

Eukaryotic genes of archaeobacterial origin are more important than the more numerous eubacterial genes, irrespective of function

James A. Cotton^{a,b,1} and James O. McInerney^{a,2}

^aDepartment of Biology, National University of Ireland, Maynooth, County Kildare, Ireland; and ^bSchool of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved August 17, 2010 (received for review January 7, 2010)

The traditional tree of life shows eukaryotes as a distinct lineage of living things, but many studies have suggested that the first eukaryotic cells were chimeric, descended from both Eubacteria (through the mitochondrion) and Archaeobacteria. Eukaryote nuclei thus contain genes of both eubacterial and archaeobacterial origins, and these genes have different functions within eukaryotic cells. Here we report that archaeobacterium-derived genes are significantly more likely to be essential to yeast viability, are more highly expressed, and are significantly more highly connected and more central in the yeast protein interaction network. These findings hold irrespective of whether the genes have an informational or operational function, so that many features of eukaryotic genes with prokaryotic homologs can be explained by their origin, rather than their function. Taken together, our results show that genes of archaeobacterial origin are in some senses more important to yeast metabolism than genes of eubacterial origin. This importance reflects these genes' origin as the ancestral nuclear component of the eukaryotic genome.

endosymbiosis | gene essentiality | eukaryote origin | protein interaction network

As one of the three domains of cellular life, the eukaryotes are typically described as the sister group to the archaeobacteria. This sister group relationship describes the evolutionary history of the “nuclear-cytoplasmic” component of eukaryotes, with mitochondria and plastids being of endosymbiotic bacterial origin (e.g., ref. 1). In this traditional scenario, the unique features of extant eukaryotes were gradually acquired in the eukaryote stem group before the endosymbiotic acquisition of the mitochondrion. Thus, the acquisition of the mitochondrion was an important, but not foundational, step in eukaryote origins, occurring subsequent to the evolution of many characteristic features of eukaryotic cell biology. Early molecular phylogenies of ribosomal RNA genes support this scenario (see refs. 2 and 3 for reviews), as do several other molecular markers. Many nuclear genes are more closely related to eubacterial homologs than to any known archaeobacterial sequence (4, 5) and appear to have been transferred to the nucleus from the ancestral mitochondrial genome by a process known as endosymbiotic gene transfer (1, 6). A similar process occurred after other symbiotic events, for example, the introduction of many chloroplast-derived genes into the nuclei of green plants (6).

An alternative view of eukaryotic nuclear-cytoplasmic origins, first suggested by Lake (7–9) is that this lineage arose from within, rather than as a sister to, the archaeobacteria. This view is supported by molecular phylogenies showing that many eukaryote genes actually derive from within the archaeobacterial domain (7–11), including a recent reanalysis of informational genes with modern phylogenetic methods (10). It also has become clear that those eukaryotes that lack mitochondria either are derived from organisms that have mitochondria or themselves host hydrogenosomes or mitosomes, which are degenerate relicts of mitochondria (3, 12). Thus, all known eukaryotes possessed mitochondria

at some point in their evolutionary history, suggesting either that the acquisition of the mitochondrion might have occurred early in eukaryote evolution (or at least that the characteristic features of extant eukaryotic cell biology arose after the initial mitochondrial endosymbiosis) or that many important lineages of primitively amitochondriate transitional “protoeukaryotes” have gone extinct. Various alternative scenarios have been proposed to explain the chimeric (archaeobacterial and eubacterial) nature of eukaryotic genomes (3, 13, 14–16), some involving symbioses or “cell fusions” quite different in character from what we call the traditional scenario (5, 14, 17). These ideas remain somewhat controversial (18, 19), but appear to be supported by a growing body of empirical evidence (12, 20).

However they arose, eukaryotic nuclei clearly contain homologs to both eubacterial and archaeobacterial genes, and a growing number of phylogenetic studies confirm that nuclear genes are derived from multiple sources (7, 12, 21, 22). Previous studies (23, 24) confirm that about half of the eukaryotic genes have homologs in prokaryotes, and that most of these homologs are eubacterial. Furthermore, archaeobacterial and eubacterial homologs are known to fulfill broadly different functions in eukaryotic cells, with eubacterial homologs largely involved in “operational” metabolic processes and archaeobacterial homologs largely involved in the “informational” processes of transcription, translation, and replication (23, 25). These different functions suggest that the different partners played different roles in the formation of the earliest eukaryotic cell. Here we reveal other fundamental differences between the contributions of the two partner genomes.

