OPEN ACCESS Freely available online

PLoS one

The Princeton Protein Orthology Database (P-POD): A Comparative Genomics Analysis Tool for Biologists

Sven Heinicke[1☯], Michael S. Livstone[1☯], Charles Lu[1☯], Rose Oughtred[1☯], Fan Kang[1], Samuel V. Angiuoli[2,3], Owen White[2], David Botstein[1], Kara Dolinski[1]*

1 Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America, 2 The Institute for Genomic Research, Rockville, Maryland, United States of America, 3 Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America

Many biological databases that provide comparative genomics information and tools are now available on the internet. While certainly quite useful, to our knowledge none of the existing databases combine results from multiple comparative genomics methods with manually curated information from the literature. Here we describe the Princeton Protein Orthology Database (P-POD, http://ortholog.princeton.edu), a user-friendly database system that allows users to find and visualize the phylogenetic relationships among predicted orthologs (based on the OrthoMCL method) to a query gene from any of eight eukaryotic organisms, and to see the orthologs in a wider evolutionary context (based on the Jaccard clustering method). In addition to the phylogenetic information, the database contains experimental results manually collected from the literature that can be compared to the computational analyses, as well as links to relevant human disease and gene information via the OMIM, model organism, and sequence databases. Our aim is for the P-POD resource to be extremely useful to typical experimental biologists wanting to learn more about the evolutionary context of their favorite genes. P-POD is based on the commonly used Generic Model Organism Database (GMOD) schema and can be downloaded in its entirety for installation on one's own system. Thus, bioinformaticians and software developers may also find P-POD useful because they can use the P-POD database infrastructure when developing their own comparative genomics resources and database tools.

Citation: Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, et al (2007) The Princeton Protein Orthology Database (P-POD): A Comparative Genomics Analysis Tool for Biologists. PLoS ONE 2(8): e766. doi:10.1371/journal.pone.0000766

INTRODUCTION

With the great explosion of biological data in the last decade, biological databases have become an essential part of today's research. The earliest online databases were the sequence repositories, such as Genbank [1] and EMBL [2], that provided the non-expert public access to the sequence data for genes, chromosomes, and eventually entire genomes, along with highly effective query and comparison tools. Soon after, several model organism databases that store and display the annotated genome sequences of well-studied organisms were developed. These databases now serve as an essential basic information source for all kinds of biological researchers.

For working biologists, some of the most important information concerns the phylogenetic relationships among proteins, which is not necessarily straightforward to recover from the basic sequence databases. Regardless of which organism one works with, much of the functional annotation of gene and protein functions is transferred, based on sequence similarity, from other organisms where more experimental information is available (for example, see the Gene Ontology annotations at http://www.geneontology.org/GO.current.annotations.shtml). It is for this reason that sequence similarity searching has become one of the most popular database tools in current use, perhaps second only to searching the published literature. To make good use of sequence similarity information, it would be very useful to have a simple, user-friendly way to visualize relationships in their phylogenetic context, particularly the relationships among the proteins in the model organisms from which most of the functional annotations are derived. It is of particular value to be able to know which proteins are (or might be) orthologous [i.e. similar to each other in sequence because they originated from a common ancestor, having been separated in evolutionary time only by speciation event(s)]. It is also useful to see these orthologous relationships in the context of the larger paralogous gene families ultimately caused by gene duplications during the course of evolution.

In this paper, we describe P-POD, which provides the user an easy way to find and visualize the orthologs to a query sequence in the eukaryotes of greatest interest to working biologists (*i.e.* the experimental model organisms and the human) in their evolutionary context, and to link these relationships with the relevant literature. Several databases that specialize in comparative genomics have recently come online. Each of these databases, including P-POD, has both useful features and problems specific to the methods or species chosen in the analysis (Table 1, reviewed in [3]); none is perfect, but each fulfills the needs of particular database users.

P-POD is meant to complement these existing databases by providing a comparative genomics analysis system readily accessible to and readable by experimentalists, containing not just computational comparative analyses of the most common

Academic Editor: Berend Snel, Utrecht University, Netherlands

Received May 18, 2007; **Accepted** July 18, 2007; **Published** August 22, 2007

Funding: This work was funded by NIH grant 5R01HG003471 awarded to DB (PI) and KD (co-investigator), by NIH grant P50 GM071508 awarded to DB, and by NIH contract NO1-AI-40038 awarded to OW.

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: kara@genomics.princeton.edu

☯ These authors contributed equally to this work.

Table 1. Comparative genomics web resources.

Name	Description	Ortholog prediction	Larger seq. families	Disease information	Curated literature
Clusters of Orthologous Groups (COGs/KOGs) [22]	Provides groups of orthologous proteins for seven eukaryotic species; the construction protocol involves manual curation	Yes	Yes	No	No
Eukaryotic Gene Orthologs (EGO) [23]	Displays predicted orthologs derived from several eukaryotic genomes based on gene alignments	Yes	No	No	No
Homologene [24]	Provides automated predictions of homologs among the genes of several eukaryotes	No	Yes	Yes	No
Inparanoid [25]	Houses pair-wise groups of orthologous proteins for multiple species	Yes	No	No	No
OrthoDisease [26]	Uses the Inparanoid algorithm to generate pair-wise orthologs between human disease genes and genes from other species	Yes	No	Yes	No
OrthoMCL-DB [4,27]	Utilizes a Markov Cluster algorithm to predict orthologous groups of proteins for multiple species simultaneously	Yes	No	No	No
Sybil (S. Angiuoli and O. White, in preparation)	Uses Jaccard clustering to group sequences based on pair-wise BLAST analysis	No	Yes	No	No
YOGY [28]	Retrieves orthologous proteins from four different resources: KOGs, Inparanoid, Homologene, and OrthoMCL-DB	Yes	No	No	Yes (only budding and fission yeast)
P-POD (This study)	Orthologs and Jaccard clusters	Yes	Yes	Yes	Yes

doi:10.1371/journal.pone.0000766.t001

experimental organisms but also literature curation and links to other databases of interest. For example, while the OrthoMCL database contains sequences from over 55 prokaryotic and eukaryotic genomes, we chose to include protein sequences from eight eukaryotic organisms for their medical value or their status as widely-studied model organisms. There are certainly users who would need the more comprehensive species set from OrthoMCL. While P-POD uses the underlying OrthoMCL algorithm, it is meant to complement the OrthoMCL online database by serving another set of users, primarily experimental biologists who wish to query with their gene of interest from a well studied model organism to quickly get the evolutionary context of that gene along with other relevant information about that gene without sorting through a very large list of other sequences.

We designed our comparative genomics analysis system so that different components could be added to and removed from the pipeline in a modular fashion; the initial version of the pipeline described here generates related protein families using two different methods to provide complementary views of phylogenetic relationships. We used OrthoMCL ([4]) to find the orthologs and a version of Jaccard Clustering [modified to find homologs across multiple genomes (S. Angiuoli and O. White, in preparation)] to provide a larger protein family context. The phylogenetic relationships among family members from each method are determined using CLUSTAL W [5] and PHYLIP and visualized as arbitrarily rooted trees. In addition, we provide relevant gene and disease information from the Online Mendelian Inheritance in Man (OMIM) [6] database and also provide information culled from the literature that can be used to indicate when functional conservation has been shown experimentally between predicted orthologs. All the data within the database are freely available through the web and by downloading the entire software and database system via the following URL: http://ortholog.princeton.edu/

Historically, genomic databases have been developed in isolation, with idiosyncratic database schemata and software. Much duplication of effort can be avoided by developing generic modular databases and software that save, especially in the long run, both time and money spent on development, maintenance, and user training. In constructing P-POD we made use of the database schema, installation and loading tools, and various software components from the Generic Model Organism Database (GMOD) project (www.gmod.org), The goal of GMOD is to develop an open and generic genomic database environment, including database schemata and required software tools.

RESULTS

The P-POD Pipeline

In the interests of both simplicity and flexibility, the P-POD pipeline employs a modular architecture. The pipeline takes FASTA-formatted protein sequences as input, performs comparative genomic analyses, and stores the results in a database. In addition, we have created web tools that allow searching and browsing of the results in a user-friendly manner. We built the initial pipeline to identify putative orthologous proteins using OrthoMCL [4]. We chose OrthoMCL over other algorithms mainly because it can be run on multiple species at once and is one of the better-performing algorithms in terms of sensitivity and specificity [7] [3]. We generated larger families of related sequences using Jaccard clustering modified to find homologs across multiple genomes; see the Materials and Methods section for algorithm details. It is important to note that we built the P-

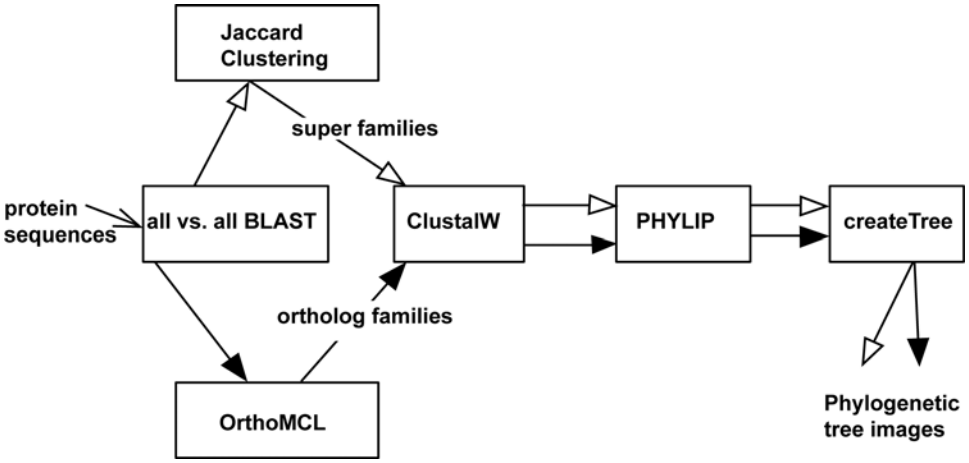

Figure 1. Steps in the analysis pipeline.
doi:10.1371/journal.pone.0000766.g001

POD system so that we can easily add or remove results from different analysis methods. We acknowledge that the first choice is not always the best choice, and as algorithms improve and/or as users request other methods, we plan to modify and expand the system as appropriate. P-POD generates phylogenetic trees from both analyses using CLUSTAL W [5] and PHYLIP; the trees are graphically displayed on the web. The overall pipeline is illustrated in Figure 1. The sources and versions of the pipeline components are listed in Table 2. The data are stored in a Generic Model Organism Database (GMOD) database schema using the freely available PostgreSQL software to make the entire system accessible to as many users as possible, not only through the web but also via download of the entire system.

