A Genome Phylogeny for Mitochondria Among α-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes

Christian Esser, Nahal Ahmadinejad,* Christian Wiegand,* Carmen Rotte,* Federico Sebastiani,* Gabriel Gelius-Dietrich,* Katrin Henze,* Ernst Kretschmann,† Erik Richly,‡ Dario Leister,‡ David Bryant,§ Michael A. Steel,|| Peter J. Lockhart,¶ David Penny,¶ and William Martin*[1]*

*Institute of Botany III, University of Düsseldorf, Düsseldorf, Germany; †European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom; ‡ Max-Planck Institut für Züchtungsforschung, Köln, Germany; §McGill Centre for Bioinformatics, Montréal, Québec, Canada; ||Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand; and ¶Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Analyses of 55 individual and 31 concatenated protein data sets encoded in *Reclinomonas americana* and *Marchantia polymorpha* mitochondrial genomes revealed that current methods for constructing phylogenetic trees are insufficiently sensitive (or artifact-insensitive) to ascertain the sister of mitochondria among the current sample of eight α-proteobacterial genomes using mitochondrially-encoded proteins. However, *Rhodospirillum rubrum* came as close to mitochondria as any α-proteobacterium investigated. This prompted a search for methods to directly compare eukaryotic genomes to their prokaryotic counterparts to investigate the origin of the mitochondrion and its host from the standpoint of nuclear genes. We examined pairwise amino acid sequence identity in comparisons of 6,214 nuclear protein-coding genes from *Saccharomyces cerevisiae* to 177,117 proteins encoded in sequenced genomes from 45 eubacteria and 15 archaebacteria. The results reveal that ~75% of yeast genes having homologues among the present prokaryotic sample share greater amino acid sequence identity to eubacterial than to archaebacterial homologues. At high stringency comparisons, only the eubacterial component of the yeast genome is detectable. Our findings indicate that at the levels of overall amino acid sequence identity and gene content, yeast shares a sister-group relationship with eubacteria, not with archaebacteria, in contrast to the current phylogenetic paradigm based on ribosomal RNA. Among eubacteria and archaebacteria, proteobacterial and methanogen genomes, respectively, shared more similarity with the yeast genome than other prokaryotic genomes surveyed.

Introduction

The current paradigm for the relatedness of eubacteria, archaebacteria, and eukaryotes is the small subunit ribosomal RNA (rRNA) tree, also called the universal tree or the tree of life. In the rRNA tree, eukaryotes are depicted as sisters to the archaebacteria (Woese, Kandler, and Wheelis 1990; Woese 2002), based on the rootings proposed with protein sequence comparisons (Gogarten et al. 1989; Iwabe et al. 1989). Yet the sister-group relationship between archaebacteria and eukaryotes implied in the rRNA tree is reflected only in some eukaryotic genes. Many genes in eukaryotes are more closely related to their eubacterial homologues than they are to their archaebacterial homologues (Doolittle and Brown 1994; Brown and Doolittle 1997; Feng, Cho, and Doolittle 1997; Brown 2003; Timmis et al. 2004). In an early evolutionary analysis of the yeast genome, Rivera et al. (1998) compared yeast proteins to the homologues from five sequenced prokaryotic genomes that were available at the time. They found that many yeast genes involved in transcription, translation, DNA maintenance, and the like ("informational" genes) were more similar to archaebacterial homologues, whereas many genes involved in biosyntheses, metabolism, and the like ("operational" genes) were more similar to eubacterial homologues.

Those studies indicated that there are many more eubacterial genes in the yeast genome (and in eukaryotic genomes in general) than would be expected on the basis of the rRNA paradigm. Although the precise number, nature, and origin of these genes have yet to be specifically pinned down, their presence is now widely accepted to indicate some kind of chimaerism during eukaryotic evolution (Brown 2003). Chimaerism poses challenging and yet unsolved problems, regarding both the classification of unicellular organisms (Doolittle 1999) and the reconstruction of early eukaryotic evolution (Knoll 2003). It has spawned models in which additional endosymbiotic partners are invoked to explain the origins of these genes in a lump-sum fashion, regardless of their specific similarity patterns (Hedges et al. 2001; Horiike et al. 2001; Hartman and Federov 2002), and models in which lateral gene transfer (LGT) is invoked to explain the origins of these eukaryotic genes on a one-acquisition-at-a-time basis (Doolittle 1998; Gogarten 2003).

Yet LGT as a vehicle to explain the excess eubacterial genes in eukaryotes involves the assumptions that the interpretation of individual gene trees is straightforward and that the reconstruction of gene trees is, at the extreme, infallible. That is, the LGT explanation for unexpected branching orders assumes not only that each gene is fully capable of accurately telling the story of its evolutionary history in the language of sequence comparisons, but furthermore that each gene does so when queried with existing phylogenetic techniques. Warnings that the resolving power of gene tree analysis has discrete limits (Meyer, Cusanovich, and Kamen 1986; Penny and Hendy 1986; Rothschild et al. 1986; Nei 1996; Embley and Hirt 1998; Philippe and Laurent 1998; Penny et al. 2001; Sober and Steel 2002; Mossel 2003) have been issued, and newer findings (Rokas et al. 2003) reinforce the older view (Penny, Foulds, and Hendy 1982) that minor topology differences among proteins sharing the same evolutionary

Key words: Endosymbiosis, Genome analysis, mitochondria, origin of eukaryotes, archaebacteria, eubacteria.

[1] E-mail: w.martin@uni-duesseldorf.de.

Mol. Biol. Evol. 21(9):1643–1660. 2004
doi:10.1093/molbev/msh160
Advance Access publication May 21, 2004

history are not surprising, rather they are to be expected even in the absence of LGT.

For the study of early cell evolution, there are only three generally accepted theories within the framework of which biologists can comfortably work: Darwinian theory (natural variation and descent with modification exists among microbes), phylogenetic theory (sequence similarity reflects in some manner evolutionary history), and endosymbiotic theory (some organelles of eukaryotic cells were once free-living prokaryotes).

In terms of endosymbiotic theory, which explains the origin of double membrane-bounded organelles in eukaryotes—chloroplasts and mitochondria, including hydrogenosomes (Embley et al. 2003; Müller 2003)—the excess eubacterial genes in eukaryotes bear on our concepts concerning the host that acquired mitochondria (Martin et al. 2001). Several models have been put forward to explain the origin of eukaryotes in a manner that could, in principle, account for the presence of too many eubacterial genes in eukaryotic genomes by virtue of the intracellular relocation of genes in the context of a symbiotic association (endosymbiotic gene transfer).

Such models generate and in some cases explicitly spell out predictions about the overall patterns of similarity that should be observable in genome sequence comparisons. For example, some models predict that eukaryotic nuclear genes should bear greatest overall similarity to their homologues from (1) methanogens and δ-proteobacteria (Moreira and Lopez-Garcia 1998), (2) actinobacteria (a group of Gram positive bacteria encompassing streptomycetes and relatives) (Cavalier-Smith 2002), (3) *Thermoplasma* and spirochaetes (Margulis, Dolan, and Guerrero 2000), (4) proteobacteria and eocytes (a group of archaebacteria also called crenarchaeotes) (Gupta 1998), or (5) methanogens and α-proteobacteria (Martin and Müller 1998).

Genome sequence sampling among those prokaryotic lineages is still quite sparse, yet even if it were dense, appropriate methods to detect or quantify overall sequence similarity at the whole genome level have not been well developed, although methods that detect overall similarity in dinucleotide frequencies have (Karlin et al. 1999). Here we report overall amino acid sequence similarity between proteins in the yeast nuclear genome and identifiable homologues in 60 prokaryotic genomes. We examine the chimaeric nature of the yeast nuclear genome and report the phylogenetic position of mitochondria among a sample of 10 α-proteobacterial genomes.

Methods

Analysis of Mitochondrion-Encoded Proteins Versus α-Proteobacterial Homologues

The 67 protein-coding genes of the *Reclinomonas americana* mitochondrial genome were compared by local FASTA search to the proteins from 48 sequenced eubacterial genomes, including the α-proteobacteria *Sinorhizobium meliloti*, *Mesorhizobium loti*, *Agrobacterium tumefaciens*, *Caulobacter crecentus*, *Brucella melitensis*, *Magnetococcus* sp. MC1, *Wolbachia w*Mel, *Rickettsia prowazeckii*, and *Rickettsia conorii*. The set of proteins common to the *Reclinomonas* and *Marchantia polymorpha* mitochondrial genomes were compared to the sequenced α-protobacterial genomes as well as to the partial genome sequence data from *Novosphingobuim aromaticivorans*, *Rhodobacter sphaeroides*, *Rhodospirillum rubrum*, and *Magnetospirillum magnetotacticum*. Sequence data from the latter four genomes were generously produced by the U.S. Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/). When more than one match per genome was detected, the best match to the mitochondrial query was used, thus allowing each α-proteobacterium to be as similar to mitochondria as possible at the level of sequence similarity, regardless of whether the sequence similarity of individual genes is due to vertical inheritance or lateral acquisition.

Sequences were aligned with ClustalW (Thompson, Higgins, and Gibson 1994) with gap open penalty 15.0, gap extension penalty 6.66, and the BLOSUM series weight matrix. Protein log-determinant (LogDet) distances (Lockhart et al. 1994) were determined with LDDist (Thollesson 2004); the fraction of invariant sites was estimated and excluded using the methods of Sidow, Nguyen, and Speed (1992) or Steel, Huson, and Lockhart (2000) as implemented in LDDist. Neighbor-Joining (NJ; Saitou and Nei 1987) was used to infer trees from distance data. Splits were detected from the distance matrix with NeighborNet (NNet; Bryant and Moulton 2004) and represented as planar graphs with SplitsTree (Huson 1998). Protein maximum likelihood trees were constructed with ProtML (Adachi and Hasegawa 1996).

For concatenated analyses, the *cox1*, *cox2*, and *cox3* genes of *R. rubrum* were not available in the partial genome data. *Magnetospirillum* homologues for *rps11* and *rps13* were also missing. However, in individual ProtML analyses, *Rhodospirillum* and *Magnetospirillum* were almost always well-supported sisters (see supplemental table S1 in the online Supplementary Material). Therefore, for the concatenated data set, *Magnetospirillum* homologues were removed from the data except in the case of *cox1*, *cox2*, and *cox3*, where the *Magnetospirillum* homologues were substituted for the missing *Rhodospirillum* sequences. Because *nad6* was missing in the available *Novosphingobium* data, *nad6* was excluded, yielding 31 genes (*atp1*, *atp6*, *atp9*, *cob*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad9*, *rpl16*, *rpl2*, *rpl5*, *rpl6*, *rps1*, *rps11*, *rps12*, *rps13*, *rps14*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *yejr*, and *yeju*) for 14 OTUs (operational taxonomical units; two mitochondria, 10 α-proteobacteria, and two outgroups: *Escherichia* and *Neisseria*).

