Comparative Genomics Evidence for a Chimeric Origin of the Eukaryotic Genome

Sujun Hua

Bioinformatics Track @ Yale Univeristy

Abstract We have made a comparison of the entire set of open reading frames in a eukaryotic genome *Saccharomyces cerevisiae* with their orthologs from thirteen bacterial genomes and seven archaeal genomes. The results shows that genes which are homologous between Archaea and Eukarya tend to regulate the genetic machinery of the cell, whereas genes that are homologous between Bacteria and Eukarya are more likely to regulate metabolic processes. The analysis provided comparative genomics evidence for a chimeric origin of the eukaryotic genome and suggested in a eukaryotic cell, the nucleus is of archaeal origin, while the cytoplasm is of bacterial origin.

Key words: genome evolution, eukaryotes origin, chimeric origin, clusters of orthologous groups, symbiosis hypotheses.

1. Introduction

The perspectives on the origin of eukaryotes are changing. Eukaryotic evolution has long been viewed through the perspective of a single molecule, rRNA. The rRNA phylogenies were used to classify all extant species into three primary domains, Archaea, Bacteria and Eukarya, with Archaea and Eukarya as sister taxa[1]. Eukaryotic cells are proposed to have evolved from an Archaea ancestor. However, recent phylogenetic analyses based on glutamate dehydrogenase[2], heat shock protein HSP70[3], glutamine synthetase[4], and others have yielded conflicting gene genealogies. These observations led to a chimeric hypothesis for the origin of eukaryotes with respect to Archaea and Bacteria[5]. For example, analysis of twenty-four protein coding genes supported the chimeric nature of eukaryotic genomes, with nine genealogies uniting Archaea and Eukarya, seven supporting the sister status of Eukarya and Bacteria, and eight genealogies being unresolved[6].

The completed sequencing of various genomes now provides a unique opportunity to analyze evolution at a higher, comprehensive level using complete genomes. Comparison of the entire set of open reading frames (ORFs) in a eukaryotic genome from *Saccharomyces cerevisiae* with their orthologs from thirteen bacterial genomes and seven archaeal genomes showed that the eukaryotic genome has a chimeric structure derived from Archaea and Bacteria.

2. Materials and methods

ORFs from these genomes and their functional annotations were retrieved from the database of Clusters of Orthologous Groups of proteins (COGs)[7, 8], which represents a phylogenetic classification of the proteins or domains encoded in completely sequenced genomes and is available at http://www.ncbi.nlm.nih.gov/COG. Each COG in the database consists of individual orthologous genes or orthologous sets of paralogs from at least three phylogenetic lineages. Any two proteins or domains from different lineages that belong to the same COG are orthologs. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events. COGs are identified using all-against-all sequence comparison of the proteins encoded in complete genomes. This comparison identifies the protein from each of the other genomes which is most similar to the target protein in a given genome

using the gapped BLAST program[9]. Each of these proteins is then considered in turn. If the comparisons reveal a reciprocal best-hit relationship between these proteins, then those that are reciprocal best-hits are identified as a COG.

ORFs from twenty-one completely sequenced genomes in the COG database were used in the analysis. These included a eukaryotic genome (*Saccharomyces cerevisiae*), thirteen bacterial genomes (*Aquifex aeolicus, Thermotoga maritime, Deinococcus radiodurans, Mycobacterium tuberculosis, Bacillus subtilis, Synechocystis sp. PCC 6803, Escherichia coli K12, Pseudomonas aeruginosa, Vibrio cholerae, Haemophilus influenzae Rd, Xylella fastidiosa, Neisseria meningitidis MC58* and *Campylobacter jejuni*) and seven archaeal genomes (*Archaeoglobus fulgidus, Halobacterium sp. NRC-1, Methanococcus jannaschii, Methanothermobacter thermautotrophicus, Thermoplasma acidophilum, Pyrococcus horikoshii* and *Aeropyrum pernix*). Parasitic bacteria, e.g., *Rickettsia prowazekii*, *Chlamydia trachomatis*, etc., were excluded from the analysis because of their biased gene composition due to massive gene loss.

