

SUPPLEMENTAL MATERIALS AND METHODS

Acquisition and processing of OrthoMCL data

Note that all files and supplementary figures are available for download at <http://yeast-phylogroups.princeton.edu>. Data defining orthologs for all yeast genes among the 149 other genomes curated by OrthoMCL was downloaded on July 18, 2011 (www.orthomcl.org). A file containing each yeast gene and corresponding numbers of orthologous genes from each assessed species was assembled ("01-OG_vs_Species-Full.txt"). All data regarding numbers of present orthologs per species was reduced to either "0" for no orthologs present, or "1" for at least one ortholog present ("02-OG_vs_Species-Binary.txt"). It is worth noting that a number of yeast genes (376 genes) had no orthology data in the OrthoMCL database; this is likely due to yeast gene annotation after the OrthoMCL data processing.

Because the 150 genomes curated by OrthoMCL have a handful of very closely related species (within the same genus), some species were removed to limit over-estimation of ortholog abundance within a diverse organism set simply based on overabundance of highly related species. In each case where multiple species from one genus were removed, at least one species was kept. The species removed due to genus over-representation were mostly from the "eukaryotic parasites" category: *Entamoeba histolytica* HM-1:IMSS, *Entamoeba invadens*-IP1, *Plasmodium yoelii yoelii* str. 17XNL, *Plasmodium vivax* SaI-1, *Plasmodium knowlesi* strain H, *Plasmodium chabaudi chabaudi*, *Plasmodium berghei* str. ANKA, *Cryptosporidium muris* RN66, *Cryptosporidium hominis* TU502, *Leishmania major* strain Friedlin, *Leishmania braziliensis*, *Leishmania Mexicana*, *Trypanosoma brucei gambiense*, *Trypanosoma congolense*, *Trypanosoma vivax*, *Giardia intestinalis* ATCC 50581, *Giardia lamblia* P15, and *Encephalitozoon cuniculi* GB M1 (*Magnaporthe oryzae* 70-15 was also removed because only its seventh chromosome was curated for the database, rather than its entire genome). After removing these species, a file was created with ortholog data derived from file 02 described above ("03-OG_vs_Species-Sorted.txt"). Clustering of these data revealed similar patterns of yeast gene conservation across defined taxonomic groups (Supplemental Figure 01). Therefore, ortholog data from individual species were collapsed into groups (archaea, bacteria, non-chordate animals, chordate animals, fungi, and eukaryotic parasites; see "Classification_of_OrthoMCL_Species.txt" file for full taxonomic analysis of each species). To collapse these data, the fraction of species with an ortholog to the corresponding yeast gene within a taxonomic group was calculated (the total number of species each containing at least one ortholog of a given gene was divided by the total number of species in the group)("04-OG_vs_Groups.txt").

GO-Slim analysis

Gene Ontology analysis was performed using the GO-Slim Mapper tool implemented in the Saccharomyces Genome Database (<http://yeastgenome.org/cgi-bin/GO/goSlimMapper.pl>). GO-Slim Mapper was used rather than standard GO Term finder due to the smaller, less redundant number of ontology terms used. Genes from each phylogroup were analyzed for enrichment/underrepresentation of three GO Slim categories (process, function, and component). P-values were calculated using the cumulative hypergeometric distribution, then corrected for multiple-hypothesis testing using the Benjamini-Yekutieli method (BENJAMINI and YEKUTIELI 2001). For display of GO term enrichment in Figure 02 and Supplemental Figure 03, GO terms were only included if at least one phylogroup had a significant enrichment (p-value of at least 1×10^{-7}). The

full set of GO Slim results is available for download from <http://yeast-phylogroups.princeton.edu>. Each table of GO terms was hierarchically clustered using Kendall's Tau (a rank-order based statistic) as the clustering metric. Average linkage was used as the linkage method. GO term leaf order was also optimized. MultiExperiment Viewer was used to perform clustering.

Yeast genome feature and phenotype information

Files containing gene feature information and gene-associated phenotype information were downloaded from Saccharomyces Genome Database (www.yeastgenome.org) in July 2011. These two files were named "2011.07.14_SGD_features.tab" and "2011.07.14_phenotype_data.tab," respectively.