The 31 homologues per genome were aligned individually and then concatenated to produce the initial 14 OTU concatenated alignment of 12,445 amino acid sites per genome, which included many gaps. Removing all gapped sites produced the 6,472-site data set. The 6,472-site data set had severe amino acid content heterogeneity as determined with PUZZLE (Strimmer and von Haeseler 1996); all 14 OTUs failed the χ^2 test for homogeneous amino acid composition at $P = 0.95$ except *Agrobacterium* and *Mesorhizobium*. By removing the most highly variable sites using the method described

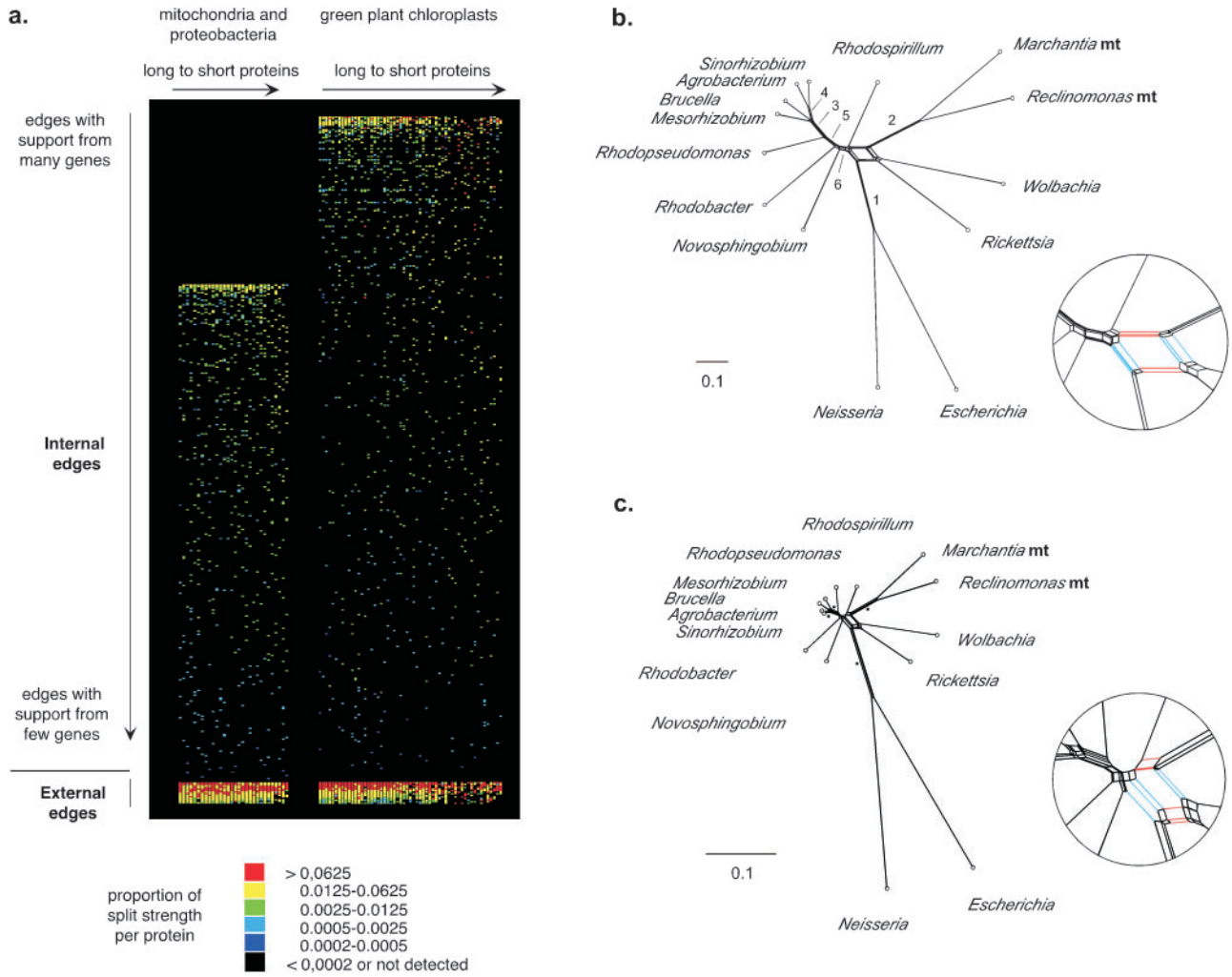


FIG. 1.—Comparisons of mitochondrial-encoded proteins with α-proteobacterial homologues. (*a*) Strength of splits in comparisons of mitochondrial encoded proteins versus α-proteobacterial homologues (left panel) and in comparisons of chloroplast-encoded proteins among green algae and land plants (right panel). The proportion of split strength (individual split strength divided by sum of split strengths per column) is color-coded. External edges are grouped at the bottom and internal edges are sorted top to bottom by row-wise sum of split strength. Proteins are sorted from left to right by the length of their corresponding alignment. Data available upon request. (*b*) NeighborNet planar graph of protein LogDet distances with invariant sites excluded for an alignment of 31 proteins common to *Reclinomonas* and *Marchantia* mitochondrial genomes and present in many α-proteobacterial genomes. All gapped sites were removed prior to analysis, leaving 6,472 amino acids sites per genome. (*c*) NeighborNet planar graph of protein LogDet distances with invariant sites excluded for the 2,500 least polymorphic positions of the alignment in (*a*), which was the longest data set found in which all sequences passed the χ^2 test for amino acid compositional homogeneity. The splits highlighted in blue and red in the inset of (*a*) and (*b*) are those that link mitochondria with free-living α-proteobacteria and to parasitic α-proteobacteria, respectively. Splits that were found in >95/100 bootstrap samples are marked with a black dot.

(Hansmann and Martin 2000), we identified the largest data set in which all sequences passed the χ^2 test as the least polymorphic 2,500 positions (the 2,500-site data).

The reference spectrum of splits in chloroplast proteins was determined for the alignments of rpoC1, psaA, psaB, rpl2, rpoA, psbB, atpA, cemA, atpB, psbC, rbcLg, ccsA, psbA, petA, rps3, rps2, atpI, petB, clpP, rps4, atpF, ycf4, petD, ycf3, rps18, rps7, rps11, rpl16, atpE, rps8, rpl20, rps12, rpl14, infA, rps14, rps19, rpl23, psbH, psbE, atpH, psaC, petL, psbZ, psbK, psaI, psbI, psaJ, psbN, psbJ, psbF, psbT, psbL, rpl36, petG, and psbM (listed as they appear from left to right in fig. 1*a*) from chloroplast genome sequences from the green algae *Chlorella vulgaris*, *Nephroselmis olivacea*, *Mesostigma viridae*, and *Chaetosphaeridium globosum*; the byrophytes *Marchantia polymorpha* and *Anthoceros formosae*; the fern *Psilotum nudum*; the gymnosperm *Pinus thunbergii*; and the angiosperms *Triticum aestivum*, *Oryza sativa*, *Zea mays*, *Castanea sativa*, *Spinacia oleracea*, and *Nicotiana tabacum* (accession numbers available from http://megasun.bch.umontreal.ca/ogmpproj.html; all alignments available upon request).

Analysis of Yeast Nuclear-Encoded Proteins Versus Prokaryotic Homologues

The set of 6,214 nuclear protein-coding genes from yeast were taken from http://www.ebi.ac.uk/proteome/. The search set (177,117 proteins) was obtained from http://www.tigr.org; it contained 143,842 proteins from 45

eubacteria and 33,275 proteins from 15 archaebacteria. In separate directories for each genome, an unambiguous species identifier was written into the sequence name following ">" in the FASTA-format files to facilitate later analyses. Sequences were converted into GCG format (Wisconsin Package version 10.3, Accelrys Inc., San Diego, Calif.) and copied into a single directory so that scores and E-values would be directly comparable. The yeast proteins were compared to the prokaryotic proteins using the Pearson-Lipman (1988) search as implemented in the FASTA program of the Wisconsin package. In each FASTA output (one per yeast query), the best scoring protein per prokaryotic genome was noted along with its E-value (the expected number of chance alignments with scores $\geq$ that observed) and the percent amino acid identity (e.g., 42.2%) in the pairwise local alignment (Smith and Waterman 1981) employed by FASTA. For the specified E-value threshold 10^{-x}, the percentage amino acid identity values (pI_x) for each pairwise comparison were written into a table with 6,214 rows specified by the yeast gene identifiers and 60 columns specified by the prokaryotic genomes. Empty elements of the matrix were written as zero. Sums of columns define total percent identity (tI) for the given genome at the E-value threshold of 10^{-x} (tI_x).

To determine whether yeast proteins were distributed more specifically among eubacterial or archaebacterial genomes, the sum of the 45 eubacterial pI_{20} values was divided by the sum of the 15 archaebacterial pI_{20} values, multiplied by 45/15 for normalization, and rows were sorted by that quotient (1 was added to zero denominators). Values of pI_{20} were colour-coded after removal of all rows containing only empty elements. Functional category assignments and gene product definitions were taken from EBI data for the yeast gene identifiers. Mitochondrial and sec-pathway targeting prediction was performed as described (Richly, Chinnery, and Leister 2003). Taxonomic designations for prokaryotic groups were taken from http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html/. Categories of yeast importers were assigned and assorted by hand from information present in the product definition line. All results are available upon request.

Results and Discussion

Mitochondrial Origins Are Unresolved by Mitochondrial Proteins, but *Rhodospirillum* Is Close

Rickettsia is often asserted to be the closest relative to mitochondria among α-proteobacteria because a few genes have produced that phylogenetic result (Kurland and Andersson 2000; Emelyanov 2003), although the genome sequence of *Wolbachia pipientis w*Mel recently revealed that *Rickettsia* is the sister of *Wolbachia*, not of mitochondria (Wu et al. 2004). In addition, genome-wide comparisons of mitochondrial-encoded proteins to their α-proteobacterial homologues employing many proteins from genome sequence data have been lacking. Using FASTA in an initial survey, we compared the 67 proteins in the *Reclinomonas americana* mitochondrial genome to all proteins from 48 completely or partially sequenced eubacterial genomes including the α-protobacteria *Bradyrhizobium*, *Sinorhizobium*, *Mesorhizobium*, *Caulobacter*, *Brucella*, two *Rickettsia* species, and *Wolbachia*. Twelve proteins did not give a match with an E-value better than 10^{-5} in more than six genomes (*atp8, rpl10, rpl18, rpl19, rpl31, orf169, orf717, orf25, orf64, rpoB, rpoC*, and *rpoD*) and were excluded from further analysis due to their poor sequence conservation. The 55 proteins that gave an E-value better than 10^{-5} in more than six genomes were aligned and investigated with protein LogDet distances and NJ trees. The *Reclinomonas* protein branched with homologues from *Rickettsia* species in 5 trees, with homologues from *Wolbachia* in 10 trees, basal to *Rickettsia* and *Wolbachia* in 5 trees, with other α-proteobacteria or groupings thereof in 16 trees, and not with homologues from any α-proteobacterium in 19 trees with bootstrap proportions less than 70% for 53 of the 55 proteins (see supplemental table S2 online). Recalling that the *Reclinomonas* mitochondrion inherited its genome from α-proteobacteria (Lang et al. 1997; Gray, Burger, and Lang 1999) rather than having acquired it through lateral acquisition from various donors, such disparate results could mean (1) that a degree of noise exists in the data (for example, due to poor conservation, as in the case of the twelve proteins that were excluded for lack of good homologues); (2) that the phylogenetic method is producing an imperfect estimation of the phylogeny, producing artifacts in some cases, but getting close to the true position in other cases; (3) that any number of problems inherent to phylogeny reconstruction, such as model misspecification or poor sampling, were present; (4) that the eubacteria sampled might be avidly exchanging these genes over time; or (5) any combination of the above.

Pinning down the relative contributions of these factors in the absence of a priori knowledge about how proteins evolve is not trivial. We took an empirical approach. Since all available evidence indicates mitochondria to have a single origin (Lang, Gray, and Burger 1999), including an additional mitochondrial genome in the sample should help, because if the two mitochondria do not branch together, something must be wrong. (In this way, endosymbiosis can be used as a control for phylogenetics.) Thus we included the homologues from the *Marchantia polymorpha* mitochondrial genome. To improve the α-proteobacterial sampling, we included data from partially sequenced or unpublished genomes (see *Methods*). We also tried to improve the alignment procedure by limiting the sample to more closely related sequences (α-proteobacterial genomes and two γ-protobacterial outgroups). Finally, we tried to improve the uniformity of the data by having approximately the same set of genomes represented in each alignment. This identified 31 proteins that are common to the *Reclinomonas* and *Marchantia* mitochondrial genomes and that are uniformly present (except *cox1-3*, see *Methods*) in data from 10 sequenced or partially sequenced α-proteobacterial genomes and two outgroups.

Individual analysis using protein LogDet and NJ for these 31 proteins revealed similarly disparate results (table 1), as in the initial analysis (supplemental table S2 online). The mitochondrial proteins branched in five trees with the group (*Rickettsia*, *Wolbachia*), in three trees with *Rickettsia*,

in four trees with *Wolbachia*, in three trees with *Rhodospirillum*, in one tree with *Rhodobacter*, in four trees basal to the free-living α-proteobacteria, in three trees with the outgroup, in three trees elsewhere, and in five trees mitochondria were not monophyletic (table 1). Although one explanation for such disparate results for these mitochondrially-encoded proteins might be LGT to mitochondria from these various sources, we suspect that difficult alignments and poor phylogenetic signal in these phylogenies of ancient events, compounded by the inadequacies of currently available phylogenetic methods (Penny et al. 2001), are the more likely cause(s). It is currently not clear how to directly show that this might be true.

Is LGT or Divergence a Better Explanation of Mitochondrial Genome Evolution?

