The Human Genome Retains Relics of Its Prokaryotic Ancestry: Human Genes of Archaebacterial and Eubacterial Origin Exhibit Remarkable Differences

David Alvarez-Ponce and James O. McInerney*

Department of Biology, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

*Corresponding author. E-mail: james.o.mcinerney@nuim.ie.

Accepted: 12 July 2011

Abstract

Eukaryotes are generally thought to stem from a fusion event involving an archaebacterium and a eubacterium. As a result of this event, contemporaneous eukaryotic genomes are chimeras of genes inherited from both endosymbiotic partners. These two coexisting gene repertoires have been shown to differ in a number of ways in yeast. Here we combine genomic and functional data in order to determine if and how human genes that have been inherited from both prokaryotic ancestors remain distinguishable. We show that, despite being fewer in number, human genes of archaebacterial origin are more highly and broadly expressed across tissues, are more likely to have lethal mouse orthologs, tend to be involved in informational processes, are more selectively constrained, and encode shorter and more central proteins in the protein–protein interaction network than eubacterium-like genes. Furthermore, consistent with endosymbiotic theory, we show that proteins tend to interact with those encoded by genes of the same ancestry. Most interestingly from a human health perspective, archaebacterial genes are less likely to be involved in heritable human disease. Taken together, these results show that more than 2 billion years after eukaryogenesis, the human genome retains at least two somewhat distinct communities of genes.

Key words: origin of eukaryotes, endosymbiosis, protein interaction network.

Introduction

The relationships among the three domains of cellular life (Eubacteria, Archaebacteria, and Eukaryotes; Woese et al. 1990) and in particular the exact phylogenetic placement of eukaryotes and the mechanisms underlying their origin (eukaryogenesis) have been the subject of ferocious debate for decades (Martin et al. 2001; Embley and Martin 2006; Kurland et al. 2006; Gribaldo et al. 2010). A number of hypotheses have been proposed, many of which posit that eukaryotes arose from a fusion event involving a eubacterium (the ancestor of the present-day mitochondrion) and an archaebacterium (Sagan 1967; Zillig et al. 1985; Rivera and Lake 2004; Pisani et al. 2007; Lane and Martin 2010). According to these hypotheses, Eukaryotes are a secondary domain, derived from the other two. Following this event, most eubacterium-derived genes may have been transferred to the nucleus (for a review, see Timmis et al. 2004). Support for a fusion event, which would have taken place more than 2 billion years ago (Brocks et al. 1999), comes from the fact that extant eukaryotic nuclear genomes contain genes that manifest sister-group phylogenetic relationships with both eubacterial and archaebacterial genes (Rivera et al. 1998; Horiike et al. 2001; Esser et al. 2004; Cotton and McInerney 2010).

Yeast genes derived from both fusion partners have been shown to play different roles in cellular metabolism, with archaebacterial genes being more likely to be involved in transcription, translation, and replication (i.e., informational processes) and eubacterial genes being preferentially involved in operational processes (Rivera et al. 1998; Cotton and McInerney 2010). Additionally, it has been recently shown that, regardless of function, yeast genes of archaebacterial origin are more highly expressed, are more essential, and encode more central proteins in the protein–protein interaction network (PIN) than eubacterium-derived genes (Cotton and McInerney 2010).

Here we test whether the different prokaryotic ancestries of human genes have an effect on phenotype, essentiality to

The Author(s) 2011. Published by Oxford University Press on behalf of the *Society for Molecular Biology and Evolution*.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

the organism, selective constraint, function, expression level and breadth, and position of the encoded products in the PIN. We also test whether the human PIN is stratified along lines of ancestry, with proteins interacting preferentially with those encoded by genes inherited from the same endosymbiotic partner.

Methods

Human Proteome

We retrieved the human proteome (21,894 proteins) from the Ensembl database release 59 (Hubbard et al. 2009). When multiple proteins were encoded by the same gene, only the longest one was used in our analyses. After eliminating proteins shorter than 50 amino acids (which are unlikely to contain enough phylogenetic information for accurate ancestry assignment), we retained 21,712 proteins. For each gene, we retrieved the following information from different sources:

Number of Paralogs

For each human gene, a list of paralogs was retrieved from Ensembl's BioMart (Kasprzyk et al. 2004). Of the studied human genes, 15,011 had at least one paralog.

Mouse Orthologs

For each human gene, we retrieved from BioMart a list of mouse orthologs. For each mouse gene, we retrieved phenotypic information from the Mouse Genome Database (Bult et al. 2008) ("MRK_Ensembl_Pheno.rpt" file downloaded on 7 October 2010). Mouse orthologs represented in this database were considered to be "lethal" if classified either as embryonic, perinatal, or postnatal lethal or as viable otherwise. A total of 2,633 human genes had at least one lethal mouse ortholog, 3,415 had only viable orthologs, and 15,664 had no orthologs or orthologs without available phenotypic information only.

For 16,471 human genes with a single mouse ortholog, we retrieved the nonsynonymous (d_N) and synonymous (d_S) divergence levels resulting from the human–mouse comparison from BioMart. The median values were calculated to be 0.070 and 0.604, respectively. We then used this information to compute the $\omega = d_N/d_S$ ratios. Highly constrained genes are expected to have low ω values, whereas genes evolving neutrally are expected to exhibit ω values close to 1.