Results

Our results are based on identifying prokaryote homologs of eukaryotic genes, examining every gene in the *Saccharomyces cerevisiae* genome. They support recent studies (23, 24) in showing that many eukaryotic genes are related to prokaryotic genes (2,460 of 6,704 genes), and that ~75% of these have eubacterial affinities. For 1,980 yeast genes, the strongest BLAST hit is to a eubacterial gene, and for 480 yeast genes, the strongest hit is archaeobacterial; 952 genes have only eubacterial homologs, showing no homology to any archaeobacterial sequence, whereas 216 genes have only archaeobacterial homologs.

Author contributions: J.A.C. and J.O.M. designed research; J.A.C. performed research; J.A.C. contributed new reagents/analytic tools; J.A.C. analyzed data; and J.A.C. and J.O.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹Present address: Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom.

²To whom correspondence should be addressed. E-mail: james.o.mcinerney@nuim.ie.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1000265107/-DCSupplemental and http://bioinf.nuim.ie/supplementary/CottonMcInerneyPNAS_2010/tableS5.pdf.

We carried out a number of phylogenetic analyses of 1,717 of these gene families, with only the very largest families not subjected to these analyses. The proportions of genes ascribed eubacterial ancestry and archaeobacterial ancestry remained similar (see *SI Results* for details). These data confirm a significant bias toward archaeobacterial homology for genes with informational functions [odds ratio (OR), 2.37; 95% confidence interval (CI), 1.59–3.52]. Although significant, this is not a clear-cut distinction, given that genes with archaeobacterial homologs are involved in most of the biological processes of the yeast cell.

These absolute numbers of homologs suggest a larger role for genes with eubacterial homologs. Absolute numbers do not necessarily tell the whole story, however, given that genes may differ in function in many different ways, such as through different patterns of expression and involvement in different metabolic pathways. To explore this functional dimension, we mapped our homologs onto data from a comprehensive gene knockout study (26), identifying each gene as having either a lethal or a viable deletion phenotype. Lethal genes are more than twice as likely to have archaeobacterial homologs than eubacterial homologs (OR, 2.23; 95% CI, 1.97–2.53). One possible explanation for this is that the informational functions of genes with archaeobacterial homology are likely to be essential to cellular viability, and indeed informational genes are more often lethal than operational genes (OR, 2.98; 95% CI, 2.03–4.40). This does not explain our result, however, because both informational and operational genes with archaeobacterial homologs are more likely to be lethal than those with eubacterial homologs. Furthermore, the greater propensity to lethality of archaeobacterial genes is very similar across the two categories (for informational genes, OR, 2.01; 95% CI, 0.92–4.41; for operational genes, OR, 1.89; 95% CI, 1.43–2.47). Although the relatively small number of informational genes means that we cannot reject the null hypothesis of no association for this subset of the data, we note that the estimated strength of this effect is actually greater for informational genes than for operational genes, confirming that the lack of significance is due to a lack of power in the test for informational genes. Counts of genes in each category are given in *SI Results*.

The foregoing results are robust to details of the data and analysis, but we emphasize that these are large-scale patterns rather than clear distinctions. Many archaeobacterial homologs have operational functions with both viable and lethal deletion phenotypes, as do many informational eubacterial homologs. Homology, function, and phenotype are also not strongly associated with the metabolic pathway in which the genes are involved (Fig. 1 and Fig. S3). Most pathways contain both eubacterial and archaeobacterial homologs, and the distribution of these within pathways shows no clear general pattern. Although we have not attempted a large-scale analysis of metabolic pathway structure or evolution, it is clear that some pathways (e.g., phospholipid and sphingolipid metabolism) are largely eubacterial, some have connected eubacterial and archaeobacterial components (e.g., sterol synthesis), and others are a complex mixture of genes of different homologies (e.g., tyrosine, tryptophan, and phenylalanine metabolism). Three other example pathways are presented in *SI Materials and Methods*.