To test the impact of LGT on the present mitochondrial data we examined the spectrum of phylogenetic signals for a given set of data and compared it to the spectrum obtained for a well-established phylogeny that contains sequences of different degrees of divergence. In doing this, phylogenetic signal could be expressed in many ways. A particularly convenient way is in terms of tree-splits (which are equivalent to edges, branches, or bipartitions). Internal splits separate OTUs (operational taxonomical units, in this case sequences) into two groups. External splits separate a single OTU from all other OTUs. Real data usually contain external splits and conflicting internal splits. For small numbers of OTUs, support for all splits under a specified model of sequence evolution can be calculated directly using a Hadamard transformation (Penny et al. 1996; Lockhart et al. 1999). For larger data sets, one heuristic approach is to use NNet, which provides a list of major splits, including conflicting splits, not just the splits that are pairwise compatible and thus fit onto a single bifurcating tree, whereby the degree of pairwise compatibility is a measure of how well the given split fits the data (Bryant and Moulton 2004).

To establish the reference spectrum, we used NNet and the LogDet correction (Lockhart et al. 1994) to identify the strongest splits in 57 proteins whose orthologues appeared in 14 green algal and higher plant chloroplast genomes (see *Methods*). We compared the strength and relative frequency of splits common to the different chloroplast proteins and found that many of the strong splits shared by the longer proteins corresponded to internal edges (fig. 1*a*, right panel). This is encouraging, because it indicates that these chloroplast proteins, which share a common evolutionary history (Martin et al. 2002), also share a detectable degree of common phylogenetic signal. Notably, many of the shorter chloroplast proteins also had some well-supported splits that were not found in the longer proteins, a finding which is likely due to sampling error inherent in short (<50 residues long) proteins.

Next, we compared our reference spectrum (the chloroplast proteins) to the spectrum obtained using the same approach for 31 protein data sets compiled from the two mitochondrial and 12 sequenced or partially sequenced proteobacterial genomes (fig. 1*a*, left panel). We reasoned that since the strength of phylogenetic signal in a given protein should decrease with time, as predicted in theory (Penny et al. 2001; Sober and Steel 2002; Mossel 2003), the spectrum for these data should show weaker support for internal tree splits than in the case of our reference spectrum. The reason for this is that the green-lineage divergence spans at most about 1 billion years of evolution, whereas the mitochondrial and proteobacterial protein comparisons span a much greater amount of evolutionary time, perhaps about 2 billion years (Knoll 2003; Martin et al. 2003). The results show that the strongest splits in the mitochondrial and proteobacterial

Table 1
LogDet-NJ Resampling Results for 31 Mitochondrial Proteins (14 OTU Data)

Gene	Sister of Mitochondria	With BP	(ric, wol) Sister BP	2 Mitochondria Sister BP	2 Outgroups Sister BP
atp1	all free-living[a]	89	—	85	100
cob	all free-living[a]	—	74	98	90
cox1	all free-living[a,b]	—	56	100	100
nad4	all free-living[a]	61	71	100	—
atp6	ric, wol	—	57	97	100
rpl5	ric, wol	—	—	62	78
nad5	ric, wol	51	83	100	—
rps2	ric, wol	—	—	81	—
yeju	ric, wol	—	51	75	—
atp9	ric, wol, rrub	—	55	79	72
cox2	rrub[b]	62	—	100	98
cox3	rrub[b]	—	—	75	93
nad1	rrub	80	82	91	—
rps12	rsph	—	—	—	82
rps3	ric	—	—	74	85
rpl2	ric	—	—	75	87
rpl6	ric	—	—	—	70
nad2	wol	—	—	99	—
rps1	wol	88	—	87	100
rps11	wol	—	—	64	93
rps13	wol	—	—	61	82
nad4l	group of six[c]	—	—	83	—
rps14	mt n.m.[d]	—	—	—	58
rps19	mt n.m.[d]	—	—	57	74
nad3	mt n.m.[d]	—	—	—	og n.m.[e]
rpl16	mt n.m.[d]	—	—	—	99
yejr	mt n.m.[d]	—	—	—	og n.m.[e]
nad9	outgroup	—	—	93	og n.m.[e]
rps4	outgroup	53	—	73	100
rps8	outgroup	—	—	76	94
rps7	eco	—	—	56	og n.m.[e]

NOTE.—Abbreviations: ric: *Rickettsia prowazekii*, wol: *Wolbachia w*Mel, rrub: *Rhodospirillum rubrum*, eco: *Escherichia coli*, rsph: *Rhodobacter sphaeroides*. BP values indicate number of times that the branch was found in 100 bootstrap samples of the sequence data; values less than 50 are indicated as a dash. The fraction of invariant sites was estimated and excluded using the method of Sidow, Nguyen, and Speed (1992).

[a] The mitochondria branch basal as the sister to all α-proteobacteria except *Rickettsia* and *Wolbachia*.

[b] The *Magnetospirillum* homologues were substituted for the missing *Rhodospirillum* sequences as explained in *Methods*; see also table S1 in the online Supplementary Material.

[c] The group *Agrobacterium*, *Sinorhizobium*, *Brucella*, *Mesorhizobium*, *Novosphingobium*, *Rhodobacter*.

[d] Mitochondria (*Reclinomonas* and *Marchantia*) not monophyletic (abbreviated as mt n.m.).

[e] Outgroup (*Escherichia* and *Neisseria* not monophyletic (abbreviated as og n.m.).

data are all among the external edges (terminal branches), and only weaker splits are seen among the internal edges.

Of the 8,177 possible internal splits for the 14 mitochondrial and proteobacterial OTUs, only 443 are observed. Six of these splits occur very frequently and are shown in the NNet graph (fig. 1*b*). By way of comparison we observe that in the chloroplast data, 10 splits occur frequently (top of each panel) for the same number of OTUs. Overall these results are encouraging given the extent of sequence divergence between proteobacterial and mitochondrial homologues, because a 14 OTU tree has only 11 internal tree splits. If LGT were more prevalent than common ancestry for these proteins across genomes, we would not expect to see a set of frequently shared splits across proteins (the splits would be randomly distributed). Only if common ancestry were widespread among these protein data sets would we expect to observe such a shared set of splits as are seen in figure 1*a*. Indeed, the probability of observing, by chance, the rather discrete set of only 443 different splits, 233 of which are shared by two or more proteins (some incompatible), across 31 genes can be estimated by standard probabilistic arguments. The total number of observed splits summed across the 31 mitochondrial genes in figure 1*a* is 738. If these 31 sets of splits were random with respect to each other (e.g., through LGT), then the probability of observing just 443 or fewer internal splits (from the 8,177 possible for each 14 OTU data set) can be estimated by the Azuma-Hoeffding inequality (Alon and Spencer 1992) applied to a bin occupancy problem. The calculated probability that we would observe this many shared internal splits by chance only is small ($P < 0.001$), providing evidence for a significant degree of common ancestry among the 31 genes under investigation with the present methods.

Furthermore, if common ancestry were widespread, but difficult to detect with LogDet distances due to conflicting signal in the data (rather than due to LGT), then we would expect to observe, in addition to a set of shared splits, a set of spurious splits as well, which should be more or less randomly distributed among proteins, just as is observed (fig. 1*a*). The six strongest internal splits observed in individual analyses of the 31 proteins (the top six splits in fig. 1*a*) are labeled in figure 1*b*. The remaining splits detected in individual analyses were either rare—210 occurring for one protein only—or were conflicting, or both.

Part of the conflicting signal is likely due to noise stemming from the many highly gapped and poorly alignable regions in the individual alignments. Hence, the proteins were concatenated into a single 14 OTU alignment with 12,445 sites and all sites that contained a gap in any sequence were removed, leaving 6,472 positions for analysis. The NNet of protein LogDet distances for the 6,472-site data (fig. 1*b*) shows good support for the monophyly of the two mitochondria, the unity of the outgroup, and several seemingly robust affiliations among members of the α-proteobacteria sampled. Furthermore, the six splits commonly detected in the individual analyses of gapped data map nicely onto the tree of concatenated sequences lacking gaps (fig. 1*b*). However, the position of the mitochondria remained unresolved, with one split linking them to *Rickettsia* and *Wolbachia* and one split linking them to the free-living forms (highlighted in the inset in red and blue, respectively).

Although LogDet can compensate for amino acid composition bias when the spatial distribution of substitutions is simple (Lockhart et al. 1999), the compositional heterogeneity in the 6,472-site data was very severe, with only two OTUs passing the χ^2 test for compositional homogeneity. Removing highly variable sites (see *Methods*) produced the compositionally homogeneous 2,500-site data, in which the NNet split associating mitochondria with the free-living α-proteobacteria increased in strength relative to the split associating mitochondria to *Rickettsia* and *Wolbachia*. The position of the mitochondria was still unresolved, although *Rhodospirillum rubrum* was slightly closer to mitochondria than the other α-proteobacteria, sharing a small split with *Marchantia* (fig. 1*c*). It is worth noting that the overall fermentative physiology of *Rhodospirillum* and related genera is quite similar in overall design to that in eukaryotes that lack mitochondria or possess anaerobic mitochondria, because the main fermentative end products in this group of α-proteobacteria are acetate, succinate, propionate, lactate, formate, H_2, and CO_2 (Imhoff and Trüper 1992), an overall physiolology that is virtually identical to that found among eukaryotes that lack mitochondria (Müller 2003) and that possess anaerobic mitochondria (Tielens et al. 2002). In the bootstrap consensus of LogDet NJ trees for the 2,500-site data, *Rhodospirillum* branched as the sister to the two mitochondria in 65/100 replicates.

With additional sampling of α-proteobacterial lineages and with improved methods of phylogenetic inference it might be possible to link mitochondria to specific members of the group using the information contained in mitochondrial genomes. Yet it might also turn out that more data per species will be necessary to clarify the origin of mitochondria. Since the current set of 31 proteins contains about as much information as mitochondrial genomes have to offer when two mitochondria are included in the analysis (Gray, Burger, and Lang 1999), the possibility to resolve the issue from mitochondrial genome information might face a fundamental limitation. Hence we asked whether the origin of mitochondria could, in principle, be addressed with data in nuclear genomes.

Comparison of Yeast Proteins to Prokaryotic Homologues

No evolutionary analysis is assumption-free. Here we assume that the origin of the prokaryotic lineages (archaebacteria and eubacteria) predates that of eukaryotic cells. The reasoning behind this premise is as follows. We accept the evidence indicating that all known eukaryotes possess a mitochondrion, a hydrogenosome (anaerobic forms of mitochondria), or a mitosome (highly reduced forms of mitochondria with apparently no direct involvement in ATP synthesis), or that they possessed one in their evolutionary past (Roger and Silbermann 2002; Embley et al. 2003; Tovar et al. 2003). Furthermore, we accept the biochemical (John and Whatley 1975) and molecular evolutionary evidence (Gray, Burger, and Lang

1999; Lang, Gray, and Burger 1999) indicating that mitochondria arose from within a group of prokaryotes called α-proteobacteria (Stackebrandt, Murray, and Trüper 1988). Therefore, the eukaryotes we know (including yeast) must have diversified subsequent to the diversification of eubacteria and probably subsequent to the diversification of α-proteobacteria from other related lineages. Regarding archaebacteria, the isotopic trace of ultralight carbon (an indicator of methanogenesis, a typically archaebacterial pathway) goes back just as far in the geochemical record as the trace of nonmethanogenic carbon fixation does (Nisbet and Fowler 1999; Nisbet and Sleep 2001), indicating in a very straightforward manner that archaebacteria are about as old as eubacteria. Furthermore, chemolithoautotrophy (the ability to make ATP via chemiosmosis with the help of redox reactions involving only inorganic electron donors and acceptors while using CO_2 as a sole carbon source) is widespread among both groups of prokaryotes but is lacking altogether in eukaryotes, all of which depend entirely upon prokaryotic CO_2 fixation pathways as a source of reduced carbon compounds. For these reasons, the origin of eukaryotic genes should postdate the origin of prokaryotes. Given that, we asked: Among yeast genes that have prokaryotic homologues by the measure of sequence similarity, to which prokaryotic homologues are they most similar?