Involvement in Human Disease

For every human gene, we retrieved from BioMart a list of the diseases in which the gene is implicated. A total of 2,694 genes with at least an "MIM Morbid Accession" assigned in the Online Mendelian Inheritance in Man database (Amberger et al. 2009) were considered to be involved in disease.

Biological Processes

For each human gene, we retrieved from BioMart a list of Gene Ontology (GO) (Ashburner et al. 2000) "biological process" terms. Genes were classified as informational if involved in "transcription," "translation," or "replication" (610 genes) or as operational otherwise (14,902 genes). The remaining 6,200 genes had no GO biological process term assigned.

Cellular Compartments

We likewise retrieved the GO "cellular component" terms to which each protein is assigned. A total of 1,358 proteins assigned to the term "mitochondrion" were classified as mitochondrial, 15,857 as nonmitochondrial, and the remaining 4,497 had no GO cellular component term assigned.

Gene Expression Level and Breadth

Gene expression data for 84 human tissues (or organs) were retrieved from Su et al. (2004) (U133A/GNF1H data set GCRMA normalized). Probes were matched to Ensembl accession IDs through BioMart (for the U133A data set) and through the annotation file provided with the data set (for the GNF1H data set). Of the genes in our data set, 16,622 could be matched to at least one probe. We used a subset of 25 nonredundant, adult noncancerous tissues (as in Alvarez-Ponce et al. 2011) for some analyses. For each probe and tissue, values were averaged across both replicates. For each probe, expression level was calculated as the average across the 25 selected tissues. For genes matching more than one probe, the one with the highest average across the 25 selected tissues was used. Expression breadth of each gene was calculated as the number of tissues (out of the 25 selected ones) in which the gene is expressed above the median across all tissues and genes.

Protein–Protein Interaction Data

We retrieved the human PIN from BioGRID version 3.0.67 (Breitkreutz et al. 2008). Only physical interactions among human proteins were considered. After removing 84 proteins that could not be matched to an Ensembl accession number and 74 that are shorter than 50 amino acids, the network (PIN1; supplementary fig. S1, Supplementary Material online) consisted of 30,528 interactions connecting 8,370 proteins. For each protein, degree was computed as the number of proteins to which it is connected, and betweenness and closeness centralities were computed using the NetworkX package (http://networkx.lanl.gov/). Genes not represented in the PIN1 network were assigned missing values.

Homology Searches

Each human protein was used as query in a homology search against a database containing the proteomes of

1,074 eubacteria and 82 archaebacteria (3,792,506 sequences in total), obtained from the National Center for Biotechnology Information in August 2010 (supplementary table S1, Supplementary Material online). Searches were carried out using context-specific iterative basic local alignment search tool (Biegert and Soding 2009) with two iterations and an *E*-value cut-off of 10^{-5}. We applied two criteria for homology assignment. In the first one, genes were classified as archaebacterial or eubacterial on the basis of the best hit retrieved, being only deemed ambiguous if there were hits of both domains sharing the best position (i.e., with the same *E* value). In an additional more conservative analysis, a human gene was considered ambiguous unless all its prokaryotic homologs belonged to the same domain or the *E* values of the best hits in both domains differed in at least 10 orders of magnitude. Throughout the main text, we use the first criterion; results using the second criterion are reported in supplementary Results (Supplementary Material online).

Statistical Tests of Association

We tested for differences in the studied parameters between pairs of gene groups using the nonparametric Mann–Whitney *U* test (for continuous variables) or the odds ratio (OR) (for categorical ones). ORs whose 95% confidence interval (CI) does not overlap the unity were considered significant. Correlations among continuous variables were evaluated using the nonparametric Spearman's rank correlation coefficient. We used partial correlation to evaluate differences between archaebacterial and eubacterial genes while controlling for potentially confounding variables. For this analysis, ancestry was encoded as a dummy variable (see Hardy 1993; Cohen et al. 2003).

Network-Level Analysis

The statistical significance of measured network parameters (e.g., number of observed interactions involving proteins with the same ancestry) was assessed on the basis of 10,000 randomized networks, each with the same nodes and the same number of interactions as the original one. Each interaction was assigned by choosing two proteins at random from a list in which each protein was represented a number of times equal to its degree, thus approximately preserving the degree of each particular node in each simulated network. Randomized networks were obtained using an in-house PERL program, which is available upon request. One-tailed *P* values (P_1) were computed as the proportion of simulations with a statistic value greater than or equal to the observed one. Two-tailed *P* values were then computed as $1 - 2 \times |0.5 - P_1|$. In addition to the PIN1 network, a number of subnetworks (supplementary fig. S1, Supplementary Material online) were used in this analysis in order to rule out potential biases introduced by various network features.

