Integration of Two Ancestral Chaperone Systems into One: The Evolution of Eukaryotic Molecular Chaperones in Light of Eukaryogenesis

David Bogumil,[1] David Alvarez-Ponce,[‡,2] Giddy Landan,[1] James O. McInerney,[2] and Tal Dagan*,[1]

[1]Institute of Microbiology, Christian-Albrechts-University of Kiel, Kiel, Germany

[2]Department of Biology, National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

[‡]Present address: Integrative and Systems Biology Laboratory, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Valencia, Spain

***Corresponding author:** E-mail: tdagan@ifam.uni-kiel.de.

Associate editor: Naoko Takezaki

Abstract

Eukaryotic genomes are mosaics of genes acquired from their prokaryotic ancestors, the eubacterial endosymbiont that gave rise to the mitochondrion and its archaebacterial host. Genomic footprints of the prokaryotic merger at the origin of eukaryotes are still discernable in eukaryotic genomes, where gene expression and function correlate with their prokaryotic ancestry. Molecular chaperones are essential in all domains of life as they assist the functional folding of their substrate proteins and protect the cell against the cytotoxic effects of protein misfolding. Eubacteria and archaebacteria code for slightly different chaperones, comprising distinct protein folding pathways. Here we study the evolution of the eukaryotic protein folding pathways following the endosymbiosis event. A phylogenetic analysis of all 64 chaperones encoded in the *Saccharomyces cerevisiae* genome revealed 25 chaperones of eubacterial ancestry, 11 of archaebacterial ancestry, 10 of ambiguous prokaryotic ancestry, and 18 that may represent eukaryotic innovations. Several chaperone families (e.g., Hsp90 and Prefoldin) trace their ancestry to only one prokaryote group, while others, such as Hsp40 and Hsp70, are of mixed ancestry, with members contributed from both prokaryotic ancestors. Analysis of the yeast chaperone–substrate interaction network revealed no preference for interaction between chaperones and substrates of the same origin. Our results suggest that the archaebacterial and eubacterial protein folding pathways have been reorganized and integrated into the present eukaryotic pathway. The highly integrated chaperone system of yeast is a manifestation of the central role of chaperone-mediated folding in maintaining cellular fitness. Most likely, both archaebacterial and eubacterial chaperone systems were essential at the very early stages of eukaryogenesis, and the retention of both may have offered new opportunities for expanding the scope of chaperone-mediated folding.

Key words: **origin of eukaryotes, molecular chaperones, protein evolution.**

Introduction

The symbiogenic model for the origin of eukaryotes posits that eukaryotes arose via a symbiotic association of two distantly related prokaryotes (Sagan 1967; Rivera and Lake 2004; Embley and Martin 2006; Pisani et al. 2007; Lane 2009; Alvarez-Ponce et al. 2013). Opinions about the precise taxonomic classification and metabolic capacities of the prokaryote involved are still divided, however there is a wide agreement among scientists that the host was an archaebacterium (Martin and Muller 1998; Cox et al. 2008; Williams et al. 2012) and the endosymbiont was an alpha-proteobacterium (Gray et al. 1999; Gabaldón and Huynen 2003; Esser et al. 2004). The eubacterial endosymbiont subsequently evolved into the mitochondrion organelle, a process that was accompanied by a massive DNA transfer from the symbiont into the host genome, the evolution of a mitochondrial protein import apparatus, a drastic miniaturization of the mitochondrial genome, and an increased complexity of the nuclear genome (Martin and Herrmann 1998; Martin 2003; Timmis et al. 2004). Phylogenomic studies show, accordingly, that eukaryotic genomes are a mosaic of genes of eubacterial and archaebacterial ancestry (Esser et al. 2004; Pisani et al. 2007; Thiergart et al. 2012; Alvarez-Ponce et al. 2013).

Evolutionary analysis of genes in the model eukaryote *Saccharomyces cerevisiae* reveals that about 37% of the genes can be traced back to either an archaebacterial or a eubacterial ancestor (Cotton and McInerney 2010). Thus, eukaryotic innovations probably account for a sizeable fraction of eukaryotic genomes. Yet, the proportion of eukaryotic genes of demonstrable prokaryotic origin is quite substantial considering the complications involved in this kind of analysis. The long divergence time elapsed since the symbiotic event limits our ability to detect prokaryotic homologs to some prokaryote-derived proteins and reduces the accuracy of phylogenetic inference for others. Furthermore, lateral gene transfer events between the eubacterial and archaebacterial lineages (e.g., Deppenmeier et al. 2002; Large and Lund 2009; Williams et al. 2010; Nelson-Sathi et al. 2012) may have obscured the genetic record of the symbiosis event, leading to an ambiguous classification of eukaryotic genes.

The chimerical origin of eukaryotic genomes is imprinted in the functional role of proteins within the cell. Many proteins that perform an informational function (e.g., replication, transcription, and translation) are of archaebacterial origin while many genes of eubacterial origin perform operational functions (e.g., metabolism, amino acid synthesis, and regulatory genes) (Rivera et al. 1998; Esser et al. 2004; Cox et al. 2008; Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011; Alvarez-Ponce et al. 2013). Eukaryotic genes of archaebacterial origin are more essential regardless of the bias towards informational functions (Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011). Furthermore, the eukaryotic protein–protein interaction network still bears the markings of a chimerical ancestry, with proteins from the same origin—archaebacterial or eubacterial—being interconnected at a frequency that is significantly above the expected by chance (Alvarez-Ponce and McInerney 2011). Thus, when considered as a whole, the eukaryotic proteome can be described as a partially integrated version of two ancestral ingredients.