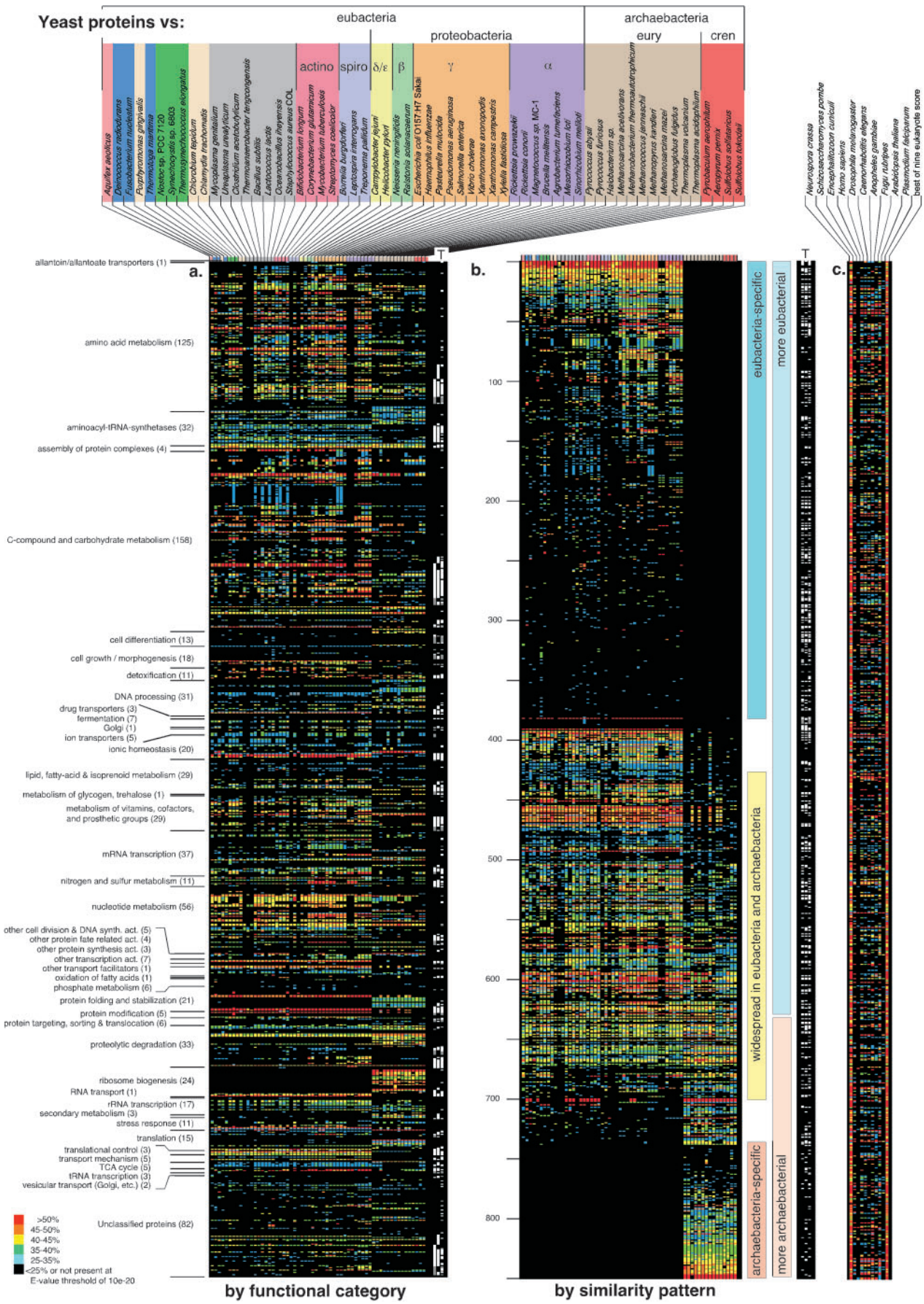
Of the 6,214 yeast proteins examined, only 850 find a match in FASTA comparison to one of the 177,117 prokaryotic proteins from 60 genomes in the search set at an E-value threshold of 10^{-20} and have at least 25% amino acid sequence identity in the Smith-Waterman pairwise alignment that FASTA performs. Figure 2*a* shows the percentage amino acid identity at this threshold (pI_{20}) for yeast proteins, ranked by functional category. A look at the figure reveals that some functional categories are mostly archaebacterial (e.g., ribosome biogenesis) or mostly eubacterial (e.g., C-compound and carbohydrate metabolism or nucleotide metabolism). However, from the standpoint of the sequence similarity of individual proteins to prokaryotic homologues, and from the standpoint of the distribution of those genes across prokaryotic genomes, all of the functional categories assigned in the yeast annotation clearly have a mixed ancestry. This finding contrasts sharply to an earlier analysis in which a lumpsum majority consensus for eubacterial or archeabacterial ancestry was inferred for each category (Horiike et al. 2001; see also Poole and Penny [2001] and Rotte and Martin [2001] for a discussion). A tab-delimited table containing all information represented in figure 2*a* is available upon request.

The mitochondrial and sec-pathway targeting predictions with three programs (TargetP, Pedotar, and iPSORT) are largely congruent ("T" lanes on the right of fig. 2*a*) and make sense for the most part. For example, the proteins of oxidative phyosphorylation are predicted to be mitochondrial (large white block in the category C-compound and carbohydrate metabolism). Yet these programs still make some evident targeting prediction errors. For example, the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase is predicted by Target P and iPSORT but not by Pedotar to be mitochondrial, although it is generally regarded as a cytosolic enzyme. Curiously, however, the highly conserved N-terminus of cytosolic GAPDH does in fact serve as a mitochondrial targeting sequence in potato (Long et al. 1996), and the enzymes of the glycolytic pathway are specifically localized to the outside of mitochondria in *Arabidopsis* (Giegé et al. 2003).

To obtain a clearer picture of the global patterns of sequence similarity in yeast proteins we summed the elements of the matrix in figure 2*a* for the eubacterial and archaebacteria matches, respectively, and sorted the genes by the resulting quotient, normalized for the smaller archaebacteria sample. Doing this had the effect of sorting homologues based on their patterns of sequence identity and, hence, likely sources of origin at the level of archaebacterial versus eubacterial ancestry (fig. 2*b*). Several aspects of the diagram are noteworthy.

Reading fig. 2*b* from the top down, 383 of these 850 yeast proteins have homologues in eubacterial genomes but not in archaebacterial genomes. We designate these proteins as eubacteria-specific. Obviously, this designation is tentative because it is dependent on taxon sampling—if an archaebacterial genome sequence becomes available that contains one of these eubacterial-specific proteins, the designation will no longer hold. From the top down, the first protein that also occurs in archeabacteria is a mitochondrial ATP synthase α chain homologue in the *Aeropyrum* genome with 29.3% identity to atpa_yeast. Reading figure 2*b* from the bottom up, 111 yeast proteins have homologues in archaebacterial genomes, but not eubacterial genomes, before the first eubacterial match appears, which is msp1_yeast (TAT-binding homolog 4), having a homologue in the *Nostoc* genome with 37.1% identity, followed by hmd1_yeast and hmd2_yeast (3-hydroxy-3-methylglutaryl-coenzyme A reductase 1 and 2), which have a homologue in the *Vibrio cholerae* genome.

From the overall sequence similarity and distribution patterns of these homologues across prokaryotic genomes, it is evident that the yeast genes listed from the top of figure 2*b* down to around position 620–650 are shared only with eubacteria or are more similar to eubacterial homologues and are more broadly distributed among eubacterial than among archaebacterial chromosomes. From about position 620–650 to the bottom of figure 2*b*, the converse is true, that is, those yeast proteins are more archaebacterial in nature. On the basis of the current prokaryotic genome sample, about three fourths (75%) of yeast proteins at the 10^{-20} E-value threshold that share at least 25% amino acid identity with any prokaryotic homologues are more similar to eubacterial homologues than they are to archaebacterial homologues. This estimate may change somewhat with time as more archaebacterial genomes become available for comparison. However, in the present sample, eubacterial genes dominate in the yeast genome by about a factor of 3:1. This reveals that at the whole genome level, yeast (as an exemplary eukaryote) is more closely related to eubacteria than to archaebacteria. Because yeast's proteins are more eubacterial than archaebacterial, the rooted rRNA tree (Woese, Kandler, and Wheelis 1990), which is often called the universal tree, has a fundamental flaw from the


Yeast proteins vs:
eubacteria
proteobacteria
archaebacteria
eury
cren
actino
spiro
δ/ε
β
γ
α
Aquifex aeolicus
Deinococcus radiodurans
Fusobacterium nucleatum
Porphyromonas gingivalis
Thermotoga maritima
Holotrichia sp. PDC 7120
Synechocystis sp. 6803
Thermosynechococcus elongatus
Chlorobium tepidum
Chlamydia trachomatis
Mycoplasma genitalium
Ureaplasma urealyticum
Clostridium acetobutylicum
Thermoanaerobacter tengcongensis
Bacillus subtilis
Lactococcus lactis
Oceanobacillus iheyensis
Staphylococcus aureus COL
Bifidobacterium longum
Corynebacterium glutamicum
Mycobacterium tuberculosis
Streptomyces coelicolor
Borrelia burgdorferi
Leptospira interrogans
Treponema pallidum
Campylobacter jejuni
Helicobacter pylori
Neisseria meningitidis
Ralstonia solanacearum
Escherichia coli O157:H7 Sakai
Haemophilus influenzae
Pasteurella multocida
Pseudomonas aeruginosa
Salmonella enterica
Vibrio cholerae
Xanthomonas axonopodis
Xanthomonas campestris
Yersinia fastidiosa
Rickettsia prowazekii
Rickettsia conorii
Magnetococcus sp. MC-1
Brucella melitensis
Agrobacterium tumefaciens
Mesorhizobium loti
Sinorhizobium meliloti
Pyrococcus abyssi
Pyrococcus furiosus
Halobacterium sp.
Methanocaldococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina mazei
Archaeoglobus fulgidus
Thermoplasma volcanium
Thermoplasma acidophilum
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus solfataricus
Sulfolobus tokodaii
Neurospora crassa
Schizosaccharomyces pombe
Encephalitozoon cuniculi
Homo sapiens
Drosophila melanogaster
Caenorhabditis elegans
Anopheles gambiae
Fugu rubripes
Arabidopsis thaliana
Plasmodium falciparum
best of nine eukaryote score
a.
b.
c.
T
allantoin/allantoate transporters (1)
amino acid metabolism (125)
aminoacyl-tRNA-synthetases (32)
assembly of protein complexes (4)
C-compound and carbohydrate metabolism (158)
cell differentiation (13)
cell growth / morphogenesis (18)
detoxification (11)
DNA processing (31)
drug transporters (3)
fermentation (7)
Golgi (1)
ion transporters (5)
ionic homeostasis (20)
lipid, fatty-acid & isoprenoid metabolism (29)
metabolism of glycogen, trehalose (1)
metabolism of vitamins, cofactors, and prosthetic groups (29)
mRNA transcription (37)
nitrogen and sulfur metabolism (11)
nucleotide metabolism (56)
other cell division & DNA synth. act. (5)
other protein fate related act. (4)
other protein synthesis act. (3)
other transcription act. (7)
other transport facilitators (1)
oxidation of fatty acids (1)
phosphate metabolism (6)
protein folding and stabilization (21)
protein modification (5)
protein targeting, sorting & translocation (6)
proteolytic degradation (33)
ribosome biogenesis (24)
RNA transport (1)
rRNA transcription (17)
secondary metabolism (3)
stress response (11)
translation (15)
translational control (3)
transport mechanism (5)
TCA cycle (5)
tRNA transcription (3)
vesicular transport (Golgi, etc.) (2)
Unclassified proteins (82)
>50%
45-50%
40-45%
35-40%
25-35%
<25% or not present at E-value threshold of 10e-20
100
200
300
400
500
600
700
800
eubacteria-specific
widespread in eubacteria and archaebacteria
archaebacteria-specific
more eubacterial
more archaebacterial
by functional category
by similarity pattern

standpoint of whole genome comparisons because it has yeast on the wrong branch.

Figure 2*b* contains not only information about yeast proteins, but it also contains information about those prokaryotic proteins that have homologues in yeast, which, all things considered, constitute a random sample of prokaryotic genes. These proteins are not randomly distributed throughout prokaryotic chromosomes; rather they have a discrete distribution. Many individual transfers between eubacteria and archeabacteria can be inferred based on the observed homologue distribution (fig. 2*b*). For example, it can be seen in this sample that the archaebacteria *Methanosarcina mazei* and *Halobacterium* sp. possess a number of genes that are otherwise specific to eubacteria (and yeast), findings which were reported in the original analyses of these complete genome sequences (Ng et al. 2000; Deppenmeier et al. 2002). Notably, presence or absence of a homologue to the yeast query varies within rows more dramatically than sequence identity does (fig. 2*a* and *b*). Thus, the observable distribution of genes in figure 2*b* suggests that lateral gene transfer—an important mechanism of natural variation in prokaryotes (Doolittle 1999)—has permutated the distribution of genes across these genomes to a considerable extent, but it has not fully randomized it.