Results

For each human protein, we performed a homology search against a database of ~3.8 million prokaryotic sequences belonging to 1,074 eubacteria and 82 archaebacteria (supplementary table S1, Supplementary Material online) and assigned its ancestry on the basis of the domain to which the best hit belongs. This resulted in 939 genes (4.3%) being classified as archaebacterial, 7,884 (36.1%) as eubacterial, 204 (0.9%) deemed ambiguous (i.e., with genes of both domains sharing the lowest *E* value), and 12,685 genes (58.4%) without detectable prokaryotic homologs. These proportions are in good agreement with previous analyses of the yeast genome (Rivera et al. 1998; Esser et al. 2004; Cotton and McInerney 2010). In an additional, more conservative analysis, a gene's ancestry was considered ambiguous unless all prokaryotic homologs belonged to the same domain or the *E* values of the best hits in both domains differed by at least 10 orders of magnitude (see supplementary Results and tables S2 and S3, Supplementary Material online).

We mapped these homology results onto gene expression (Su et al. 2004), protein–protein interaction (Breitkreutz et al. 2008) (supplementary figs. S1 and S2, Supplementary Material online), comparative genomics, functional (Ashburner et al. 2000), and phenotypic (Bult et al. 2008; Amberger et al. 2009) data. Human genes with prokaryotic homologs are significantly more highly (Mann–Whitney *U* test, $P = 2.51 \times 10^{-10}$) and broadly ($P = 1.38 \times 10^{-9}$) expressed, are twice as likely to be involved in human disease (OR = 2.01, 95% CI = 1.86–2.18), are more likely to have mouse orthologs that are lethal upon inactivation (OR = 1.14, 95% CI = 1.12–1.25), have a higher number of paralogs ($P = 1.21 \times 10^{-91}$), are more frequently involved in informational processes (OR = 2.15, 95% CI = 1.81–2.55), are more selectively constrained (as evidenced from the lower ω and d_N values obtained from the human–mouse comparison; $P = 3.55 \times 10^{-34}$), and encode longer proteins ($P < 10^{-92}$) than those without prokaryotic homologs (supplementary table S4, Supplementary Material online).

Among genes with prokaryotic homologs, archaebacterium-like genes have a significantly higher expression level across human tissues ($P = 0.047$), are more likely to have lethal mouse orthologs (OR = 1.37, 95% CI = 1.05–1.79), have a lower number of paralogs ($P = 7.83 \times 10^{-34}$), are more frequently involved in informational processes (OR = 6.45, 95% CI = 5.16–8.07), and encode proteins that occupy a more central position in the human PIN (degree: $P = 0.003$; betweenness: $P = 0.037$; closeness: $P = 3.42 \times 10^{-4}$) (table 1). These results mirror previous observations in the yeast genome (Rivera et al. 1998; Cotton and McInerney 2010), suggesting a generality of these patterns across a broad range of eukaryotes. Archaebacterial genes are more highly expressed than eubacterial genes in all 84

Table 1
Comparison of Human Archaebacterial and Eubacterial Genes

	Eubacterial				Archaebacterial				
	n	Median	Average	SD	*n*	Median	Average	SD	*P* Value[a]
Expression level	6,735	15.70	89.68	439.13	776	17.29	203.62	919.07	0.047*
Expression breadth	6,735	12.00	12.68	10.92	776	17.00	13.78	11.04	0.014*
d_N	6,612	0.06	0.09	0.19	764	0.05	0.08	0.09	3.17×10^{-4}***
d_N/d_S	6,612	0.10	0.13	0.11	764	0.09	0.12	0.11	0.006**
Degree	3,342	3.00	7.01	12.81	489	4.00	8.06	11.36	0.003**
Betweenness	3,342	2.07×10^{-5}	4.10×10^{-4}	2.21×10^{-3}	489	4.03×10^{-5}	3.74×10^{-4}	9.95×10^{-4}	0.037*
Closeness	3,342	0.22	0.21	0.05	489	0.23	0.22	0.04	3.42×10^{-4}***
Protein length	7,884	540.00	707.41	607.69	939	496.00	665.19	627.95	3.26×10^{-7}***
#Paralogs	7,884	3.00	4.31	5.77	939	1.00	2.86	4.92	7.83×10^{-34}***
	n		Percent		*n*		Percent		*P* Value
Lethal mouse orthologs[b]	2,588		44.3		247		52.2		<0.05*
Involved in human disease[b]	7,884		17.3		939		12.2		<0.05*
Informational[b]	6,515		3.4		795		18.6		<0.05*
Mitochondrial[b]	6,798		11.5		809		6.4		<0.05*

Note.—In total, 7,884 eubacterial and 939 archaebacterial genes were compared; *n* refers to the number of genes used in each particular comparison. SD, standard deviation.
[a] The Mann–Whitney test was used to compare both groups except for categorical variables, for which ORs were used.
[b] Categorical variables. Tests were considered significant if the 95% CI for the ORs did not overlap 1.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

tissues represented in Su et al. (2004), with a statistically significant difference in 82 tissues (supplementary table S5, Supplementary Material online). Furthermore, archaebacterial genes are more selectively constrained ($P = 0.006$ for ω; $P = 3.17 \times 10^{-4}$ for d_N), encode shorter proteins ($P = 3.26 \times 10^{-7}$), and, surprisingly, are less likely to be involved in human disease (OR = 0.67, 95% CI = 0.55–0.82) than eubacterial ones (table 1).