In this study, we have set forth to examine the evolution of the eukaryotic protein folding pathway in light of the symbiogenic model. Molecular chaperones are proteins that assist the folding and unfolding of other proteins, as well as the complex assembly and stabilization of protein and nucleic acids interactions (Hartl and Hayer-Hartl 2009; Large et al. 2009). Chaperones often function in assembly-line-like pathways where various chaperones interact consecutively with the same substrate driving the transition of the newly synthetized peptide into a functional protein (Young et al. 2004). Chaperones are essential in all living organisms and have been shown to play a role as capacitors of phenotypic variation (Rutherford and Lindquist 1998; Queitsch et al. 2002) and drivers of increased fitness within organisms facing a high mutational load (Fares et al. 2002; Maisnier-Patin et al. 2005). Furthermore, their function as biochemical mediators of protein assembly played an important role in shaping genomic landscapes (Bogumil and Dagan 2010; Williams and Fares 2010; Bogumil et al. 2012). The utility of molecular chaperones is thought to be constrained by a delicate balance between their help in mitigating the effects of protein misfolding and the slower rate of protein production and maturation of their substrate (Bogumil and Dagan 2012). Archaebacteria and eubacteria harbor slightly different repertoires of chaperone families. The Hsp40 and Hsp70 chaperone families are present in both domains (Macario et al. 1991; Macario et al. 1993), whereas other chaperone systems, such as chaperonins, differ in their composition and assembly.

Here we study the extent to which the chimeric origin of eukaryotes is detectable in the eukaryotic protein folding pathway of contemporary genomes. We infer the ancestry of yeast chaperones and their substrates, examine the yeast chaperone repertoire, and use a network approach to study the relationship between chaperones and their substrates in light of their origin.

Results

Prokaryotic Ancestry of *S. cerevisiae* Proteins

To determine the prokaryotic origin of yeast proteins, we searched for their prokaryotic homologs among 82 archaebacterial and 1,074 eubacterial genomes. A total of 1,230 yeast proteins had detectable homologs in one or more prokaryotic genomes. The remaining proteins did not manifest detectable homology with prokaryotic proteins, and we therefore consider them to be eukaryotic innovations. A total of 689 phylogenetic trees were reconstructed for yeast proteins having more than three homologs belonging to both archaebacteria and eubacteria. Yeast proteins were classified according to the prokaryotic domain within which they branch. Our analysis revealed 289 proteins of archaebacterial ancestry, 803 of eubacterial ancestry, and 138 of an unresolved prokaryotic ancestry. All phylogenetic trees are provided in supplementary tables S1 and S2, Supplementary Material online.

The Mosaic Structure of the *S. cerevisiae* Chaperone Repertoire

Of the 64 known yeast molecular chaperones, 46 had homologs in prokaryotic genomes. These were classified based on their tree topology into 11 chaperones of archaebacterial ancestry and 25 chaperones of eubacterial ancestry. The ancestry of the remaining ten chaperones could not be resolved from the data (fig. 1). The Hsp90 family in yeast includes two paralogs whose sequences are highly similar (96% identity at the amino acid level). Both paralogs are homologous to eubacterial htpG sequences exclusively, and hence the yeast Hsp90 is clearly of eubacterial origin. The prefoldin (PFD) chaperones transfer target proteins to the chaperone-containing T-complex polypeptide 1 (CCT) system for further folding (Vainberg et al. 1998). The yeast genome encodes six PFD paralogs whose protein sequences are 15.2 ± 3.8% identical. Three of the six PFDs have homologs in prokaryotic genomes, all of which are archaebacterial. The remaining three paralogs had no detectable homologs in prokaryotic genomes applying the sequence similarity threshold used in this study (>25% identical amino acids). This indicates that PFD is an archaebacterial contribution to eukaryotic genomes, and the family further diversified within eukaryotes. All five small heat shock proteins (sHsp) were inferred to be of eubacterial ancestry. Hsp26 is homologous to eubacterial sequences only, and the four paralogous genes Hsp31, Hsp32, Hsp33, and Sno4 clearly branch within the eubacterial clade, although homologs in halophilic and methanogenic archaebacteria were found as well. Members of the Hsp100 chaperone family (Clp) play a role in protein disaggregation (Parsell et al. 1994). Of the three Hsp100 proteins in yeast, one is localized in the mitochondria and two are cytosolic (van Dyck et al. 1998). The mitochondrial Clp protein Mcx1 was inferred to be of eubacterial origin. The cytosolic Hsp104 was inferred to have been derived from an archaebacterial AAA+ ATPase, while the second cytosolic Hsp78 is of ambiguous ancestry. The Hsp40 and