There are many gene distribution patterns evident in figure 2*b* that can be examined in greater detail on the basis of the tab-delimited table. For example, at position 465 six highly conserved genes that are almost ubiquitous in eubacteria are seen to also be present in the genomes of four euryarchaeotes (*Halobacterium*, *Methanobacterium*, *Methanosarcina mazei*, and two *Thermoplasma* species). In the table these are revealed as the heat shock proteins hs71_yeast to hs76_yeast. Conversely, at position 700 there are three highly conserved proteins present in all archaebacteria sampled that have a sparse distribution among eubacteria. These are the SNZ proteins, of which SNZ1 is involved in pyridoxalphospahte biosynthesis (Stolz and Vielriecher 2003). In the yeast annotation, SNZ1-3 are assigned to three different categories: other cell division and DNA synthesis, vitamin biosynthesis, and stress response, respectively. They are hardly visible in figure 2*a*, but they stand out in figure 2*b* by virtue of a visible, distinct, and shared distribution across genomes.

This Is Not "You Are What You Eat," but It Might Be the Iceberg Below the Tip

One possibility to explain the predominance of eubacterial genes in the yeast genome would be that yeast specifically acquired these genes by lateral transfer from a myriad of individual donors that were ingested as food bacteria and thus donated genes to the nucleus over evolutionary time ("you are what you eat") (Doolittle 1998). Three lines of evidence argue quite clearly against that suggestion in the present context.

First, yeast is not phagotrophic, nor is any fungus phagotrophic, for that matter (Martin et al. 2003). Fungi are heterotrophic osmotrophs; they gain energy through the oxidative breakdown if reduced carbon compounds that they sequester are not from food vacuoles, but from their surroundings with the help of substrate importers in their plasma membrane.

Second, if these are yeast-specific acquisitions, they should not be present in other eukaryotic genomes, but they are (fig. 2*c*). We could only identify six genes among the 850 sampled here that did not occur in another eukaryotic genome on the basis of Blast searching. Those six genes are yei0_yeast, yjv7_yeast, q03036, yg1f_yeast, q08347, and yd39_yeast. Thus, yeast-specific LGT might be responsible for 6/850 (0.7%) of these prokaryotic genes in the yeast genome, but it is also possible that additional sampling will uncover these genes, too, in other eukaryotic genomes, as in the case of the 400 genes in the human genome that were originally claimed to be lateral transfers but turned out, upon closer inspection, not to be LGT after all (Salzberg et al. 2001; Stanhope et al. 2001).

Third, if LGT were at work delivering genes to the yeast genome from various prokaryotic donors over time, then one would expect to see recent transfers with glaring sequence similarities, not just ancient transfers, as are usually inferred from phylogenies. Indeed, evidence for recent transfers from organelle genomes (chloroplast and mitochondria) to nuclear genomes is abundant (Timmis et al. 2004). In such cases, recently transferred organelle DNA sequences in eukaryotic chromosomes may have $\geq$99% identity to their organelle counterparts at the nucleotide level. By contrast, the greatest extent of amino acid identity that we observed between yeast and any prokaryotic protein in the 6,214 $\times$ 177,117 (1.1 billion) FASTA comparisons was 76.8% between atpb_yeast, an important component of the mitochondrial ATP-synthase, and its homologue from the α-proteobacterium *Agrobacterium*. These two *atpb* nucleotide sequences are 66% identical. If we assume that this *atpb* gene was acquired by outright LGT, rather than by endosymbiotic gene transfer from mitochondria (Timmis et al. 2004), then it would be the most recent transfer in the yeast genome relative to the prokaryotic sample investigated here. Using a dramatically

←

FIG. 2.—Amino acid sequence identity in Smith-Waterman alignments for the 850 yeast proteins that produce a match with an E-value of 10^{-20} or better in FASTA comparisons to all proteins from the prokaryotic genomes listed at the top of the figure. Color-coding of the percentage identity values is shown at lower left. (*a*) Yeast proteins grouped by functional category. Lane T at right indicates the targeting prediction (white, mitochondria; grey, sec-pathway) using (from left-to-right) Target P, Pedotar, and iPSORT. Prokaryotic groups are designated; abbreviations are: Actino, actinobacteria; spiro, spirochaetes; eury, euryarchaeotes; cren, crenarchaeotes (also called eocytes). (*b*) Yeast proteins sorted by the quotient [15·(sum of eubacterial identities)]/[45·(sum of archaebacterial identities)]; zero quotients were replaced by one. The scale bar at left indicates the number of the gene in the corresponding table, to facilitate identification of specific genes of interest. The 383 eubacterial-specific proteins, 111 archaebacterial-specific proteins, and 263 proteins widespread among both groups are indicated by colored bars. Lane T at right is as in (*a*). (*c*) Pairwise amino acid identity between yeast homologues and eukaryotic homologues in Blast searches (Altschul et al. 1997), showing that the yeast proteins are not lateral acquisitions specific to the yeast lineage.

oversimplified (but also over conservative) molecular clock calculation and assuming an (extreme) pseudogene rate of roughly 5×10^{-9} per site per year in both lineages (Graur and Li 2000), this most recent transfer would have occurred 34 MYA, and the use of any slower rate would make this most recent transfer even more ancient. In other words, the natural lateral acquisition rate for protein coding-genes in the yeast lineage appears to be much, much less than one gene per 34 Myr.

More recently, Doolittle et al. (2003) have asked, "How big is the iceberg of which organellar genes in nuclear genomes are but the tip?" Figure 2 shows that the iceberg might be quite large, comprising possibly 75% of all nuclear genes in yeast, if we assume that the fraction of genes with a eubacterial ancestry in yeast is the same among those 850 genes that reveal their ancestry by virtue of primary sequence conservation (fig. 2*a*) as it is among those that do not, and if we entertain the possibility that these eubacterial genes could, in principle, all stem from the mitochondrion. At the very low E-value threshold of 10^{-4}, 2,073 yeast genes have $\geq$25% sequence identity to at least one prokaryotic homologue, 699 are eubacterial-specific, 198 are archaebacterial-specific, 1,457 are more eubacterial, and 616 are more archaebacterial in the present sample (see supplemental fig. S1 online).

Which Genes Belong to the Eubacterial- and Archaebacterial-Specific Groups?

The eubacterial- and archaebacterial-specific genes of yeast are shown in more detail in figure 3. Fully consistent with the findings that Rivera et al. (1998) incisively inferred from the analysis of only five prokaryotic genomes, the eubacterial-specific genes are mostly involved in metabolic and biosynthetic processes (operational genes), whereas the archaebacterial-specific genes are mostly involved in information processing (informational genes). However, there are some exceptions; for example, some aminoacyl tRNA synthetases (informational) are among the eubacterial-specific genes, and some amino acid biosynthetic (operational) genes are among the archaebacterial-specific ones. Nonetheless, our analyses very strongly support Rivera et al.'s (1998) distinction of gene classes, but they reveal it in somewhat greater depth, breadth, and detail. The eubacterial/archaebacterial dichotomy in eukaryotic genes was also apparent from the study of individual enzymes involved in ATP synthesis (Martin and Müller 1998).

The left portion of figure 3 shows the 50 most highly conserved yeast proteins that are specific to eubacterial and archaebacteria, respectively. Among the archaebacterial-specific genes several ribosomal proteins, DNA metabolic enzymes, and proteasome subunits are prominent. Core carbon metabolic, core biosynthetic, and glycolytic enzymes are prominent among the eubacterial genes. The latter finding is of interest because it has been claimed that eukaryotes do not possess eubacterial glycolytic enzymes (Canback, Andersson, and Kurland 2002). However, in the present taxon sample there are numerous glycolytic enzymes and other enzymes of core carbon metabolism among the 383 genes that do not occur in 15 sequenced archaebacterial genomes, including glyceraldehyde 3-phosphate dehydrogenase, triosephosphate isomerase, phosphoglycerate mutase (2,3-BPG-dependent), fructose-1,6-bisphosphatase, phosphoglucomutase, fructose-bisphosphate aldolase, glucose-6-phosphate isomerase, glucose-6-phosphate 1-dehydrogenase, phosphoenolpyruvate carboxykinase (ATP-dependent), NAD-dependent malic enzyme, glycerol-3-phosphate dehydrogenase, malate dehydrogenase, ribulose-phosphate 3-epimerase, transketolase, transaldolase, pyruvate decarboxylase, glycerol kinase, malate dehydrogenase, and invertase.

Which Prokaryotic Genomes Are Most Similar to That of Yeast?

In phylogenetic comparisons of genes or proteins, observed site patterns are assumed to be independent and are compared individually; the overall similarity of site patterns provides a measure of overall similarity (or overall difference) between the sequences. Gaps are usually not counted, because of the uncertainty of modeling insertion and deletion events. By analogy, in an evaluation of the extent of similarity or difference between genomes, one could consider genes as being equivalent to site patterns, and comparisons could focus on the character state (presence/absence; if present, extent of identity) of comparable genes in different genomes.

We have used the approach just described in the present study, taking the sum of pairwise amino acid sequence identity across all genes shared by yeast and a prokaryote with at least 25% amino acid identity at a given E-value threshold of 10^{-x} as a measure of the overall similarity of the two respective genomes. This measure at a particular threshold (tI_x) that we calculate takes into account both gene presence/absence ("gaps") and amino acid identity between genes (state similarity at the "position"). We use straight amino acid identity rather than any estimate of similarity for this measure to avoid introducing additional assumptions and uncertainty concerning the general applicability of LogOdds scoring matrices for such anciently diverged sequences. Of course, sequence similarity is not always a good predictor of neighborliness in phylogenetic trees (Koski and Golding 2001); hence, figure 4*a* is not a substitute for a phylogenetic tree. However, the values of tI_x do provide a measure of overall genome similarity that takes both amino acid identity and gene presence or absence into account; few such measures have yet been explored (Lake and Rivera 2004).

Measures of overall genome similarity for tI_4, tI_{20} to tI_{100}, and tI_{150} are shown in figure 4*a* for 60 prokaryotes in comparison to yeast. At low E-value thresholds, the archaebacteria have higher scores of similarity than many eubacteria with small-genomes. However, at higher E-value thresholds, the inference of a close relationship between yeast and archaebacteria disappears altogether. The only apparent relationship at high stringency levels is one between yeast and eubacterial genomes. This striking finding is particularly at odds with the placement of eukaryotes as sisters of archaebacteria in the rooted