Genes involved in informational processes are more highly ($P = 2.73 \times 10^{-9}$) and broadly ($P = 8.33 \times 10^{-15}$) expressed, are less likely to be involved in human disease (OR = 0.61, 95% CI = 0.47–0.79) and more likely to have lethal mouse orthologs (OR = 2.50, 95% CI = 1.75–3.57), have less paralogs ($P = 7.40 \times 10^{-88}$), and encode shorter proteins ($P = 0.001$) that are more central in the PIN (degree: $P = 7.75 \times 10^{-6}$; betweenness: $P = 0.001$; closeness: $P = 3.74 \times 10^{-8}$) than those involved in operational processes (table 2). This, combined with the fact that archaebacterial genes are more likely to have informational functions (see above and table 1), could, at least partially, account for the observed differences between archaebacterial and eubacterial genes. To discard this possibility, we obtained two subsets of our data set, one containing only informational genes ($n = 610$) and the other only operational ones ($n = 14{,}902$) and evaluated the differences between archaebacterial and eubacterial genes within each subset separately (supplementary table S6, Supplementary Material online). Except for expression breadth, degree, betweenness, and the likelihood of having a lethal mouse ortholog, there is a significant difference between archaebacterial and eubacterial genes for all studied parameters within at least one of the two subsets, indicating that the differences observed between archaebacterial and eubacterial genes are independent of function. Despite the lack of significance for the previously mentioned parameters, the group with the highest value is generally the same as in the full data set, suggesting that the lack of significance for these parameters may be the result of the reduction in sample size introduced by partitioning the data set.

Most of the variables considered in the present analysis correlate to each other (for a review, see Koonin and Wolf 2006), raising the possibility that some of the differences observed between archaebacterial and eubacterial genes might in fact be a by-product of these correlations. For instance, the ω values are known to correlate with expression level and breadth (Duret and Mouchiroud 2000; Subramanian and Kumar 2004), protein length (Subramanian and Kumar 2004), and number of paralogs (Lynch and Conery 2000). Therefore, the lower ω values observed in archaebacterial genes could be the result of these factors differing between archaebacterial and eubacterial genes (see above and table 1). In order to rule out this possibility, we used partial correlation analysis to evaluate the association between ancestry and ω while controlling for these variables, with significant results ($P = 0.0022$). Therefore, the effect of gene ancestry on ω is independent of these factors. Similarly, centrality in the PIN correlates with number of paralogs (Liang and Li 2007), protein length (Lemos et al. 2005), and expression level (Bloom and Adami 2003; Lemos et al. 2005). However, the

Table 2
Comparison of Human Informational and Operational Genes

	Informational				Operational				
	n	Median	Average	SD	*n*	Median	Average	SD	*P* Value[a]
Expression level	486	24.11	431.94	1,352.64	12,605	15.89	95.48	400.00	2.73×10^{-9}***
Expression breadth	486	25.00	16.56	10.44	12,605	11.00	12.68	10.92	8.33×10^{-15}***
d_N	486	0.06	0.08	0.08	12,211	0.06	0.09	0.13	0.098
d_N/d_S	486	0.10	0.13	0.11	12,211	0.10	0.13	0.12	0.498
Degree	380	5.00	9.64	15.39	7,052	3.00	7.45	12.62	7.75×10^{-6}***
Betweenness	380	5.02×10^{-5}	5.17×10^{-4}	2.30×10^{-3}	7,052	2.66×10^{-5}	3.98×10^{-4}	1.83×10^{-3}	0.001**
Closeness	380	0.23	0.23	0.03	7,052	0.22	0.22	0.05	3.74×10^{-8}***
Protein length	610	406.00	576.08	630.36	14,902	449.00	598.43	615.94	0.001**
#Paralogs	610	0.00	1.14	2.24	14,902	3.00	3.73	4.60	7.40×10^{-88}***

	n	Percent	*n*	Percent	*P* Value[a]
Lethal mouse orthologs[b]	139	66.2	5,580	44.1	<0.05*
Involved in human disease[b]	610	10.8	14,902	16.7	<0.05*
Mitochondrial[b]	592	21.1	14,002	7.5	<0.05*

Note.—In total, 610 informational and 14,902 operational genes were compared; *n* refers to the number of genes used in each particular comparison. SD, standard deviation.
[a] The Mann–Whitney test was used to compare both groups except for categorical variables, for which ORs were used.
[b] Categorical variables. Tests were considered significant if the 95% CI for the ORs did not overlap 1.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

association between ancestry and centrality measures remains significant even when controlling for these factors (degree: $P = 0.0002$; betweenness: $P = 0.0054$; closeness: $P = 0.0002$).