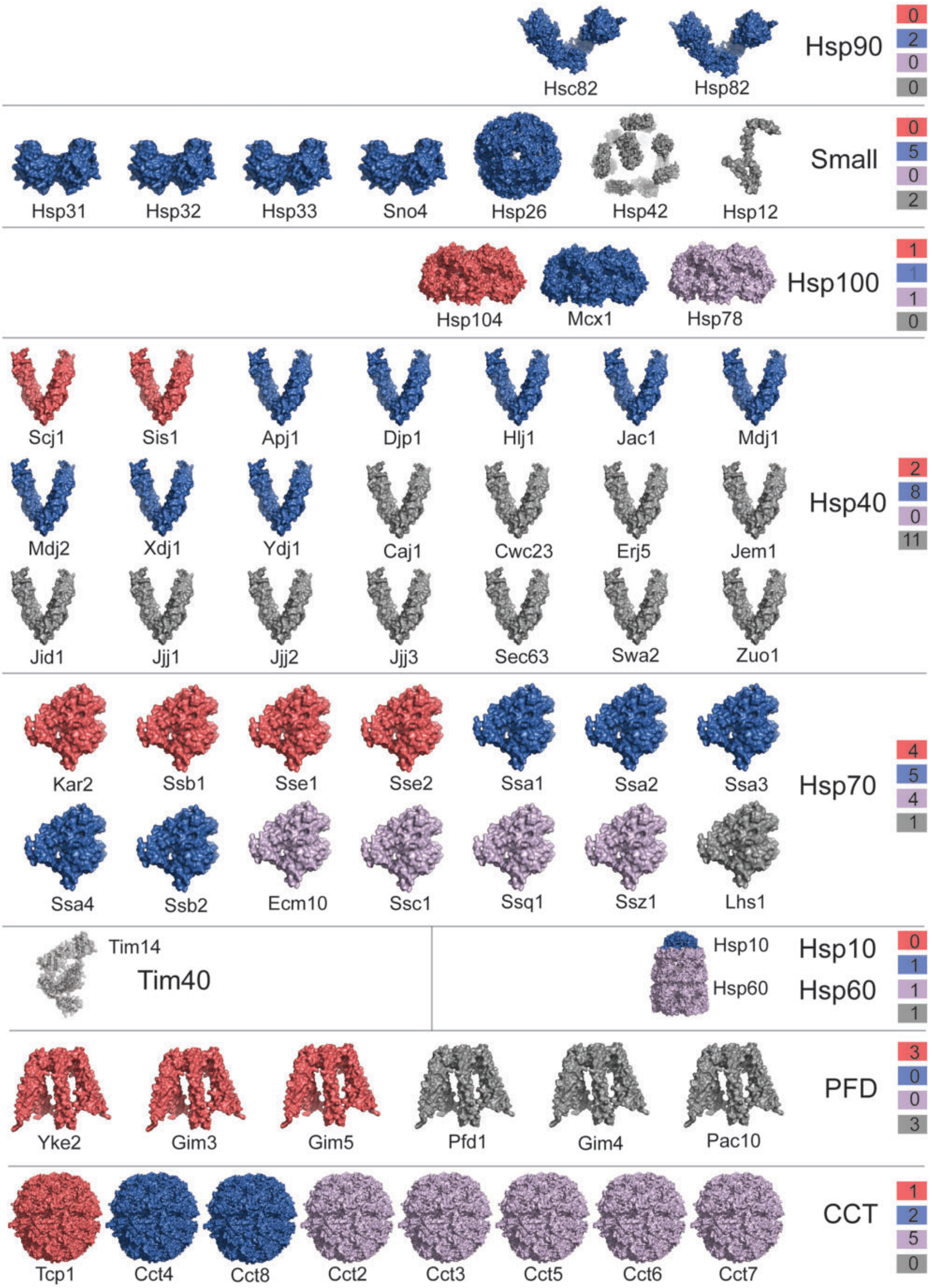


Fig. 1. Yeast chaperones and their reconstructed ancestries. Archaebacterial ancestry is shown in red and eubacterial ancestry in blue. Chaperones with ambiguous ancestry or no homology to prokaryotic proteins are colored in purple and gray, respectively. Here we use the same structural model for all members of the same family; Note that paralogs may deviate in their protein structures. Molecule plots were generated using the PyMOL Molecular Graphics System, version 1.5.0.4 (Schrödinger, LLC).

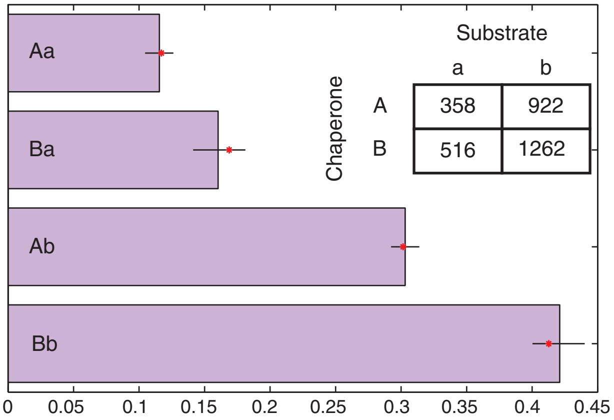
Hsp70 families include chaperones with eubacterial as well as archaebacterial ancestry, although the majority of chaperones from these particular families are of eubacterial descent.

Eukaryotic genomes typically encode two chaperonin systems: the type I mitochondrial Hsp60/Hsp10 system (GroEL/ES-like) and the type II chaperonin (CCT-like). The type I chaperonin system is usually viewed as a eubacterial set of chaperones; however, it is also encoded in the genomes of several methanogenic and halophilic archaebacteria (e.g., Deppenmeier et al. 2002). The yeast Hsp60 branched in between a purely archaebacterial clade and a purely eubacterial clade. Consequently, it was classified as of ambiguous prokaryotic ancestry. The cochaperone Hsp10 is clearly of eubacterial origin. This classification fits well with its localization in the mitochondrion. The type II eukaryotic chaperonins comprise eight different protein subunits (Archibald et al. 1999; Valpuesta et al. 2002). These chaperones are usually viewed as archaebacterial; however, several Clostridia species encode type II chaperonins as well (Techtmann and Robb 2010; Williams et al. 2010). An archaebacterial ancestry was inferred for Tcp1 and a eubacterial origin was inferred for Cct4 and Cct8. The other five CCT genes were classified as ambiguous as they branch between clostridial and archaebacterial homologs.