The P-POD database contains protein sequences from eight eukaryotic organisms with fully sequenced genomes chosen for either their medical value or their status as widely-studied model organisms. They include a yeast (*Saccharomyces cerevisiae*), a nematode worm (*Caenorhabditis elegans*), a fruit fly (*Drosophila melanogaster*), a flowering plant (*Arabidopsis thaliana*), a fish (*Danio rerio*), a mouse (*Mus musculus*), and human (*Homo sapiens*). These are the leading experimental organisms for modern biologists, and among them span much of the evolutionary tree of the eukaryotes. Also included is the malaria parasite *Plasmodium falciparum*, an organism that, although it is a eukaryote, has a relatively exotic parasitic lifestyle. Sources for each protein set are listed in Table 3. Also stored in the system are results from each step of the pipeline, gene and disease information from OMIM, and curated information from the literature describing experimental tests of functional conservation (see Figure 2).

The pipeline generated a total of 25,271 OrthoMCL families and 15,050 Jaccard Clustering families that contain a total of 165,970 proteins (154,736 and 152,799 for each method, respectively) from eight different organisms. There are 984 OrthoMCL families that contain at least one protein from each of the species, with 112 of them containing exactly one protein from each. We used the GO Term Mapper tool available at SGD to determine the distribution of GO annotations for the 112 yeast proteins in these families; we chose the yeast proteins because complete GO annotation is available for the entire yeast genome [8]. Not surprisingly, these proteins are involved in core biological processes that are common across eukaryotes, including translation, transport, cell cycle regulation, and cytoskeleton organization. These genes are also well characterized; only four of the 112 genes were annotated to "biological process unknown." We also used the GO Term Finder [9] implementation at Princeton (http://go.princeton.edu/) to look for enrichment of GO terms among the 112 genes. Again unsurprisingly, the most significant shared term is "ribosome biogenesis and assembly" (corrected P-value = 5.85e-18) along with other terms related to translation and basic metabolic processes, all processes common among the eukaryotes.

The complete species distribution of each family is available via the web (http://ortholog.princeton.edu/organismdist.html), and the number of proteins found in families and orphan proteins

Table 2. Components of the analysis pipeline.

Program	Version	Source
GMOD::Loader		This study
WU-BLAST	2.0MP-WashU 10-May-2005	http://blast.wustl.edu/
OrthoMCL [4]	Version 1.2 14-March-2005	http://sourceforge.net/projects/orthomcl/
MCL [29]	Version 1.005, 05-118	http://micans.org/mcl/
Jaccard Clustering	NA	S. Angiuoli and O. White (in preparation)
Clustal W [5]	Version 1.83	ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw
PHYLIP	Version 3.64	http://evolution.genetics.washington.edu/phylip.html
createTree		This study

doi:10.1371/journal.pone.0000766.t002

Table 3. Sources and numbers of sequences analyzed.

Organism	Proteins	Database	Filename
S. cerevisiae	6704	SGD	orf_trans_all.fasta.gz
H. sapiens	33869	ENSEMBL	Homo_sapiens.NCBI35.nov.pep.fa.gz
M. musculus	36471	ENSEMBL	Mus_musculus.NCBIM34.nov.pep.fa
D. rerio	32143	ENSEMBL	Danio_rerio.ZFISH5.nov.pep.fa
D. melanogaster	19178	FlyBase	dmel-all-translation-r4.2.1.fa
C. elegans	22858	WormBase	wormpep150.fa
A. thaliana	30690	TAIR	TAIR6_pep_20051108.fa
P. falciparum	5363	PlasmoDB	Pfa3D7_WholeGenome_Annotated_PEP_2005.2.11.fa

doi:10.1371/journal.pone.0000766.t003

(those not found in an OrthoMCL or Jaccard family) from all the species is found in Table 4.

The percentage of orphans is generally strikingly low, with the percent orphaned in a given species 13% or lower, with two exceptions, yeast (32%) and *Plasmodium* (33%). These numbers confirm the high conservation of proteins across eukaryotes, with the notable exception the *Plasmodium* outlier. The high percentage of yeast orphans is due to the fact that we did the analysis with the complete protein set, including over 800 ORFs flagged as "Dubious" by SGD; these are not likely to actually encode proteins, and when they are excluded the percentage of orphans in yeast drops to about 20%.

P-POD includes 1,895 human proteins that are associated with human diseases (based on protein-OMIM disease files downloaded from ENSEMBL), 1,852 of which were found in either an OrthoMCL or Jaccard family; in each of these cases, links to the relevant OMIM records are provided online.

Manually Curated Information

P-POD also includes curated literature that contains information relevant to the yeast proteins in the database. The source of the literature is the *Saccharomyces* Genome Database (SGD). SGD provides a Literature Guide tool that categorizes yeast literature into different topics, two of which, "Cross-species expression" and "Disease-gene related," are particularly relevant to the data in P-POD; we believe that this set of papers, which is continually updated and curated, contains most, if not quite all, of the experimental data testing functional conservation between yeast and other organisms. All papers associated with these topics were downloaded from the SGD FTP site and loaded into the database (see Materials and Methods). They are then displayed on the web interface, with links to PubMed, so that users can compare experimentally determined functional conservation and computationally predicted orthology. This set of papers does not, of course, address proteins without a yeast ortholog. A way of dealing with this limitation is under study; a likely development will be the inclusion of papers from the literatures of other model organisms. For disease-related genes, we provide OMIM links that at least partially fill this gap for the human.

In addition, we manually curated the "Cross-species expression" papers to indicate explicitly when functional conservation was experimentally determined. These cross-species expression experiments test whether expressing a putative ortholog from one organism will restore wildtype function to the corresponding inactivated gene in another organism (almost always *S. cerevisiae*). Table 5 summarizes this curated information for only the yeast proteins in the disease-related families to illustrate how this information can be compared to computational results, but P-POD contains experimental results for all yeast proteins for which curated information is available. The orthologs predicted by OrthoMCL often exhibit conserved function. Of the 643 curated complementation experiments between yeast genes and their putative orthologous sequences from other organisms, 395 showed functional conservation and were also identified as orthologs by OrthoMCL; 50 did not complement and were also not predicted to be orthologs by OrthoMCL. Thus, in most cases (445/643), the computational determination of orthology was consistent with experimental results of functional conservation. However, in 153 experiments, complementation was observed, but the proteins were not in the same OrthoMCL family, and in 45 experiments, complementation did not occur, but OrthoMCL predicted an orthologous relationship between the two proteins. These experimental results can be used as a rudimentary assessment of the computational predictions but it must be noted that the definition of orthology does not require functional conservation [10], and there are actual cases (*e.g.* actin) where *in vivo* complementation fails for biological reasons, even for true orthologs that can function *in vitro* [11].

The P-POD User Interface: Orthologs, Families and Diseases

We designed a simple web interface that allows users to search and browse the data in several ways (Figure 2). Results can be queried by various peptide identifiers or gene names, choosing from any of eight model organisms for the query protein and a particular analysis method, or they can be searched or browsed by Online Mendelian Inheritance in Man (OMIM) ID.

Searches generate result pages that contain:

- a hyperlinked phylogenetic tree of predicted orthologs generated by OrthoMCL or of more distantly-related proteins generated by Jaccard clustering,
- a list of diseases and genes associated with the human ortholog(s) as documented in OMIM,
- a manually curated list of papers with cross-complementation experiments involving the yeast ortholog(s), and
- a downloadable ClustalW alignment of family members.

Using P-POD to Compare Methods: Jaccard and OrthoMCL

To illustrate the usefulness of being able to store multiple analyses in a single database, we further compared the results between the

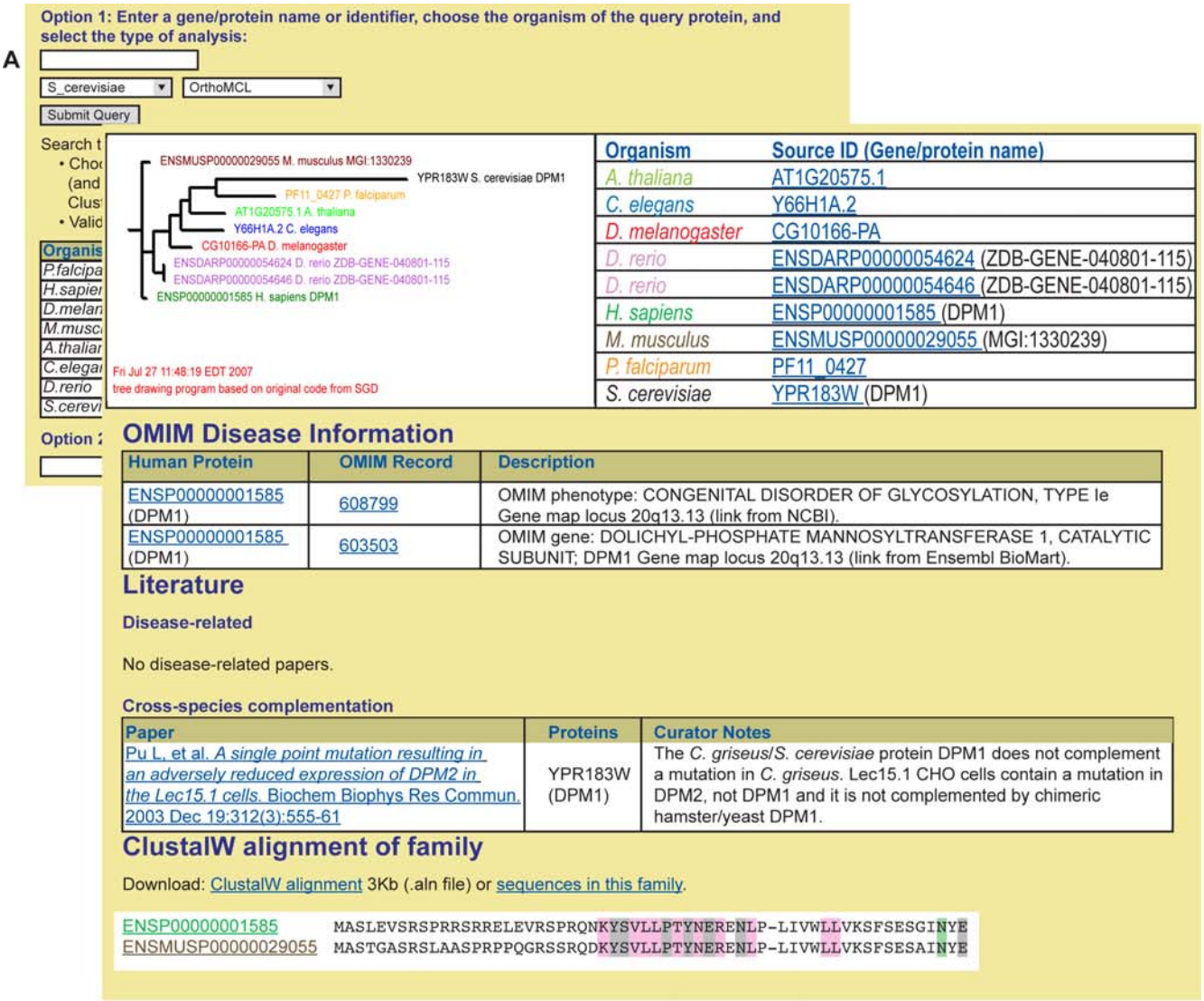

Figure 2. Screenshots of the P-POD web interface. (A) A portion of the results page for the *DPM1* OrthoMCL family is shown superimposed on the search form. Results from OrthoMCL are provided, and a link to the larger Jaccard family (B) is also available. Disease information from OMIM is displayed, as well as any relevant disease or cross-complementation literature.
doi:10.1371/journal.pone.0000766.g002

Table 4. Number of proteins in each organism found in OrthoMCL or Jaccard families.