We also wished to examine whether the human interactome is stratified in ways that correlate with ancestry. We therefore carried out an analysis of adjacent nodes in the network. The human PIN contains 489 archaebacterial proteins (5.84%), 3,342 eubacterial ones (39.93%), 4,445 encoded by genes without prokaryotic homologs (i.e., eukaryotic-specific proteins, ESPs) (53.11%), and 94 classified as ambiguous (1.12%) (supplementary table S7 and fig. S2, Supplementary Material online). An average archaebacterial protein interacts with a total of 8.06 proteins: 1.17 archaebacterial (14.54%), 2.79 eubacterial (34.64%), 3.91 ESPs (48.55%), and 0.18 ambiguous (2.26%). Comparison of these numbers with the composition of the entire network suggests an excess of interactions involving two archaebacterial proteins. In addition, eubacterial proteins interact on average with 7.01 proteins: 3.07 eubacterial (43.78%), 0.41 archaebacterial (5.83%), 3.45 ESPs (49.27%), and 0.08 proteins considered ambiguous (1.12%), pointing again to an excess of interactions between proteins encoded by genes with the same ancestry.

Out of the 30,528 interactions contained in the human PIN (PIN1; supplementary figs. S1 and S2, Supplementary Material online), 15,036 connect two proteins with the same ancestry (308 involve two proteins encoded by archaebacterium-like genes, 5,323 link eubacterium-like ones, 9,389 connect ESPs, and 16 connect proteins deemed ambiguous; supplementary table S7, Supplementary Material online). Comparison of these numbers with those obtained from a set of 10,000 randomized networks (see Methods) shows that they are significantly higher than expected at random ($P < 2 \times 10^{-4}$ for the archaebacterial–archaebacterial, eubacterial—eubacterial, and ESP–ESP interactions; supplementary table S7, Supplementary Material online; fig. 1). Therefore, the human PIN is enriched in interactions between proteins encoded by genes inherited from the same endosymbiotic partner. Conversely, the number of interactions involving proteins encoded by genes with different ancestries is significantly lower than expected at random ($P < 2 \times 10^{-4}$ for the archaebacterial–eubacterial, archaebacterial–ESP, and eubacterial–ESP classes; supplementary table S7, Supplementary Material online; fig. 1). These results mirror previous observations in yeast that proteins with similar phylogenetic profiles tend to interact with each other (Qin et al. 2003).

It has been reported that PINs are strongly enriched in self-interactions (i.e., interactions between proteins encoded by the same gene) (Ispolatov et al. 2005), as well as in interactions involving proteins encoded by paralogous genes (Ispolatov et al. 2005; Pereira-Leal et al. 2007). Our analyses show that these kinds of interactions are indeed overrepresented in our data set (fig. 2). Given the possibility that these patterns could be inflating the numbers of interactions connecting proteins encoded by genes within the same ancestry group, we carried out a sequential trimming of the data set to rule out these potential biases. We filtered the PIN1 network by removing self-interactions (giving rise