In an effort to explain the greater essentiality of archaeobacteria-related genes, we examined other data that might shed light on the differing cellular functions of these genes and their protein products. Using data from RNAseq experiments (27), we found significantly greater expression of genes with archaeobacterial homologs (Table 1). The average number of tags that could be attached to genes of archaeobacterial origin was 164.64 (95% CI, 131.0–198.5), compared with 73.81 (95% CI, 61.03–86.46) for eubacteria. This is a >2-fold difference on average. No significant differences are seen between the expression levels of operational and informational gene categories (Table S4).

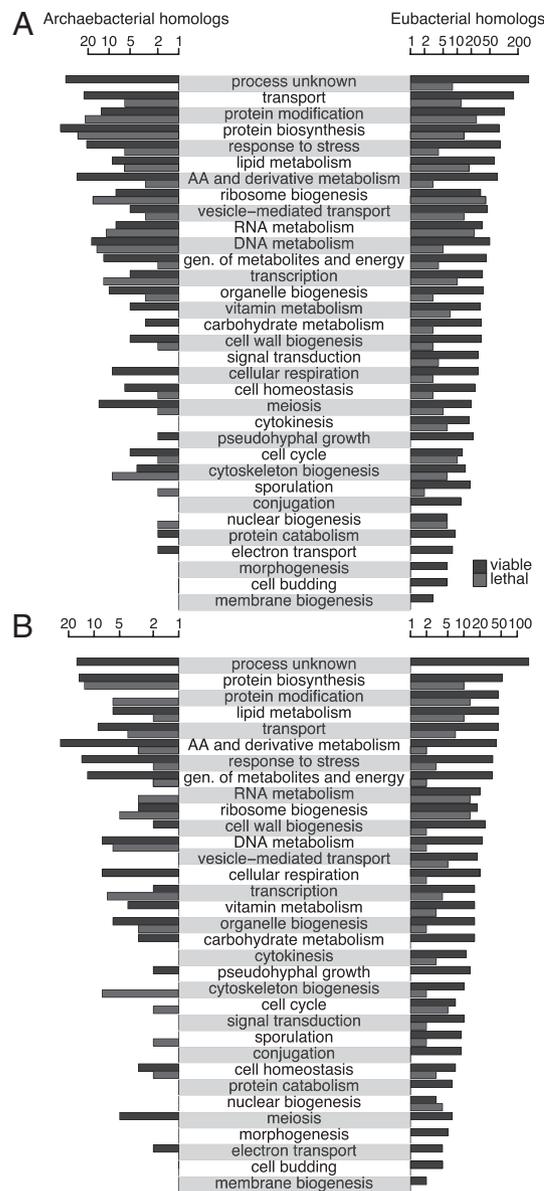


Fig. 1. Distribution of homologs for yeast genes. Homologs are listed by homology domain, functional category, and deletion phenotype. (A) Best-hit domain. (B) Unambiguous hits, with homology only to one of the two prokaryotic domains. Light bars represent lethal genes and dark bars represent viable genes in each domain. Note that the number of genes is plotted on a log axis.

Genes with archaeobacterial homologs are more central and more highly connected in the yeast protein interaction network (28–30) (Fig. S4 and Table 1; see *SI Materials and Methods* for details on data and methods), which has been shown to reflect greater essentiality (29, 31). This difference is partly explained by the greater centrality and connectedness of informational genes, but a statistically significant difference is still observed for operational genes alone (Table S4). Furthermore, operational genes whose products interact directly with the products of genes with informational functions are more likely to have a lethal knockout phenotype compared with other operational genes; however, because this effect is similar for both archaeobacterial and eubacterial homologs, the pattern of protein-protein interactions does not explain our main result (*SI Results*).

Table 1. Functional correlates of prokaryote homology for yeast genes

Data	Eubacterial	Archaeobacterial	All	P value (arch ≤ bact)
Expression level: number of tags	73.81 (61.03–86.46)	164.64 (131.0–198.5)	85.89 (78.80–93.09)	< 0.0001
Closeness centrality in interaction network	0.314 (0.312–0.316)	0.324 (0.321–0.327)	0.316 (0.315–0.317)	< 0.0001
Degree in interaction network	15.91 (15.20–16.62)	20.90 (19.33–22.48)	18.02 (17.60–18.48)	< 0.0001
Number of paralogs in yeast genome	13.13 (12.09–14.16)	8.02 (6.89–9.22)	7.58 (7.14–8.04)	1

Values are listed by domain of best BLAST hit, showing 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 archaeobacterial homologs being less than or equal to the mean in eubacterial homologs, based on 10,000 replicates.