Connectivity in the Chaperone Interaction Network and Protein Ancestry

The chaperone–substrate interaction (CSI) network is based on a large-scale screening for proteins that interact with 64 chaperones encoded in *Saccharomyces cerevisae* (Gong et al. 2009). The CSI network contains 4,340 substrate proteins that interact with at least one chaperone and a total of 21,428 CSIs. Interactions in the CSI network are unweighted and do not reflect their relative prevalence. We reduced the data set to include only those chaperones and substrates for which prokaryotic ancestry could be determined. This network contained 36 chaperones and 790 substrates. A total of 3,058 interactions included in the network were classified into four classes based on the ancestry of both the chaperones and substrates (inset in fig. 2).

The network connectivity pattern is not biased toward a higher number of interactions between chaperones and substrates of the same ancestry (χ^2 test; $P = 0.52$, inset in fig. 2). This type of network data may sometimes be biased by nodes having extreme connectivity or gene expression levels. To guard against such a possibility, we classified both chaperones and substrates into high/low categories according to the following properties: network connectivity degree, mRNA expression, and protein expression. We repeated the analysis with subsets of the network defined by these contrasts and observed the same pattern as in the full network, indicating that the result is robust (see supplementary table S3, Supplementary Material online). Moreover, this conclusion still holds when considering only substrates that interact with at least two chaperones or more (χ^2 test; $P = 0.49$). Although the mean connectivity degree of substrates of archaebacterial ancestry (5.33) is higher than that of substrates of eubacterial ancestry (4.81), this difference is not statistically significant (Wilcoxon rank-sum test, $P = 0.07$). To further test for possible biases in the network connectivity pattern, we examined the ratio of eubacterial to archaebacterial interaction partners for each chaperone and substrate, and tested for differences in the distributions of ratios in the two ancestry groups. We found no significant difference in the distributions of the chaperone ancestry ratio between archaeal and eubacterial substrates (Wilcoxon rank-sum test, $P = 0.62$), and no significant difference in the distributions of the substrate ancestry ratio between archaeal and eubacterial chaperones (Wilcoxon rank-sum test, $P = 0.18$). We further tested whether any of the four chaperone–substrate ancestry combinations is enriched in the network by conducting a network randomization test with 10,000 randomization replicates (fig. 2). This analysis shows that none of the four interaction types is found at a frequency that is significantly different from the random expectation (at a false discovery rate [FDR] of 0.01).



		Substrate	
		a	b
Chaperone	A	358	922
	B	516	1262

Fig. 2. Prokaryotic origin and connectivity distribution. Asterisks indicate the observed percentage of edges in the network, and bars show the mean expected frequency from randomization simulations. Lines indicate the 1–99 percentile range. Abbreviations: A, archaebacterial, B, eubacterial; uppercase indicates chaperones and lowercase indicates substrates.

Protein Ancestry and Protein Function

Substrates in the network were further classified into two major functional categories according to their annotation in the Gene Ontology database (GO, Ashburner et al. 2000). Substrates whose annotation includes the terms “translation,” “transcription,” “DNA-dependent DNA replication,” or their subterms were classified as proteins performing an informational function. The remaining substrates were classified as operational proteins (Rivera et al. 1998; Cotton and McInerney 2010). Combining the functional classification with prokaryotic ancestry reconstruction revealed that 59% of the 216 archaebacterial substrates and 15% of the 528 eubacterial substrates found in GO perform informational functions. Hence, substrates of archaebacterial origin are enriched for informational functions ($P < 10^{-16}$, using χ^2 test), confirming the known correlations between prokaryotic ancestry and protein function (Esser et al. 2004; Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011; Alvarez-Ponce et al. 2013). In addition, we found that informational substrates interact with a larger number of different chaperones than operational substrates (Wilcoxon rank-sum test, $P < 10^{-16}$).

Prokaryotic Ancestry and Protein Physicochemical Properties

A comparison of protein physicochemical properties between the two ancestry groups revealed several significant differences. The differences are manifest in proteins that interact with chaperones as well as in proteins that are chaperon independent. Interestingly, the differences observed in chaperone independent proteins are significantly larger than those in the chaperone substrates (fig. 3).

Eubacterial substrates were found to be longer on average, in agreement with previous studies (Alvarez-Ponce and McInerney 2011). In addition, eubacterial substrates are also enriched in hydrophobic and aromatic amino acids in comparison to archaebacterial substrates. Archaebacterial substrates are more conserved, more highly expressed, and are encoded by higher proportions of preferred codons than eubacterial substrates (fig. 3). Biases in the three latter properties fit well with the known correlation among evolutionary rates, expression level, and codon usage bias (Grantham et al. 1981; Sharp and Li 1987; Pál et al. 2001; Drummond et al. 2005; Pál et al. 2006). In addition, substrates of archaebacterial origin were enriched for positively charged amino acids as well as arginine, lysine, and valine. On the other hand, substrates of eubacterial origin are significantly enriched in cysteine, histidine, isoleucine, leucine, phenylalanine, proline, serine, and tryptophane (fig. 3).