Organism	OrthoMCL	Jaccard	Orphan (% of total proteome)
S. cerevisiae	4,333	3,660	2,176 (32%)
H. sapiens	27,606	29,315	3,193 (9%)
M. musculus	29,214	31,388	3,902 (11%)
D. rerio	27,602	28,968	1,903 (6%)
D. melanogaster	16,015	15,048	2,503 (13%)
C. elegans	18,070	16,308	4,078 (7%)
A. thaliana	27,987	25,819	2,279 (13%)
P. falciparum	3,909	2,293	1,284 (33%)

doi:10.1371/journal.pone.0000766.t004

OrthoMCL and Jaccard Clustering methods. A query for yeast *TUB1* using only OrthoMCL reveals the alpha tubulins from yeast and other organisms (Figure 3), but not the important paralogous relationships to the beta and gamma tubulins [12] [13], which are observed in the *TUB1* Jaccard cluster (not shown). These three main classes of tubulins are related to the bacterial FtsZ protein and diverged prior to the divergence of the eukaryotes [12]. Many such examples are found, especially among the ancient gene families that go back to the common ancestors of all eukaryotes. The Jaccard clustering provides this larger evolutionary context.

While OrthoMCL identifies predicted orthologs, the Jaccard clustering algorithm should build broader families of more distantly related sequences. Accordingly, one might initially expect that each OrthoMCL family would be a subset of a corresponding Jaccard cluster. Of course, because each algorithm defines homologs quite differently, in practice it would be reasonable to expect a certain degree of disagreement between the OrthoMCL and Jaccard clustering results. Of the 25,271 OrthoMCL families, 17,340 (69%) are subsets of Jaccard clusters. A certain amount of the "loss" of family members is due to stochastic effects; 72% of the 22,216 OrthoMCL families with ten or fewer members remain intact as subsets of Jaccard clusters, compared to only 49% of the 3,055 larger families. Fully 91% of the peptides assigned to OrthoMCL families also lie in Jaccard clusters. 82% of the OrthoMCL families have 80% or more of their peptides in a single Jaccard cluster; 93% have 50% or more.

Another possible source of inconsistency between the OrthoMCL and Jaccard results is that these analyses were run with different parameter settings. In particular, an alignment constraint was used for the Jaccard clustering alone because the default and recommended settings for OrthoMCL do not include an alignment constraint (see http://orthomcl.cbil.upenn.edu/ORTHOMCL/). The Jaccard clustering software was configured to ignore BLAST hits that did not align over 50% of the length of both peptides. For example, yeast *MET3* and *MET14* respectively encode ATP sulfurylase and adenylylsulfate kinase, which catalyze the first two steps of a sulfate assimilation pathway. *A. thaliana* retains this distinction, but *C. elegans*, *D. melanogaster*, *D. rerio*, human, and mouse have bifunctional proteins containing both activities. The OrthoMCL family contains all of these peptides (Figure 4B), but *MET14* and the four *Arabidopsis* adenylylsulfate kinases form their own Jaccard cluster (Figure 4A). At 202 amino acids, Met14p is less than half the length of the other OrthoMCL family members and therefore fails to satisfy the 50% alignment constraint used in the Jaccard clustering algorithm.

Again, having both sets of results within the same database made comparison of the two methods and detection of possible issues relatively straightforward. We expect that this will be a useful feature for database developers and/or bioinformaticians who may download the entire P-POD system for local installation to use as a development base for their algorithms of choice.

Other Uses for P-POD

We provide several examples of how P-POD might be used by experimental biologists, and not necessarily those expert in phylogenomics. In addition, we illustrate how providing results from different analysis methods can help to identify issues characteristic of the different methods.

The P-POD system can be used in a simple way to learn something global about the genes and/or proteins of an organism. As an illustration, we studied the conservation of essential genes, *i.e.* genes that are required for viability, across yeast and mammals. Among the 929 OrthoMCL families with unambiguous orthologs from yeast, mouse, and human (*i.e.* exactly one member from each of these species), phenotype data were available for the yeast and mouse genes in 107 cases. In 28 cases, the yeast gene was essential, and in 24 of these families (86%), the mouse gene was also essential. The entire analysis can be found at http://ortholog.princeton.edu/essential_analysis.html.

P-POD can be used to estimate whether essential yeast genes are more likely to be conserved and/or related to a human disease gene. There are 1100 essential and 4670 non-essential yeast genes, respectively. 853 essential yeast genes (77.5%) are found in an OrthoMCL family, while 247 (22.5%) are not. Of the non-essential genes, 2968 (63.6%) are found in families, while 1702 (36.4%) are not. These data suggest that essential genes are more conserved than non-essential genes ($\chi^2 = 78$, $p = 1.1e\text{-}18$). When examining essentiality among the 954 yeast genes found in disease-related families, 191 of them are essential (20% of the disease-related genes, 17% of all essential genes), while 691 of them are non-essential (72% of disease-related genes, 14.8% of all non-essential genes); phenotype data are not available for the remaining 72 yeast genes. Thus, there does not appear to be enrichment of essential genes among the disease-related yeast genes ($\chi^2 = 4.5$, $p = 0.03$). The lack of enrichment of essential genes among disease-related genes is initially surprising; however, this result can be explained if genes required for viability in yeast are also required for viability of human cells, thus making it impossible for the mammal to fully develop into even a diseased organism.

P-POD simplifies the study of the relationships among families of proteins with related functions. One example is the DNA-dependent RNA polymerase family (Figure 5A, B, C). Transcription of genes in eukaryotes is generally performed by three RNA polymerases (I, II, and III), each of which is composed of more than 10 subunits [14], Searching on a selection of individual yeast RNA polymerase subunits (*RPO21, RPO31, RPA190, RPB2, RPB4, RPB5, RPA135*, and *RET1*) resulted in separate phylogenetic tree displays for each protein, demonstrating that they had been effectively resolved into distinct ortholog clusters. Within each cluster, there were mainly one-to-one orthologous relationships between the proteins from each species, except for *RPA135*, and *RET1*, which include orthologs from each species examined except for *D. rerio* (Figure 5A, B).

For some subunits, in particular *RPO21, RPA190*, and *RPA135*, there appear to be more than one mouse or human paralog; however, upon further investigation, it was determined that the separate peptides were encoded by a single mouse or human gene (Figure 5A). Therefore, for the most part, each protein from each species appeared to be orthologous to the others, as would be expected for proteins functioning in a core biological process [14].

Table 5. Functional conservation vs. ortholog prediction: comparing experimental results with the OrthoMCL ortholog predictions for disease-related families.