Finally, eubacterial homologs show more duplicate copies (paralogs) within the yeast genome, suggesting that a greater degree of genetic redundancy is protecting the cell against deletion of eubacterial homologs. Although there is a significant difference in the number of duplicate copies between operational genes and informational genes, the significant difference in the number of duplicates between archaeal and eubacterial homologs is consistent for both functional groups. However, unlike our other findings, this result is sensitive to the dataset used (*SI Results*), and other studies have found little evidence of a relationship between duplication and redundancy (32), which may vary with the function and mode of duplication of the genes and even between genomes (33).

Discussion

Genes of different origins play significantly different roles in eukaryotic cells that cannot be explained by the functional (operational vs. informational) distinction between sets of genes. Genes of archaeobacterial origin and those of eubacterial origin differ significantly in many aspects, including essentiality, expression level, and centrality in protein interaction networks. This complex pattern suggests that this is a signal of the history of the yeast genome.

Our methods do not allow us to estimate the timing or exact source of the genes that we identify as having homology to genes from different prokaryotic domains. These genes could be found in the yeast genome as a result of more recent lateral gene transfer (LGT), rather than being a relict of mitochondrial endosymbiosis. Both pre-eukaryogenesis LGT events among and between groups of prokaryotes (34, 35) and LGT from either group to eukaryotes (21) could have affected some of our data. Although there are plenty of examples of prokaryote-to-eukaryote LGT, there is limited evidence of LGT being an important mode of genome evolution in most eukaryotes (36). Extensive investigation has found no conclusive evidence of prokaryotic genes in the human genome, and there appears to be little evidence of prokaryotic gene transfer into the yeast genome (37, 38), although there may be methodological problems with these studies (36). The statistically significant results of our analyses are even more surprising in light of these processes. Although it seems likely that recently acquired genes would occupy peripheral roles in cellular metabolism or regulation, we know of no proposed mechanism to explain the very different lethality of genes from archaeobacterial and eubacterial sources if recent LGT is responsible for many of the prokaryotic homologs that we observe, unless there is some systematic difference in the timing of LGT from the two domains.

If most of the prokaryotic homologs that we observe are descended from the fusion of a eubacterium and archaeobacterium to form the first eukaryotic cell, then our results can be interpreted in terms of the different roles of the two ancestors. In this scenario, genes from the archaeobacterial host formed the original eukaryote nucleus and so have been cointeracting for a longer time and form a core part of metabolism. Incoming eubacterial genes, from genome fusion or from subsequent endosymbiotic gene transfer (our data cannot distinguish between

the two scenarios), have more peripheral roles in the network of protein interactions that controls metabolism, because the archaeobacterial genes that performed essential functions might have been more difficult to displace by the influx of eubacterial genes. Although genes of eubacterial affinity seem to have replaced large parts of this ancestral metabolism, our findings suggest that much of eukaryotic metabolism may have been built on an ancestral foundation that still plays a central role in the eukaryotic cell. Our results also support other ideas about genome evolution. For instance, the complexity hypothesis proposes that genes that encode proteins in large complexes are highly connected and thus less likely to experience LGT (39). Our findings add to the evidence indicating that the protein interaction network of yeast shows significant historical structure (40), confirming that subsequent evolution has not completely erased the effect of ancient evolutionary history on eukaryotic genomes.

Whatever the source of the prokaryote homologs that we have identified, our results demonstrate that whereas eubacteria have made a greater quantitative contribution to yeast metabolism, the archaeobacteria made a different, arguably more important contribution. These results are compatible with previous findings (12, 20) and with some ideas about the origin of the eukaryotic cell (13, 41).

It is not clear that the historical process of eukaryogenesis should be able to help us understand the biology of modern eukaryotic cells, given that > 2.5 billion years (42) of evolution have shuffled genes between pathways, changed expression levels, and altered the interactions between gene products. For example, only half of eukaryotic genes have an identifiable prokaryotic homolog, and no large functional category consists solely of genes with homology to a sole prokaryotic domain. Rapid genomic changes are likely to have followed eukaryogenesis, as they did when genomes fused more recently (43), so it is remarkable that some of the original partners' contributions might have persisted for > 1 billion years of evolution.