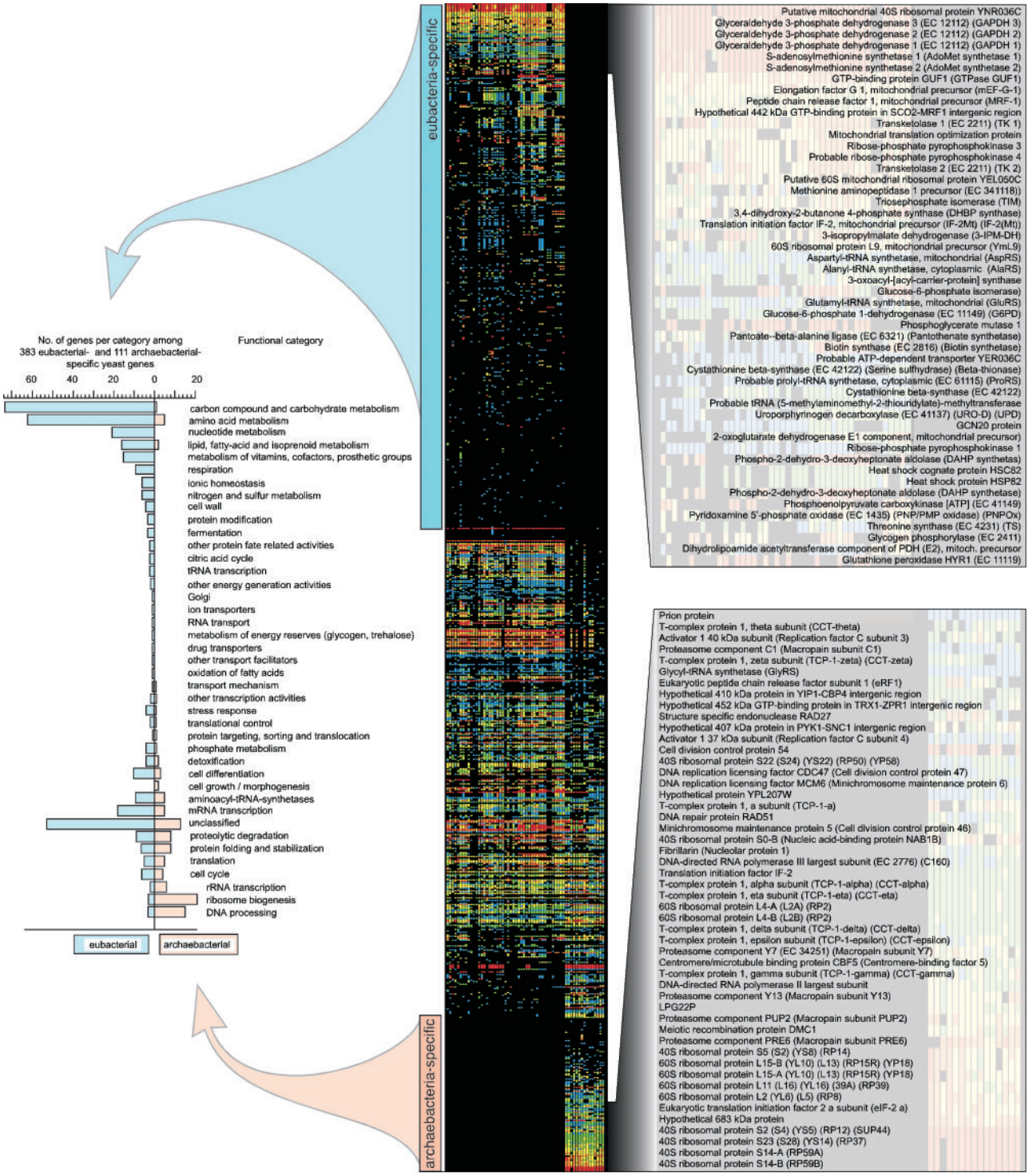
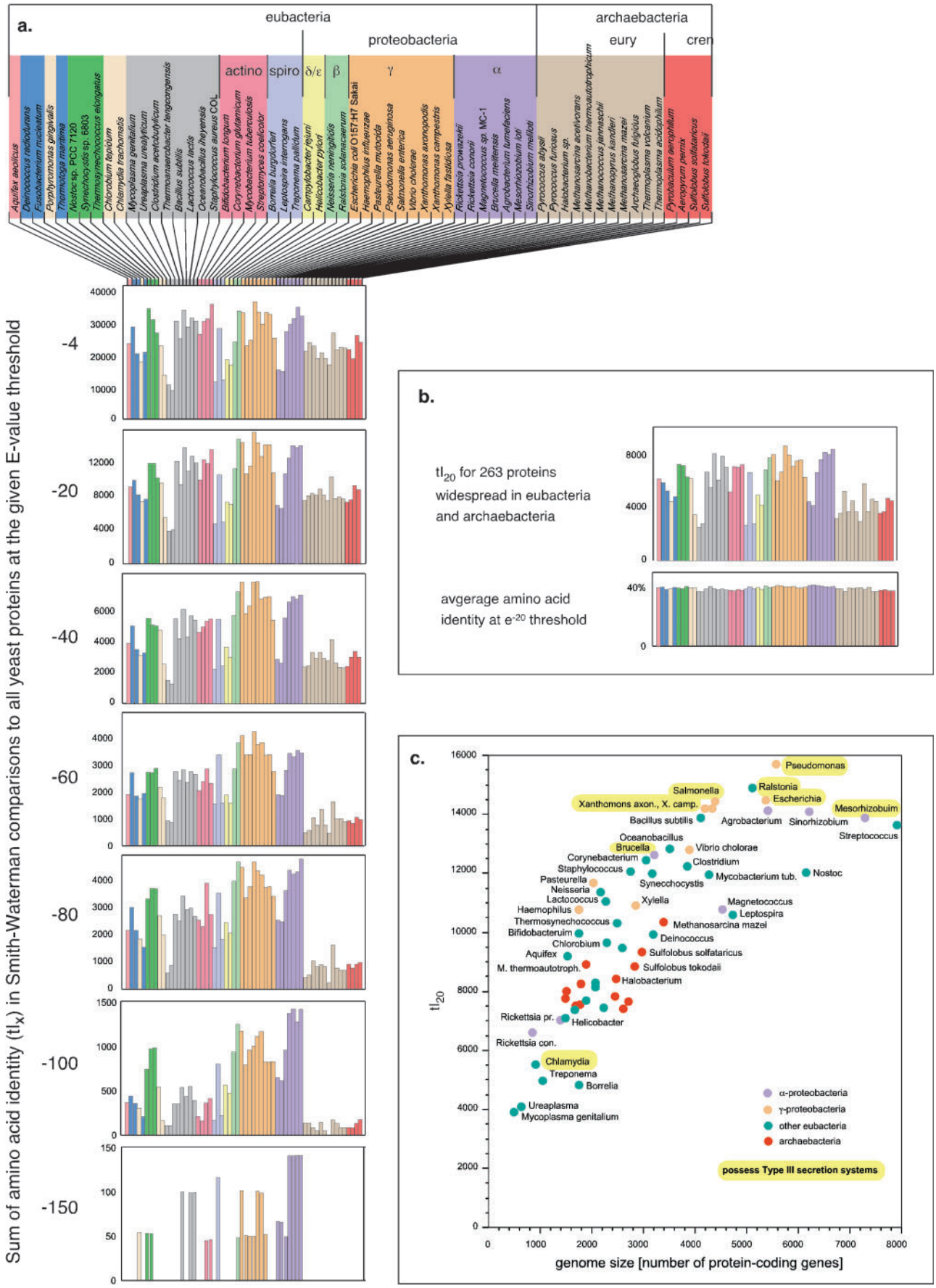


FIG. 3.—Numbers of proteins per functional category for the eubacterial- and archaebacterial-specific yeast proteins (left) and gene definition lines for the 50 most eubacterial- and archaebacterial-specific proteins (right). The central panel from figure 2 is shown for clarity.

versions of the rRNA tree, which is found in many textbooks.

At tI_4 to tI_{60} the γ-proteobacterium *Pseudomonas aeruginosa* bears the greatest overall similarity with yeast among prokaryotes sampled. At tI_{80} and tI_{100} the α-proteobacterium *Sinorhizobium meliloti* becomes the most similar in this sample. Of course, the "winners" in such a comparison will change as more genomes become available. However, the method should be applicable to larger genome samples and to other eukaryotic genomes. The *Rhodobacter*, *Novosphingobium*, and *Rhodospirillum* are not complete and hence were not included in this sample, but it is noteworthy that all three species were less distant to mitochondria in figure 1*b* and *c* than *Sinorhi-*


a.
eubacteria
archaebacteria
proteobacteria
eury
cren
actino
spiro
δ/ε
β
γ
α
Aquifex aeolicus
Deinococcus radiodurans
Fusobacterium nucleatum
Porphyromonas gingivalis
Thermotoga maritima
Nostoc sp. PCC 7120
Synechocystis sp. 6803
Thermosynechococcus elongatus
Chlorobium tepidum
Chlamydia trachomatis
Mycoplasma genitalium
Ureaplasma urealyticum
Clostridium acetobutylicum
Thermoanaerobacter tengcongensis
Bacillus subtilis
Lactococcus lactis
Oceanobacillus iheyensis
Staphylococcus aureus COL
Bifidobacterium longum
Corynebacterium glutamicum
Mycobacterium tuberculosis
Streptomyces coelicolor
Borrelia burgdorferi
Leptospira interrogans
Treponema pallidum
Campylobacter jejuni
Helicobacter pylori
Neisseria meningitidis
Ralstonia solanacearum
Escherichia coli O157:H7 Sakai
Haemophilus influenzae
Pasteurella multocida
Pseudomonas aeruginosa
Salmonella enterica
Vibrio cholerae
Xanthomonas axonopodis
Xanthomonas campestris
Xylella fastidiosa
Rickettsia prowazekii
Rickettsia conorii
Magnetococcus sp. MC-1
Brucella melitensis
Agrobacterium tumefaciens
Mesorhizobium loti
Sinorhizobium meliloti
Pyrococcus abyssi
Pyrococcus furiosus
Halobacterium sp.
Methanosarcina acetivorans
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanopyrus kandleri
Methanosarcina mazei
Archaeoglobus fulgidus
Thermoplasma volcanium
Thermoplasma acidophilum
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus solfataricus
Sulfolobus tokodaii
Sum of amino acid identity (tl$_X$) in Smith-Waterman comparisons to all yeast proteins at the given E-value threshold
-4
-20
-40
-60
-80
-100
-150
b.
tl$_{20}$ for 263 proteins widespread in eubacteria and archaebacteria
avgerage amino acid identity at e^{-20} threshold
40%
c.
tl$_{20}$
genome size [number of protein-coding genes]
Pseudomonas
Salmonella
Ralstonia
Escherichia
Xanthomons axon., X. camp.
Mesorhizobuim
Bacillus subtilis
Agrobacterium
Sinorhizobium
Streptococcus
Oceanobacillus
Brucella
Vibrio cholorae
Corynebacterium
Clostridium
Staphylococcus
Nostoc
Mycobacterium tub.
Pasteurella
Synecchocystis
Neisseria
Lactococcus
Xylella
Magnetococcus
Haemophilus
Leptospira
Thermosynechococcus
Methanosarcina mazei
Bifidobacteruim
Deinococcus
Chlorobium
Sulfolobus solfataricus
Aquifex
M. thermoautotroph.
Sulfolobus tokodaii
Halobacterium
Rickettsia pr.
Helicobacter
Rickettsia con.
Chlamydia
Treponema
Borrelia
Ureaplasma
Mycoplasma genitalium
α-proteobacteria
γ-proteobacteria
other eubacteria
archaebacteria
possess Type III secretion systems

zobium was. It will be of interest to extend the present comparison to additional α-proteobacterial genomes.

Among the archaebacteria, the genome that is most similar to yeast at all thresholds is that of the methanogen *Methanosarcina mazei*. However, the nature of *Methanosarcina*'s evident similarity to yeast is founded largely in the fact that this methanogen has acquired about 30% eubacterial genes, which are involved in its ability to metabolize a moderately broad spectrum of C1 compounds, such as methyamines and methanol, in addition to CO_2 as a sole carbon source (Deppenmeier et al. 2002), attributes (and genes) that autotrophic methanogens in this sample lack. At lower thresholds the *Sulfolobus* species come in second. At higher thresholds, however, it is again a methanogen, the autotroph *Methanobacterium thermoautotrophicum*, that scores well, as does the aerobic heterotroph *Halobacterium*, which might be a derived methanogen that became an aerobic heterotrophic through gene acquisition and gene loss.

That the methanogens score well among this extremely narrow sample of archaebacteria might seem surprising at first sight. It is in line, however, with the predictons of the hydrogen hypothesis (Martin and Müller 1998) and of the syntrophic hypothesis (Moreira and Lopez-Garcia 1998), because both models implicate a methanogen-like metabolism for the archaebacterial partner presumed to have been involved at the symbiogenic origin of eukaryotes. That *Methanosarcina mazei* shows the highest overall similarity to yeast in the present sample is likely due to convergence, but the circumstance that this methanogen is able to acquire and express eubacterial genes for carbon importers, carbon metabolism, protein folding, and other functions (Deppenmeier et al. 2002) bears out a prediction of the hydrogen hypothesis that such acquisitions and expression should be possible.

Whereas the hydrogen hypothesis predicts the strongest signals from methanogens and α-proteobacteria, which is observed at several thresholds in the present analysis (fig. 4), the syntrophic hypothesis predicts the strongest signals from methanogens and δ-proteobacteria (plus a presumably smaller α-proteobacterial signal). The only two representatives from the δ/ε group of proteobacteria in this sample are *Campylobacter* and *Helicobacter*, both of which fare poorly in the present comparison, but the sample is quite small.

The model of Margulis, Dolan, and Guerrero (2000) presumes a *Thermoplasma*-like host and a spirochaete at the origin of eukaryotes, but neither group fares particularly well in the present highly restricted sample. The model of Cavalier-Smith (2002) predicts a strong signal from the actinobacteria, which is in fact present (*Streptomyces*) at low thresholds but, in contrast to the α-proteobacterial signal, dwindles at higher thresholds. Yet, again, the present sample is quite small and there is much room for additional comparisons. The model of Gupta (1998) predicts a strong signal from proteobacteria and from the group of archaebacteria known as eocytes (Lake 1988), also known as crenarchaeotes (Woese, Kandler, and Wheelis 1990). Indeed, members of the γ- and β-proteobacteria have the highest overall tI_{20} values (and at several other thresholds), and *Sulfolobus* (an eocyte) also scores quite well at several thresholds. Clearly, additional sampling is needed.

If we look at the 263 proteins that are widespread among both prokaryotic groups, the proteobacteria battle it out tightly, and *Methanosarcina* remains at the forefront among archaebacteria. Importantly, the values of tI_x are predominantly a function of gene content in the prokaryotic genomes, because the average sequence identity of non-zero values is almost completely constant at 40% across genomes (fig. 4*b*).

