Supporting Information

Cotton and McInerney 10.1073/pnas.1000265107

SI Materials and Methods

Genes Showing Homology to a Single Domain. To confirm that our results could not be affected by the best BLAST hit not being the most closely related sequence, due to variable rates of evolution, we performed the same analyses using only those genes that have significant BLAST hits to either archaebacteria or eubacteria. Because these genes show unambiguous homology to a single domain, phylogenetic analysis would show ancestry only within that domain. With these data, the OR for informational genes to be archaebacterial versus operational genes to be archaebacterial is 2.66 (95% CI, 1.60–4.40). Lethal genes are 2.13 times more likely than viable genes to be archaebacterial (95% CI, 1.46–3.12), and informational genes are 3.27 times more likely than operational genes to be lethal (95% CI, 1.90–5.64). Within just informational genes, lethal genes are 1.70 times more likely than viable genes to be archaebacterial (95% CI, 0.589–4.86), and within operational genes, they are 1.77 times more likely than viable genes to be archaebacterial (95% CI, 1.15–2.73). There are few informational genes with unambiguous BLAST hits to a single domain (a total of 59), so the 95% CI for the OR within this category is too wide for there to be a significant difference between the two probabilities; nonetheless, the pattern of greater lethality of archaebacterial genes compared with eubacterial genes is very similar in the two categories (OR, 1.70 for informational and 1.77 for operational).

The only notable difference between the two datasets is that archaebacterial genes show a higher mean number of duplicates in the single-domain hit data, whereas the opposite was true in the best-hit domain data. This is probably due to the effect of a few genes with particularly large numbers of duplicates being only weakly assigned as eubacterial. Repeating this analysis using median values rather than the mean, which has much lower statistical power but is more robust to these extreme values, supports this result for both datasets. Repeating our other tests using medians supports our findings, although this test has insufficient power to give significant P values for many of the comparisons (Table S3).

Phylogenetic Analysis. To check whether our BLAST-based results are consistent with results from phylogenetic analysis, we designed a phylogenetic analysis pipeline to test hypotheses of phylogenetic relationships of particular yeast genes and their homologs in other genomes. Building robust phylogenetic trees for individual genes that diverged very anciently is difficult (1–3), and the large size of many of our trees (up to 2,009 taxa) also makes it difficult to correctly identify optimal trees; any heuristic approach is likely to misplace some taxa when working at this scale. Thus, we designed a pipeline that aims to make robust inference about the relationships by attempting to identify which relationships each alignment could significantly reject, rather than relying on correctly inferring a single tree in any case. This hypothesis-testing approach should be more robust than relying on a single tree topology, but still may be sensitive to assumptions made in the substitution model. Although we have tested alternative empirical substitution matrices for every locus, we have not attempted to test the overall fit of any model or to fit more complex heterogeneous models, which is computationally impractical for such a large dataset.

We used RaxML version 7.0.4 (4) to perform both model selection and maximum likelihood (ML) phylogenetic inference for all 1,717 yeast ORFs for which we obtained significant hits from more than one prokaryotic domain in our PSI-BLAST search. For each ORF, an alignment of the yeast protein, any eukaryotic seed sequences used in the PSI-BLAST search, and all prokaryotic hits was generated using MUSCLE version 3.7 (5). For each alignment, we ran a pipeline that:

(*a*) Found the ML tree topology under the PROTCATWAG model.
(*b*) Calculated the likelihood for this tree under the PROTCAT versions of all of the empirical AA substitution models supported by RaxML (WAG, DAYHOFF, DCMUT, JTT, MTREV, RTREV, CPREV, VT, BLOSUM62, and MTMAM) both with and without invariant sites and empirical base frequencies, both singly and together.
(*c*) Found the best fitting of these models under the Akaike information criterion, corrected for sample size, for subsequent analysis.
(*d*) Found the unconstrained ML tree under this optimal model using the fast heuristic search algorithm of RaxML.
(*e*) Found ML trees under four different constraints:

  (*i*) Monophyly of eukaryotes
  (*ii*) Reciprocal monophyly of eukaryotes, archaea, and eubacteria
  (*iii*) Presence of (eukaryote + archaea) clade
  (*iv*) Presence of (eukaryote + eubacteria) clade.

(*f*) Tested whether any of these constraints can be rejected by the data using the approximately unbiased (AU) test (6) as implemented in Consel version 0.1i (7).

Note that the four constraints together allow us to test three different possibilities for the relationships of each yeast locus. We assume a priori that a gene for which eukaryote monophyly cannot be rejected has a single origin in this domain. A gene for which both constraints (*ii*) and (*iii*) can be rejected shows significant support for a clade of eukaryote sequences nested within the eubacterial radiation, whereas rejection of (*ii*) and (*iv*) suggests that a eukaryotic clade is nested within an archaebacterial radiation. Failure to reject hypothesis (*ii*) indicates that for this gene, we cannot reject the traditional three-domain tree of life. Trees rejecting both (*iii*) and (*iv*) must show a more complex evolutionary history in which neither archaeabacterial nor eubacterial sequences are monophyletic, indicative of lateral transfer among prokaryotes.

Note that if none of the hypotheses can be rejected significantly, it is likely to be because of a lack of statistical power.