3. Results

COGs in the database were assigned to a hierarchical functional category. We focused on three main functional categories including metabolism, cellular processes, and information storage and processing. For each functional subcategory in the three main categories (the details can be seen in the supplementary material), we counted the number of COGs coexisting in yeast *S. cerevisiae* and each of the other genomes (seven Archaea and thirteen Bacteria). For example, the number of COGs present in both yeast and each species of Archaea and Bacteria that were related to the functional subcategories of transcription (K) and carbohydrate transport and metabolism (G) are shown in Fig. 1(A) and Fig. 1(B), respectively. The letters (K and G, etc.) are the function codes in the COG database. The statistical results for other functional subcategories are listed in the supplementary material which is available at http://bioinfo.mbb.yale.edu/~sujun/COG/. Fisher tests (F-tests) at the 5% significance level were used to determine whether the number of COGs coexisting in yeast and Archaea belonging to a special functional subcategory was significantly different from that co-occurring in yeast and Bacteria. The results of the F-tests for each yeast ORF group classified by category or subcategory showed that the yeast ORF groups related to information storage and processing which included subcategories of transcription (K), translation, ribosomal structure and biogenesis (J) had more homology with archaeal ORFs than with bacterial ORFs. Furthermore, the yeast ORF groups related to the functions of metabolism and cellular processes including amino acid transport and metabolism (E), nucleotide transport and metabolism (F), carbohydrate transport and metabolism (G), coenzyme transport and metabolism (H), lipid metabolism (I), cell division and chromosome partitioning (D), cell envelope and outer membrane biogenesis (M), cell motility and secretion (N), posttranslational modification, protein turnover and chaperones (D), and inorganic ion transport and metabolism (P) had more homology with bacterial ORFs than with archaeal ORFs.

Apparently, most ORFs with functions related to information storage and processing are nucleus-related while most of those with functions related to metabolism and cellular process are cytoplasm related. Our genomic analyses strongly support the chimeric nature of the eukaryotic genome which states that the eukaryotic nuclear genome, instead of having descended directly from a common ancestor shared with Archaea, is an evolutionary chimera that incorporates substantial contributions from both Archaea and Bacteria progenitors. The results are compatible with previous phylogenetic tree analyses[6, 10]. Recently, Rivera et al.[11] gave evidence for two distinct functional gene superclasses of eukaryotes, i.e., the informational genes which function in genetic

information processing (transcription, translation, etc.) and the operational genes which function in cell operation and metabolism (amino acid synthesis, intermediary metabolism, etc.), by comparing all the genes from a eukaryote, a cyanobacterium, a proteobacterium and a methanogen using a simple distance metric. Our analysis of more complete genomes strengthened the evidence of Rivera and his collogues. More recently, similar results have been obtained using the homology-hit analysis with multiple similarity thresholds using the yeast ORF group classified by functional categories[12]. Our analysis of more genomic sequences has clarified the conclusions that genes which are homologous between Archaea and Eukarya tend to regulate the genetic machinery of the cell, whereas genes that are homologous between Bacteria and Eukarya are more likely to regulate metabolic processes.

4. Discussion

The chimeric nature of eukaryotic nuclear genomes can be partially explained by the endosymbiont hypothesis of mitochondria (or chloroplast in plants)[13, 14] which proposes that a large proportion of organellar genes were either lost or transferred to the nuclear genome during symbiotic evolution. In fact, some genes now found in the nucleus encode proteins that are transported back into the mitochondria. An experimental system to study and quantify the transfer of genetic sequences between mitochondria and nuclear genomes in the yeast was developed by Thorsness and Fox[15]. However, our results clearly show that the bacterial component of the nuclear genome includes massive genes unrelated to mitochondrial biogenesis and function which appear to be considerably greater than that usually attributed to specific gene transfer from the evolving mitochondrial genome. The results