Most of the above differences in substrate physicochemical properties are observed when contrasting informational and operational proteins, as expected from the

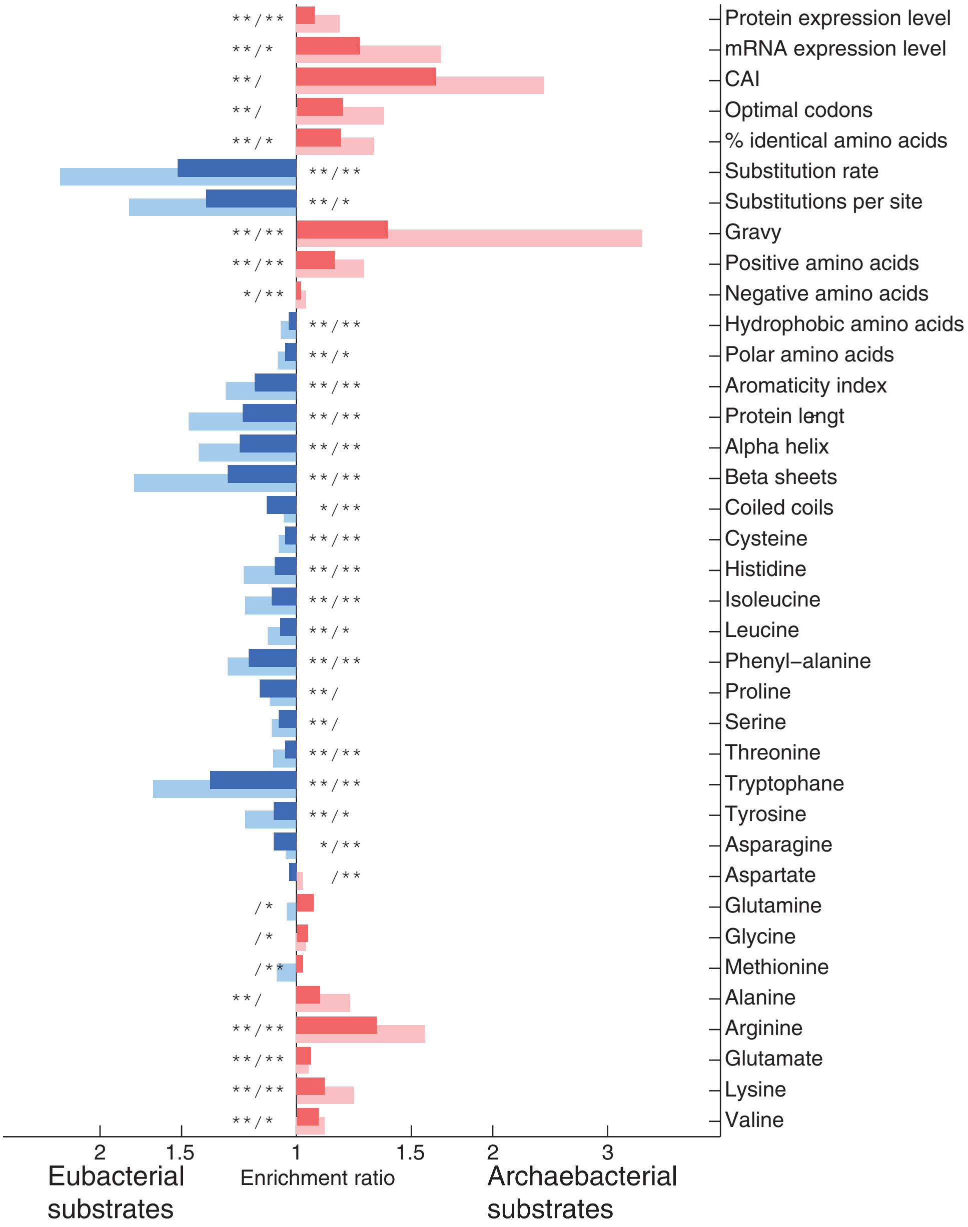


Fig. 3. Differences in protein physicochemical properties between proteins of eubacterial and archaebacterial origin. Enrichment in proteins of eubacterial origin is on the left and shown as blue shades and that of proteins of archaebacterial origin on the right and shown as red shades. Chaperone substrates are in dark shades and proteins not connected to chaperones are in light shades. Asterisks denote statistical significance (Kolmogorov–Smirnov tests); * denotes 5% FDR and ** 1% FDR; Asterisks to the left of slash refers to tests contrasting protein ancestries and asterisks to the right of slash refers to tests contrasting substrates with chaperone-independent proteins. Bar lengths indicate the enrichment ratio in log 10 scale.

congruence between the ancestral and functional classifications.

Discussion

Our evolutionary reconstruction of the ancestry of chaperones involved in the yeast protein, folding pathway reveals that chaperones of different descent are used in a coordinated fashion to fold common substrates. For example, the Hsp40/Hsp70 system in yeast comprises a total of 21 Hsp40 and 14 Hsp70 genes from diverse origins including archaebacterial, eubacterial, and eukaryotic-specific proteins (ESPs). Interestingly, the Hsp40 family, with 11 ESPs, has diversified within eukaryotes to a larger extent in comparison to the Hsp70 family that includes only one ESP. The difference between the two families can be explained by their mode of function. Chaperones of the Hsp40 family are the drivers of Hsp70 substrate activity and specificity (Cyr and Douglas 1994; Kampinga and Craig 2010). Thus, the diversification of Hsp40 family within eukaryotes probably enabled the whole Hsp40–Hsp70 system to increase its operational potential. A mosaic of ancestries is observed in all chaperone families that are present in both archaebacteria and eubacteria. It is noteworthy that in contrast to cytosolic chaperones, yeast chaperones that are localized in the mitochondria are an exception. All mitochondrial chaperones that could be classified by their tree topology are inferred to be of eubacterial ancestry, underlining the role of the mitochondrion as a functional eubacterial unit within the eukaryotic cell (Esser et al. 2004).

