**Protein domains of unknown function are essential in yeast**

Norman Goodacre1 & Peter Uetz2

1 FDA, xxx

2 Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA

1 corresponding author: ngoodacre@gmail.com

**Abstract**

**Introduction**

Yeast, both *Saccharomyces cerevisiae* (baker’s yeast)and *Schizosaccharomyces pombe* (fission yeast), are among the most valuable model organisms for biological research. *S. cerevisiae* is arguably the best-understood of all organisms but certainly among eukaryotes, having played a figurative role in the development of various high-throughput technologies, including microarrays (Lashkari et al., 1997; Eisen et al., 1998; Spellman et al., 1998), genotyping (Giaever et al 2002) and proteomics (Uetz et al., 2000; Ito et al., 2001; Ho et al., 2002; Krogan et al., 2006, Gavin et al. 2006). The findings from such experiments have generally been made available to the public, and are collected in the Saccharomyces Genome Database (SGD) (Cherry et al., 2011). Nearly 1,000 *S. cerevisiae* genes are members of orthologous gene families linked to disease in humans (Heinicke et al., 2007) approximately half of which are mitochondrial (Steinmetz et al., 2002). Both species of yeast have been leveraged to explore essentiality of eukaryotic genes, using genome-wide knockouts (Kim et al., 2010) and genetic interaction experiments (Costanzo et al., 2010; Baryshnikova et al., 2010, Costanzo et al., 2016), which should help shed light on the role of these gene families in disease. At least 85% of *S. cerevisiae* genes have been experimentally characterized, more than for *S. pombe* or any other eukaryote. Given that complete genome essentiality screens have been performed for *S. cerevisiae (REFs)*, essentiality information exists even for the 15% of genes that remain uncharacterized, some encoding domains of unknown function (DUFs).

While *S. cerevisiae* is useful as a well-characterized model organism, *S. pombe* is attractive to the molecular biology community as a result of its greater similarity to human cells in many aspects. While *S. cerevisiae* has undergone a whole genome duplication event (REF Ashbya) there seems to be no evidence for such an event in *S. pombe* (Wolfe and Shields, 1997, Wixon 2002). *S. pombe* also has more in common with higher eukaryotes, such as cell-cycle control elements (G2/M), contractile ring in cytokinesis, repeating centromeres, more complex heterochromatin, more complex splicing regulation, and RNA interference ([Sabatinos and Forsburg, 2010](http://www.sciencedirect.com/science/article/pii/S0092867412005739" \l "bib46); [Rhind et al., 2011](http://www.sciencedirect.com/science/article/pii/S0092867412005739#bib42)). These cell cycle processes likely arose after selective pressure to maintain genome stability following the ancestral duplication.

The significance of differences in essential genes between *S. cerevisiae* and *S. pombe* has not been studied in detail. A greater proportion of genes appear to be essential in *S. pombe* (Database of Essential Genes or DEG, Luo et al 2013). Analysis of co-fitness profiles from Costanzo et al. (2010) revealed widespread differences between gene pairs in *S. pombe* and orthologous pairs in *S. cerevisiae*, which may provide evidence of functional repurposing, or the co-opting of a gene’s function for a different biological process in one organism compared to another (Frost et al., 2012). Since domains are the most fundamental unit of protein function, an obvious way to study these differences in gene essentiality is to compare the domain architecture of essential proteins between both species of yeast.

The present study has three goals: (1) to discover essential yeast domains of unknown function (yeDUFs), (2) to evaluate the differences in yeDUFs and their apparent characteristics between *S. cerevisiae* and *S. pombe,* and (3) to prioritize these proteins and domains for experimental analysis. As shown in earlier work (Goodacre et al., 2013), bacterial eDUFs (beDUFs) can be found in a number of common model bacterial organisms, including pathogens. Although the majority of beDUFs were not found to be highly conserved, some beDUFs were found in almost every known bacterial family and even in multiple kingdoms. Here we show that some yeDUFs are well-conserved, and that yeDUFs specific to *S.cerevisiae* or *S. pombe* contain clues about the evolution of molecular functions specific to eukaryotes, especially as regards the cell cycle. Focusing on yeDUFs may also provide more valuable clues to the synthetic biology community about which component of genes can be removed in *S. cerevisiae* or *S. pombe*.

**Methods**

**Data sources**

Both single-gene knockout (SGKO) and genetic interaction (GI) data were used as gene essentiality information for the inference of yeDUFs in the present study. These two types of essentiality information were used for distinct purposes: SGKO was used as the basis for inference of yeDUFs, while GI was used to explore possible conditional essentiality of inferred yeDUFs. SGK essentiality information was obtained from the Database of Essential Genes (DEG) version 13.3 (Luo et al., 2014). DEG contains both essentiality and non-essentiality information for *S. pombe*, and essentiality information for *S. cerevisiae*. DEG genes were mapped to Uniprot accessions using the Uniprot ID mapping tool (UniProt Consortium, 2014) with gene identifiers (GI numbers) as input or gene names where GI numbers were missing. GI essentiality information is described further in the section “YeDUFs in Hillenmeyer, 2008 conditionally-essential genes”, below.

Protein domains of unknown function were defined as in (Goodacre et al., 2014), as any domain in the Pfam database (Punta et al., 2011) with “DUFxxxx” as a name or “unknown function” or “uncharacterized/uncharacterized” + “protein/domain” in the description. Domain content of yeast proteins was obtained from Uniprot annotation.

**Inference of eDUFs**

Essential DUFs were inferred using both an adaptation of the rule-based method from (Goodacre et al., 2014) and the maximum likelihood (ML)-based statistical method from (Lu et al., 2015). For clarification of these methodologies, a flowchart and example is shown in **Figure 1**. The union of both sets of inferred eDUFs for both species of yeast (**Figure 2**) was used for subsequent analyses.

*Domains*

For all inferences algorithms, the set of length *L* domains from protein products of all genes, whether from essential genes / interactions or not, was defined as *K* = *{k1*, *k2*… *kL*}. Because domains are generally present in multiple proteins, it was necessary to define them both independently (*K*) and in their protein context.