Yeast metabolism, and presumably eukaryotic metabolism in general, is a complex tapestry of prokaryotic threads and eukaryotic innovations. Our analysis of the features of eukaryotic genes that have a prokaryotic history shows that a protein's group of origin plays an important role in defining its expression profile, likelihood of lethality, and position and connectivity in a protein interaction network independent of the actual function of the protein. This suggests that the roles of genes from the various partners in the eukaryotic cell differ in ways beyond the simple split between operational and informational functions.

Materials and Methods

Homology Search. To produce a homology search that would be both sensitive and specific, we built a profile alignment of the amino acid sequence of a range of eukaryotic homologs for each yeast gene, then used PSI-BLAST (44) to search against a database of 197 eubacterial and 22 archaeobacterial genome sequences. To build the profile alignments for PSI-BLAST, each protein-coding gene in the *Saccharomyces cerevisiae* genome sequence [downloaded from the Cogent database (45)] was compared with the protein-coding gene content of six other eukaryotic genomes (*Caenorhabditis elegans*, *Arabidopsis thaliana*, *Schizosaccharomyces pombe*, *Neu-*

rospora crassa, *Ashbya gossypii*, and *Trypanosoma cruzi*) downloaded from the same source. For each yeast gene, a multiple sequence alignment of the yeast gene and the best (i.e., lowest e-value) hit with $e < 0.001$ from each of these genomes was constructed using the alignment program MUSCLE (46) with default settings. This alignment, of between one and seven sequences (depending on how many eukaryotic genomes had a hit with $e < 0.001$ for the yeast gene) was used as a seed profile for a PSI-BLAST search against the combined database of prokaryotic protein sequences, with an e-value cutoff of 1×10^{-6} . Genes were classified as homologs to the prokaryotic domains in two different ways. In the least stringent case, genes were assigned to whichever domain their best BLAST hit sequence belonged, being considered ambiguous only if they had equally good best hits in both domains (Results and Fig. 1A). In the second case, genes were considered ambiguous unless all BLAST hits with an e-value below the cutoff were in the same domain.

Functional Comparisons. Comparisons of domain homology and knockout phenotype, functional category, expression level, and interaction network position were carried out using Perl scripts (available from the authors on request). Genes annotated with Gene Ontology (GO) (47) terms “transcription,” “transcription,” “DNA-dependent DNA replication” or any of their subterms were considered informational; all other genes were considered operational. Interaction network statistics were calculated using the Pajek

(48) package. GO mappings were downloaded from the Saccharomyces Genome Database (49), RNAseq data were obtained from Nagalakshmi et al. (27), knockout phenotype data were downloaded from the comprehensive yeast genome database (50), and protein interaction data were obtained from BioGRID (30).

Statistical Analysis. We describe the strength of associations between factors using ORs (51); for example, the odds of being archaeobacterial for informational genes is calculated as the probability of an informational gene having an archaeobacterial homolog, divided by the probability of the gene having a eubacterial homolog. We can similarly calculate the odds of being archaeobacterial for operational genes. The OR is the ratio of these two odds. Thus, this statistic is not affected by the absolute sizes of the different categories. To test the significance of associations, we constructed a 95% CI for the OR under a normal approximation to the log OR (51). A significant association is one for which this CI does not overlap unity.

ACKNOWLEDGMENTS. We thank the three anonymous referees for their comments, which greatly improved the manuscript. This research was funded by Science Foundation Ireland; the Irish Research Council for Science, Engineering and Technology; and a Research Councils UK Academic Fellowship. The computation was facilitated in part by the Irish Centre for High End Computing (ICHEC) and the NUI Maynooth computing centre.