The tI_{20} values are correlated with genome size, as shown in figure 4*c*, but they are not strictly a function of genome size. For example, *Streptomyces* has a low specific similarity to yeast whereas *Brucella* (an α-proteobacterium) and *Bacillus* (a Gram positive) have comparatively high tI_{20} values for their respective genome sizes.

Many of the top-scoring proteobacteria are pathogens or otherwise interact intimately with eukaryotic cells. Accordingly, many of them possess type III secretion systems (yellow shading in fig. 4*c*), which allow pathogens to inject proteins into their eukaryotic hosts, thereby often interfering with their host's ability to detect infection or respond to it (Gauthier, Thomas, and Finlay 2003). Among the prokaryotes that lack Type III secretion systems in our sample, *Agrobacterium* and *Sinorhizobium* fare best at the 10^{-20} threshold (fig. 4*c*). Pathogens are overrepresented in the present eubacterial genome sample. Complete sequence data from additional nonpathogenic eubacteria are needed.

Horiike et al. (2001) studied the yeast genome using Blast comparisons and found that several functional categories of yeast genes were on average more similar to eubacterial or archaebacterial homologues, respectively. However, Horiike et al. (2001) embraced the a priori assumption that those eubacterial genes in the yeast genome encoding mitochondrion-specific proteins stem from the α-proteobacterial ancestor of mitochondria, and those eubacterial proteins that are not mitochondrion-specific stem from a different source—in their view a eubacterial host that acquired an archaebacterial symbiont, the latter of which became the nucleus. Hedges et al. (2001) assumed that the excess eubacterial genes in eukaryotes stem from a symbiont that arose prior to the mitochondrion. Both Hedges et al. (2001) and Horiike et al. (2001), following Gupta's argument (1998), attributed the excess eubacterial genes to a single eubacterial partner at the origin of eukaryotes that was distinct from the mitochondrial endosymbiont. This assumption is also contained in the model of Hartman and Fedorov (2002), in the much earlier suggestion of Zillig et al. (1989), and in the more recent suggestions of Emelyanov (2003). All six

←

FIG. 4.—Sums of amino acid identity between yeast proteins and prokaryotic homologues. (*a*) Values of tI_x for prokaryotic genomes at several E-value thresholds. (*b*) Values of tI_{20} for subsets of the data indicated and average amino acid identity for non-zero values at the 10^{-20} E-value threshold. (*c*) Values of tI_{20} plotted against number of proteins per genome. Species that possess type III secretion systems are highlighted in yellow.

models presume that there was an additional symbiotic partner in the evolution of eukaryotes that preceded the mitochondrial symbiont, and the former five suggest that some amitochondriate eukaryotes, in particular *Giardia intestinalis*, are primitively amitochondriate. However, as some might have expected (Roger and Silbermann 2002; Embley et al. 2003), *Giardia* possesses mitochondria after all (Henze and Martin 2003; Tovar et al. 2003), so models that derive the *Giardia* lineage prior to the acquisition of mitochondria can currently be excluded.

In our view, it is not yet clear whether the data really require the supposition of an additional eubacterial symbiont as the source of these "too many" eubacterial genes in yeast. An α-proteobacterial symbiont (the ancestor of mitochondria) with a broad diversity of genes in its genome would suffice to account for the excess eubacterial genes in eukaryotes. The circumstance that many genes of mitochondrial origin in eukaryotes are not targeted to the mitochondrion is difficult to explain or not at all addressed in some models (Hedges et al. 2001; Horiike et al. 2001; Hartman and Fedorov 2002; Emelyanov 2003), but it is directly predicted under others, in which gene transfer from endosymbiont to host is viewed as a eukaryote-specific mechanism of natural variation that existed before the origin of the mitochondrial protein import apparatus (Martin and Müller 1998; Timmis et al. 2004).

Eukaryotic Substrate Importers

An explicit prediction of some models (Martin and Müller 1998) and an implicit prediction of others (Moreira and Lopez-Garcia 1998; Cavalier-Smith 2002) is that eukaryotes should have eubacterial importers for reduced carbon compounds in their plasma membrane. Yet importers (used here synonymously with all proteins involved in the movement of substrates from one side of a membrane to the other) are generally poorly conserved in comparison to glycolytic enzymes or some ribosomal proteins, for example. This is mostly because transmembrane domains are rich in nonpolar amino acids but can easily accept the replacement of one nonpolar residue by another at many sites. Among the eubacterial-specific carbon importers identified at the E-value threshold of 10^{-20} are the hexose transporters HXT10, HXT11, HXT13, HXT15, HXT16, HXT17, HXT8, HXT9; the high-affinity glucose transporters HXT2, SNF3, HXT6; the low-affinity glucose transporters HXT1, HXT3, HXT4; and the sugar transporter STL1. To examine importers more broadly, we had to lower the E-value threshold. At the very low 10^{-4} threshold, overall sequence similarity between yeast and prokaryotes is low and individual sequence identities exceeding 35% are rare (fig. 5). The current sample indicates that homologues of the importers possessed by yeast are more widespread among eubacteria than among archaebacteria; this is particularly noticeable in the class of unspecified importers.

What About the Yeast Proteins that Detect No Prokaryotic Homologues Here?

The present findings indicate that about 3/4 of the nuclear protein-coding genes in *Saccharomyces cerevisiae* that detect homologues in sequenced prokaryotic genomes are more similar to eubacterial homologues than they are to archaebacterial homologues, and they indicate, furthermore, that at high stringency the archaebacterial component of similarity in the yeast genome disappears almost entirely, whereas the eubacterial component does not. These findings, founded in genome comparisons, are irreconcilable with a supposed sister-group relationship between archaebacteria and eukaryotes, which is the current paradigm and which is founded mostly in the analysis of a single gene (small subunit ribosomal RNA) as rooted with protein trees (Woese, Kandler, and Wheelis 1990).

However, only about 15% (850/6,214) of yeast's genes share at least 25% amino acid identity with homologues detected at the E-value threshold of at least 10^{-20} in this sample of 177,117 prokaryotic proteins. This raises the question, where do the other 85% come from? In principle, there are three possibilities, which can be labeled as "mystery host," "sequence divergence," and "descent with modification."

The suggestion of "mystery host" (exemplified in Hartman and Federov [2002]), supposes that eukaryotic genes lacking detectable homologues in prokaryotes constitute direct evidence for a third kind of cell that existed early in evolution but was in supply for a limited time only. It was neither a eubacterium nor an archaebacterium. Instead, that cell (called the "cronocyte" in some formulations) is to be envisaged as a free-living cytoskeleton with abundant calcium signaling pathways but lacking genes for ATP synthesis and core genetic apparatus (Hartman and Federov 2002), because those kinds of genes are found in prokaryotes (fig. 2). In this variant of endosymbiotic theory, the "mystery host" serves as a preformed eukaryotic cytosol incertae sedis into which the nucleus and mitochondria may penetrate as endosymbionts (Hartman and Federov 2002). Where the cronocyte comes from is not an issue for the theory (Hartman and Federov 2002). The postulated existence of such a cell is essential to uphold many prominent theories, because without it "then the three cellular domains, Eukarya, Archaea, and Bacteria, would collapse into two cellular domains" (Hartman and Federov 2002, pp. 1420). The "mystery host" explanation for eukaryotic-specific genes attributes their origin to an inheritance, by eukaryotes, from an imaginary form of life and is thus unfalsifiable, for which reason it can be set aside for the time being.

The second possibility is "sequence divergence." This explanation for the paucity of sequence conservation between eukaryotic and prokaryotic proteins operates with a known mechanism popular among proponents of the New Synthesis: point mutation. Unradically, it posits that prokaryotes arose before eukaryotes, that the ancestral set of eukaryotic genes therefore had prokaryotic counterparts, and that many mutations have accumulated in the brunt of both prokaryotic and eukaryotic genes subsequent to the origin eukaryotes, such that a good portion of eukaryotic genes therefore no longer have detectable primary sequence similarity with their prokaryotic counterparts (Martin et al. 2002).

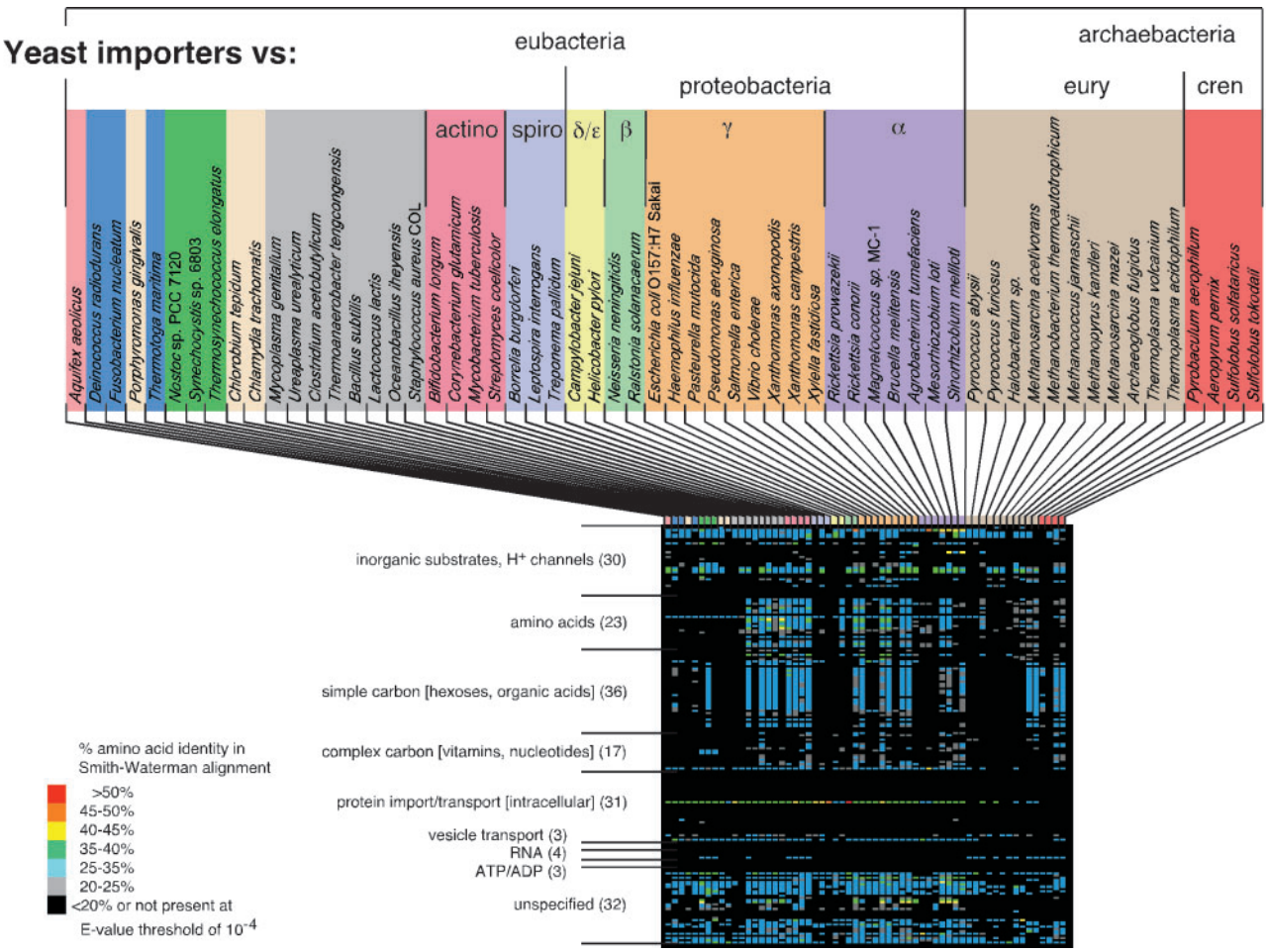


FIG. 5.—Amino acid sequence identity in Smith-Waterman alignments for the 176 yeast membrane-transport proteins that produce a match with an E-value of 10^{-4} or better in FASTA comparisons to all proteins from the prokaryotic genomes listed at the top of the figure. Color-coding of the percentage identity values is shown at lower left. Proteins were grouped into the substrate categories shown on the basis of information in the database annotations.

The third possibility is "descent with modification," a well-established evolutionary principle that is applicable to genes. Sequence divergence is a special case of descent with modification, because the former takes only point mutations into account, whereas the latter would also include recombination, insertion/deletion, duplication, optimization, and functional specialization, during which processes proteins would become increasingly dissimilar to their prokaryotic progenitors, while the original genetic starting material was becoming suited, via natural variation and natural selection, to ensure the survival of the earliest eukaryotic progeny. Descent with modification would allow the possibility that eukaryotes might have invented some genes from preexisting prokaryotic starting material and that such genes might have subsequently come under strong functional constrains so as to evolve in a very conserved manner within the eukaryotic lineage, without ever having arisen in prokaryotes.