Because of time constraints imposed by the computing facility that we used, each alignment was run with a limit of 84 h of CPU time to complete the pipeline; 1,247 of 1,717 jobs completed within this time. As we expected, these were mainly the smaller alignments in our dataset [median number of sequences, 159 (range, 8–1,329) in completed jobs vs. 692.5 (range, 95–2,009) in uncompleted jobs].

SI Results

To identify the phylogenetic relationships of yeast genes as displayed on the ML tree under the best-fitting model from our pipeline, we used a Perl script that identifies the smallest (i.e., least inclusive) cluster (or clan; ref. 8) on each tree that includes the yeast gene and at least one prokaryotic sequence. These prokaryotic sequences then form the closest noneukaryotic sister group to the eukaryotic sequences under most possible rootings of our unrooted gene trees. We then tested whether this cluster included just archaebacterial sequences, just eubacterial se-

quences, or sequences from both domains, and whether this cluster included all of the sequences from a particular domain that were present on the tree, indicative of a tree displaying the three-domain relationship.

We found a total of 143 trees in which the yeast gene is most closely related only to archaebacterial sequences, 717 trees showing this relationship to only eubacterial sequences, 48 loci for which the three-domain tree is most likely, and 283 loci in which the closest sister group contains sequences from both prokaryotic domains. These results are ambiguous, presumably indicating that lateral gene transfer has influenced the phylogeny for this gene.

Comparing these results with our BLAST-based analysis, we find that, of 620 genes assigned as eubacterial in the best-hit analysis that we could analyze phylogenetically, in 25 cases the yeast gene clustered instead with archaebacterial homologs, contradicting the BLAST result, and an additional 17 showed the three-domain tree. However, for genes identified by best-BLAST hit as archaebacterial, a much higher proportion (114 out of 266) were contradicted by the ML tree, and 37 showed the three-domain tree.

Although these results underscore the difficulty of accurately identifying the evolutionary relationships of individual genes, our main results are robust to these differences (Table S2). When using the phylogenetic results for assigning all of those genes with homology to both domains, the OR for informational genes to be archaebacterial versus operational genes to be archaebacterial is 2.50 (95% CI, 2.22–2.81). Lethal genes are 2.91 times more likely than viable genes to be archaebacterial (95% CI, 2.15–3.94), and informational genes are 2.57 times more likely than operational genes to be lethal (95% CI, 1.65–4.01). Within just informational genes, lethal genes are 2.65 times more likely than viable genes to be archaebacterial (95% CI, 0.96–7.29), and within operational genes they are 2.45 times more likely than viable genes to be archaebacterial (95% CI, 1.74–3.44). The results of all of these tests closely match those from our best-hit data and indeed demonstrate the effect more strongly than our BLAST analysis results in all cases except the increased lethality of informational genes, which is slightly weaker in this analysis (but still significant). These results suggest that the BLAST approach is essentially reliable, but may be adding some noise to our results.

We would caution against taking our phylogenetic results as any kind of gold standard for assigning domain identity for the loci that we have investigated, given that the relationships within our ML trees are probably not entirely reliable and certainly are rather poorly supported in many cases. This is emphasized by the results of our AU tests for these data, which reveal that most of the alignments that we analyzed lack the statistical power to unambiguously assign the evolutionary origin of most eukaryotic genes. For example, of a total of 1,247 analyzed alignments, monophyly of the eukaryotic sequences was significantly rejected in 189 (all AU tests are at an α level of $P < 0.01$). These alignments were removed from subsequent analyses, because any inference about the origins of these genes would be ambiguous. Of the remaining 1,058 alignments, 553 rejected the three-domain constraint, of which 25 also rejected a eubacterial affinity for the eukaryotic sequences [constraint (*iv*) above], 154 rejected an archaeal affinity for archaebacterial sequences [constraint (*iii*)], 345 rejected both of these possibilities, and 29 rejected neither possibility. Of the 1,058 alignments, 498 could not reject the three-domain models, the vast majority of which (478) could not reject any of the constraints, perhaps indicating a lack of power for these loci. Of the remainder, 17 rejected only constraint (*iii*), and 3 rejected only constraint (*iv*).

In summary, our main findings are supported by our phylogenetic results, but our results underline the difficulty of accurately and unambiguously reconstructing the sequence of evolutionary events that occurred in the distant past.

1. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM (2008) The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci USA* 105:20356–20361.
2. Pisani D, Cotton JA, McInerney JO (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol* 24:1752–1760.
3. Rodríguez-Espeleta N, et al. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol* 56:389–399.
4. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood–based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
5. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
6. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508.
7. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
8. Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM (2007) Of clades and clans: Terms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol* 22: 114–115.
9. Duarte NC, Herrgård MJ, Palsson BO (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* 14:1298–1309.