Previous studies showed that there is a significant preference for proteins to interact with partners of the same ancestry rather than across the archaebacterial–eubacterial divide (Alvarez-Ponce and McInerney 2011). Such preference can be expected if the proteins participating in specific cellular pathways are usually of a single ancestry. Because protein connectivity is higher within pathways than across pathways, common ancestry of pathway proteins will result in an overall trend for same ancestry interactions. Thus, same ancestry preference, while demonstrable on average, may still be violated when considering specific systems. Our results suggest that the general trend does not hold for the CSI network, where no preference for interaction of chaperones and substrates of the same ancestry could be observed. This indicates that the protein folding pathways have been reorganized and integrated to a larger extent in comparison to the overall protein–protein interactions within the cell.

Yeast proteins originating from the two endosymbiosis partners are distinct in their physicochemical properties profile (Alvarez-Ponce and McInerney 2011). These differences, while still significant, are much smaller among proteins that utilize chaperones in their folding pathway than among chaperone-independent proteins. The molecular features that enable substrates to interact with chaperones, while not yet well understood, are likely to place constraints on the various physicochemical properties and can thus result in greater similarity of chaperone substrates when compared with other proteins. Moreover, adaptation to chaperone assisted folding is likely to affect these same features, actively driving substrates away from their ancestral profile and toward a common eukaryotic profile. For example, archaebacterial substrates are expressed in significantly higher levels than eubacterial substrates, yet these differences are still significantly smaller than those observed for proteins that do not interact with chaperones. In the crowded cell environment, successful competition for chaperones is likely to be linked to expression levels, thus putting eubacterial proteins in a disadvantage. Thus, a narrower expression range for substrate proteins may provide a competition field that is more balanced. This can be seen as a homogenizing effect of chaperones on their substrates, and from this perspective, chaperones can be viewed as inducers of the eukaryotic integrated state. Thus, chaperones have a cumulative impact on eukaryotic genome evolution.

What makes molecular chaperones a class of proteins that is more amenable to integration? Chaperones are highly versatile proteins that increase the probability of their substrates to attain a functional conformation and by that can contribute significantly to the organismal fitness. Chaperones are essential in both prokaryotic domains (Hartl and Hayer-Hartl 2002; Calloni et al. 2012); hence, at the very origin of the eukaryotic cytosol, there was an absolute need for chaperones of both ancestries to assist in the folding of their respective substrates. Some molecular chaperones are very versatile, and in vitro they can assist folding of substrates from unrelated organisms, even from another prokaryotic domain (e.g., Yam et al. 2008). Moreover, similar chaperones may have similar substrate specificity and interact with similar sets of proteins. Therefore, eubacterial and archaebacterial chaperones might have had overlapping substrate sets at the initial steps of eukaryogenesis. In vivo, however, this capacity may not be sufficient, as the organism must sustain a balanced stoichiometric and energetic profile, which requires chaperones and substrate expression to be coordinated by a common regulatory regime. Thus, a nonspecific interaction pattern allows chaperones to acquire new clients without the need for intensive sequence modification or adaptation, and the evolution of a completely integrated system is expected to also include the regulatory context governing coexpression of chaperones and their substrates as well as optimizing the competitive binding of substrates and their dedicated chaperones.

The effects of combining two ancestral chaperone systems may have conferred an even larger fitness benefit than was possible by either of the ancestral systems on its own. Moreover, the apparent redundancy in the chaperone repertoire may reflect not only the demands of protein folding pathways but the possibility that some chaperones are involved in other functions. Chaperones have been reported to posses such moonlighting functions (e.g., Wuppermann et al. 2008, see Henderson et al. 2013 for a review). Moonlighting may also explain the expansion and diversification observed in several of the larger eukaryotic chaperone families.

Nonetheless, retaining two chaperone systems would have entailed an additional energetic cost for the cell as chaperone synthesis and operation is expensive in terms of ATP usage.

In the context of eukaryogenesis, this would not have posed an insurmountable problem, since the formation of mitochondria as an intracellular organelle resulted in a dramatic increase in the available energy for all cellular processes (Lane and Martin 2010). Nevertheless, energetic considerations might still play a role in the evolution of CSIs (Bogumil and Dagan 2012).

In summary, in contrast with other proteins that still show a tendency to form network communities reflecting their ancestries, molecular chaperones have been able to cross the divide between the ancestral prokaryotic domains. The central role of chaperone-assisted folding in maintaining cellular fitness is reflected in the high degree of integration of an archaebacterial and a eubacterial chaperone systems into one at the origin of eukaryotes.

Materials and Methods

Data

Yeast protein sequences, amino acid usage data, functional assignments, chromosomal locations, frequencies of optimal codons, codon adaptation index, gravy scores (hydropathy index), and aromaticity scores were downloaded from the *Saccharomyces* Genome Database (Cherry et al. 1998). Chaperone–protein interaction data were obtained from Gong et al. (2009). The secondary structure of all proteins was inferred using PsiPred (Jones 1999), applying a threshold of 70% for the calculation of secondary structure probability. Quantitative protein expression data were obtained from Ghaemmaghami et al. (2003). The mRNA levels data were obtained from Wang et al. (2002). For the statistical analysis of protein expression levels, natural log transformation was applied. Proteins for which expression levels were not available (107 in total) or with zero expression level (1,665 proteins) were excluded from the analysis. All statistical analyses were performed using the MatLab Statistics toolbox.

Evolutionary Rate

Positional orthology assignments among 20 fungal genomes were obtained from Wapinski et al. (2007). Proteins lacking orthologs in any genome (282 in total) were excluded from the analysis. Multiple sequence alignments of all yeast open reading frames with orthologous sequences were generated with MAFFT (Katoh et al. 2005). Phylogenetic trees were reconstructed with PhyML v3.0_360-500 M (Guindon and Gascuel 2003) using the best-fit model as inferred by ProtTest 3 (Darriba et al. 2011) according to the Akaike information criterion measure (Akaike 1974). Distances from the *S. cerevisiae* proteins to their orthologs were calculated as the sum of branch lengths. To calculate the relative amino acid substitution rates of substrates, the distances to the 20 proteomes were first Z-transformed separately and then averaged over all orthologs (Bogumil and Dagan 2010).

