/D%5Ferio/>]
  41. Eichinger L, Pachebat JA, Glöckner G, Rajandream MA, Sugang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hempill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Babu MM, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail MA, Urushihara H, Hernandez J, Rabinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shauly G, Schleicher M, Weinstock G, Rosenthal A, Cox EC, Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel AA, Barrell B, Kuspa A: **The genome of the social amoeba Dictyostelium discoideum.** *Nature* 2005, **435(7038)**:43-57.
  42. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PV, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burbulis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng X, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirkas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodard T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome**

- sequence of *Drosophila melanogaster*. *Science* 2000, **287(5461)**:2185-95.
43. Martin F, Aerts A, Ahrén D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buée M, Brokstein P, Canbäck B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbé J, Lin YC, Legué V, Tacon FL, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Secq MPOL, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kües U, Lucas S, de Peer YV, Podila GK, Polle A, Pukila PJ, Richardson PM, Rouzé P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV: **The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis**. *Nature* 2008, **452(7183)**:88-92.
  44. King N, Westbrock MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzov R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Sequencing JGI, Bork P, Lim WA, Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev IV, Rokhsar D: **The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans**. *Nature* 2008, **451(7180)**:783-8.
  45. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans N, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyraes E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson SL, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, James Kent W, Kirby A, Kolbe DL, Korfi I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger J, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie RW, McLaren S, McLay K, McPherson JD, Meldrum J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pezner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Raymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Brian S, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhauser AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420(6915)**:520-62.
  46. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization**. *Science* 2007, **317(5834)**:86-94.
  47. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, Jindo T, Kobayashi D, Shimada A, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto SI, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin-i T, Takeda H, Morishita S, Kohara Y: **The medaka draft genome and insights into vertebrate genome evolution**. *Nature* 2007, **447(7145)**:714-9.
  48. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtler K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, de Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Albà MM, Abril JF, Guigó R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hübner N, Ganten D, Goesele C, Hummel O, Kreidler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyraes E, Searle SM, Cooper GM, Batzoglu S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, López-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pezner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F, Consortium RGSP: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. *Nature* 2004, **428(6982)**:493-521.
  49. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes**. *Science* 1996, **274(5287)**:563-7.
  50. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, Hornsby T, Howarth S, Huckle EJ, Hunt S, Jagels K, James K, Jones L, Jones M, Leather S, McDonald S, McLean J, Mooney P, Moule S, Mungall K, Murphy L, Niblett D, Odell C, Oliver K, O'Neil S, Pearson D, Quail MA, Rabinowitz E, Rutherford K, Rutter S, Saunders D, Seeger K, Sharp S, Skelton J, Simmonds M, Squares R, Squares S, Stevens K, Taylor K, Taylor RG, Tivey A, Walsh S, Warren T, Whitehead S, Woodward J, Volckaert G, Aert R, Robben J, Grymonprez B, Weltjens I, Vanstreels E, Rieger M, Schäfer M, Müller-Auer S, Gabel C, Fuchs M, Dusterhöft A, Fritz C, Holzer E, Moestl D, Hilbert H, Borzym K, Langer I, Beck A, Lehrach H, Reinhardt R, Pohl TM, Eger P, Zimmermann W, Wedler H, Wambutt R, Purnelle B, Goffeau A, Cadieu E, Dréano S, Gloux S, Lelaure V, Mottier S, Galibert F, Aves SJ, Xiang Z, Hunt C, Moore K, Hurst SM, Lucas M, Rochet M, Gaillardin C, Tallada VA, Garzon A, Thode G, Daga RR, Cruzado L, Jimenez J, Sánchez M, del Rey F, Benito J, Domínguez A, Revuelta JL, Moreno S, Armstrong J, Forsburg SL, Cerutti L, Lowe T, McCombie WR, Paulsen I, Potashkin J, Shpakovski GV, Ussery D, Barrell BG, Nurse P, Cerrutti L: **The genome sequence of *Schizosaccharomyces pombe***. *Nature* 2002, **415(6874)**:871-80.
  51. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoeff F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G,

- Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes***. *Science* 2002, **297(5585)**:1301-10.
52. **The Tetraodon Sequencing Project at Broad Institute of MIT and Harvard** [<http://www.broad.mit.edu>]
53. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS: **The Trichoplax genome and the nature of placozoans**. *Nature* 2008, **454(7207)**:955-60.
54. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schüssler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkman-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lücking R, Büdel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R: **Reconstructing the early evolution of Fungi using a six-gene phylogeny**. *Nature* 2006, **443(7113)**:818-22.
55. Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR: **Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes**. *Nature* 1987, **326(6111)**:411-4.
56. Keeling P, Fast N: *Ecology and evolution of fungal endophytes and their roles against insects* Oxford Univ. Press, Oxford; 2005.
57. Felsenstein J: **Cases in which Parsimony or Compatibility Methods Will be Positively Misleading**. *Syst Zool* 1978, **27(4)**:401-410.
58. Keeling PJ, Luker MA, Palmer JD: **Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi**. *Mol Biol Evol* 2000, **17(1)**:23-31.
59. Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, Embley TM: **Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins**. *Proc Natl Acad Sci USA* 1999, **96(2)**:580-5.
60. Germot A, Philippe H, Guyader HL: **Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae***. *Mol Biochem Parasitol* 1997, **87(2)**:159-68.
61. Hirt RP, Healy B, Vossbrinck CR, Canning EU, Embley TM: **A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria**. *Curr Biol* 1997, **7(12)**:995-8.
62. Peyretailade E, Broussolle V, Peyret P, Méténier G, Gouy M, Vivarès CP: **Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin**. *Mol Biol Evol* 1998, **15(6)**:683-689.
63. de Peer YV, Ali AB, Meyer A: **Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi**. *Gene* 2000, **246(1-2)**:1-8.
64. Pennisi E: **Modernizing the tree of life**. *Science* 2003, **300(5626)**:1692-7.
65. Thomas-Chollier M, Ledent V: **Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*: comment**. *BMC Genomics* 2008, **9**:35.
66. Schaeffer B: **Deuterostome monophyly and phylogeny**. *Evol Biol* 1987, **21**:179-235.
67. Delsuc F, Brinkmann H, Chourrout D, Philippe H: **Tunicates and not cephalochordates are the closest living relatives of vertebrates**. *Nature* 2006, **439(7079)**:965-8.
68. Gerlach D, Wolf M, Dandekar T, Müller T, Pokorný A, Rahmann S: **Deep metazoan phylogeny**. In *Silico Biol* 2007, **7(2)**:151-4.
69. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits**. *Bioinformatics* 2008, **24(3)**:319-24.
70. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing performance of orthology detection strategies applied to eukaryotic genomes**. *PLoS ONE* 2007, **2(4)**:e383.
71. Hulsen T, Huynen MA, de Vlieg J, Groenen PMA: **Benchmarking ortholog identification methods using functional genomics data**. *Genome Biol* 2006, **7(4)**:R31.
72. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content**. *Genome Res* 2002, **12**:17-25.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