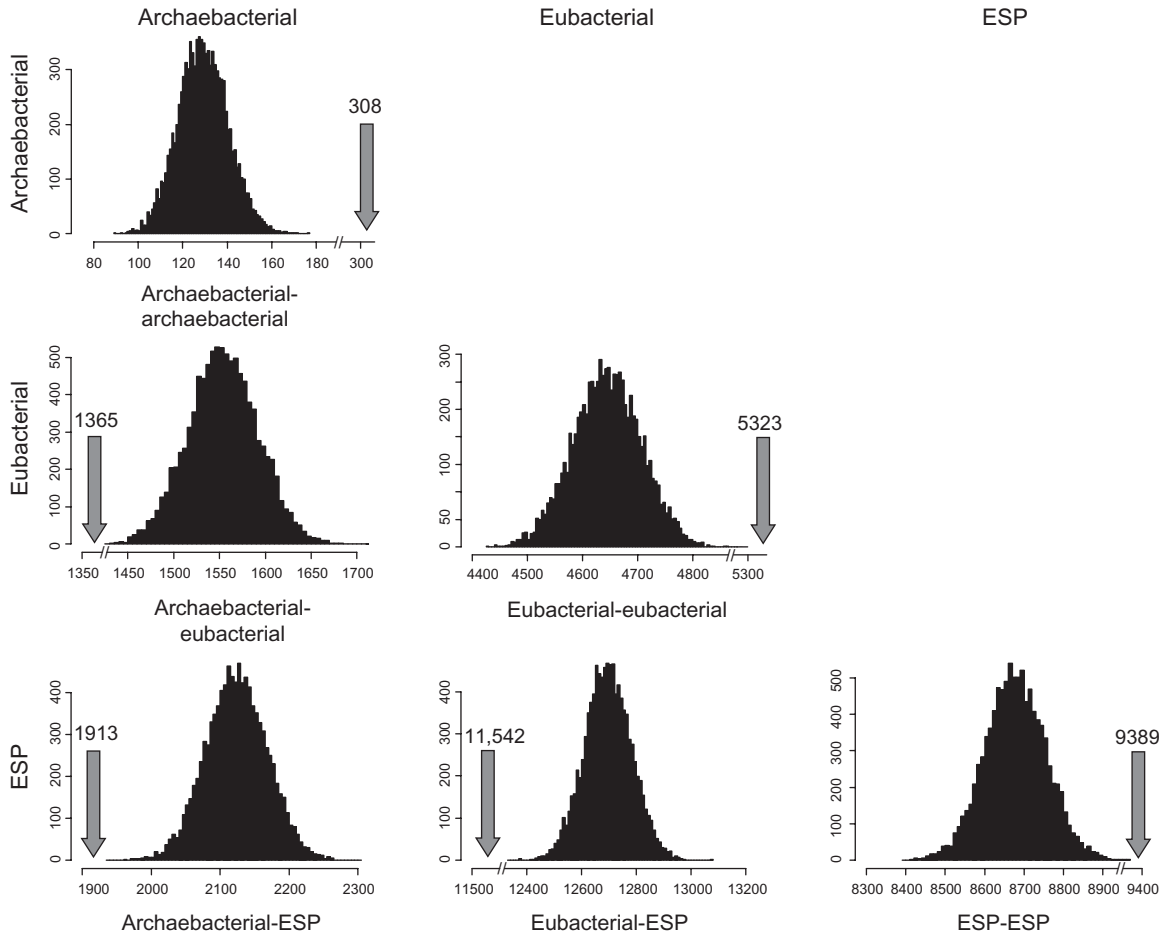

Fig. 1.—Comparison of the numbers of interactions in each ancestry category with those expected from a random network. Arrows represent the observed statistics in the PIN1 network. Histograms represent the empirical distribution of each statistic obtained from 10,000 randomizations of the network (see Methods).

to the PIN2 network, consisting of 8,280 proteins connected by 29,611 interactions; supplementary fig. S1, Supplementary Material online) and both self-interactions and interactions among proteins encoded by paralogous genes (PIN3, 8,189 proteins and 29,059 interactions; supplementary fig. S1, Supplementary Material online). Analysis of these subnetworks shows that they are also enriched in interactions among proteins encoded by genes with the same origin (supplementary table S7, Supplementary Material online), indicating that the observed trend is not a by-product of the aforementioned network features.

Eubacterial proteins are more likely than archaebacterial to be targeted to the mitochondrion (OR = 1.89, CI = 1.41–2.50; table 1). This, together with the fact that mitochondrion-targeted proteins tend to interact to each other (supplementary table S8, Supplementary Material online; fig. 2), could potentially account for the observed trend. To discard this possibility, we generated two subnetworks of PIN3, one containing only mitochondrion-targeted proteins (PIN3mit, 246 proteins and 255 interactions) and the other containing only proteins not targeted to this subcellular compartment (PIN3nonmit, 6,695 proteins and 22,368 interactions; supplementary fig. S1, Supplementary Material online). The number of interactions between proteins within the same ancestry category is significantly higher than expected at random in both subnetworks (supplementary table S7, Supplementary Material online), indicating that the observed trend is not a by-product of eubacterial proteins being preferentially targeted to the mitochondrion.

Our analyses show that proteins also tend to interact with those within the same functional category (i.e., informational interact with informational and operational with operational) in the human PIN (supplementary table S9, Supplementary Material online; fig. 2), in agreement with previous observations that genes tend to interact with those involved in the same biological processes (von Mering et al. 2002; Rual et al. 2005). Because human genes of