*Rule-based inference (Goodacre et al., 2014)*

The set of 6,740 proteins for *S. cerevisiae*, Sc\_P P\_Sc, was sub-divided into essential and non-essential components; likewise, the set of 5,147 proteins for *S. pombe*, Sp\_P. The set of 1,129 essential proteins for *S. cerevisiae* was defined as *Sc\_Pe* = {sc\_*pe1, sc\_pe2* … sc\_*pe1129*}; likewise the set of 1,234essential genes for *S. pombe* was defined as *Sp\_Pe* = {sp\_*pe1, sp\_pe2* … sp\_*pe1234*}. Non-essential proteins for *S. cerevisiae* and *S. pombe* were taken to be all except the essential ones. Although explicit non-essentiality information existed in DEG13.3 for 1,937 *S. pombe* proteins, in order to make inferences comparable between the two strains the same methodology was applied to both. The set of 5,611 non-essential proteins for *S. cerevisiae* was defined as *Sc\_Pne* = {sc\_*pne1, sc\_pne2* … sc\_*pne5611*}; likewise the set of 3,913 non-essential proteins for *S. pombe* was defined as *Sp\_Pne* = {sp\_*pne1, sp\_pne2* … sp\_*pne3913*}. For each protein *i* the set of length *j* of domains defined as *Di = {di1, di2 … dij}.*

Essential domains in yeast were inferred using three cases (rule-based inference, **Fig. 1**):

Case 1 (single-domain essential protein) – domain *k* is an essential domain if there exists ≥ 1 *pei*with a domain set *Di* = {*k*}.

Case 2 (addition of domain to non-essential protein domain architecture makes protein essential) – domain *k* is an essential domain if there exists ≥ 1 essential gene *pei* with domain set *Di* and ≥ 1 non-essential protein *pnej* from the same species with domain set *Dj,* such that *Di  \ Dj= {k}*.

Case 3 (domain exclusively and uniquely present in essential proteins) - domain *k* is an essential domain if there exists ≥ 1 essential protein *pei* with domain set *Da*containing *k*. and 0 non-essential genes with domain sets containing *k*. Further, for each essential protein *pei* containing *k*, none of the other domains in domain set *Da* meet the prior criteria.

*Estimation-maximization-based inference (Lu et al., 2015)*

Lu et al. (2015) use genes in their formal definition, however, here we substitute the protein products. With this substitution, the Lu et al. (2015) methodology defines the set of *n* proteins in an organism as *P* = {*p1, p2* … *pn*}, where each value *pi*, *i*  (1,*n*) is either 1 (essential) or 0 (non-essential). For each protein, the set of domains of length *Li*, *D(pi)= {Di1, Di2 … DLi}* was defined similarly (1 if essential, 0 if non-essential). Thus, essentiality was represented for the set of proteins and for each proteins’s set of domains, as a vector of binary values.

Estimation maximization is a statistical technique used to find the maximum likelihood parameters for a probabilistic model, when standard maximum likelihood estimation techniques cannot be used. This is the case when the optimization equations, which are derivatives of the probability function, cannot be solved directly, which generally occurs when there is missing data (e.g., domain essentiality). The general form of the estimation maximization function is *L(Ө,X) = .* Lu et al. (2015), with proteins substituted for genes, defined *X* and *Z* as *P* and *D,* respectively, while the parameter *Ө* is the collection of domain essentiality assignments. Thus, the likelihood function implemented was:

*L(Ө,P) = L(P,D|Ө) = L(D|Ө)L(P|D, Ө)* (adaption of EM likelihood function for domain essentiality, Lu et al. 2015)

*L(D|Ө)* =, where *Lk = δ* (essential domain) or 1-δ (non-essential domain) (essentiality labels generally change during optimization). *L(G|D, Ө)* = , where *Li = 1-FNR, FER, FNR, or 1-FER;* the FNR and FER are the false negative rate and the false expectation rate, respectively, for the observed gene essentiality, given the domain essentiality assignments. At least one domain must be designated essential in order for a gene to be predicted essential (but see Discussion). The four possible values of, above, correspond to true positive, false positive, false negative, and false negative observations, respectively, given the domain essentiality assignments.

During the expectation step, *L(D|Ө)* and *L(P|D, Ө)* are calculated for each domain during and a modified their ratio, *L(G|D, Ө) /* *L(D|Ө)*, gives a new value for δ. In the maximization step, the parameter *Ө* is modified by re-assigning essentiality of domains based on δ. In the present study, a threshold of δ ≥ 0.6 was used, whereas a stricter threshold of δ ≥ 0.9 was used in (Lu et al., 2015). The expectation and maximization stages are executed until the parameter *Ө* converges. Full details of the expectation and maximization equations can be found in (Lu et al., 2015).

Lists of all *S. cerevisiae*, and *S. pombe* genes are provided in Supplementary Table Sx.

**Conservation analysis of yeDUFs**

Uniprot representative proteomes (i.e. proteomes of fully-sequenced organisms), consisting of 5,862 proteomes at the time of writing of this paper (4,334 bacteria, 210 archaea, 813 eukaryotes, and 505 viruses), each linked to a distinct NCBI taxonomic identifier, were downloaded from Uniprot (xxx) and mapped to interactive tree of life (iTOL) leaves, using common species membership. Of the 191 organisms represented as leaves in iTOL, 20 lacked reference proteomes in Uniprot and therefore complete non-reference proteomes were used instead. Three additional organisms lacked any proteome in Uniprot. For two of these (*Gemmata obscuriglobus* and *Cryptosporidium hominis*) no near relative within iTOL could be found, and so they were removed. For the third, *Agrobacterium tumefaciens WashU*, counts from the related strain *Agrobacterium tumefaciens WashU* were used. The *Drosophila* endosymbiont, represented in iTOL as *Wolbachia sp. wMel*, could only be mapped to the nearest complete Uniprot proteome (*Wolbachia pipientis wMel Pop*) at the genus level. These proteomes contained 24,824,673 proteins or 36% of Uniprot (Swiss-Prot + TrEMBL). The 217,339 proteins from this list containing yeDUFs were then extracted using domain annotation from Pfam-A.full (v28). YeDUF representation was calculated for each representative organism, resulting in a non-redundant list of yeDUFs / taxonomic identifiers.The yeDUFs were divided into groups based on presence and essentiality in both Sc and Sp. There were five groups:

1. essential in Sc and absent in Sp (SceSp-)
2. essential in Sc and present in Sp (SceSp+)
3. essential in both Sc and Sp (SceSpe)
4. present in Sc and essential in Sp (Sc+Spe), and
5. absent in Sc and essential in Sp (Sc-Spe).