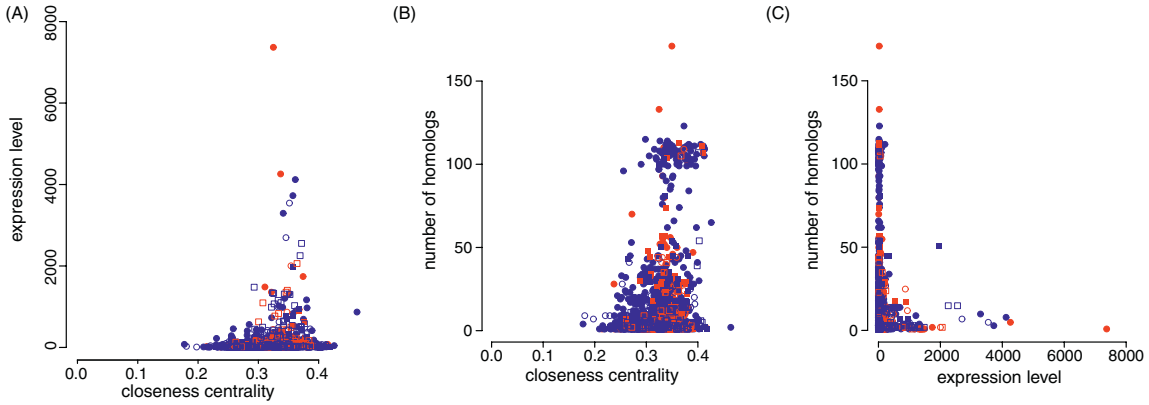

Fig. S1. Expression level, protein–protein interaction (closeness centrality in the interaction network), and number of yeast homologs. Each point is a single yeast gene. Blue points represent genes with a viable deletion phenotype; red points, genes with a lethal deletion phenotype. Circles represent operational genes; squares, informational genes. Filled points represent genes with eubacterial homology; open points, genes with archaebacterial homology, under the best-hit criterion. Because closeness centrality and degree in the interaction network are correlated, only the closeness statistic is presented.

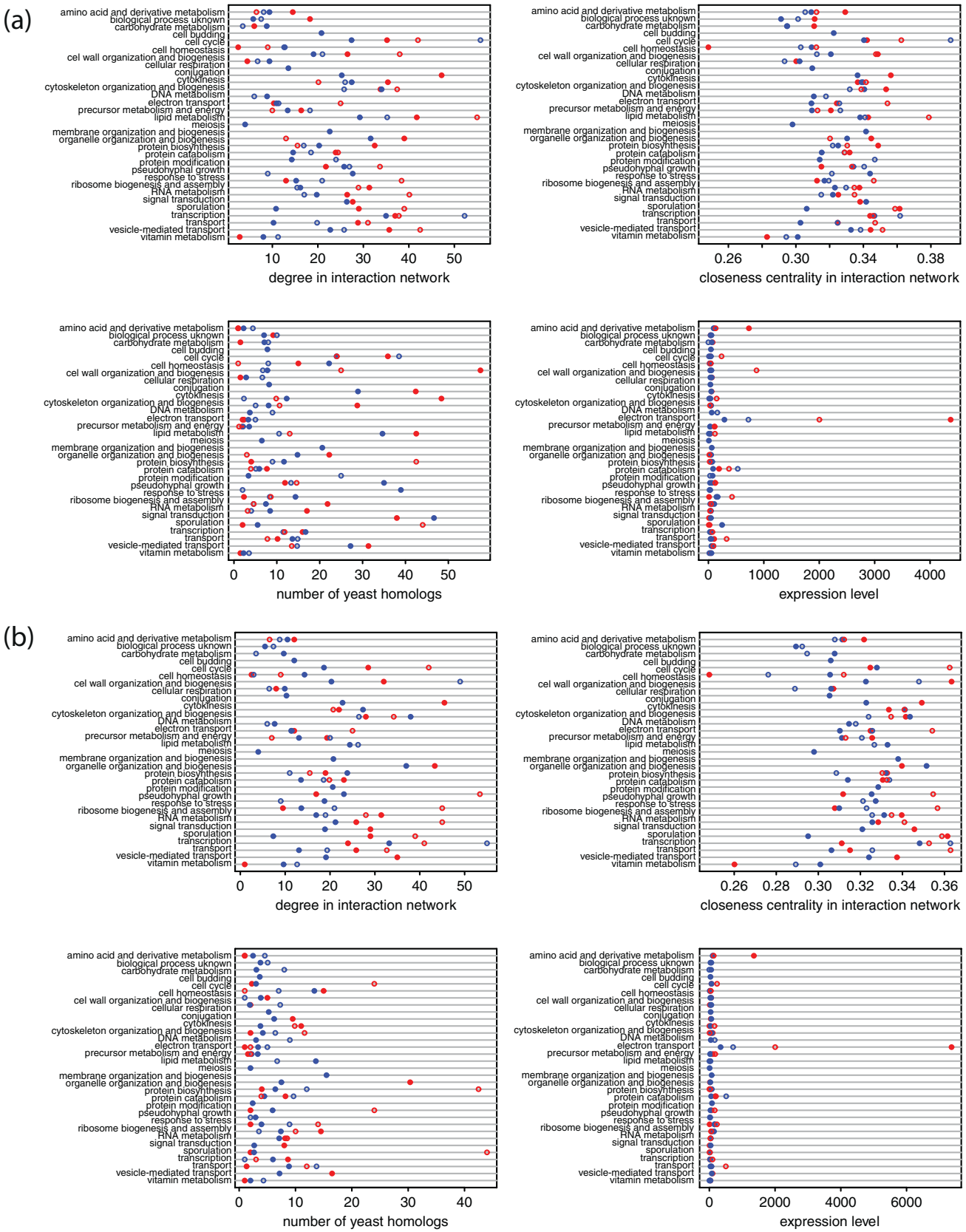

Fig. S2. Expression level, protein–protein interaction (closeness centrality and degree in the interaction network) and number of yeast homologs per functional category. Each point is the mean of the values for genes in a category with a particular deletion phenotype and with homology to a particular domain. Blue points represent genes with a viable deletion phenotype, red points represent genes with a lethal deletion phenotype, filled circles represent genes with eubacterial homology, and open circles represent genes with archaebacterial homologs under the best-hit criterion (*A*) and the single domain hit criterion (*B*).