Defining the set of yeast genes

To define our list of yeast genes using "2011.07.14_SGD_features.tab", non-protein coding genome features were removed (mitochondrial genes were retained). This list includes all features except for "ORF" (*not physically mapped, rRNA, autonomously replicating sequence, not in systematic sequence, centromere, external transcribed spacer region, 5' UTR, insertion, gene cassette, intron, long terminal repeat, +1 translational frameshift, pseudogene, repeat region, retrotransposon, telomere, telomeric repeat, transposable element gene, tRNA, snRNA, snoRNA, ncRNA, mating locus, multigene locus, non-transcribed region, noncodon exon, dubious, W region, X element combinatorial repeats, X region, Y region, Y' element, Z1 region, and Z2 region*). Additionally, because many dubious ORFs overlap existing genes, their inclusion into our analysis would artificially duplicate yeast copy numbers of ortholog groups. Therefore, the list of 6,604 open reading frames was further refined by removal of the 806 dubious ORFs to define a set of 5,798 genes. This resulting list of genes can be found in the file "06-Results_Summary.txt."

Defining a set of yeast genes with unknown function

Among the 5,798 genes remaining are 4,931 that are listed in SGD as "verified" and 867 listed as "uncharacterized." For some of the analyses described in this manuscript, we were interested in defining a set of yeast genes whose biological role is unclear – genes with unknown function. To determine the total number of genes with unknown function in our gene set, we searched the SGD-provided gene descriptions of the "verified" ORFs for the phrase "unknown function." This resulted in 239 of 4,931 genes (each gene description was manually examined to confirm that the gene function was uncharacterized). Therefore, the list of 1,222 uncharacterized genes in our data set includes those annotated as uncharacterized (867 genes) and those that are verified but are described as having an unknown function (239 genes). This resulting designation for each gene can be found in the "Unknown Function" column of the file "06-Results_Summary.txt."

Assessment of gene deletion viability

The "2011.07.14_phenotype_data.tab" file was parsed by first removing all non-ORFs, as described above for the "SGD_features.tab" file. For our analysis, we were interested in whether or not a complete gene deletion has been annotated as inviable or viable (e.g. whether or not the gene is essential under normal growth conditions). We therefore only kept data for the "null" mutant type (*this included removal of the following mutant types: activation, conditional, dominant*

negative, gain of function, misexpression, overexpression, reduction of function, repressible, and unspecified). Because null alleles have been engineered in multiple strain backgrounds, we opted for calling a gene essential if its deletion in any background resulted in inviability under normal growth conditions. For genes with an “inviable” null phenotype, we have included the genetic background information in parentheses. This resulting designation for each gene can be found in the “Null Phenotype” column of the file “06-Results_Summary.txt.”

Among our defined set of yeast genes, we identified 4,250 gene deletions that are viable, 1,109 gene deletions that are inviable, and 325 gene deletions that have not been tested. The untested category consists mostly of genes that have been annotated after construction of the original systematic deletion collection, and also some genes encoded by the mitochondrial genome (GIAEVER *et al.* 2002). Interestingly, 114 genes in the untested category are present in the commercially available haploid deletion collection, suggesting that the gene deletion is viable under normal conditions (OpenBiosystems). A list of these genes is available in the file: “Untested_gene_deletions_in_haploid_collection.txt.”

Data organization, processing, and visualization

Data were organized and processed using a combination of Microsoft Excel and R (www.r-project.org). Data was visualized using R or MultiExperiment Viewer (MeV_4_7, version 10.2; www.tm4.org/mev/). Both R and MultiExperiment Viewer are free, open-source software packages.

Identification of unreported gene deletions

In attempting to identify phenotypes (essential or non-essential) for the complement of protein-coding genes in yeast, we found a set of genes for which no gene deletion information is available in SGD (Saccharomyces Genome Database). This group of 325 genes contains genes that genuinely have not been published as deleted, but it also contains genes that are present in the haploid deletion collection from OpenBiosystems, and can be thus categorized as non-essential with as much confidence as any gene deletion in a large-scale collection. However, commercial availability does not constitute a curatable data source for SGD; a primary literature source is required (SGD personnel, personal communication). We have included this latter set of genes as downloadable data to provide such a data source.