The log10(#yeDUFs) was plotted on the interactive Tree of Life (iTOL), for all yeDUFs as well as the 5 sub-categories. YeDUFs conserved in at least 10 prokaryotic iTOL species are referred to as prokaryotic yeDUFs or “pyeDUFs”, while the remainder of yeDUFs are referred to as eukaryotic yeDUFs or “eyeDUFs”.

**Functional analysis**

Gene ontology (GO) terms were extracted from all yeDUF-containing proteins in Sc and Sp. Species-specific gene ontology slim .obo files were used to obtain GO terms for Sc, Sp individually – resulting in two sets of annotation, while the primary gene ontology file – go\_basic.obo – was used to climb from these basic GO terms to GO slim terms. GO terms for yeDUF-containing proteins in Sc and Sp were mapped to GO slim terms using annotation specific for their respective organisms. These GO slims terms were then allocated to the 5 yeDUF categories (SceSp-, SceSp+, SceSpe, Sc+Spe, Sc-Spe), ranked by frequency among the respective yeDUFs, and evaluated to infer major biological processes. A yeDUF with at least 2 GO slim terms belonging to the same major biological process was assessed to have a possible functional link to that process.

**YeDUFs in Hillenmeyer, 2008 conditionally-essential genes**

Hillenmeyer et al. 2008 provided results of genome-wide heterozygous and homozygous single-gene knockout experiments across a range of different chemical and nutrient conditions. The findings indicate that the essentiality of many genes can only be observed under one or more such limiting conditions, and further suggest that clustering of conditional essentiality profiles, binary vectors wherein essentiality under each condition is represented as either a 1 or a 0, for essential or non-essential, can lead to clues about common function between genes.

Chemical fitness profiles were downloaded from Hillenmeyer et al. 2008, including both homozygous knockout and heterozygous knockout datasets. The latter implies a stricter definition of conditional essentiality, since a growth change must be observed after only one allele, rather than both, have been deleted. In addition to co-fitness quantities (on a scale of -1 to 1), both homozygous and heterozygous knockout datasets contain an r-value, or ratio of growth before / after gene knockout, for each gene under a range of 636 different chemical and nutrient challenge conditions comprising 316 distinct chemical compounds. The 636 conditions varied by chemical compound as well as by dosage, time of exposure, and experimental equipment. A higher r-value indicates a greater essentiality of the protein for the specific condition. Proteins were assumed to be conditionally essential if they had an r-value ≥ 2.0. Conditionally-essential proteins were examined for yeDUFs. A “fitness profile” consisting of a vector of 0s and 1 was generated for each yeDUF found in at least 1 conditionally-essential gene, for at least 1 chemical condition. Fitness profiles from yeDUFs in heterozygous-knockout conditionally-essential proteins were clustered hierarchically and visualized in R. Because a single-allele knockout they represented a stricter definition of conditional essentiality (knockout

**Results**

**Inference of yeDUFs**

A total of 199 yeDUFs were inferred for yeast, 120 for *Saccharomyces cerevisiae* (Sc) and 148 for *Schizosaccharomyces pombe* (Sp). Approximately one-third of yeDUFs (69 / 199) were found to be essential in both species of yeast (**Figure 2, bottom**). More yeDUFs were inferred using the expectation-maximization (EM) method than were inferred using the rule-based method. However, most rule-based inferences were redundant with EM inferences (only 2 and 8 yeDUFs were inferred exclusively using rules, for Sc and Sp, respectively).

**Conservation analysis of yeDUFs**

The inferred yeDUFs were found to be highly-conserved among eukaryotes (**Figure 3**). Interestingly, a small number of yeDUFs ) were also broadly conserved across all kingdoms of life. Of the 192 bacterial and archaeal species, sub-species and strains (i.e. leaves of tree) represented in iTOL, 146 encoded at least one yeDUF in their genomes, with 116 encoding at least 10 (maximum of 24, for *Methanosarcina acetivorans* and *Streptomyces avermitilis*). Roughly one-fourth of yeDUFs (48 / 199) were found among the species meeting the second criterion (>= 10 yeDUFs). SceSpe yeDUFs were the most conserved sub-group, while SceSp+ were the least conserved (**Figure 3**).

**YeDUFs in Hillenmeyer, 2008 conditionally-essential genes**

About half (93 out of 199) yeDUFs were found to be present in conditionally-essential genes, representing 232 of 636 conditions. In all, 133 of the 316 chemicals (i.e. 232 of the 636 conditions), tested by Hillenmeyer et al. 2008 study involved essential genes containing 93 yeDUFs. Although more *conditions* involved essentiality of *S. pombe* yeDUFs than *S. cerevisiae* yeDUFs (127 vs 103), the number of *yeDUFs* involved was lower for *S. pombe* (68 vs. 62), despite the fact that *S. pombe* yeDUFs outnumber *S. cerevisiae* yeDUFs (148 vs 120). This is not surprising, given that the experiment was performed in *S. cerevisiae*. Interestingly, of the 73 yeDUFs found to be essential in one species of yeast and present but not essential in the other (SceSp+, Sc+Spe), a majority were also found to be present in conditionally-essential genes from Hillenmeyer, 2008 (44/73, 60%), which represents a modest increase over the total proportion of yeDUFs overlapping (47%). If the additional intersection with byeDUFs is taken, an even greater proportion are present in Hillenmeyer yeDUFs (14/18, 78%). From the GO Slim analysis, these 14 may have functions related to ion binding, ion transport, transmembrane transport, and mitochondrial processes, as well as vacuole organization/fusion, membrane transport, and metabolism.

The heatmap of conditions vs. yeDUFs shows clustering of yeDUFs profiles for minimal media, drofenine hydrochloride, clozapine, trunculin, hydroxyurea, helenine, caspofungin, atorvastatin, benomyl + nocodazole, ketoconazole, itriconazole, among others (**Figure 5**).