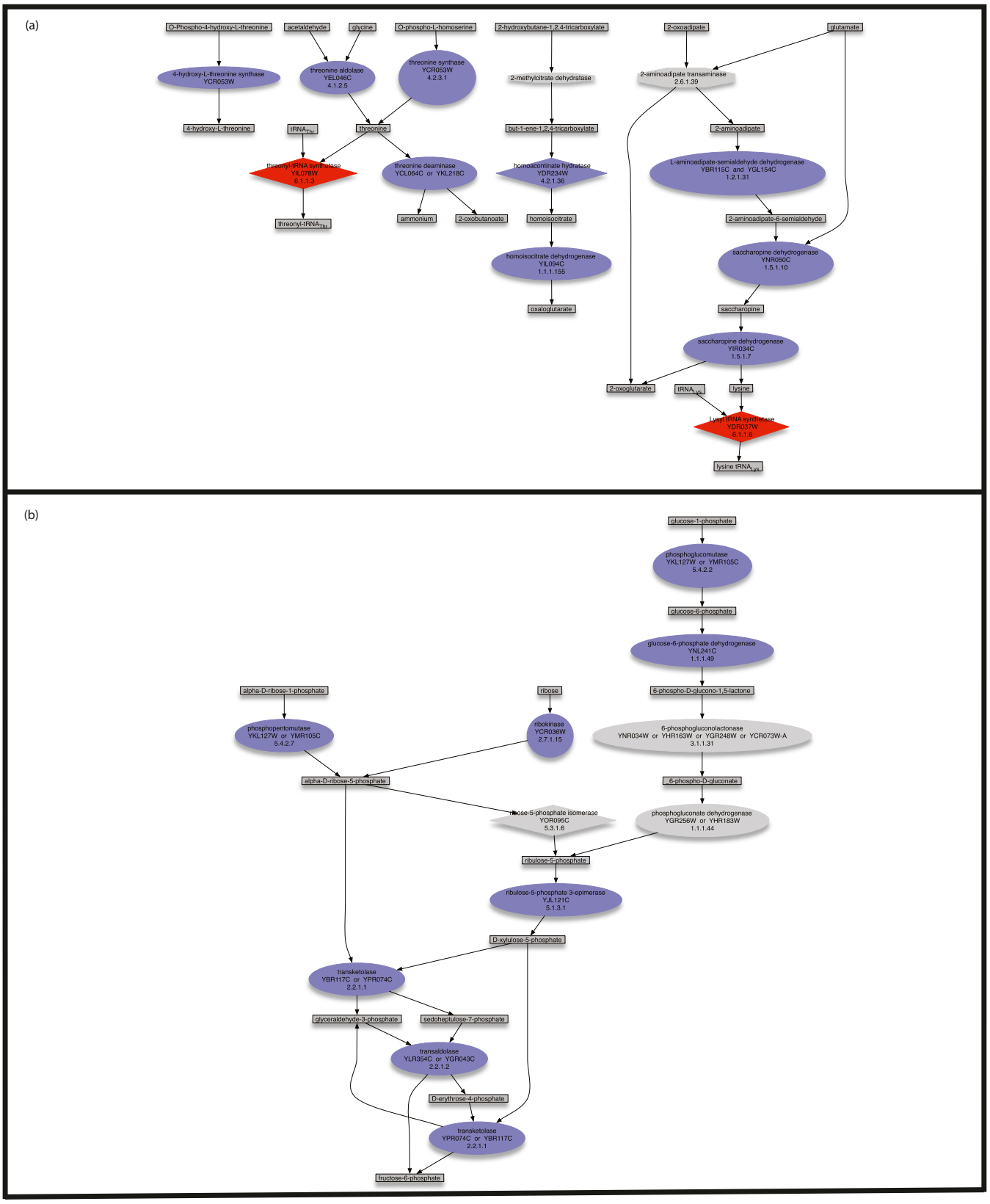

Fig. S3. (Continued)