Reconstruction of Prokaryotic Ancestries

We classified each of the 5,880 yeast protein-coding genes into archaebacterial, eubacterial, ambiguous prokaryotic ancestry, or eukaryote-specific, based on its phylogenetic affinities. Each yeast protein sequence was used as a query in a homology search against a database containing the proteomes of 82 archaebacteria and 1,074 eubacteria (3,792,506 proteins in total). Homology searches were carried out using position specific iterated–basic local alignment search tool (PSI-BLAST) (Altschul et al. 1997) without filtering. Global pairwise alignments of BLAST-hits were calculated using the EMBOSS package (Needleman and Wunsch 1970; Rice et al. 2000). Prokaryotic sequences with less than 25% identity were considered as having no significant similarity to the particular yeast query. Of the yeast genes, 161 had significant similarity to archaebacterial sequences exclusively (and were thus classified as being of archaebacterial ancestry), 383 had significant similarity to eubacterial sequences only (and were thus deemed as eubacterium-derived), and 686 had homologs in both prokaryotic domains. The remaining genes had no detectable prokaryotic homologs at the specified thresholds and were thus considered eukaryote-specific.

To ascertain the ancestry of the 686 yeast genes with both archaebacterial and eubacterial homologs, we conducted a phylogenetic analysis. For each of these genes, a multiple sequence alignment including the 15 best BLAST hits from each prokaryotic domain was generated using MAFFT v6.843 b (Katoh and Toh 2008), and the quality of the alignment was tested with guidance (Penn et al. 2010). To be conservative in our analysis, columns with a confidence score <0.93 were removed. Phylogenetic trees were reconstructed with PhyML v3.0_360-500 M (Guindon and Gascuel 2003) using the best-fit model as inferred by ProtTest 3 (Darriba et al. 2011) according to the Akaike information criterion (Akaike 1974).

We next rooted each tree on the branch that maximized the separation of archaebacterial and eubacterial sequences. The internal branch yielding the maximum ratio of archaebacteria to eubacteria content in the resulting clades was determined with the MRP function implemented in CLANN 3.2.2 (Creevey and McInerney 2005) using Spearman's rank correlation coefficient. The yeast gene was classified as of eubacterial or archaebacterial ancestry depending on the clade within which it branched (see supplementary fig. S1, Supplementary Material online, for illustrative trees). Yeast genes were considered of ambiguous ancestry if no branch yielded a clear separation into an archaebacterial and eubacterial clades, if multiple branches separated the archaebacterial and eubacterial sequences equally well and resulted in conflicting ancestry assignments, or if the yeast gene branched between the archaebacterial and eubacterial clades. In such ambiguous cases, we repeated the analysis with a larger sample of homologous sequences, first with the 30 best BLAST hits from each domain, and if still ambiguous, with the 45 best BLAST hits from each domain. This analysis shifted 125 genes from the ambiguous to the unambiguous class. Of the 686 yeast genes with both archaebacterial and eubacterial homologs, 128 were classified as of archaebacterial ancestry, 420 as of eubacterial ancestry, and 138 as ambiguous. All phylogenetic trees are provided in supplementary tables S1 and S2, Supplementary Material online.

In total, we inferred 289 proteins to be of archaebacterial ancestry, 803 of eubacterial ancestry, and 138 proteins with an unresolvable prokaryotic ancestry. The remaining yeast proteins did not show significant similarity with any prokaryotic protein.

Network Randomization

Randomization of the CSI network was carried out using the switching methodology (Stone and Roberts 1990; Artzy-Randrup and Stone 2005) implemented in an in-house MatLab script.

Supplementary Material

Supplementary figure S1 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Robin Koch, Ovidiu Popa, and Thorsten Thiergart for their helpful comments on the manuscript and acknowledge support from the European Research Council (grant no. 281357).

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr.* 19:716–723.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Alvarez-Ponce D, McInerney JO. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biol Evol.* 3:782–790.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci U S A.* 110:1594–1603.

Archibald JM, Logsdon JM, Doolittle WF. 1999. Recurrent paralogy in the evolution of archaeal chaperonins. *Curr Biol.* 9:1053–1056.

Artzy-Randrup Y, Stone L. 2005. Generating uniformly distributed random networks. *Phys Rev E.* 72:056708.

Ashburner M, Ball CA, Blake JA, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25:25–29.

Bogumil D, Dagan T. 2010. Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biol Evol.* 2:602–608.

Bogumil D, Dagan T. 2012. Cumulative impact of chaperone-mediated folding on genome evolution. *Biochemistry* 51:9941–9953.

Bogumil D, Landan G, Ilhan J, Dagan T. 2012. Chaperones divide yeast proteins into classes of expression level and evolutionary rate. *Genome Biol Evol.* 4:618–625.

Calloni G, Chen T, Schermann SM, Chang H-C, Genevaux P, Agostini F, Tartaglia GG, Hayer-Hartl M, Hartl FU. 2012. DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Rep.* 1: 251–264.

Cherry J, Adler C, Ball C, et al. (12 co-authors). 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* 26:73–79.

Cotton JA, McInerney JO. 2010. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc Natl Acad Sci U S A.* 107: 17252–17255.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A.* 105: 20356–20361.