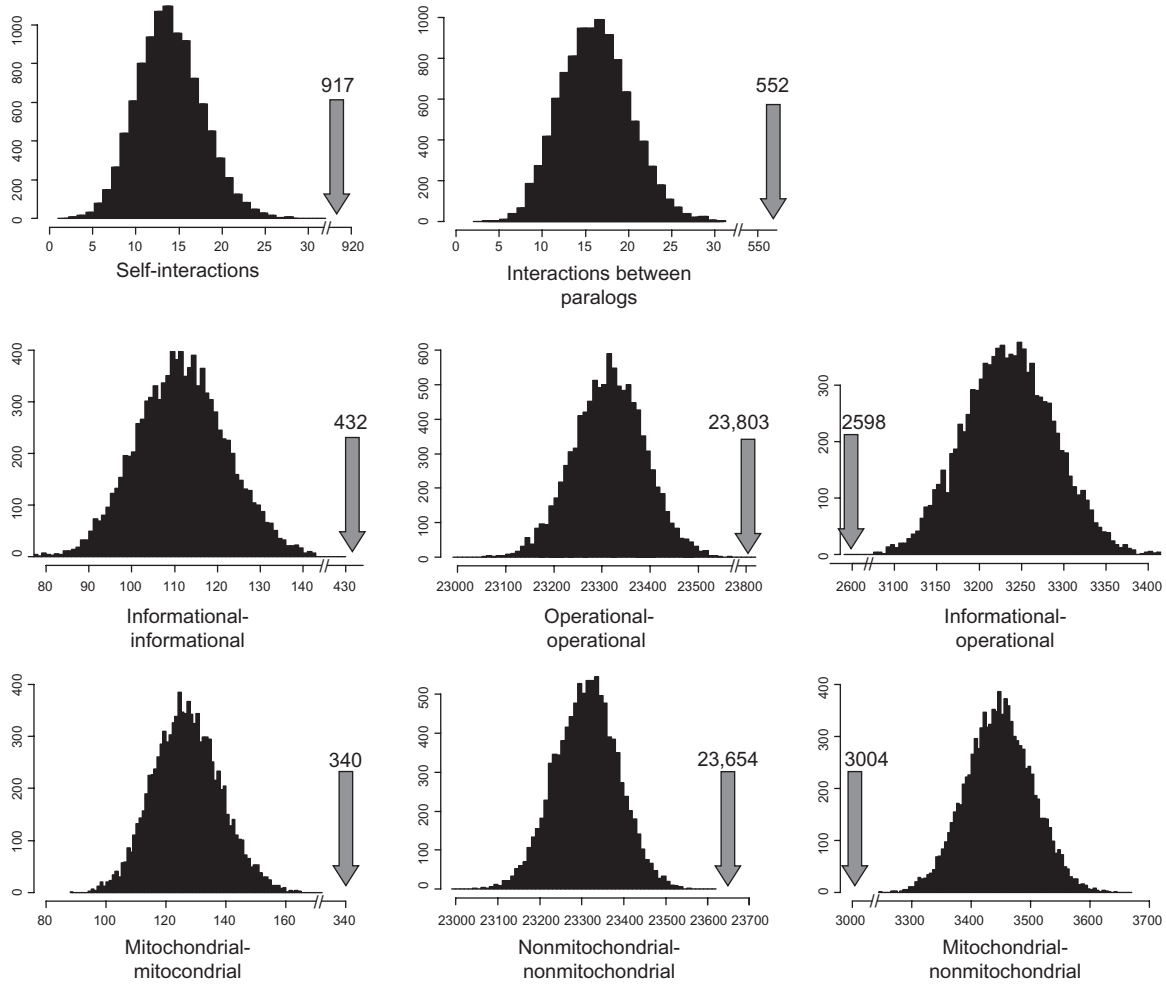

Fig. 2.—Comparison of observed statistics with those expected from a random network. Arrows represent the observed statistics in the PIN1 network. Histograms represent the empirical distribution of each statistic obtained from 10,000 randomizations of the network (see Methods).

archaebacterial and eubacterial origin tend to have informational and operational functions, respectively (see above and table 1), the observed clustering in the network of proteins within the same functional category might contribute to the observed tendency of proteins to interact with those encoded by genes of the same ancestry. We therefore generated two subsets of PIN3 containing either informational (PIN3inf, with 179 proteins connected by 387 interactions) or operational genes only (PIN3op, 6,683 proteins and 22,477 interactions; supplementary fig. S1, Supplementary Material online). The tendency of proteins to interact with those with the same ancestry, rather than proteins of different ancestry, is significant in the PIN3op subnetwork (supplementary table S7, Supplementary Material online), indicating that it is not a by-product of proteins preferentially interacting with those with the same function. The lack of significance for the PIN3inf subnetwork may be caused by a reduction in statistical power due to its smaller size.

Finally, it has been reported that proteins tend to interact with those encoded by genes of the same age (Qin et al. 2003; Rual et al. 2005). This might account for the observed enrichment of the human PIN in archaebacterial–archaebacterial, eubacterial–eubacterial, and ESP–ESP interactions (supplementary table S7, Supplementary Material online; fig. 1). As archaebacterial and eubacterial genes can be considered to have the same age, in the sense that they have been in the eukaryotic cell for the same length of time, the preferential interaction of proteins with those of the same age would also involve an enrichment in archaebacterial–eubacterial interactions. However, our analyses show that this class of interactions is not only not overrepresented but in fact significantly underrepresented in the human PIN (supplementary table S7, Supplementary Material online). For confirmation, we generated a subnetwork of PIN3 containing only proteins encoded by genes with prokaryotic homologs (PIN3prok, 3,031 proteins and 6,666 interactions; supplementary

fig. S1, Supplementary Material online). This network has also more intradomain and less interdomain interactions than expected at random (supplementary table S7, Supplementary Material online).

Discussion

Taken together, results presented here draw a picture of the human cell in particular and of the eukaryotic cell in general as a chimera with genes inherited from the archaebacterial and eubacterial ancestors that remain distinguishable even after more than 2 billion years of evolution (Brocks et al. 1999). In this chimera, archaebacterium-derived genes, despite being fewer in number, occupy more important positions, being more likely to be lethal upon inactivation and more highly and broadly expressed and encoding proteins that occupy more central positions in the PIN. Furthermore, natural selection preserves the amino acid sequences of archaebacterial genes more strongly, as inferred from the lower ω values observed within this category (table 1). This greater selective constraint acting on archaebacterial genes points to a greater functional importance of this gene repertoire. Less intuitive is our observation that archaebacterial genes are less likely to be involved in human disease. A possible explanation for this observation would be that, because mutations in these genes are likely to produce the death of affected individuals at an early stage, these genes may be less likely to be detected as involved in disease. However, the proportion of disease-involved genes is higher for human genes with lethal mouse orthologs than for genes without lethal orthologs (OR = 1.59, 95% CI = 1.42–1.78), making this explanation unlikely.

A total of 12,685 human genes have no detectable prokaryotic homologs. A number of hypotheses have been proposed for the existence of such genes in eukaryotic genomes (for a review, see Esser et al. 2004). First, they might be genes of archaebacterial or eubacterial origin that, owing to a faster evolutionary rate, have lost any detectable similarity with their prokaryotic relatives. Second, they could have been contributed by a third, noneubacterial, and non-archaebacterial prokaryote without living descendants. Third, they might be eukaryotic innovations. We observed that human genes without prokaryotic homologs are less selectively constrained (and hence evolve faster) than those with archaebacterial or eubacterial homologs (table 1; supplementary table S4, Supplementary Material online). This is consistent with the first hypothesis, although our results do not allow us to rule out the competing possibilities.