OrthoMCL	Experimental	Yeast gene	Protein(s) tested	Citation
No	No	*YJL095W: BCK1*	*H. sapiens: ENSP00000306124*	[31]
No	No	*YJR040W: GEF1*	*M. musculus: ENSMUSP00000035964*	[32]
No	No	*YMR190C: SGS1*	*H. sapiens: ENSP00000298139*	[33]
No	No	*YOL090W: MSH2*	*H. sapiens: ENSP00000265081, ENSP00000234420*	[34]
Yes	Yes	*YAL016W: TPD3*	*A. thaliana: AT1G25490.1*	[35]
Yes	Yes	*YBR110W: ALG1*	*H. sapiens: ENSP00000262374*	[36] [37]
Yes	Yes	*YBR140C: IRA1*	*H. sapiens: ENSP00000351015, ENSP00000348498*	[38]
Yes	Yes	*YBR140C: IRA1*	*H. sapiens: ENSP00000351015, ENSP00000352435, ENSP00000348498*	[39]
Yes	Yes	*YBR254C: TRS20*	*H. sapiens: ENSP00000310153*	[40]
Yes	Yes	*YCR075C: ERS1*	*H. sapiens: ENSP00000046640*	[41]
Yes	Yes	*YDL120W: YFH1*	*H. sapiens: ENSP00000297735*	[42,43]
Yes	Yes	*YDL126C: CDC48*	*A. thaliana: AT3G09840.1*	[44]
Yes	Yes	*YDR270W: CCC2*	*H. sapiens: ENSP00000242839, ENSP00000342559*	[45] [46] [47]
Yes	Yes	*YDR270W: CCC2*	*C. elegans: Y76A2A.2*	[48]
Yes	Yes	*YDR270W: CCC2*	*H. sapiens: ENSP00000343026, ENSP00000345728*	[49] [50]
Yes	Yes	*YDR363W-A: SEM1*	*M. musculus: ENSMUSP00000040741*	[51]
Yes	Yes	*YDR363W-A: SEM1*	*H. sapiens: ENSP00000248566*	[52]
Yes	Yes	*YER095W: RAD51*	*M. musculus: ENSMUSP00000028795*	[53]
Yes	Yes	*YER120W: SCS2*	*H. sapiens: ENSP00000217602, ENSP00000345656*	[54]
Yes	Yes	*YER171W: RAD3*	*H. sapiens: ENSP00000221481*	[55] [56]
Yes	Yes	*YFL018C: LPD1*	*H. sapiens: ENSP00000205402*	[57]
Yes	Yes	*YFR019W: FAB1*	*M. musculus: ENSMUSP00000079926*	[58]
Yes	Yes	*YFR053C: HXK1*	*H. sapiens: ENSP00000338009, ENSP00000223366, ENSP00000350996*	[59]
Yes	Yes	*YGL001C: ERG26*	*M. musculus: ENSMUSP00000033715*	[60]
Yes	Yes	*YGL006W: PMC1*	*A. thaliana: AT2G41560.1*	[61]
Yes	Yes	*YGL006W: PMC1*	*A. thaliana: AT3G21180.1*	[62]
Yes	Yes	*YGL115W: SNF4*	*A. thaliana: AT1G09020.1*	[63,64]
Yes	Yes	*YGL125W: MET13*	*A. thaliana: AT3G59970.1, AT2G44160.1*	[65]
Yes	Yes	*YGL125W: MET13*	*H. sapiens: ENSP00000315965*	[66]
Yes	Yes	*YGL167C: PMR1*	*H. sapiens: ENSP00000306816, ENSP00000329664, ENSP00000352665*	[67,68]
Yes	Yes	*YGL167C: PMR1*	*H. sapiens: ENSP00000306816, ENSP00000329664, ENSP00000349901, ENSP00000352580, ENSP00000352665*	[69]
Yes	Yes	*YGL253W: HXK2*	*H. sapiens: ENSP00000338009, ENSP00000223366, ENSP00000350996*	[59]
Yes	Yes	*YGR240C: PFK1*	*H. sapiens: ENSP00000345771, ENSP00000352842*	[70,71]
Yes	Yes	*YGR267C: FOL2*	*H. sapiens: ENSP00000352686, ENSP00000254299*	[72,73]
Yes	Yes	*YHR037W: PUT2*	*H. sapiens: ENSP00000290597, ENSP00000336944*	[74,75]
Yes	Yes	*YIL143C: SSL2*	*A. thaliana: AT5G41360.1*	[76]
Yes	Yes	*YJL059W: YHC3*	*H. sapiens: ENSP00000353116, ENSP00000353116, ENSP00000346650*	[77]
Yes	Yes	*YJL101C: GSH1*	*D. melanogaster: CG2259-PA, CG2259-PB*	[78]
Yes	Yes	*YJR104C: SOD1*	*H. sapiens: ENSP00000270142*	[79]
Yes	Yes	*YJR117W: STE24*	*H. sapiens: ENSP00000196805*	[80,81]
Yes	Yes	*YJR135W-A: TIM8*	*H. sapiens: ENSP00000247385*	[82,83]
Yes	Yes	*YKL209C: STE6*	*M. musculus: ENSMUSP00000041204*	[84]
Yes	Yes	*YKL209C: STE6*	*M. musculus: ENSMUSP00000041204, ENSMUSP00000088389*	[85]
Yes	Yes	*YKR079C: TRZ1*	*H. sapiens: ENSP00000337445*	[86]
Yes	Yes	*YLR142W: PUT1*	*A. thaliana: AT5G38710.1*	[87]
Yes	Yes	*YML021C: UNG1*	*H. sapiens: ENSP00000242576, ENSP00000337398*	[88]
Yes	Yes	*YMR190C: SGS1*	*H. sapiens: ENSP00000347232, ENSP00000349859*	[33,89,90]
Yes	Yes	*YMR205C: PFK2*	*H. sapiens: ENSP00000345771, ENSP00000352842*	[70,71]
Yes	Yes	*YNL219C: ALG9*	*H. sapiens: ENSP00000316397*	[36]

Table 5. Cont.

OrthoMCL	Experimental	Yeast gene	Protein(s) tested	Citation
Yes	Yes	*YNR030W: ALG12*	*H. sapiens: ENSP00000333813*	[91]
Yes	Yes	*YNR041C: COQ2*	*H. sapiens: ENSP00000310873*	[92]
Yes	Yes	*YNR041C: COQ2*	*A. thaliana: AT4G23660.1*	[93]
Yes	Yes	*YOL049W: GSH2*	*H. sapiens: ENSP00000216951*	[94]
Yes	Yes	*YOL081W: IRA2*	*H. sapiens: ENSP00000351015, ENSP00000348498*	[38,95]
Yes	Yes	*YOR204W: DED1*	*H. sapiens: ENSP00000310870*	[96]
Yes	Yes	*YOR204W: DED1*	*D. melanogaster: CG9748-PA*	[97]
Yes	Yes	*YPL022W: RAD1*	*A. thaliana: AT5G41150.1*	[98]
Yes	Yes	*YPL153C: RAD53*	*H. sapiens: ENSP00000329178, ENSP00000329012*	[99]
Yes	Yes	*YPL218W: SAR1*	*A. thaliana: AT1G56330.1*	[100]
Yes	Yes	*YPR183W: DPM1*	*S. cerevisiae: DPM1*	[101]
No	Yes	*YBR018C: GAL7*	*H. sapiens: ENSP00000338703*	[102]
No	Yes	*YBR289W: SNF5*	*A. thaliana: AT3G17590*	[103]
No	Yes	*YDR135C: YCF1*	*A. thaliana: AT3G13080.1*	[104,105]
No	Yes	*YGL006W: PMC1*	*H. sapiens: ENSP00000306816, ENSP00000329664, ENSP00000352665*	[68]
No	Yes	*YGL167C: PMR1*	*A. thaliana: AT1G07810.1*	[106]
No	Yes	*YGL167C: PMR1*	*A. thaliana: AT2G41560.1*	[61]
No	Yes	*YGL167C: PMR1*	*A. thaliana: AT3G21180.1*	[62]
No	Yes	*YHL007C: STE20*	*A. thaliana: AT4G08500.1*	[107]
No	Yes	*YJR040W: GEF1*	*M. musculus: ENSMUSP00000030879*	[32]
No	Yes	*YJR104C: SOD1*	*H. sapiens: ENSP00000307870*	[108]
No	Yes	*YNL098C: RAS2*	*H. sapiens: ENSP00000309845*	[109]
No	Yes	*YOR101W: RAS1*	*H. sapiens: ENSP00000309845*	[109]
No	Yes	*YOR130C: ORT1*	*A. thaliana: AT1G79900.1*	[110]
No	Yes	*YPL111W: CAR1*	*A. thaliana: AT4G08900.1*	[111]
Yes	No	*YDR529C: QCR7*	*H. sapiens: ENSP00000287022*	[112]
Yes	No	*YER148W: SPT15*	*H. sapiens: ENSP00000230354*	[113]
Yes	No	*YNL280C: ERG24*	*D. melanogaster: CG17952-PC*	[114]
Yes	No	*YOL090W: MSH2*	*H. sapiens: ENSP00000233146*	[34]
Yes	No	*YPR183W: DPM1*	*H. sapiens: ENSP00000001585*	[115]

In all but one of these experiments, the yeast gene was mutated and the gene from the other organism was tested for the ability to complement the mutant phenotype. In the one exception, the yeast gene *DPM1* was expressed in mouse. In the OrthoMCL column, "Yes" indicates that the OrthoMCL algorithm placed the two proteins in the same ortholog family, while "No" indicates it did not. In the Experimental column, "Yes" indicates functional complementation, while "No" indicates none. Thus, when both columns are the same, the OrthoMCL prediction is consistent with the experimental result i.e. in the cases where both are "Yes," the predicted orthologs are functionally conserved, and when both are "No," the proteins are not predicted to be orthologs, and they are not functionally conserved.
doi:10.1371/journal.pone.0000766.t005

Interestingly, experimental evidence shows that although all eukaryotes have RNA polymerases I, II, and III, plants are unique in that they have subunits for a fourth polymerase, Pol IV. The closely related genes, AT3G18090.1 (NRPD2B) and AT3G23780.1 (NRPD2A), have been found to encode the second largest subunit of plant Pol IV, with most of the NRPD2 transcripts coming from NRPD2A. These atypical second largest subunits occurring only in plants are most similar in sequence to the RNA polymerase II second largest subunits in other eukaryotes such as yeast *RPB2* [15,16]. Despite this sequence similarity, they were effectively resolved away from the OrthoMCL-generated ortholog cluster containing yeast *RBP2* into their own distinct two-member family. The Jaccard clustering method, on the other hand, correctly grouped these unique Pol IV plant subunits with the other second largest RNA polymerase subunit families, as shown in Figure 5D.

As another illustration, we examined thirty yeast ER proteins involved in asparagine-linked glycosylation, a pathway which is well-conserved between yeast and humans in its early steps and diverges soon after glycosylated proteins enter the Golgi (Table 6). Of these, 27 are known from the literature to have human homologs. This analysis shows that 26 lie in ortholog families, with the majority having orthologs in *Homo sapiens* (26), *D. melanogaster* (24), *A. thaliana* (24), *M. musculus* (23), *C. elegans* (23), and *D. rerio* (21). The four proteins that do not lie in ortholog families are subunits of the yeast oligosaccharyltransferase complex. Deleterious mutations in ten of the human homologs cause congenital disorders of glycosylation. Interestingly, only nine of the thirty yeast ER proteins have orthologs in *P. falciparum*. N-linked glycosylation has been detected only at very low levels in *P. falciparum* [17], and ensuring appropriate glycosylation in heterologously-expressed *P.*