**Discussion**

All organisms encode numerous proteins and domains of unknown function (Bateman et al. 2010 Acta Cryst.). However, it became clear only recently that many of these unknown domains are essential, at least in bacteria (Goodacre et al. 2014). This is especially surprising, given that many essential domains (eDUFs), are either restricted to relatively few species or clades, or are essential only in certain species. However, this finding matches the behavior of proteins, which may be essential in one species but not another, depending on the specific conditions these bacteria live in.

Interestingly, 124 yeDUFs are also shared with humans and thus demarcate a high-priority set for future experimental characterization (Table / Fig. X).

**Functional analysis**

SceSpe yeDUFs may play roles in protein folding, protein complex assembly, protein targeting / modification, ribosome biogenesis, or translation (Figure 4). SceSp- yeDUFs may be involved in cell wall organization and chromatin organization / remodeling, while SceSp+ yeDUFs may be involved in alternative carbohydrate and lipid pathways in the mitochondrion. S. pombe yeDUFs are also more likely to be conserved in eukaryotes. Conversely, Sc-Spe yeDUFs may be involved in vacuole organization and transport, while Sc+Spe yeDUFs may be involved in actin cytoskeletal organization and cell cycle control.

The yeDUFs conserved in bacterial species were enriched for transcription and translation-related processes (e.g. and DNA-binding), while they were depleted in membrane/ lipid-binding, protein folding, structural, and transport-related processes.

Many **yeDUFs are part of well-understood proteins**. For instance, the 50-60 amino acid yeDUF PF08354 (DUF1729) is found in fatty acid synthase beta subunits together with the MaoC-like domain (PF01575) and the acyltransferase domain (PF00698) (Figure XXX) [PMID:9693066]. The domain is conserved across fungi and bacteria. Nevertheless, the role of DUF1729 remains unclear. Pubmed searches do not find the domain names (PF08354, DUF1729) in the past literature.

A number of studies have analyzed **domain combinations**. For instance, shortly after the first eukaryotic genomes became available, Apic et al. found that half of 221 domain combinations in eukaryotes are shared by archaea and eubacteria while the other half is specific to eukaryotic genomes. However, the number of domain combinations was relatively small (PMID: 11472996).

Although the majority (186/199) of yeDUFs were found in combination with at least one other, non-yeDUF domain,we found only relatively few domain combinations involved multiple (2+) yeDUFs (84 combinations, involving 43 yeDUFs) in 2640 proteins from 649 organisms with representative proteomes in Uniprot. Interestingly, a large number (505) of these multi-yeDUF proteins had the same combination of 3 yeDUFs (PF06012, PF06025, PF14377). Of these 505 proteins containing the same trio of yeDUFs, 479 also contained the ubiquitin-transferase HECT domain PF00632, 279 also contained the ubiquitin-associated (UBA) domain PF00627, 260 contained both PF00632 and PF00627, and 98 contained the WWE domain PF02825, which serves as an interaction module in ubiquitination. Ubiquitination is linked to regulation and turnover of a number of cellular processes, including DNA damage repair, ribosome biogenesis. Interestingly, many of the 505 tri-yeDUF-containing proteins were also annotated with rRNA processing gene ontology terms. It may be possible that these yeDUFs are involved in the regulation of rRNA splicing and ribosome biogenesis. The three domains were also found by themselves as single-domain proteins, and less commonly in combination with one another to the exclusion of other domains. 28, 29, and 24 proteins were found to contain PF06012, PF06025, or PF14377 as single-domain proteins, respectively. All three of the yeDUF trio had such proteins annotated with ubiquitin ligase activity and ubiquitin transferase activity; PF06012 and PF14377 contained such proteins annotated as integral components of a membrane. Therefore, it is conceivable that the yeDUF trio function as components of a membrane-bound E3 ubiquitin-ligase protein involved in regulation of ribosome biogenesis. Also of note, many of the organisms with the greatest number of multi-yeDUF proteins were fungi (4 of the top 10). It has been found that fungi require E3 ubiquitin ligases for pathogenicity (Liu and Xue, 2011). Further study about the substrate specificity of PF06012-PF06025-PF14377 – containing E3 ligases may result in clues about fungal mechanisms of pathogenicity.

One major challenge of protein function analysis is that proteins or domains are not consistently essential – their “essentiality” is often **conditional**. We made use of the data produced by Hilenmeyer et al (2010) to investigate which DUFs are essential under certain conditions and found XXX. It is surprising that such a large proportion of yeDUFs from the present study were also found to be present in conditionally essential genes. It is possible that requiring at least one domain to be essential in an essential protein, a criterion for both rule-based and EM-based yeDUF methods, resulted in the inference of some conditionally-essential yeast DUFs. Such DUFs may only be essential when in combination with other specific domains. In particular, EM-based inference may have inferred some conditionally-essential yeast DUFs, since inference was based on a probability rather than strict criteria, and since the situation of multiple essential domains present in the same protein was not explicitly ignored (as in rule-based inference). However, rather than viewing this as a limitation of the study, we propose that EM-based inference could be a useful tool for inferring conditionally-essential domains, especially if it were expanded to include domain combinations, in addition to individual domains, as units of essentiality.

Species-specific essentiality also appears to be prevalent, since the majority (130/199) of yeDUFs were essential in only one species; of these, the majority were present in the other species, but not found to be essential (73/130). Interestingly, these 73 species-conditionally essential yeDUFs were enriched in overlap with yeDUFs in Hillenmeyer, 2008 conditionally-essentially genes (44/73), compared to total yeDUFs (92/199 overlapping). Despite being poorly conserved in bacteria (only 18/73 found in 10 or more bacterial species from the Interactive Tree of Life), those that were conserved in bacteria were nearly all (14/18) also found in Hillenmeyer, 2008 conditionally-essential genes. These 14 yeDUFs may play roles in the export of toxins, since they were predominantly annotated (in Sc and Sp proteins) with ion-binding, ion-transport, and transmembrane transport functions. It is tempting to speculate that certain conditionally-essential functions, such as export / metabolism of toxins, have been highly-conserved, although it is not clear why they were found to be essential in single-gene knockout screens for Sc and Sp.