Our observations that human proteins tend to interact with those encoded by genes inherited from the same ancestor (fig. 1; supplementary table S7, Supplementary Material online) can be interpreted in terms of endosymbiotic theory. Each of the endosymbiotic partners had its own PIN, which merged during eukaryogenesis. It is likely that, immediately after this fusion event, proteins encoded by genes contributed by both endosymbiotic partners acted as relatively isolated communities, with few interactions involving proteins of both ancestries. The intervening 2 billion years of rewiring have clearly merged both networks, but not in a seamless way, as we can still observe the relics of the two ancestral networks in the current human PIN.

As an alternative to endosymbiotic theory, it has been proposed that eukaryotes appeared first and that archaebacteria and eubacteria arose secondarily from eukaryotic life forms through parallel genome reduction (Doolittle 1980; Forterre and Philippe 1999; Kurland et al. 2006). This model is difficult to reconcile with the differences observed between eukaryotic genes with archaebacterial and eubacterial homologs (Rivera et al. 1998; Cotton and McInerney 2010; current work). Under the eukaryotes-early model, the archaebacterial lineage would have somehow preferentially retained those genes that were informational, more essential, more highly and broadly expressed, more selectively constrained, and encoded more central proteins in the eukaryotic ancestor, whereas eubacteria would have preferentially retained operational, dispensable, lowly and narrowly expressed, less constrained, and peripheral eukaryotic genes. This seems to be a very unparsimonious scenario, which would argue against the eukaryotes-early hypothesis.

Taken together, findings reported here, combined with previous observations in yeast that also point to different functions, expression levels, and network positions for genes of archaebacterial and eubacterial origin (Rivera et al. 1998; Cotton and McInerney 2010), suggest that eukaryotic cells are composed of a tight community of at least two gene repertoires.

Supplementary Material

Supplementary results, figures S1 and S2, and tables S1–S9 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We are grateful to two anonymous referees for helpful comments on the manuscript. This work was funded by a Science Foundation Ireland Research Frontiers Programme grant (09/RFP/EOB25I0) to J.O.M.

Literature Cited

Alvarez-Ponce D, Aguadé M, Rozas J. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. Genome Biol Evol. 3:87–101.

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 37:D793–D796.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Biegert A, Soding J. 2009. Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A. 106:3770–3775.
Bloom JD, Adami C. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol. 3:21.
Breitkreutz BJ, et al. 2008. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. 36:D637–D640.
Brocks JJ, Logan GA, Buick R, Summons RE. 1999. Archean molecular fossils and the early rise of eukaryotes. Science. 285:1033–1036.
Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA. 2008. The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res. 36:D724–D728.
Cohen J, Cohen P, West SG, Aiken LS. 2003. Applied multiple regression/correlation analysis for the behavioral sciences. Mahwah (NJ): Lawrence Erlbaum Associates.
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.
Doolittle WF. 1980. Revolutionary concepts in evolutionary cell biology. Trends Biochem Sci. 5:146–149.
Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.
Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature. 440:623–630.
Esser C et al. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 21:1643–1660.
Forterre P, Philippe H. 1999. Where is the root of the universal tree of life? Bioessays. 21:871–879.
Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. 2010. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? Nat Rev Microbiol. 8:743–752.
Hardy MA. 1993. Regression with dummy variables. London: Sage Publications.
Horiike T, Hamada K, Kanaya S, Shinozawa T. 2001. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. Nat Cell Biol. 3:210–214.
Hubbard TJ et al. 2009. Ensembl 2009. Nucleic Acids Res. 37: D690–D697.
Ispolatov I, Yuryev A, Mazo I, Maslov S. 2005. Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res. 33:3629–3635.
Kasprzyk A et al. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res. 14:160–169.
Koonin EV, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 17:481–487.
Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. Science. 312:1011–1014.
Lane N, Martin W. 2010. The energetics of genome complexity. Nature. 467:929–934.
Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345–1354.
Liang H, Li WH. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet. 23:375–378.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.
Martin W, Hoffmeister M, Rotte C, Henze K. 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.
Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. 2007. Evolution of protein complexes by duplication of homomeric interactions. Genome Biol. 8:R51.
Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24:1752–1760.
Qin H, Lu HH, Wu WB, Li WH. 2003. Evolution of the yeast protein interaction network. Proc Natl Acad Sci U S A. 100:12820–12824.
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.
Rual JF, et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 437:1173–1178.
Sagan L. 1967. On the origin of mitosing cells. J Theor Biol. 14:255–274.
Su AI, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101:6062–6067.
Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 168:373–381.
Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123–135.
von Mering C, et al. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 417:399–403.
Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 87:4576–4579.
Zillig W, Schnabel R, Stetter KO. 1985. Archaebacteria and the origin of the eukaryotic cytoplasm. Curr Top Microbiol Immunol. 114:1–18.

Associate editor: Martin Embley