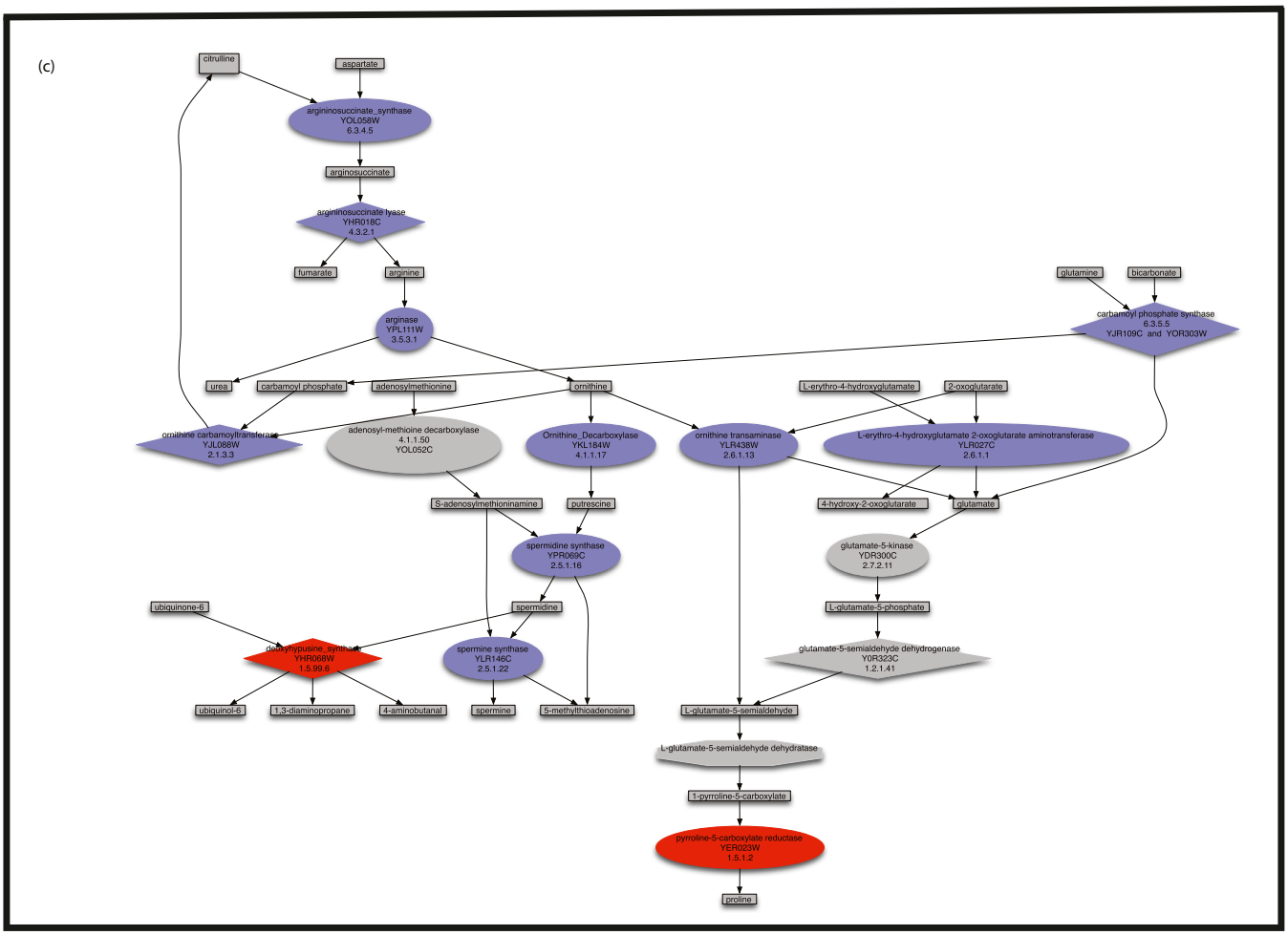

Fig. S3. Three metabolic pathways annotated with homology domain and knockout phenotype for genes in the pathway. The gray, rectangular boxes represent major metabolites, with common cofactors and intermediates removed for clarity. Other boxes represent enzymes. Circular or ellipsoid boxes represent enzymes encoded by genes with eubacterial homologs; diamond-shaped boxes, those encoded by genes with archaebacterial homologs. Octagonal boxes show steps for which no gene is annotated in the model used. Red boxes represent genes with lethal knockout phenotype; blue boxes, viable knockout phenotype; gray boxes, those for which this data are not available or are ambiguous because the different genes possibly encoding this activity vary in phenotype. Pathways are from the iND750 model of yeast metabolism (9). Pathways shown are for threonine and lysine metabolism (*A*), pentose phosphate metabolism (*B*), and arginine metabolism (*C*). Note that these examples contain both archaebacterial and eubacterial homologs showing both lethal and viable deletion phenotypes, but that the proportions of these different categories reflect those found across the whole yeast genome. For example, two out of three archaebacterial genes involved in threonine and lysine metabolism are lethal, whereas all eubacterial genes are viable, and both of these lethal genes are aminoacyl-tRNA synthases, with informational functions. Only a single gene in the operational pentose phosphate pathway has archaebacterial homology, and all genes in this pathway are viable or have an unknown deletion phenotype. Arginine metabolism contains examples of genes with both lethal and viable phenotypes of both archaebacterial and bacterial homology, although it contains a greater proportion of genes with archaebacterial homology than is typical.