- Sagan L (1967) On the origin of mitosing cells. *J Theor Biol* 14:255–274.
- Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271.
- Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440:623–630.
- Brown JR, Doolittle WF (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* 61:456–502.
- Pühler G, et al. (1989) Archaeobacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc Natl Acad Sci USA* 86:4569–4573.
- Martin W, et al. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99:12246–12251.
- Lake JA (1985) Evolving ribosome structure: Domains in archaeobacteria, eubacteria, eocytes and eukaryotes. *Annu Rev Biochem* 54:507–530.
- Lake JA (1987) Prokaryotes and archaeobacteria are not monophyletic: Rate-invariant analysis of rRNA genes indicates that eukaryotes and eocytes form a monophyletic taxon. *Cold Spring Harb Symp Quant Biol* 52:839–846.
- Lake JA (1988) Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* 331:184–186.
- Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM (2008) The archaeobacterial origin of eukaryotes. *Proc Natl Acad Sci USA* 105:20356–20361.
- Pisani D, Cotton JA, McInerney JO (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol* 24:1752–1760.
- van der Giezen M, Tovar J, Clark CG (2005) Mitochondrion-derived organelles in protists and fungi. *Int Rev Cytol* 244:175–225.
- Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392:37–41.
- Hartman H (1984) The origin of the eukaryotic cell. *Speculations Sci Technol* 7:77–81.
- Sogin ML (1991) Early evolution and the origin of eukaryotes. *Curr Opin Genet Dev* 1:457–463.
- Searcy DG (1992) Origins of mitochondria and chloroplasts from sulfur based symbiosis. *The Origin and Evolution of the Cell*, eds Hartman H, Matsuno K (World Scientific, Singapore), pp 47–78.
- Doolittle WF (1995) Some aspects of the biology of cells and their evolutionary significance. *Evolution of Microbial Life*, eds Roberts DM, Sharp P, Alderson G, Collins M (Cambridge Univ Press, Cambridge, UK), Vol 54, pp 1–21.
- Kurland CG, Collins LJ, Penny D (2006) Genomics and the irreducible nature of eukaryote cells. *Science* 312:1011–1014.
- Cavalier-Smith T (2009) Predation and eukaryote cell origins: A coevolutionary perspective. *Int J Biochem Cell Biol* 41:307–322.
- Rivera MC, Lake JA (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155.
- Doolittle WF, et al. (2003) How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos Trans R Soc Lond B Biol Sci* 358:39–57, discussion 57–58.
- Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV (2008) The deep archaeal roots of eukaryotes. *Mol Biol Evol* 25:1619–1630.
- Esser C, et al. (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol* 21:1643–1660.
- Horiike T, Hamada K, Kanaya S, Shinozawa T (2001) Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. *Nat Cell Biol* 3:210–214.
- Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95:6239–6244.
- Giaever G, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418:387–391.
- Nagalakshmi U, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349.
- Uetz P, et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623–627.
- Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411:41–42.
- Stark C, et al. (2006) BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–D539.
- Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein–interaction networks. *Mol Biol Evol* 22:803–806.
- Wagner A (2000) Robustness against mutations in genetic networks of yeast. *Nat Genet* 24:355–361.
- Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. *Trends Genet* 25:152–155.
- Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 102:14332–14337.
- Esser C, Martin W, Dagan T (2007) The origin of mitochondria in light of a fluid prokaryotic chromosome model. *Biol Lett* 3:180–184.
- Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* 9:605–618.
- Dujon B, et al. (2004) Genome evolution in yeasts. *Nature* 430:35–44.
- Hall C, Brachat S, Dietrich FS (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell* 4:1102–1115.
- Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806.
- Qin H, Lu HHS, Wu WB, Li WH (2003) Evolution of the yeast protein interaction network. *Proc Natl Acad Sci USA* 100:12820–12824.
- Searcy DG (2006) Rapid hydrogen sulfide consumption by *Tetrahymena pyriformis* and its implications for the origin of mitochondria. *Eur J Protistol* 42:221–231.
- Brocks JJ, Logan GA, Buick R, Summons RE (1999) Archean molecular fossils and the early rise of eukaryotes. *Science* 285:1033–1036.
- Soltis DE, Soltis PS (1999) Polyploidy: Recurrent formation and genome evolution. *Trends Ecol Evol* 14:348–352.
- Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
- Janssen PJ, et al. (2003) COmplete GENome Tracking (COGENT): A flexible data environment for computational genomics. *Bioinformatics* 19:1451–1452.
- Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
- Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene Ontology: Tool for the unification of biology. *Nat Genet* 25:25–29.
- Batagelj V, Mrvar A (1998) Pajek—Program for Large Network Analysis. *Connections* 21:47–57.
- Nash R, et al. (2007) Expanded protein information at SGD: New pages and proteome browser. *Nucleic Acids Res* 35(Database issue):D468–D471.
- Güldener U, et al. (2005) CYGD: The Comprehensive Yeast Genome Database. *Nucleic Acids Res* 33(Database issue):D364–D368.
- Agresti A (2002) *Categorical Data Analysis* (Wiley, Hoboken, NJ), 2nd Ed.