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* 21:390–392.

Cyr DM, Douglas MG. 1994. Differential regulation of Hsp70 subfamilies by the eukaryotic DnaJ homologue YDJ1. *J Biol Chem.* 269: 9798–9804.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27: 1164–1165.

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

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 102:14338–14343.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* 440:623–630.

Esser C, Ahmadinejad N, Wiegand C, et al. (15 co-authors). 2004. A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol.* 21:1643–1660.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. GroEL buffers against deleterious mutations. *Nature* 417:398.

Gabaldón T, Huynen MA. 2003. Reconstruction of the proto-mitochondrial metabolism. *Science* 301:609.

Ghaemmaghami S, Huh W, Bower K, Howson R, Belle A, Dephoure N, O'Shea E, Weissman J. 2003. Global analysis of protein expression in yeast. *Nature* 425:737–741.

Gong Y, Kakihara Y, Krogan N, Greenblatt J, Emili A, Zhang Z, Houry WA. 2009. An atlas of chaperone–protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell. *Mol Syst Biol.* 5:1–14.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* 9:43—74.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. *Science* 283:1476–1481.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Hartl FU. 2002. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* 295:1852–1858.

Hartl FU, Hayer-Hartl M. 2009. Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol.* 16:574–581.

Henderson B, Fares MA, Lund PA. 2013. Chaperonin 60: a paradoxical, evolutionarily conserved protein family with multiple moonlighting functions. *Biol Rev Camb Philos Soc.* 88:955–987.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 292:195–202.

Kampinga HH, Craig EA. 2010. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat Rev Mol Cell Biol.* 11:579–592.

Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–518.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinformatics.* 9:286–298.

Lane N. 2009. Life ascending: the ten greatest inventions of evolution. London: Profile Books. p. 344.

Lane N, Martin W. 2010. The energetics of genome complexity. *Nature* 467:929–934.

Large AT, Goldberg MD, Lund PA. 2009. Chaperones and protein folding in the archaea. *Biochem Soc Trans.* 37:46.

Large AT, Lund PA. 2009. Archaeal chaperonins. *Front Biosci.* 14: 1304–1324.

Macario AJ, Dugan CB, Clarens M, Conway de Macario E. 1993. dnaJ in Archaea. *Nucleic Acids Res.* 21:2773.

Macario AJ, Dugan CB, Conway de Macario E. 1991. A dnaK homolog in the archaebacterium *Methanosarcina mazei* S6. *Gene* 108:133–137.

Maisnier-Patin S, Roth JR, Fredriksson Å, Nyström T, Berg OG, Andersson DI. 2005. Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nat Genet.* 37:1376–1379.

Martin W. 2003. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc Natl Acad Sci U S A.* 100:8612–8614.

Martin W, Herrmann R. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol.* 118:9–17.

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

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 48:443–453.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci U S A.* 109:20537–20542.

Parsell DA, Kowal AS, Singer MA, Lindquist S. 1994. Protein disaggregation mediated by heat-shock protein Hsp104. *Nature* 372:475–478.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7:337–348.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* 38:W23–W28.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol.* 24: 1752–1760.

Queitsch C, Sangster T, Lindquist S. 2002. Hsp90 as a capacitor of phenotypic variation. *Nature* 417:618–624.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16:276–277.

Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A.* 95:6239–6244.

Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155.

Rutherford SL, Lindquist S. 1998. Hsp90 as a capacitor for morphological evolution. *Nature* 396:336–342.

Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol.* 14:255–274.

Sharp PM, Li WH. 1987. The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281—1295.

Stone L, Roberts A. 1990. The checkerboard score and species distribution. *Oecologia* 85:74–79.

Techtmann SM, Robb FT. 2010. Archaeal-like chaperonins in bacteria. *Proc Natl Acad Sci U S A.* 107:20269—20274.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol.* 4:466–485.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet.* 5:123–135.

Vainberg I, Lewis S, Rommelaere H, Ampe C, Vandekerckhove J, Klein H, Cowan N. 1998. Prefoldin, a chaperone that delivers unfolded proteins to cytosolic chaperonin. *Cell* 93:863–873.

Valpuesta J, Martin-Benito J, Gomez-Puertas P, Carrascosa J, Willison K. 2002. Structure and function of a protein folding machine: the eukaryotic cytosolic chaperonin CCT. *FEBS Lett.* 529:11–16.

van Dyck L, Dembowski M, Neupert W, Langer T. 1998. Mcx1p, a ClpX homologue in mitochondria of *Saccharomyces cerevisiae*. *FEBS Lett.* 438:250–254.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A.* 99:5860–5865.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449: 54–61.

Williams TA, Codoñer FM, Toft C, Fares MA. 2010. Two chaperonin systems in bacterial genomes with distinct ecological roles. *Trends Genet.* 26:47–51.

Williams TA, Fares MA. 2010. The effect of chaperonin buffering on protein evolution. *Genome Biol Evol.* 2:609–619.

Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc Roy Soc B.* 279: 4870–4879.

Wuppermann FN, Molleken K, Julien M, Jantos CA, Hegemann JH. 2008. *Chlamydia pneumoniae* GroEL1 protein is cell surface associated and required for infection of HEp-2 cells. *J Bacteriol.* 190: 3757–3767.

Yam AY, Xia Y, Lin H-TJ, Burlingame A, Gerstein M, Frydman J. 2008. Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nat Struct Mol Biol.* 15:1255–1262.

Young J, Agashe V, Siegers K, Hartl F. 2004. Pathways of chaperone-mediated protein folding in the cytosol. *Nat Rev Mol Cell Biol.* 5: 781–791.