**References**

Aslett, M. and Wood, V., 2006. Gene Ontology annotation status of the fission yeast genome: preliminary coverage approaches 100%. *Yeast*,*23*(13), pp.913-919.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*,*4*(1), 2.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., ... & Myers, C. L. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature methods*, *7*(12), 1017-1024.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., ... & Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091-1093

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., ... & Wong, E. D. (2011). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research*, gkr1029.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., ... & Kaiser, C. A. (2010). The genetic landscape of a cell. *science*,*327*(5964), 425-431

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), pp.14863-14868.

Frost, A., Elgort, M. G., Brandman, O., Ives, C., Collins, S. R., Miller-Vedam, L., ... & Weissman, J. S. (2012). Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. *Cell*, *149*(6), 1339-1352.

Heinicke, S., Livstone, M.S., Lu, C., Oughtred, R., Kang, F., Angiuoli, S.V., White, O., Botstein, D. and Dolinski, K., 2007. The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*, *2*(8), p.e766

Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., Onge, R.P.S., Tyers, M., Koller, D. and Altman, R.B., 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes.*Science*, *320*(5874), pp.362-365.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K. and Yang, L., 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, *415*(6868), pp.180-183.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, *98*(8), pp.4569-4574

Kim, D. U., Hayles, J., Kim, D., Wood, V., Park, H. O., Won, M., ... & Hoe, K. L. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nature biotechnology*, *28*(6), 617-623.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P. and Punna, T., 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*,*440*(7084), pp.637-643.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O. and Davis, R.W., 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, *94*(24), pp.13057-13062.

Li, Z., Vizeacoumar, F. J., Bahr, S., Li, J., Warringer, J., Vizeacoumar, F. S., ... & Boone, C. (2011). Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nature biotechnology*, *29*(4), 361-36

Liu, T.B. and Xue, C., 2011. The ubiquitin-proteasome system and F-box proteins in pathogenic fungi. *Mycobiology*, *39*(4), pp.243-248.

Lu, Y., Lu, Y., Deng, J., Lu, H., & Lu, L. J. (2015). Discovering Essential Domains in Essential Genes. *Gene Essentiality: Methods and Protocols*, 235-245

Luo, H., Lin, Y., Gao, F., Zhang, C. T., & Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic acids research*, *42*(D1), D574-D580.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., ... & Finn, R. D. (2011). The Pfam protein families database. *Nucleic acids research*, gkr1065.

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., ... & Levin, H. (2011). Comparative functional genomics of the fission yeasts.*Science*, *332*(6032), 930-936.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-2504

UniProt Consortium. (2014). Activities at the universal protein resource (UniProt). *Nucleic acids research*, *42*(D1), D191-D198.

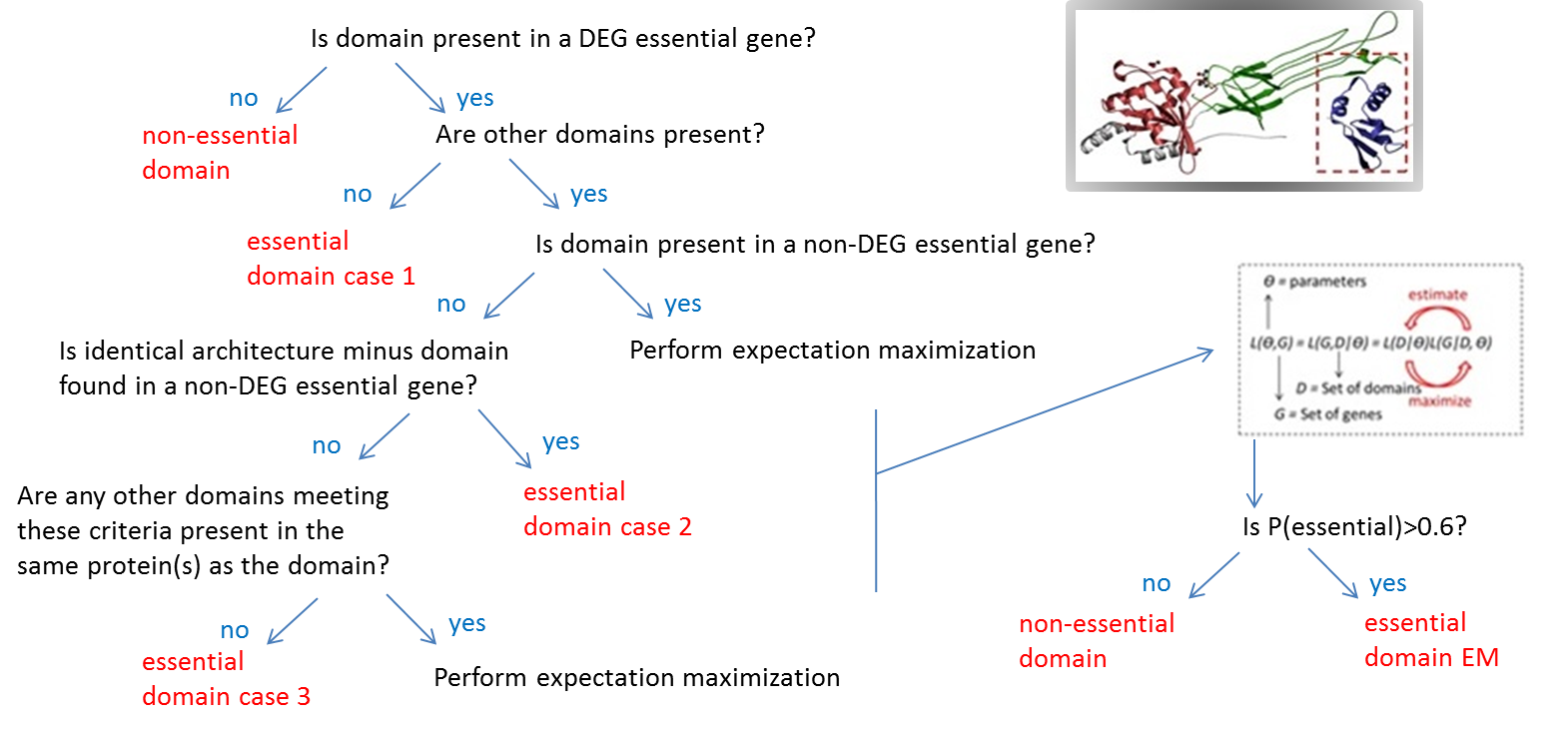
Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*(6634), 708-713.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B., 1998. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, *9*(12), pp.3273-3297.