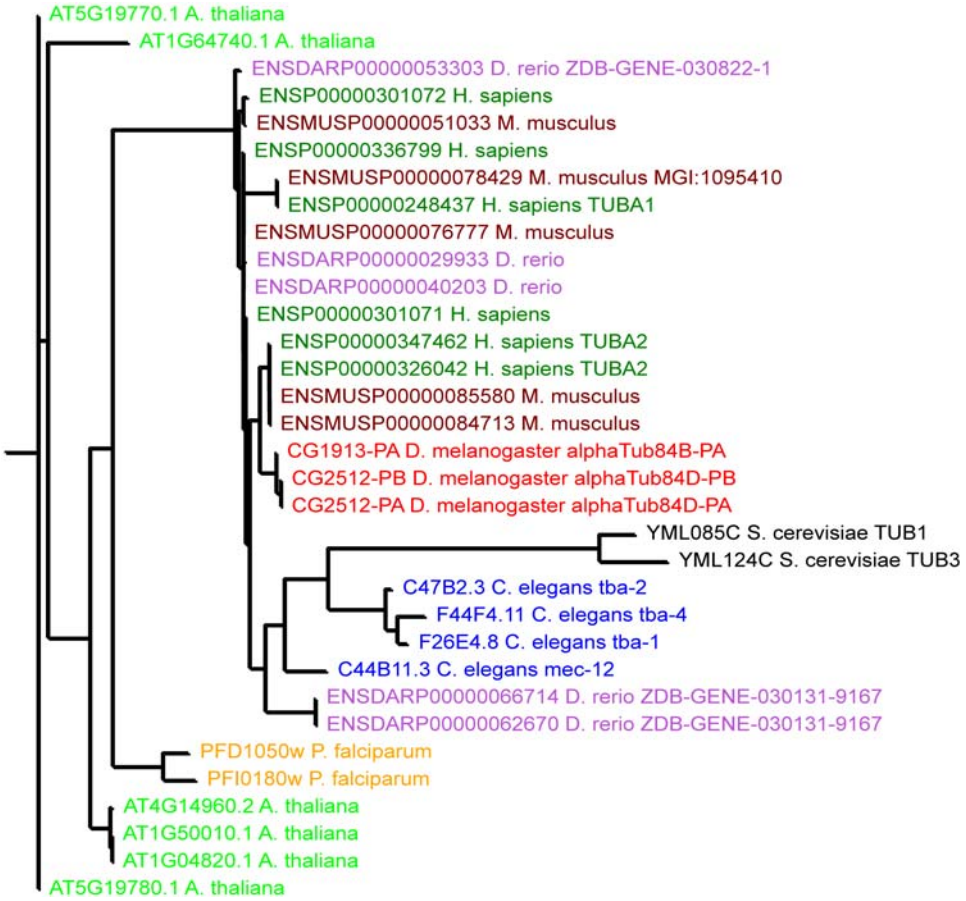

Figure 3. OrthoMCL family of the alpha tubulins. This OrthoMCL family contains only the alpha tubulins, while the tubulin family generated by the Jaccard family (too large to be shown here) contains the alpha, beta, and gamma tubulins.
doi:10.1371/journal.pone.0000766.g003

falciparum proteins has been a technical challenge in the development of malaria vaccines [18,19].

DISCUSSION

The database system (P-POD) we constructed shows users predicted orthologs of query proteins alone (using OrthoMCL) and in their broader evolutionary context (using Jaccard clustering). It consists of a comparative genomics analysis pipeline whose results are stored in a generic, modular database schema (GMOD/chado) using a freely available database system (PostgreSQL). P-POD is meant not to replace but rather to complement the currently available comparative genomics databases. To our knowledge, no other comparative genomics database provides experimental evidence of conservation curated from the primary literature.

We envision at least three sets of users of our database system. First, molecular biologists can query the database over the web to browse orthology data, both computational and experimental, for their favorite proteins. Another set of users consists of model organism database developers, who will quickly be able to provide comparative genomics tools with their species of interest by implementing our system. Finally, we expect that computational biologists who are developing novel comparative genomics algorithms will find the curated information and computational data from other methods extremely useful in assessing their approach. In addition, by using our system, they will save time in implementation and will be able to more readily distribute their algorithms.

It is important to emphasize that while computational methods to identify orthologs are extremely useful, they are by no means perfect. While OrthoMCL does reasonably well in creating putative orthologous groups, like all computational methods, in many cases it fails, either leaving out true orthologs or inappropriately including paralogs [7]. If one's main goal is to use such an algorithm solely to identify strict orthologs, then the selection of species is critical, and the inclusion of two mammals along with the distantly related *Plasmodium* certainly will increase the number of families that contain extraneous paralogs. Our goal, however, is to provide a database that can serve not only computational or evolutionary biologists but also the day-to-day needs of biologists who work on the common model organisms. P-POD provides a way for biologists to query directly for their gene of interest from their species of study, even though in some cases the phylogenetic trees must be manually examined to determine true orthologs because of the occasional inclusion of paralogs. As more refined methods for automatic detection of orthology are developed (for example, [20,21]) we plan to incorporate them

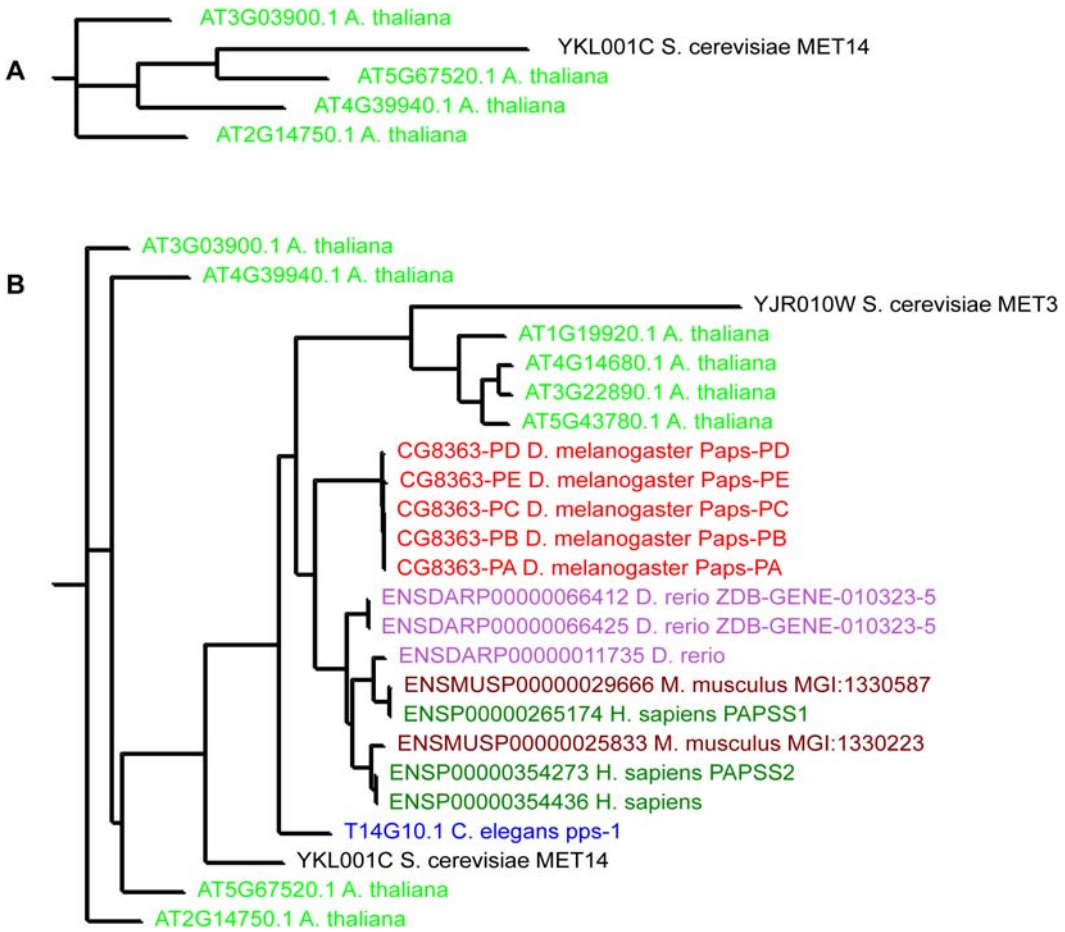

Figure 4. The *MET3/MET14* families. (A) *MET14* Jaccard family, and (B) *MET3/MET14* OrthoMCL family.
doi:10.1371/journal.pone.0000766.g004

into the P-POD tool, taking advantage of our modular design scheme.

We plan to provide regular updates to the data contained within the database. At the time of writing, we are running the analysis pipeline with the latest versions of the genomes. In addition, we will add new features to the web interface and will expand upon the amount of data stored within the database. We will also continue to provide curated literature describing experimental confirmation of orthology. All the data within the database are freely and publicly available through the web and by downloading the entire database system via the URL http://ortholog.princeton.edu/.

MATERIALS AND METHODS

The overall analysis pipeline is illustrated in Figure 1. The sources and versions of the pipeline components are listed in Table 2.

WU-BLAST

The same WU-BLAST results were used as input to both OrthoMCL and Jaccard algorithms described below. WU-BLAST (version 2.0MP-WashU) was run with the default BLASTP settings: matrix = BLOSUM62, Expectation Threshold = 10, ctxfactor = 1.0, no filtering.

OrthoMCL and Jaccard Algorithms

OrthoMCL (v. 1.2, 14-March-2005 [4]) compares the all-against-all BLASTP scores from a set of genomes, first identifying putative orthologs as reciprocal best hits between pairs of genomes, then identifying candidate recent paralogs as proteins within the same species that are more similar to each other than to any sequence in the other species. All orthologs and recent paralogs are then converted into a graph where the nodes represent the proteins and the edges represent their relationships. A normalization step is then used to correct for systematic biases when comparing pairs of genomes. Finally, the ortholog families are resolved by application of the Markov Cluster algorithm (MCL v. 1.005, 05-118). Since this procedure maximally includes in a family only those proteins at least as closely related as between-species reciprocal best hits, the resultant OrthoMCL group can be considered a set of putative orthologs in that every protein in the group is likely orthologous to at least one other group member. Some groups, however, consist solely of proteins from a single species; obviously, such groups only contain recent paralogs, but this information is often of great importance to experimental biologists.

We used the following OrthoMCL parameters. P-value cutoff: $1e^{-5}$, percent identity and percent match cutoffs: 0, maximum weight: 100.