Conclusion

At the level of overall amino acid sequence identity and gene presence or absence, proteobacterial genomes were found to be the most similar to the yeast genome among eubacteria surveyed, whereas among archaebacteria surveyed, the genome of the methanogen *Methanosarcina mazei* was the most similar to yeast. The similarity of the yeast genome to that of *Methanosarcina* is likely due to convergence, because that has acquired and expresses many eubacterial genes for carbon metabolism and carbon importers in a process that surely occurred independently from any putatively analogous acquisitions in eukaryotes. The analysis of proteins encoded in mitochondrial genomes reveals that the position of mitochondria is unresolved with the present sample of data from α-proteobacterial genomes, although *Rhodospirillum* comes as close to mitochondria as any α-proteobacterium sampled. That about 75% of yeast's nuclear genes that detect prokaryotic homologues are more similar to eubacterial than to archaebacterial homologues and are furthermore present in other eukaryotes suggests (1) that the common ancestor of eukaryotes surveyed here also may have possessed a majority of eubacterial genes, though it is still unclear how many of these ultimately come from the ancestral mitochondrial genome, and (2) that lineage-specifc lateral acquisitions in the yeast lineage account for $<1\%$ of the observed gene distribution. The approaches

described here to genome comparison may hold promise for discrimination between alternative hypotheses for the origins of eukaryotes and mitochondria.

Literature Cited

Adachi, J., and M. Hasegawa. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

Alon N., and J. H. Spencer. 1992. The probabilistic method. Wiley, New York.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Brown, J. R. 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. **4**:121–132.

Brown, J. R., and W. F. Doolittle. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. **61**:456–502.

Bryant, D., and V. Moulton. 2004. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. Mol. Biol. Evol. **21**:255–265.

Canback, B., S. G., Andersson, and C. G. Kurland. 2002. The global phylogeny of glycolytic enzymes. Proc. Natl. Acad. Sci. USA **99**:6097–6102.

Cavalier-Smith, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int. J. Syst. Evol. Microbiol. **52**:297–354.

Deppenmeier, U., A. Johann, T. Hartsch et al. (22 co-authors). 2002. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archae a. J. Mol. Microbiol. Biotechnol. **4**:453–461.

Doolittle, W. F. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. **14**:307–311.

———. 1999. Phylogenetic classification and the universal tree. Science **284**:2124–2128.

Doolittle, W. F., Y. Boucher, C. L. Nesbo, C. J. Douady, J. O. Andersson, and A. J. Roger. 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**:39–58.

Doolittle, W. F., and J. R. Brown. 1994. Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA **91**:6721–6728.

Embley, T. M., and R. P. Hirt. 1998. Early branching eukaryotes? Curr. Opin. Genet. Devel. **8**:655–661.

Embley, T. M., M. van der Giezen, D. S. Horner, P. L. Dyal, and P. Foster. 2003. Hydrogenosomes and mitochondria: phenotypic variants of the same fundamental organelle. Philos. Trans. R. Soc. Lond. B. **358**:191–203.

Emelyanov, V. V. 2003. Mitochondrial connection to the origin of the eukaryotic cell. Eur. J. Biochem. **270**:1599–1618.

Feng, D.-F., G. Cho, and R. F. Doolittle. 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94**:13028–13033.

Gauthier, A., N. A. Thomas, and B. B. Finlay. 2003. Bacterial injection machines. J. Biol. Chem. **278**:25273–62527.

Giegé, P., J. L. Heazlewood, U. Roessner-Tunali, A. H. Millar, A. R. Fernie, C. J. Leaver, and L. J. Sweetlove. 2003. Enzymes of glycolysis are functionally associated with the mitochondrion in *Arabidopsis* cells. Plant Cell **15**: 2140–2151.

Gogarten, J. P. 2003. Gene swapping craze reaches eukaryotes. Curr. Biol. **13**:R53–54.

Gogarten, J. P., H. Kibak, P. Dittrich et al. (13 coauthors) 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86**: 6661–6665.

Graur, D., and Li, W.-H. 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Gray, M. W., G. Burger, and B. F. Lang. 1999. Mitochondrial evolution. Science **283**:1476–1481.

Gupta, R. S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. **62**:1435–1491.

Hansmann, S., and W. Martin. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes. Int. J. Syst. Evol. Microbiol. **50**:1655–1663.

Hartman, H., and A. Fedorov. 2002. The origin of the eukaryotic cell: a genomic investigation. Proc. Natl. Acad. Sci. USA **99**:1420–1425.

Hedges, S. B., H. Chen, S. Kumar, D. Y. Wang, A. S. Thompson, and H. Watanabe. 2001. A genomic timescale for the origin of eukaryotes. BMC Evol. Biol. **1**:4.

Henze, K., and W. Martin. 2003. Essence of mitochondria. Nature **426**:127–128.

Horiike, T., K. Hamada, S. Kanaya, and T. Shinozawa. 2001. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria revealed is revealed by homology hit analysis. Nat. Cell Biol. **3**:210–214.

Huson, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**:68–73.

Imhoff, J. F., and H. G. Trüper. 1992. The genus *Rhodospirillum* and related genera. Pp. 2141–2159 *in* A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K.-H. Schleifer, eds. The prokaryotes, 2nd ed., Vol. III. Springer-Verlag, New York.

Iwabe, N., K.-I. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**:9355–9359.

John, P., and F. R. Whatley. 1975. *Paracoccus denitrificans* and the evolutionary origin of the mitochondrion. Nature **254**: 495–498.

Karlin, S., L. Brocchieri, J. Mrazek, A. M. Campbell, and A. M. Spormann. 1999. A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. Proc. Natl. Acad. Sci. USA **96**:9190–9195.

Knoll, A. H. 2003. Life on a young planet: the first three billion years of evolution on Earth. Pp. 122–160. Princeton University Press, Princeton, N.J.

Koski, L. B., and G. B. Golding. 2001. The closest Blast hit is often not the nearest neighbor. J. Mol. Evol. **52**:540–542.

Kurland, C. G., and S. G. Andersson. 2000. Origin and evolution of the mitochondrial proteome. Microbiol. Mol. Biol. Rev. **64**:786–820.

Lake, J. A. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature **331**:184–186.

Lake, J. A., and M. C. Rivera. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. Mol. Biol. Evol. **21**:681–690.

Lang, B. F., G. Burger, C. J. O'Kelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, and M. W. Gray. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature **387**:493–497.

Lang, B. F., M. W. Gray, and G. Burger. 1999. Mitochondrial genome evolution and the origin of eukaryotes. Ann. Rev. Genet. **33**:351–397.

Lockhart, P. J., C. J. Howe, A. C. Barbrook, A. W. D. Larkum, and D. Penny. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol. Biol. Evol. **16**:573–576.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

Long, M., S. J. de Souza, C. Rosenberg, and W. Gilbert. 1996. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome *c*1 precursor. Proc. Natl. Acad. Sci. USA **93**:7727–7731.

Margulis, L., M. F. Dolan, and R. Guerrero. 2000. The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. Proc. Natl. Acad. Sci. USA **97**:6954–6959.

Martin, W., M. Hoffmeister, C. Rotte, and K. Henze. 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biol. Chem. **382**:1521–1539.

Martin, W., and M. Müller. 1998. The hydrogen hypothesis for the first eukaryote. Nature **392**:37–41.

Martin, W., C. Rotte, M. Hoffmeister, U. Theissen, G. Gelius-Dietrich, S. Ahr, and K. Henze. 2003. Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. IUBMB Life **55**:193–204.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. 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.

Meyer, T. R., M. A. Cusanovich, and M. D. Kamen. 1986. Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. Proc. Natl. Acad. Sci. USA **83**:217–220.

Moreira, D., and P. Lopez-Garcia. 1998. Symbiosis between methanogenic archaea and δ-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. J. Mol. Evol. **47**:517–530.

Mossel, E. 2003. On the impossibility of reconstructing ancestral data and phylogenies, J. Comput. Biol. **10**:669–676.

Müller, M. 2003. Energy metabolism. Part I: anaerobic protozoa. Pp. 125–139 *in* J. Marr, ed. Molecular medical parasitology. Academic Press, London.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Ann. Rev. Genet. **30**:371–403.

Ng, W. V., S. P. Kennedy, G. G. Mahairas et al. (42 co-authors). 2000. Genome sequence of *Halobacterium* species NRC-1. Proc. Natl. Acad. Sci. USA **97**:12176–12181.

Nisbet, E. G., and C. M. R. Fowler. 1999. Archaean metabolic evolution of microbial mats. Proc. R. Soc. Lond. B. **266**:2375–2382.

Nisbet, E. G., and N. H. Sleep. 2001. The habitat and nature of early life. Nature **409**:1083–1091.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**:2444–2448.

Penny, D., L. R. Foulds, and M. D. Hendy. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature **297**:197–200.

Penny, D., and M. Hendy. 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. **3**:403–417.

Penny, D., M. D. Hendy, P. J. Lockhart, and M. A. Steel. 1996. Corrected parsimony, minimum evolution, and Hadamard conjugations. Syst. Biol. **45**:596–606.

Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. **53**:711–723.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. **8**:616–623.

Poole, A., and D. Penny. 2001. Endosymbiosis does not explain the origin of the nucleus. Nat. Cell Biol. **8**:E173.

Richly, E. P. F., Chinnery, and D. Leister. 2003. Evolutionary diversification of mitochondrial proteomes: implications for human disease. Trends Genet. **19**:356–362.

Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake. 1998. Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA **95**:6239–6244.

Roger, A. J., and J. D. Silbermann. 2002. Mitochondria in hiding. Nature **418**:827–828.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425**:798–804.

Rothschild, L. J., M. A. Ragan, A. W. Coleman, P. Heywood, and S. A. Gerbi. 1986. Are rRNA sequence comparisons the rosetta stone of phylogenetics? Cell **47**:640.

Rotte, C., and W. Martin. 2001. Endosymbiosis does not explain the origin of the nucleus. Nat. Cell Biol. **8**:E173–174.

Saitou, N., and M. Nei. 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Salzberg, S. L., O. White, J. Peterson, and J. A. Eisen. 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science **292**:1903–1906.

Sidow, A., T. Nguyen, and T. P. Speed. 1992. Estimating the fraction of invariable codons with a capture–recapture method. J. Mol. Evol. **35**:253–260.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

Sober, E., and M. Steel. 2002. Testing the hypothesis of common ancestry. J. Theor. Biol. **218**:395–408.

Stackebrandt, E., R. G. E. Murray, and H. G. Trüper. 1988. Proteobacteria classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives." Int. J. Syst. Bact. **38**:321–325.

Stanhope, M. J., A. Lupas, M. J. Italia, K. K. Koretke, C. Volker, and J. R. Brown. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature **411**:940–944.

Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst. Biol. **49**:225–232.

Stolz, J., and M. Vielreicher. 2003. Tpn1p, the plasma membrane vitamin B6 transporter of *Saccharomyces cerevisiae*. J. Biol. Chem. **278**:18990–18996.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

Thollesson, M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics **20**:416–418.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple

sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Tielens, A. G. M., C. Rotte, J. J. van Hellemond, and W. Martin. 2002. Mitochondria as we don't know them. Trends Biochem. Sci. **27**:564–572.

Timmis, J. N., M. A. Ayliffe, C. Y. Huang, and W. Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. **5**:123–135.

Tovar, J., G. León-Avila, L. B. Sánchez, R. Sutak, J. Tachezy, M. van der Giezen, M. Hernández, M. Müller, and J. M. Lucocq. 2003. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. Nature **426**:172–176.

Woese, C. R. 2002. On the evolution of cells. Proc. Natl. Acad. Sci. USA **99**:8742–8747.

Woese, C., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc. Natl. Acad. Sci. USA **87**:4576–4579.

Wu, M., L. V. Sun, J. Vamathevan et al. (30 coauthors). 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis w*Mel: a streamlined genome overrun by mobile genetic elements. PloS Biol. **2**:327–341.

Zillig, W., H.-P. Klenk, P. Palm, H. Leffers, G. Pühler, F. Gropp, and R. A. Garrett. 1989. Did eukaryotes originate by a fusion event? Endocyt. C. Res. **6**:1–25.

Martin Embley, Associate Editor

Accepted May 3, 2004