Steinmetz, L.M., Scharfe, C., Deutschbauer, A.M., Mokranjac, D., Herman, Z.S., Jones, T., Chu, A.M., Giaever, G., Prokisch, H., Oefner, P.J. and Davis, R.W., 2002. Systematic screen for human disease genes in yeast.*Nature genetics*, *31*(4), pp.400-404.

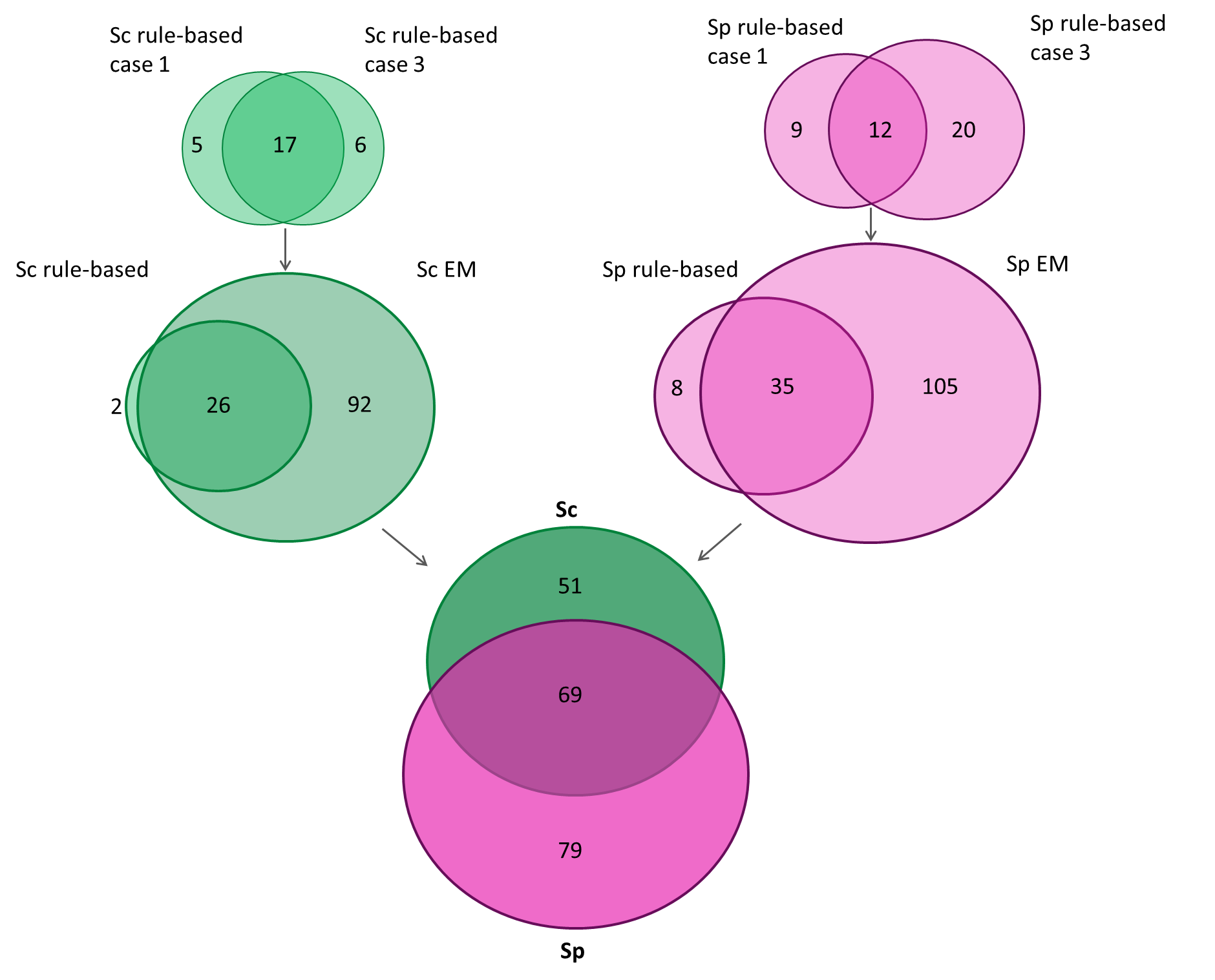
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. and Qureshi-Emili, A., 2000. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, *403*(6770), pp.623-627.