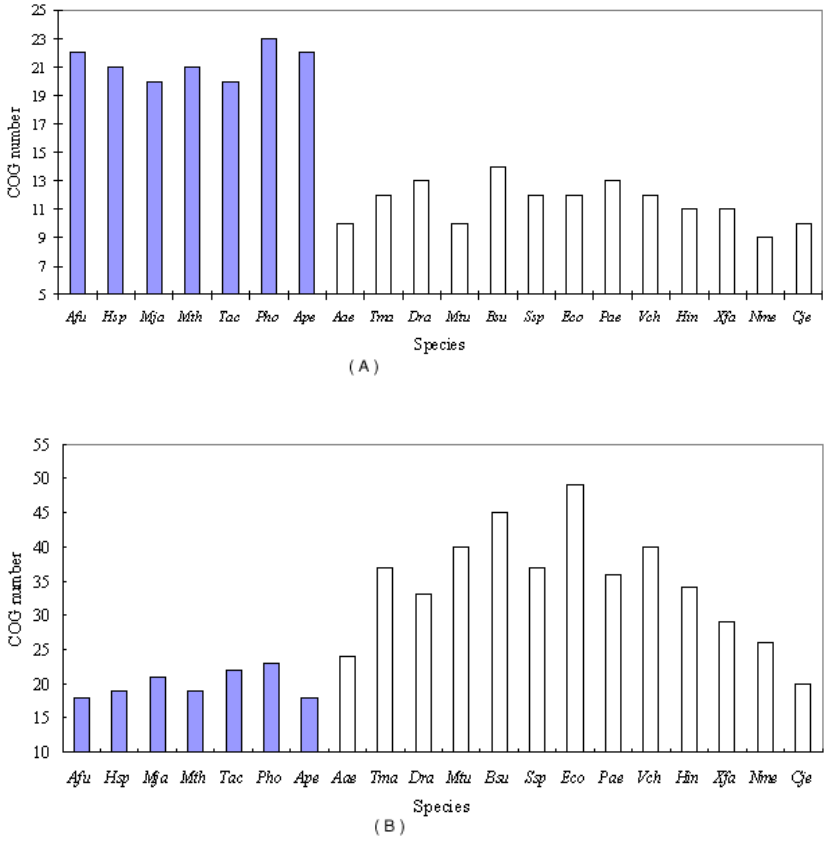

Fig. 1. Number of COGs coexisting in yeast *S. cerevisiae* and each of the other genomes (seven Archaea and thirteen Bacteria) related to various biological functions. The filled bars indicate Archaeal species; while the open bars are for bacterial species. The statistical results related to the functional subcategory of transcription are shown in (A) with the results related to carbohydrate transport and metabolism shown in (B). The results indicate that most genes in the eukaryotic nuclear genome related to genetic information storage and processing are of Archaeal origin while most genes related to metabolism and cellular processes are of bacterial origin. The analysis includes seven Archaeal species (Afu, *Archaeoglobus fulgidus*; Hsp, *Halobacterium sp. NRC-1*; Mja, *Methanococcus jannaschii*; Mth, *Methanothermobacter thermautotrophicus*; Tac, *Thermoplasma acidophilum*; Pho, *Pyrococcus horikoshii* and Ape, *Aeropyrum pernix*) and thirteen bacterial species (Aae, *Aquifex aeolicus*; Tma, *Thermotoga maritime*; Dra, *Deinococcus radiodurans*; Mtu, *Mycobacterium tuberculosis*; Bsu, *Bacillus subtilis*; Ssp, *Synechocystis sp. PCC 6803*; Eco, *Escherichia coli K12*; Pae, *Pseudomonas aeruginosa*; Vch, *Vibrio cholerae*; Hin, *Haemophilus influenzae Rd*; Xfa, *Xylella fastidiosa*; Nme, *Neisseria meningitidis MC58* and Cje, *Campylobacter jejuni*).

indicate that massive gene transfer between genomes might occur during eukaryogenesis. Indeed, a growing body of evidence suggests that the extent of horizontal gene transfer (HGT) is far greater than previous recognized[16-19] and may be a principle force in the early evolution of genomes.