OrthoMCL family size can be adjusted by changing the inflation index (1.5 in this study), but this does not loosen the

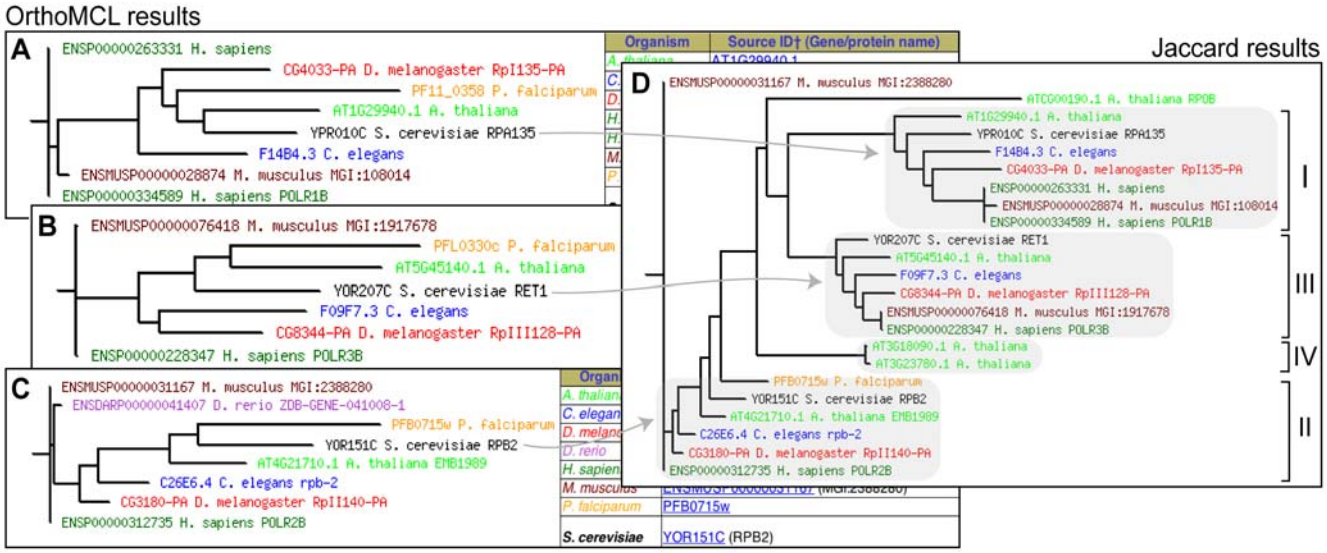

Figure 5. OrthoMCL and Jaccard clustering results for the second largest RNA polymerase subunit families of *S. cerevisiae.* The second largest subunits of RNA polymerase I, II, and III in yeast are named *RPA135, RPB2*, and *RET1*, respectively. (A) Phylogenetic tree display of OrthoMCL results showing individual yeast subunit *RPA135* and its predicted orthologs resolved into a distinct family. OrthoMCL results showing yeast RNA polymerase subunits *RET1* (B) and *RPB2* (C) resolved into separate families of orthologs. (D) Jaccard clustering results showing a "super family" of related RNA polymerase subfamilies. Arrows from each OrthoMCL family on the left point to the separate subfamilies in the Jaccard results. I to IV on the right of each tree indicates RNA polymerase subfamily. The second largest subunits for a fourth RNA polymerase, Pol IV, unique to plants were resolved into their own distinct two-member family by the OrthoMCL program (not shown), and were appropriately clustered with this superfamily by the Jaccard clustering method. (Adapted from figure 2 of [15])
doi:10.1371/journal.pone.0000766.g005

fundamental restriction that the algorithm begins with a list of putative orthologs and paralogs. To get larger families showing more distant relationships, we wanted to remove this restriction and include proteins that exhibit significant sequence similarity over a large portion of their lengths. We chose to perform Jaccard clustering and to apply a more broadly-defined set of criteria, namely that members of the same family should have significant BLAST scores over at least half of their length. This last point is important to reduce the chance of grouping two sequences together based on the presence of short promiscuous domains.

In the Jaccard clustering analysis, two proteins are grouped into the same family if they share a significant number of homologs, calculated as follows. First, a list of homologs for each sequence, consisting of those whose relative BLASTP scores are less than $1e^{-5}$ over a total of at least 50% of the length of each, is generated for each protein. Then the Jaccard index for each pair is calculated; this is the ratio of the magnitude of the intersection of their homolog sets vs. the union, or $|A \cap B| / |A \cup B|$. Final clusters are generated by linking proteins whose mutual Jaccard index is above a pre-determined cutoff. We evaluated the impact of varying the cutoff over a range of 0.3 to 0.8 for several well-characterized protein families, such as actins, tubulins, RNA polymerases, and several proteins containing RING finger or SH3 domains. We chose a Jaccard index of 0.4 since it most broadly permitted the inclusion of expected members of the families while excluding obvious non-members. For example, at a cutoff of 0.5, the family containing yeast actin (*ACT1*) inappropriately omitted the human and mouse actin-related proteins *ACTR8* and *Actr8*, while a cutoff of 0.3 was clearly too low and yielded many families with hundreds of extraneous members.

Generation of phylogenetic trees

P-POD generates phylogenetic trees of the OrthoMCL and Jaccard families using CLUSTAL W [5] and PHYLIP (Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle), using ProML with global rearrangements turned on. CLUSTAL W was run with the default settings: matrix = BLOSUM, Gapopen = 10, Gapext = 0.2, Gapdist = 8, Max div. = 40, ENDGAPS, NOPGAPS and NOHGAPS off, PWMATRIX = BLOSUM, PWGAPOPEN = 10, PWGAPEXT = 0.1, Distance = Kimura, TOSSGAPS = ON, Output = PHYLIP.

Literature

During literature curation at SGD for its "Literature Guide" resource, papers may be associated with yeast genes and various topics that describe what the paper addresses. A list of all papers associated with the topics "Cross-species expression" or "Disease-related" was downloaded from the SGD FTP site and loaded into the P-POD database, along with links to the yeast genes as made by the SGD curators. These papers are displayed on the P-POD interface whenever a family that contains the relevant yeast genes is viewed; each paper displayed is hyperlinked to the PubMed database. For the papers associated with the "Cross-species expression" topic, we manually read each paper to extract which gene(s) from which organism(s) were tested, and whether functional complementation was demonstrated. These results are stored in the database and displayed on the P-POD interface.

Database schema and software

P-POD uses the Generic Model Organism Database (GMOD) database package using PostgreSQL software. Information and documentation about the GMOD schema (also known as the "chado" schema) can be found on the GMOD web site (www.gmod.org). In addition, Supplemental Table 1 (http://ortholog.

Table 6. Conservation of yeast proteins involved in N-linked glycosylation.

Function	Yeast gene	Human gene	CDG (OMIM)	*At*	*Ce*	*Dm*	*Dr*	*Mm*	*Pf*
Dolichol synthesis and modification	*RER2*	DHDDS		x		x	x	x	
	SEC59	TMEM15		x	x	x	x		x
	DPM1	DPM1	Ie (608799)	x	x	x	x	x	x
	ALG5	ALG5		x	x	x	x	x	
	CAX4	DOLPP1		x			x	x	x
Assembly of core oligo-saccharides	*ALG7*	DPAGT1	Ij (608093)	x	x	x	x	x	x
	ALG13	GLT28D1		x	x	x	x	x	x
	ALG14	unnamed		x	x	x	x	x	x
	ALG1	ALG1	Ik (608540)	x	x	x	x	x	
	ALG2	ALG2	Ii (607906)	x	x	x		x	
	ALG11	unnamed		x	x	x	x	x	
	RFT1	RFT1		x	x	x		x	
	ALG3	ALG3	Id (601110)	x	x	x		x	
	ALG9	ALG9	Il (608776)	x	x	x	x	x	
	ALG12	ALG12	Ig (607143)		x	x	x	x	
	ALG6	ALG6	Ic (603147)	x	x	x	x		
	ALG8	ALG8	Ih (608104)	x	x	x	x	x	
	DIE2/ALG10	ALG10/KCR1		x	x	x			
Oligo-saccharyl-transferase complex	*OST1*	RPN1		x	x	x		x	x
	OST2	DAD1		x	x	x	x	x	
	OST3	TUSC3					x	x	
	STT3	ITM1		x	x	x	x	x	x
	WBP1	DDOST		x	x	x	x	x	x
Trimming of outer saccharides	*CWH41/GLS1*	GCS1	IIb (606056)	x	x	x	x	x	
	ROT2/GLS2	GANAB		x	x	x	x	x	
	MNS1	MAN1B1		x	x	x	x	x	

Genes are broadly categorized by function. Human genes are identified by name when possible and the corresponding congenital disorders of glycosylation (CDG, with OMIM ID) are shown. For *A. thaliana*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *P. falciparum*, boxes marked with "x" indicate that a peptide from this organism was placed in the same OrthoMCL family with the yeast gene. Not shown: *SWP1* is homologous to human ribophorin II [30], and *SWP1*, *OST4*, *OST5*, and *OST6* do not lie in ortholog families.

doi:10.1371/journal.pone.0000766.t006

princeton.edu/help.html#schema) provides details about our particular implementation of the GMOD schema, including how data from our analysis (FASTA files, OrthoMCL results, etc.) are mapped to the GMOD database tables.

ACKNOWLEDGMENTS

We acknowledge John Wiggins and Mark Schroeder for excellent technical support and Mike Cherry (SGD), Shuai Weng (SGD), Eurie Hong (SGD), Laurie Kramer (Princeton) and John Matese (Princeton) for valuable discussions.

Author Contributions

Conceived and designed the experiments: DB KD SH CL. Performed the experiments: SH CL. Analyzed the data: KD RO SH ML. Contributed reagents/materials/analysis tools: OW SA FK. Wrote the paper: DB KD RO ML.

REFERENCES

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. Nucleic Acids Res 35: D21–25.
2. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, et al. (2007) EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res 35: D16–20.
3. Alexeyenko A, Lindberg J, Perez-Bercoff A, Sonnhammer ELL (2006) Overview and comparison of ortholog databases. Drug Discov Today 11: 137–143.
4. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.
5. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
6. Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, et al. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. Nucleic Acids Res 34: D599–601.

7. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2: e383.
8. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, et al. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res 30: 69–72.
9. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.
10. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338.
11. Kron SJ, Drubin DG, Botstein D, Spudich JA (1992) Yeast actin filaments display ATP-dependent sliding movement over surfaces coated with rabbit muscle myosin. Proc Natl Acad Sci U S A 89: 4466–4470.
12. Keeling PJ, Doolittle WF (1996) Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol 13: 1297–1305.
13. Dutcher SK (2003) Long-lost relatives reappear: identification of new members of the tubulin superfamily. Curr Opin Microbiol 6: 634–640.
14. Archambault J, Friesen JD (1993) Genetics of eukaryotic RNA polymerases I, II, and III. Microbiol Rev 57: 703–724.
15. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC (2005) RNA polymerase IV directs silencing of endogenous DNA. Science 308: 118–120.
16. Onodera Y, Haag JR, Ream T, Nunes PC, Pontes O, et al. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell 120: 613–622.
17. Gowda DC, Gupta P, Davidson EA (1997) Glycosylphosphatidylinositol anchors represent the major carbohydrate modification in proteins of intraerythrocytic stage Plasmodium falciparum. J Biol Chem 272: 6428–6439.
18. Kedees MH, Azzouz N, Gerold P, Shams-Eldin H, Iqbal J, et al. (2002) Plasmodium falciparum: glycosylation status of Plasmodium falciparum circumsporozoite protein expressed in the baculovirus system. Exp Parasitol 101: 64–68.
19. Kocken CH, Withers-Martinez C, Dubbeld MA, van der Wel A, Hackett F, et al. (2002) High-level expression of the malaria blood-stage vaccine candidate Plasmodium falciparum apical membrane antigen 1 and induction of antibodies that inhibit erythrocyte invasion. Infect Immun 70: 4471–4476.
20. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22: e9–15.
21. Jothi R, Zotenko E, Tasneem A, Przytycka TM (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. Bioinformatics 22: 779–788.
22. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
23. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, et al. (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res 12: 493–502.
24. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2006) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 34: D173–180.
25. O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476–480.
26. O'Brien KP, Westerlund I, Sonnhammer EL (2004) OrthoDisease: a database of human disease orthologs. Hum Mutat 24: 112–119.
27. Chen F, Mackey AJ, Stoeckert CJ, Jr, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34: D363–368.
28. Penkett CJ, Morris JA, Wood V, Bahler J (2006) YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. Nucleic Acids Res 34: W330–334.
29. Samuel Lattimore B, van Dongen S, Crabbe MJ (2005) GeneMCL in microarray analysis. Comput Biol Chem 29: 354–359.
30. Kelleher DJ, Gilmore R (1994) The Saccharomyces cerevisiae oligosaccharyltransferase is a protein complex composed of Wbp1p, Swp1p, and four additional polypeptides. J Biol Chem 269: 12908–12917.
31. Nomoto S, Watanabe Y, Ninomiya-Tsuji J, Yang LX, Nagai Y, et al. (1997) Functional analyses of mammalian protein kinase C isozymes in budding yeast and mammalian fibroblasts. Genes Cells 2: 601–614.
32. Kida Y, Uchida S, Miyazaki H, Sasaki S, Marumo F (2001) Localization of mouse CLC-6 and CLC-7 mRNA and their functional complementation of yeast CLC gene mutant. Histochem Cell Biol 115: 189–194.
33. Yamagata K, Kato J, Shimamoto A, Goto M, Furuichi Y, et al. (1998) Bloom's and Werner's syndrome genes suppress hyperrecombination in yeast sgs1 mutant: implication for genomic instability in human diseases. Proc Natl Acad Sci U S A 95: 8733–8738.
34. Clark AB, Cook ME, Tran HT, Gordenin DA, Resnick MA, et al. (1999) Functional analysis of human MutSalpha and MutSbeta complexes in yeast. Nucleic Acids Res 27: 736–742.
35. Garbers C, DeLong A, Deruere J, Bernasconi P, Soll D (1996) A mutation in protein phosphatase 2A regulatory subunit A affects auxin transport in Arabidopsis. Embo J 15: 2115–2124.
36. Frank CG, Grubenmann CE, Eyaid W, Berger EG, Aebi M, et al. (2004) Identification and functional analysis of a defect in the human ALG9 gene: definition of congenital disorder of glycosylation type IL. Am J Hum Genet 75: 146–150.
37. Schwarz M, Thiel C, Lubbehusen J, Dorland B, de Koning T, et al. (2004) Deficiency of GDP-Man:GlcNAc2-PP-dolichol mannosyltransferase causes congenital disorder of glycosylation type Ik. Am J Hum Genet 74: 472–481.
38. Ballester R, Marchuk D, Boguski M, Saulino A, Letcher R, et al. (1990) The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. Cell 63: 851–859.
39. Poullet P, Lin B, Esson K, Tamanoi F (1994) Functional significance of lysine 1423 of neurofibromin and characterization of a second site suppressor which rescues mutations at this residue and suppresses RAS2Val-19-activated phenotypes. Mol Cell Biol 14: 815–821.
40. Gecz J, Shaw MA, Bellon JR, de Barros Lopes M (2003) Human wild-type SEDL protein functionally complements yeast Trs20p but some naturally occurring SEDL mutants do not. Gene 320: 137–144.
41. Gao XD, Wang J, Keppler-Ross S, Dean N (2005) ERS1 encodes a functional homologue of the human lysosomal cystine transporter. Febs J 272: 2497–2511.
42. Cavadini P, Gellera C, Patel PI, Isaya G (2000) Human frataxin maintains mitochondrial iron homeostasis in Saccharomyces cerevisiae. Hum Mol Genet 9: 2523–2530.
43. Desmyter L, Dewaele S, Reekmans R, Nystrom T, Contreras R, et al. (2004) Expression of the human ferritin light chain in a frataxin mutant yeast affects ageing and cell death. Exp Gerontol 39: 707–715.
44. Feiler HS, Desprez T, Santoni V, Kronenberger J, Caboche M, et al. (1995) The higher plant Arabidopsis thaliana encodes a functional CDC48 homologue which is highly expressed in dividing and expanding cells. Embo J 14: 5626–5637.
45. Hsi G, Cullen LM, Moira Glerum D, Cox DW (2004) Functional assessment of the carboxy-terminus of the Wilson disease copper-transporting ATPase, ATP7B. Genomics 83: 473–481.
46. Bussey H, Storms RK, Ahmed A, Albermann K, Allen E, et al. (1997) The nucleotide sequence of Saccharomyces cerevisiae chromosome XVI. Nature 387: 103–105.
47. Portmann R, Solioz M (2005) Purification and functional reconstitution of the human Wilson copper ATPase, ATP7B. FEBS Lett 579: 3589–3595.
48. Sambongi Y, Wakabayashi T, Yoshimizu T, Omote H, Oka T, et al. (1997) Caenorhabditis elegans cDNA for a Menkes/Wilson disease gene homologue and its function in a yeast CCC2 gene deletion mutant. J Biochem (Tokyo) 121: 1169–1175.
49. Mercer JF, Barnes N, Stevenson J, Strausak D, Llanos RM (2003) Copper-induced trafficking of the cU-ATPases: a key mechanism for copper homeostasis. Biometals 16: 175–184.
50. Payne AS, Gitlin JD (1998) Functional expression of the menkes disease protein reveals common biochemical mechanisms among the copper-transporting P-type ATPases. J Biol Chem 273: 3765–3770.
51. Jantti J, Lahdenranta J, Olkkonen VM, Soderlund H, Keranen S (1999) SEM1, a homologue of the split hand/split foot malformation candidate gene Dss1, regulates exocytosis and pseudohyphal differentiation in yeast. Proc Natl Acad Sci U S A 96: 909–914.
52. Sone T, Saeki Y, Toh-e A, Yokosawa H (2004) Sem1p is a novel subunit of the 26 S proteasome from Saccharomyces cerevisiae. J Biol Chem 279: 28807–28816.
53. Morita T, Yoshimura Y, Yamamoto A, Murata K, Mori M, et al. (1993) A mouse homolog of the Escherichia coli recA and Saccharomyces cerevisiae RAD51 genes. Proc Natl Acad Sci U S A 90: 6577–6580.
54. Loewen CJ, Levine TP (2005) A highly conserved binding site in vesicle-associated membrane protein-associated protein (VAP) for the FFAT motif of lipid-binding proteins. J Biol Chem 280: 14097–14104.
55. Guzder SN, Sung P, Prakash S, Prakash L (1995) Lethality in yeast of trichothiodystrophy (TTD) mutations in the human xeroderma pigmentosum group D gene. Implications for transcriptional defect in TTD. J Biol Chem 270: 17660–17663.
56. Sung P, Bailly V, Weber C, Thompson LH, Prakash L, et al. (1993) Human xeroderma pigmentosum group D gene encodes a DNA helicase. Nature 365: 852–855.
57. Lanterman MM, Dickinson JR, Danner DJ (1996) Functional analysis in Saccharomyces cerevisiae of naturally occurring amino acid substitutions in human dihydrolipoamide dehydrogenase. Hum Mol Genet 5: 1643–1648.
58. McEwen RK, Dove SK, Cooke FT, Painter GF, Holmes AB, et al. (1999) Complementation analysis in PtdInsP kinase-deficient yeast mutants demonstrates that Schizosaccharomyces pombe and murine Fab1p homologues are phosphatidylinositol 3-phosphate 5-kinases. J Biol Chem 274: 33905–33912.
59. Mayordomo I, Sanz P (2001) Human pancreatic glucokinase (GlkB) complements the glucose signalling defect of Saccharomyces cerevisiae hxk2 mutants. Yeast 18: 1309–1316.
60. Lucas ME, Ma Q, Cunningham D, Peters J, Cattanach B, et al. (2003) Identification of two novel mutations in the murine Nsdhl sterol dehydrogenase gene and development of a functional complementation assay in yeast. Mol Genet Metab 80: 227–233.