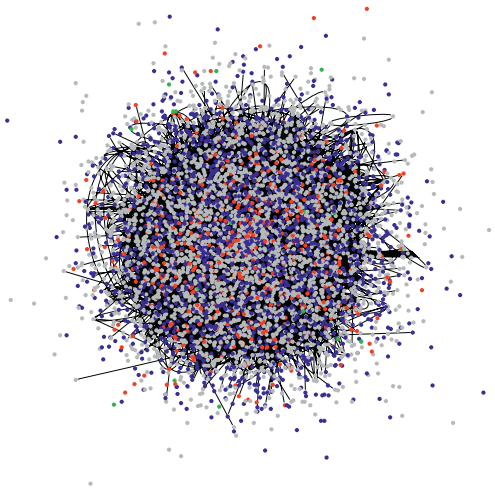

Fig. S4. The yeast protein–protein interaction network. Each vertex is a single *Saccharomyces* gene, with edges connecting genes whose protein products are known to interact. Vertices are colored by the prokaryote domain of best BLAST-hit homology for each gene (blue for eubacteria, red for archaebacteria, green for equal or nearly equally good hits to both domains, gray for genes showing no significant homology to either prokaryote domain).

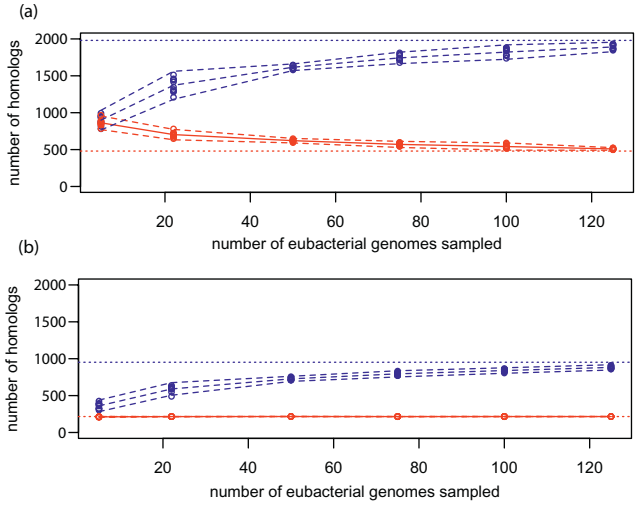

Fig. S5. Testing the impact of taxonomic sampling. To test the sensitivity to the particular set of prokaryote genomes used, we repeated our BLAST experiment on databases consisting of all 22 archaebacterial genomes used in our full dataset together with randomly chosen subsets of the 197 eubacterial genomes of different sizes. We ran 10 replicates each with subsets of 5, 22 (equal to the archaebacterial count), 50, 75, 100, and 125 genomes. For each replicate, we recorded the number of yeast genes showing homology to archaebacteria or eubacteria under our two different criteria, taking the domain of the best BLAST hit and only counting genes that show homology to just one of the two prokaryote domains. This figure shows the results of this analysis. The results suggest that the results are fairly consistent for any reasonably large (≥50) sample of eubacterial genomes, and thus the exact taxonomic sample chosen is not critical. A corollary of this is that we would not expect our results to be significantly different if additional prokaryotic genomes were included in the full dataset, so our conclusions should remain valid as, for example, more and more prokaryote genomes are sequenced and assembled. In particular, we note that those genes identified as archaebacterial homologs are largely robust to taxonomic sampling as long as at least 22 eubacterial genomes are included, and are almost entirely robust for samples of size ≥50. Eubacterial identity is slightly more labile but shows a similar pattern. These findings confirm that for any samples of more than about 50 eubacterial genomes, and for most of the samples with only 22 eubacterial genomes, the difference between these results and our results from the full dataset is much too small to alter the main result of the paper; for this, an ~2-fold change in the numbers of genes assigned to archaebacterial and eubacterial categories would be needed. The data including only those genes showing homology to a single prokaryotic domain are particularly robust to taxonomic sampling.

Table S1. Functional correlates for single domain hit data

Data type	Eubacteria	Archaebacteria	All	*P* value
Expression level: number of tags	79.51 (58.45–100.90)	172.21 (112.5–232.0)	85.89 (78.80–93.09)	5×10^{-4}
Closeness centrality in interaction network	0.312 (0.310–0.315)	0.321 (0.317–0.326)	0.316 (0.315–0.317)	0.0007
Degree in interaction network	15.05 (14.17–15.98)	18.17 (16.05–20.18)	18.02 (17.60–18.48)	0.0039
Number of homologs in yeast genome	5.75 (5.02–6.54)	7.66 (6.69–8.73)	7.58 (7.14–8.04)	0.001

Values are means and 95% bootstrap percentile CIs for the mean of each parameter (calculated using the nonparametric bootstrap). *P* values are bootstrap probabilities for the mean of the statistic in archaebacteria being less than or equal to the mean in eubacteria, based on 10,000 replicates.

Table S2. Genes showing archaebacterial and eubacterial homology, with lethal and viable deletion phenotypes, for both informational and operational functional categories, for best-hit domain, for single-domain hit data, and for genes showing homology to both domains based on our phylogenetic assignment

	Lethal deletion phenotype					Viable deletion phenotype				
	Eubacteria	Archaebacteria	No hit	Ambiguous	Missing	Eubacteria	Archaebacteria	No hit	Ambiguous	Missing
Best-hit domain										
Informational genes	20	35	100	0	0	39	18	127	0	0
Operational genes	210	102	444	2	0	1,226	257	1,565	8	0
Unknown function	7	0	19	0	0	341	41	745	0	0
All genes*	237	137	630	2	0	1,610	316	2,912	8	2
Single-domain hit										
Informational genes	11	18	100	26	0	19	11	127	27	0
Operational genes	89	37	444	188	0	595	118	1,565	778	0
Unknown function	0	0	19	7	0	164	15	745	203	0
All genes*	100	55	630	221	0	781	144	2,912	1,009	2
Genes showing homology to both domains										
Informational genes	21	20	100	14	0	38	7	127	12	0
Operational genes	188	62	444	64	0	1,127	127	1,565	237	0
Unknown function	3	0	19	4	0	311	23	745	48	0
All genes*	212	82	630	827	0	1,480	157	2,912	297	2

"No hit" indicates genes that have no significant homology to any sequence in the prokaryotic genome data used here.
*All gene counts include genes for which no Gene Ontology data are available; thus, this row is not the sum of the rows above.