Two symbiosis hypotheses for the origin of eukaryotes have recently been proposed, the hydrogen hypothesis[20] and the syntrophy hypothesis[21]. Both suggest that eukaryotes arose through metabolic symbiosis between Bacteria and methanogenic Archaea thus giving a chimeric origin for eukaryotes. In this scenario, the eukaryotes emerged from a symbiotic event involving methanogenic Archaea and Bacteria (α -proteobacterium in the hydrogen hypothesis but δ -proteobacterium in the syntrophy hypothesis) in an anaerobic context. During eukaryogenesis, progressive cellular and genomic cointegration of both types of partners occurred with bacterial-to-archaeal preferential gene transfer and eventual replacement. The bacterial genome was greatly reduced and may even have disappeared. Emerging eukaryotes would have inherited most of the archaeal DNA-processing information systems while the cellular metabolism systems would have mainly come from the versatile bacterial organotrophy. The symbiosis hypothesis also effectively accounts for the chimeric nature of eukaryotic genomes as was also revealed by our analysis. In addition, the syntrophy hypothesis tried to explain the form of the cellular membrane system and hypothesized that the plasmic membrane was mostly derived from a bacterial membrane. This hypothesis is supported by our results concerning the bacterial origin of genes related to the cell envelope and outer membrane biogenesis (M).

In conclusion, our analysis gives comparative genomics evidence for a chimeric origin of the eukaryotic genome. The bacterial component of the eukaryotic nuclear genome beyond the mitochondria-related genes contribution would be well explained by the symbiosis hypothesis for the eukaryotes origin. The results allow the conclusion that in a eukaryotic cell, the nucleus is of archaeal origin, while the cytoplasm is of bacterial origin.

References

[1] Woese, C.R., K&ler, O. & Wheelis, M.L. (1990) Proc. Natl. Acad. Sci. USA 87, 4576-4579.

[2] Benachenhou-Lahfa, N., Forterre, P. & Labedan, B. (1994) J. Mol. Evol. 36, 335-346.

[3] Gupta, R. & Golding, G.B. (1993) J. Mol. Evol. 37, 573-582.

[4] Brown, J., Masuchi, Y., Robb, F.T. & Doolittle, W.F. (1994) J. Mol. Evol. 38, 566-576.

[5] Katz, L.A. (1998) Trends Ecol. Evol. 13, 493-497.

[6] Golding, G.B. & Gupta, R. (1995) Mol. Biol. Evol. 12, 1-6.

[7] Tatusov, R.L., Koonin, E.V. & Lipman, D.J. (1997) Science 278, 631-637.

[8] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. & Koonin, E.V. (2001) Nucleic Acids Res. 29, 22-28.

[9] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389-3402.

[10] Ribeiro, S. & Golding, G.B. (1998) Mol. Biol. Evol. 15, 779-788.

[11] Rivera, M.C., Jain, R., Moore, J.E. & Lake, J.A. (1998) Proc. Natl. Acad. Sci. USA 95, 6239-6244.

[12] Horiike, T., Hamada, K., Kanaya, S. & Shinozawa, T. (2001) Nat Cell Biol. 3, 210-214.

[13] Gray, M.W. (1992) Int. Rev. Cytol. 141, 233-357.

[14] Gray, M.W., Burger, G. & Lang, B.F. (1999) Science 283, 1476-1481.

[15] Thorsness, P.E. & Fox, T.D. (1990) Nature 346, 376-379.

[16] Doolittle, W.F. (1999) Science 284, 2124-2129.

[17] Eisen, J.A. (2000). Curr. Opinion Genet. Devel. 10, 606-611.

[18] Salzberg, S.L., White, O., Peterson, J. & Eisen, J.A. (2001) Science 292, 1903-1906.

[19] Stanhope, M.J., Lupas, A., Italia, M.J., Koretke, K.K., Volker, C. & Brown, J.R. (2001) Nature 411, 940-944.

[20] Martin, W. & Muller, M. (1998) Nature 392, 37-41.

[21] Moreira D. & Lopez-Garcia, P. (1998) J. Mol. Evol. 47, 517-530.