**Figures**



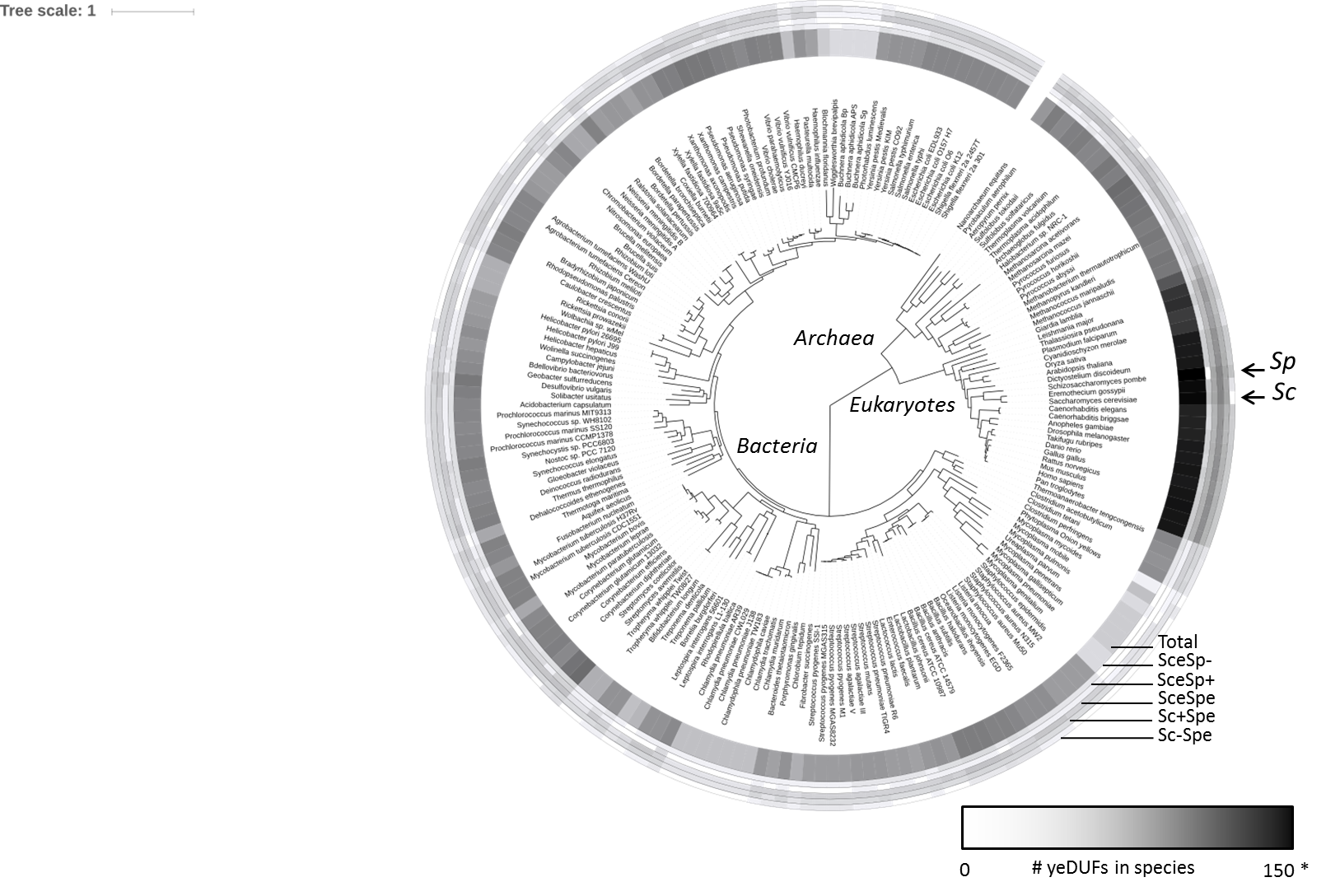
**Figure 1. Strategy to infer yeDUFs using a combined rule-based and expectation-maximization approach.**

The essentiality of protein domains of unknown function was inferred using a hybrid rule-based and expectation-maximization approach, which was performed separately for *S. cerevisiae* and *S. pombe*. Three rules were used to infer essential domains: **case 1** were domains in single-domain essential proteins, **case 2** were domains occurring as the difference in the set of domains between essential and non-essential multi-domain proteins, and **case 3** were domains always and only present in essential proteins, when no other such domains were found in the same protein. Additional essential domains were inferred using an expectation maximization process that maximized the probability of observed gene essentiality, given the assigned domain essentiality parameters. It is important to note that, while the methodology is depicted here as a decision tree for clarity, each form of inference was performed independently. *Inset:* PF09269 (blue, dotted red box) is present in over 2,000 bacterial and eukaryotic species, where it is often essential. PF09269 is often found in GTPases involved in ribosome biogenesis, although its precise molecular function remains unknown. Also shown: PF01018 / OBG fold (red), PF01926 / 50S ribosome-binding GTPase (green), bound GTP.

****

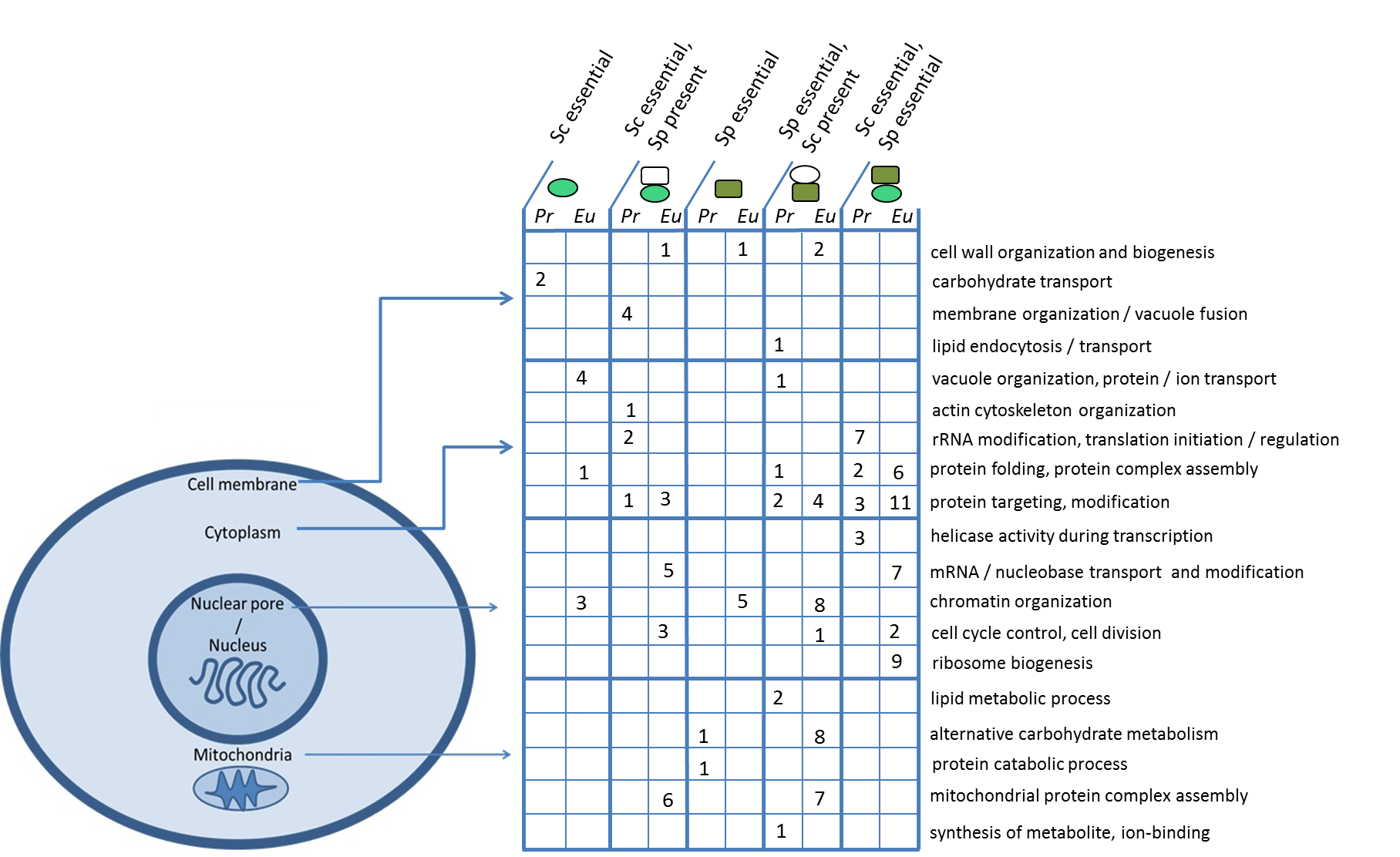
**Figure 2. *Saccharomyces cerevisiae (Sc)* and *Schizosaccharomyces pombe (Sp)* encode 199 yeDUFs.**