Table S3. Functional correlates for yeast genes, based on best-hit domain and single domain hit data, using medians rather than means

Data type	Eubacteria	Archaebacteria	All	*P* value
Best-hit domain				
Expression level: number of tags	27 (25.02–29.10)	41 (32.55, 48.46)	26 (24.63–26.77)	<0.0001
Closeness centrality in interaction network	0.317 (0.315–0.319)	0.327 (0.325–0.330)	0.326 (0.311–0.318)	<0.0001
Degree in interaction network	10 (9.38–11.42)	15 (12.65–17.41)	12 (11.50–13.46)	<0.0001
Number of homologs in yeast genome	3 (2.88–3.11)	4 (3.43–4.93)	2 (2–2)	0.175
Single-domain hit data				
Expression level: number of tags	28 (24.98–30.45)	47 (31.25–60.24)	26 (24.63–26.77)	0.0006
Closeness centrality in interaction network	0.315 (0.312–0.318)	0.326 (0.320–0.331)	0.326 (0.311–0.318)	0.0002
Degree in interaction network	9 (7.71–10.09)	13 (9.95–16.34)	12 (11.50–13.46)	0.0087
Number of homologs in yeast genome	2 (1.66–2.28)	5 (3.75–5.92)	2 (2–2)	<0.0001

Values are medians and 95% bootstrap percentile CIs for the median of each parameter (calculated using the nonparametric bootstrap). *P* values are bootstrap probabilities for the median of the statistic in archaebacteria being less than or equal to the median in eubacteria, based on 10,000 replicates.

Table S4. Functional correlate data for operational and informational genes

		Informational genes			Operational genes			Both domains		
		Archaebacteria	Eubacteria	*P* (arch. < eub.)	Archaebacteria	Eubacteria	*P* (arch. < eub.)	Informational	Operational	*P* (inf. > op.)
Expression level number of tags	Single	267.9 (56.2–482.4)	97.37 (35.06–159.58)	0.0521	165.30 (100.9–229.7)	89.05 (61.78–115.80)	0.0098	87.27 (61.50–113.21)	100.55 (91.1–110.2)	0.167
	Best	176.2 (53.2–298.7)	89.49 (45.34–133.39)	0.0837	175.663 (137.0 214.5)	81.67202 (65.38–97.88)	<0.0001			
Closeness centrality in interaction network	Single	0.336 (0.326–0.346)	0.331 (0.321–0.342)	0.252	0.322 (0.317–0.327)	0.316 (0.314–0.319)	0.036	0.329 (0.326–0.332)	0.322 (0.321–0.323)	>0.9999
	Best	0.333 (0.327–0.341)	0.335 (0.328–0.343)	0.607	0.326 (0.323–0.330)	0.318 (0.316–0.320)	<0.0001			
Degree in interaction network	Single	27.931 (21.99–33.81)	23.866 (16.53–31.23)	0.192	17.37 (15.00–19.70)	16.65 (15.51–17.80)	0.290	27.92 (25.82–29.99)	20.76 (20.21–21.29)	>0.9999
	Best	31.208 (26.65–35.79)	28.932 (23.58–34.25)	0.262	21.102 (19.32–22.88)	17.586 (16.76–18.41)	0.0002			
Number of homologs in yeast genome	Single	5.655 (4.170–7.130)	8.367 (5.367–11.365)	0.948	8.367 (7.033–9.709)	4.905 (4.320–5.490)	<0.0001	5.79 (4.886–6.700)	7.98 (7.41–8.56)	<0.0001

Values are means and 95% bootstrap percentile CIs for the mean of each parameter (calculated using the nonparametric bootstrap). The *P* values in columns 5 and 8 are bootstrap probabilities for the mean of the statistic in archaebacteria being less than or equal to the mean in eubacteria, based on 10,000 replicates. Those in column 11 are for the mean for operational genes being less than or equal to that for informational genes, across hits to both domains. Alternate rows show single-domain hit data and best-hit domain data, respectively.