61. Geisler M, Frangne N, Gomes E, Martinoia E, Palmgren MG (2000) The ACA4 gene of Arabidopsis encodes a vacuolar membrane calcium pump that improves salt tolerance in yeast. Plant Physiol 124: 1814–1827.
62. Schiott M, Romanowsky SM, Baekgaard L, Jakobsen MK, Palmgren MG, et al. (2004) A plant plasma membrane Ca2+ pump is required for normal pollen tube growth and fertilization. Proc Natl Acad Sci U S A 101: 9502–9507.
63. Kleinow T, Bhalerao R, Breuer F, Umeda M, Salchert K, et al. (2000) Functional identification of an Arabidopsis snf4 ortholog by screening for heterologous multicopy suppressors of snf4 deficiency in yeast. Plant J 23: 115–122.
64. Lumbreras V, Alba MM, Kleinow T, Koncz C, Pages M (2001) Domain fusion between SNF1-related kinase subunits during plant evolution. EMBO Rep 2: 55–60.
65. Roje S, Wang H, McNeil SD, Raymond RK, Appling DR, et al. (1999) Isolation, characterization, and functional expression of cDNAs encoding NADH-dependent methylenetetrahydrofolate reductase from higher plants. J Biol Chem 274: 36089–36096.
66. Raymond RK, Kastanos EK, Appling DR (1999) Saccharomyces cerevisiae expresses two genes encoding isozymes of methylenetetrahydrofolate reductase. Arch Biochem Biophys 372: 300–308.
67. Ton VK, Mandal D, Vahadji C, Rao R (2002) Functional expression in yeast of the human secretory pathway Ca(2+), Mn(2+)-ATPase defective in Hailey-Hailey disease. J Biol Chem 277: 6422–6427.
68. Ton VK, Rao R (2004) Expression of Hailey-Hailey disease mutations in yeast. J Invest Dermatol 123: 1192–1194.
69. Kellermayer R (2005) Hailey-Hailey disease as an orthodisease of PMR1 deficiency in Saccharomyces cerevisiae. FEBS Lett 579: 2021–2025.
70. Heinisch JJ (1993) Expression of heterologous phosphofructokinase genes in yeast. FEBS Lett 328: 35–40.
71. Raben N, Exelbert R, Spiegel R, Sherman JB, Nakajima H, et al. (1995) Functional expression of human mutant phosphofructokinase in yeast: genetic defects in French Canadian and Swiss patients with phosphofructokinase deficiency. Am J Hum Genet 56: 131–141.
72. Garavaglia B, Invernizzi F, Carbone ML, Viscardi V, Saracino F, et al. (2004) GTP-cyclohydrolase I gene mutations in patients with autosomal dominant and recessive GTP-CH1 deficiency: identification and functional characterization of four novel mutations. J Inherit Metab Dis 27: 455–463.
73. Mancini R, Saracino F, Buscemi G, Fischer M, Schramek N, et al. (1999) Complementation of the fol2 deletion in Saccharomyces cerevisiae by human and Escherichia coli genes encoding GTP cyclohydrolase I. Biochem Biophys Res Commun 255: 521–527.
74. Geraghty MT, Vaughn D, Nicholson AJ, Lin WW, Jimenez-Sanchez G, et al. (1998) Mutations in the Delta1-pyrroline 5-carboxylate dehydrogenase gene cause type II hyperprolinemia. Hum Mol Genet 7: 1411–1415.
75. Hu CA, Lin WW, Valle D (1996) Cloning, characterization, and expression of cDNAs encoding human delta 1-pyrroline-5-carboxylate dehydrogenase. J Biol Chem 271: 9795–9800.
76. Morgante PG, Berra CM, Nakabashi M, Costa RM, Menck CF, et al. (2005) Functional XPB/RAD25 redundancy in Arabidopsis genome: characterization of AtXPB2 and expression analysis. Gene 344: 93–103.
77. Pearce DA, Sherman F (1998) A yeast model for the study of Batten disease. Proc Natl Acad Sci U S A 95: 6915–6918.
78. Saunders RD, McLellan LI (2000) Molecular cloning of Drosophila gamma-glutamylcysteine synthetase by functional complementation of a yeast mutant. FEBS Lett 467: 337–340.
79. Srinivasan C, Liba A, Imlay JA, Valentine JS, Gralla EB (2000) Yeast lacking superoxide dismutase(s) show elevated levels of "free iron" as measured by whole cell electron paramagnetic resonance. J Biol Chem 275: 29187–29192.
80. Agarwal AK, Fryns JP, Auchus RJ, Garg A (2003) Zinc metalloproteinase, ZMPSTE24, is mutated in mandibuloacral dysplasia. Hum Mol Genet 12: 1995–2001.
81. Schmidt WK, Tam A, Michaelis S (2000) Reconstitution of the Ste24p-dependent N-terminal proteolytic step in yeast a-factor biogenesis. J Biol Chem 275: 6227–6233.
82. Hofmann S, Rothbauer U, Muhlenbein N, Neupert W, Gerbitz KD, et al. (2002) The C66W mutation in the deafness dystonia peptide 1 (DDP1) affects the formation of functional DDP1.TIM13 complexes in the mitochondrial intermembrane space. J Biol Chem 277: 23287–23293.
83. Rothbauer U, Hofmann S, Muhlenbein N, Paschen SA, Gerbitz KD, et al. (2001) Role of the deafness dystonia peptide 1 (DDP1) in import of human Tim23 into the inner membrane of mitochondria. J Biol Chem 276: 37327–37334.
84. Raymond M, Gros P, Whiteway M, Thomas DY (1992) Functional complementation of yeast ste6 by a mammalian multidrug resistance mdr gene. Science 256: 232–234.
85. Boyum R, Guidotti G (1997) Effect of ATP binding cassette/multidrug resistance proteins on ATP efflux of Saccharomyces cerevisiae. Biochem Biophys Res Commun 230: 22–26.
86. Chen Y, Beck A, Davenport C, Chen Y, Shattuck D, et al. (2005) Characterization of TRZ1, a yeast homolog of the human candidate prostate cancer susceptibility gene ELAC2 encoding tRNase Z. BMC Mol Biol 6: 12.
87. Peng Z, Lu Q, Verma DP (1996) Reciprocal regulation of delta 1-pyrroline-5-carboxylate synthetase and proline dehydrogenase genes controls proline levels during and after osmotic stress in plants. Mol Gen Genet 253: 334–341.
88. Chatterjee A, Singh KK (2001) Uracil-DNA glycosylase-deficient yeast exhibit a mitochondrial mutator phenotype. Nucleic Acids Res 29: 4935–4940.
89. Lillard-Wetherell K, Combs KA, Groden J (2005) BLM helicase complements disrupted type II telomere lengthening in telomerase-negative sgs1 yeast. Cancer Res 65: 5520–5522.
90. Neff NF, Ellis NA, Ye TZ, Noonan J, Huang K, et al. (1999) The DNA helicase activity of BLM is necessary for the correction of the genomic instability of bloom syndrome cells. Mol Biol Cell 10: 665–676.
91. Grubenmann CE, Frank CG, Kjaergaard S, Berger EG, Aebi M, et al. (2002) ALG12 mannosyltransferase defect in congenital disorder of glycosylation type lg. Hum Mol Genet 11: 2331–2339.
92. Forsgren M, Attersand A, Lake S, Grunler J, Swiezewska E, et al. (2004) Isolation and functional expression of human COQ2, a gene encoding a polyprenyl transferase involved in the synthesis of CoQ. Biochem J 382: 519–526.
93. Okada K, Ohara K, Yazaki K, Nozaki K, Uchida N, et al. (2004) The AtPPT1 gene encoding 4-hydroxybenzoate polyprenyl diphosphate transferase in ubiquinone biosynthesis is required for embryo development in Arabidopsis thaliana. Plant Mol Biol 55: 567–577.
94. Willingham S, Outeiro TF, DeVit MJ, Lindquist SL, Muchowski PJ (2003) Yeast genes that enhance the toxicity of a mutant huntingtin fragment or alpha-synuclein. Science 302: 1769–1772.
95. Xu GF, Lin B, Tanaka K, Dunn D, Wood D, et al. (1990) The catalytic domain of the neurofibromatosis type 1 gene product stimulates ras GTPase and complements ira mutants of S. cerevisiae. Cell 63: 835–841.
96. Mamiya N, Worman HJ (1999) Hepatitis C virus core protein binds to a DEAD box RNA helicase. J Biol Chem 274: 15751–15756.
97. Johnstone O, Deuring R, Bock R, Linder P, Fuller MT, et al. (2005) Belle is a Drosophila DEAD-box protein required for viability and in the germ line. Dev Biol 277: 92–101.
98. Vonarx EJ, Howlett NG, Schiestl RH, Kunz BA (2002) Detection of Arabidopsis thaliana AtRAD1 cDNA variants and assessment of function by expression in a yeast rad1 mutant. Gene 296: 1–9.
99. Shaag A, Walsh T, Renbaum P, Kirchhoff T, Nafa K, et al. (2005) Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population. Hum Mol Genet 14: 555–563.
100. Takeuchi M, Tada M, Saito C, Yashiroda H, Nakano A (1998) Isolation of a tobacco cDNA encoding Sar1 GTPase and analysis of its dominant mutations in vesicular traffic using a yeast complementation system. Plant Cell Physiol 39: 590–599.
101. Tomita S, Inoue N, Maeda Y, Ohishi K, Takeda J, et al. (1998) A homologue of Saccharomyces cerevisiae Dpm1p is not sufficient for synthesis of dolichol-phosphate-mannose in mammalian cells. J Biol Chem 273: 9249–9254.
102. Lai K, Elsas LJ (2000) Overexpression of human UDP-glucose pyrophosphorylase rescues galactose-1-phosphate uridyltransferase-deficient yeast. Biochem Biophys Res Commun 271: 392–400.
103. Brzeski J, Podstolski W, Olczak K, Jerzmanowski A (1999) Identification and analysis of the Arabidopsis thaliana BSH gene, a member of the SNF5 gene family. Nucleic Acids Res 27: 2393–2399.
104. Song WY, Martinoia E, Lee J, Kim D, Kim DY, et al. (2004) A novel family of cys-rich membrane proteins mediates cadmium resistance in Arabidopsis. Plant Physiol 135: 1027–1039.
105. Tommasini R, Vogt E, Fromenteau M, Hortensteiner S, Matile P, et al. (1998) An ABC-transporter of Arabidopsis thaliana has both glutathione-conjugate and chlorophyll catabolite transport activity. Plant J 13: 773–780.
106. Liang F, Cunningham KW, Harper JF, Sze H (1997) ECA1 complements yeast mutants defective in Ca2+ pumps and encodes an endoplasmic reticulum-type Ca2+-ATPase in Arabidopsis thaliana. Proc Natl Acad Sci U S A 94: 8579–8584.
107. Covic L, Lew RR (1996) Arabidopsis thaliana cDNA isolated by functional complementation shows homology to serine/threonine protein kinases. Biochim Biophys Acta 1305: 125–129.
108. Schmidt PJ, Ramos-Gomez M, Culotta VC (1999) A gain of superoxide dismutase (SOD) activity obtained with CCS, the copper metallochaperone for SOD1. J Biol Chem 274: 36952–36956.
109. Kataoka T, Powers S, Cameron S, Fasano O, Goldfarb M, et al. (1985) Functional homology of mammalian and yeast RAS genes. Cell 40: 19–26.
110. Catoni E, Desimone M, Hilpert M, Wipf D, Kunze R, et al. (2003) Expression pattern of a nuclear encoded mitochondrial arginine-ornithine translocator gene from Arabidopsis. BMC Plant Biol 3: 1.
111. Krumpelman PM, Freyermuth SK, Cannon JF, Fink GR, Polacco JC (1995) Nucleotide sequence of Arabidopsis thaliana arginase expressed in yeast. Plant Physiol 107: 1479–1480.
112. van Wilpe S, Boumans H, Lobo-Hajdu G, Grivell LA, Berden JA (1999) Functional complementation analysis of yeast bc1 mutants. A study of the mitochondrial import of heterologous and hybrid proteins. Eur J Biochem 264: 825–832.
113. Schaffar G, Breuer P, Boteva R, Behrends C, Tzvetkov N, et al. (2004) Cellular toxicity of polyglutamine expansion proteins: mechanism of transcription factor deactivation Functional complementation analysis of yeast bc1 mutants. A

study of the mitochondrial import of heterologous and hybrid proteins. Mol Cell 15: 95–105.
114. Wagner N, Weber D, Seitz S, Krohne G (2004) The lamin B receptor of Drosophila melanogaster. J Cell Sci 117: 2015–2028.
115. Colussi PA, Taron CH, Mack JC, Orlean P (1997) Human and Saccharomyces cerevisiae dolichol phosphate mannose synthases represent two classes of the enzyme, but both function in Schizosaccharomyces pombe. Proc Natl Acad Sci U S A 94: 7873–7878.