YeDUFs were inferred for each species of yeast separately, using rule-based inference (top), as well as expectation maximization (EM, middle). EM-inferred yeDUFs were more numerous and mostly contained their respective rule-based predictions as a subset. Approximately one-third of all inferred yeDUFs were shared between Sc and Sp (bottom). See Supplementary Table Sx for a list of yeDUFs.

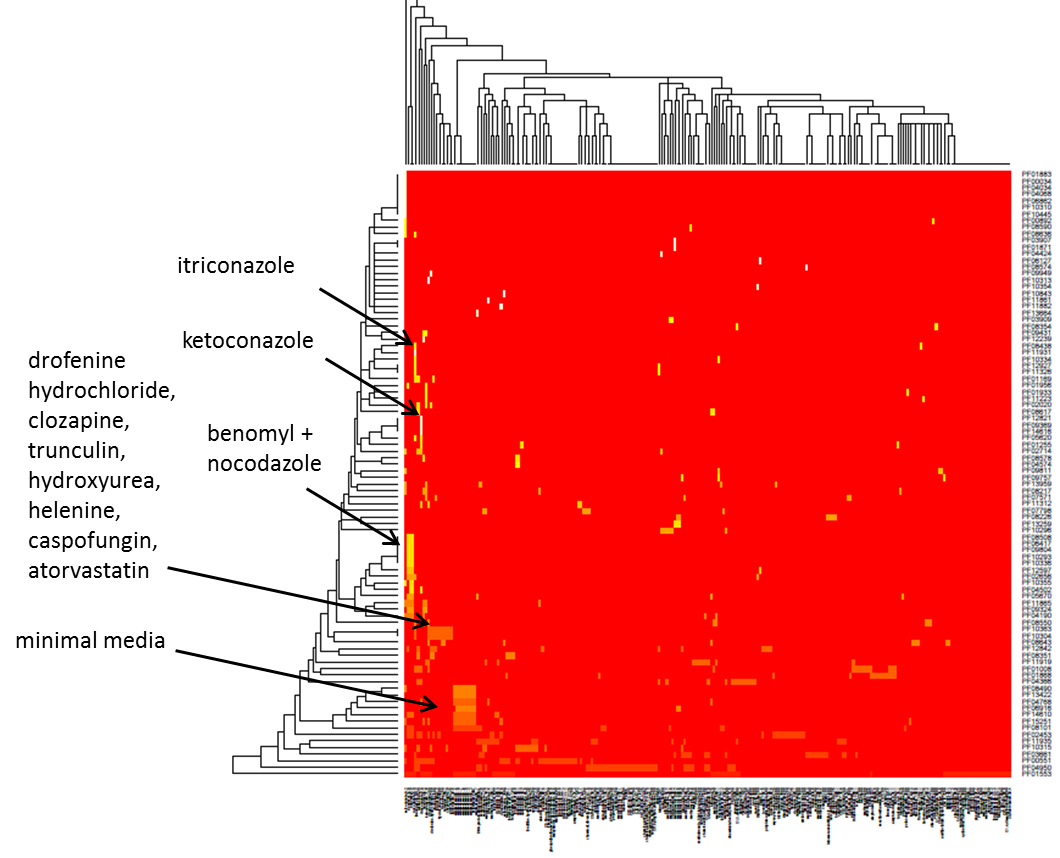
****

**Figure 3. YeDUFs are highly-conserved in eukaryotes, and broadly conserved across all kingdoms of life.**

Conservation of total yeDUFs is shown in the innermost ring, while the outer rings (moving outward) show conservation for SceSp- (essential in *S. cerevisiae*, absent in *S. pombe*), SceSp+ (essential in S. cerevisiae, present in *S. pombe*), SceSpe (essential in S. cerevisiae, essential in S. pombe), Sc+Spe (present in *S. cerevisiae*, essential in S. pombe), Sc-Spe (absent in *S. cerevisiae*, essential in *S. pombe*) sub-categories of yeDUFs, respectively. YeDUFs The greatest number of yeDUFs found in a single organism was 148\*, for Sp. SceSpe yeDUFs were the most highly-conserved among prokaryotes, while SceSp+ yeDUFs were the least highly-conserved among prokaryotes.

****

**Figure 4. The majority of yeDUFs are present in both species of yeast, but only essential in one. Pr = Prokaryotes, Eu = Eukaryotes**



**Figure 5. yeDUFs chemical fitness profiles cluster, suggesting collaboration or shared functions.** Some yeDUFs are essential for survival in multiple chemical conditions. Yellow areas correspond to XXX. A zoomable version is shown in Suppl. Fig. 5. Data from Hillenmeyer et al. 2011.