Table S5. OR results

For all data

Test of informational/operational bias

	Info.	Oper.
Archaebacterial	53	359
Eubacterial	59	1436

P(arch|info) = 53/53+59 = 53/112
P(arch|oper) = 359/359+1436 = 359 /1795
OR = 2.366071
ASE = ASE (log odds) = sqrt (1/53 + 1/359 + 1/59 + 1/436) = 0.202228
Log OR = log(2.366071) = 0.8612308
95% CI = 0.8612308 + 1.96*0.202228 = 1.257598
0.8612308–1.96*0.202228 = 0.4648639
95% CI for OR (out of log space): 1.591798–3.51692

Test of archaebacterial lethality versus archaebacterial viable phenotype

	Lethal	Viable
Archaebacterial	137	316
Eubacterial	237	1610

P(arch|lethal) = 137/374
P(arch|viable) = 316/1926
OR = 2.232637; log OR = 0.8031834
ASE = ASE(log odds) = sqrt(1/137 + 1/316 + 1/237 + 1/1610) = 0.1237108
95% CI for log OR = 0.5607102–1.045657
95% CI for OR = 1.751916–2.845267

Test of lethality of informational genes versus lethality of operational genes

	Info.	Oper.
Lethal	55	312
Viable	57	1483

P(lethal|informational) = 55/112
P(lethal|operational) = 312/1795
OR = 2.979338; log OR = 1.091701
ASE = sqrt(1/55 + 1/57 + 1/312 + 1/1483) = 0.1990103
95% CI for log OR = 1.481761–0.7016408
95% CI for OR = 2.017060–4.400689

Test of lethality of archaebacterial genes versus lethality of eubacterial genes for informational genes only

	Lethal	Viable
Arachaebacterial	35	18
Eubacterial	20	39

P(arch|lethal) = 35/55
P(arch|viable) = 18/57
OR = 2.015152; log OR = 0.7006944
ASE = sqrt(1/35 + 1/18 + 1/20 + 1/39) = 0.3997099
95% CI for log OR = 1.484126–0.082737
95% CI for OR = 0.9205932–4.411108

Test of lethality of archaebacterial genes versus lethality of eubacterial genes for operational genes only

	Lethal	Viable
Archaebacterial	102	257
Eubacterial	210	1226

P(arch|lethal) = 102/312
P(arch|viable) = 257/1483
OR = 1.886486, log OR = 0.6347159
ASE = sqrt(1/102 + 1/210 + 1/257 + 1/1226) = 0.1388256
95% CI for log OR = 0.3626177,0.906814E
95% CI for OR = 1.437086,2.476420)

For informational hits data

Test of informational/operational bias

	Info.	Oper.
Archaebacterial	29	155
Eubacterial	30	684

P(arch|info) = 29/59
P(arch|oper) = 155/839
OR = 2.660580; log OR = 0.9785441
ASE = sqrt(1/29 + 1/155 + 1/30 + 1/684) = 0.2571903

Table S5. Cont.

95% CI for log OR =0.4744511–1.482637
95% CI for OR = 1.607132–4.404545

Test of archaebacterial lethality versus archaebacterial viable phenotype

	Lethal	Viable
Arachaebacterial	55	129
Eubacterial	100	614

P(arch|lethal) = 55/184
P(arch|viable) = 100/714
OR = 2.134239; log OR = 0.7581102
ASE = sqrt(1/55 + 1/129 + 1/100 + 1/614) = 0.1938103
95% CI for log OR = 0.378242–1.137978
95% CI for OR = 1.459716–3.120454

Test of lethality of informational genes versus lethality of operational genes

	Info.	Oper.
Lethal	29	126
Viable	30	713

P(lethal|info) = 29/59
P(lethal|oper) = 126/839
OR = 3.272935; log OR = 1.185687
ASE = sqrt(1/29 + 1/30 + 1/126 + 1/713) = 0.2777681
95% CI for log OR = 0.6412615–1.730112
95% CI for OR = 1.898875–5.641288

Test of lethality of archaebacterial genes versus lethality of eubacterial genes for informational genes only

	Lethal	Viable
Archaebacterial	18	11
Eubacterial	11	19

P(arch|lethal) = 18/29
P(arch|viable) = 11/30
OR = 1.69279; log OR = 0.526378
95% ASE = sqrt(1/18 + 1/11 + 1/11 + 1/19) = 0.5385214
95% CI for log OR = -0.5291239–1.58188
95% CI for OR = 0.5891208–4.864091

Test of lethality of archaebacterial genes versus lethality of eubacterial genes for operational genes only

	Lethal	Viable
Archaebacterial	37	118
Eubacterial	89	595

P(arch|lethal) = 37/126
P(arch|viable) = 118/713
OR = 1.774348; log OR = 0.5734328
95% ASE = sqrt(1/37 + 1/118 + 1/89 + 1/595) = 0.2200414
95% CI for log OR = 0.1421517–1.004714
95% CI for OR = 1.152751–2.731126

Each calculation presents first the numbers of genes involved in the calculation as a 2×2 table. Then the two probabilities (odds) are calculated separately. Then the OR is calculated, followed by the SE and 95% CI.

Table S6. Genes in each homology, function, and lethality category, with ORF names, gene names, GO cellular process annotation, and descriptions from the *Saccharomyces* Genome Database (SGD)

http://bioinf.nuim.ie/supplementary/CottonMcInerneyPNAS_2010/tableS5.pdf

An asterisk in the "GO cellular process" column indicates that there are multiple GO terms in this category attached to this gene and we have reported the most commonly used term, as reported by the SGD. Shaded rows are those genes that exhibited significant similarity to sequences to both prokaryotic domains. These are genes that are present in the "best hit" data set but removed in the data set that is used for calculations based on hits to only one of the two prokaryotic groups. This table can be downloaded from http://bioinf.nuim.ie/supplementary/CottonMcInerneyPNAS_2010/tableS5.pdf.