The origin of the eukaryotic cell: A genomic investigation

Hyman Hartman[†‡] and Alexei Fedorov[§]

[†]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [§]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

Communicated by Carl R. Woese, University of Illinois at Urbana–Champaign, Urbana, IL, December 10, 2001 (received for review October 25, 2001)

We have collected a set of 347 proteins that are found in eukaryotic cells but have no significant homology to proteins in Archaea and Bacteria. We call these proteins eukaryotic signature proteins (ESPs). The dominant hypothesis for the formation of the eukaryotic cell is that it is a fusion of an archaeon with a bacterium. If this hypothesis is accepted then the three cellular domains, Eukarya, Archaea, and Bacteria, would collapse into two cellular domains. We have used the existence of this set of ESPs to test this hypothesis. The evidence of the ESPs implicates a third cell (chronocyte) in the formation of the eukaryotic cell. The chronocyte had a cytoskeleton that enabled it to engulf prokaryotic cells and a complex internal membrane system where lipids and proteins were synthesized. It also had a complex internal signaling system involving calcium ions, calmodulin, inositol phosphates, ubiquitin, cyclin, and GTP-binding proteins. The nucleus was formed when a number of archaea and bacteria were engulfed by a chronocyte. This formation of the nucleus would restore the three cellular domains as the Chronocyte was not a cell that belonged to the Archaea or to the Bacteria.

Recently, Horiike *et al.* (1) proposed that the eukaryotic nucleus was derived from the symbiosis of Archaea in Bacteria. Their results were based on a search for homologies between proteins found in the yeast genome and those found in the genomes of Archaea and Bacteria. However, the use of homologies to determine relationships between the three main cellular domains is misleading because of the extensive horizontal transfer of genes between the Archaea and the Bacteria (2) as well as between Archaea, Bacteria, and the Eukarya. An alternative to searching for protein homologies in studying the evolution of cellular domains is to search for proteins unique to one domain with no significant homology to proteins in the other domains. This approach has been used by Woese's group to search for signature proteins in Archaea that are unique to the Archaea and that are absent from the Bacteria and Eukarya (3).

In this paper we set out to find the signature proteins that would delineate the Eukarya from the Archaea and Bacteria. We characterize the set of eukaryotic signature proteins (ESPs) as those proteins that have homologs in all main branches of eukaryotes (animals, plants, fungi, and protozoa), but do not have any homologs in the Archaea and Bacteria. Our ESP set was derived from the completely sequenced genomes of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Giardia lamblia*, and all the 44 available completed genomes of Archaea and Bacteria in GenBank. To obtain the ESP set, we started from the complete sample of 6,271 yeast proteins derived from the sequenced genome of *S. cerevisiae*. After the removal of proteins that do not have homologs in the genomes of *D. melanogaster*, *C. elegans*, and *A. thaliana*, we obtained a set of 2,136 proteins. Comparing this set with all proteins derived from the completed genomes of Bacteria and Archaea and removing those that do have homologs in Archaea and/or Bacteria narrowed the list of eukaryote-specific proteins to 914. However, because fungi, plants, nematodes, and insects are late branches on the eukaryotic tree, we compared these 914 proteins with one of the most deeply divergent eukaryotic cells—the protozoan *G. lamblia*. The resulting 347 proteins we call ESPs. We chose *G. lamblia* because both the small ribosomal RNA and a host of proteins have identified this cell as being an extremely early-diverging eukaryote (4). The loss of about 567 proteins as one brings *Giardia* into the search for ESPs is caused by the absence of a mitochondrion in *Giardia* and to the simplification of cellular structures in the parasitic protozoan *Giardia*. Thus the set of 347 ESPs we obtain is a minimal set of early-divergent proteins that can be used to test the hypotheses of eukaryotic origins.

The obtained set of 347 ESPs was divided into 180 nonredundant protein groups including cytoplasmic proteins involved with the cytoskeleton, endocytosis, and phagocytosis, and protein synthesis and degradation (91 proteins, Table 1); internal signaling proteins (108 proteins, Table 2); proteins in the nucleus such as histones, nuclear pore proteins, and spliceosomal proteins (47 proteins, Table 3); and enzymes and unknown proteins (101 proteins, Table 4). The ESP set was used to test the existing theories of the origin of the Eukarya, especially the origin of the nucleus.

A major problem in the formation of the eukaryotic cell is the origin and evolution of the nucleus. Mereschowsky proposed in 1910 that the nucleus was formed from bacteria that had found a home in an entity that was composed of ''amoebaplasm'' and was not a bacterium (5).

At present, there are two major competing theories for the endosymbiotic origin of the nucleus. The first theory claims that the eukaryotic cell is a fusion of an archaeon with a bacterium. One can symbolize this relationship as E = A + B or in words Eukarya = Archaea + Bacteria. We call this the AB hypothesis. There are several variants of this AB conjecture with different proposals for the host cell in which the nucleus became an endosymbiont (1, 6–8). Horiike *et al.* (1) claimed that the host cell was a bacterium.

A major difficulty with the AB conjecture is that the prokaryotic host cell must have been able to engulf its future symbiont. The engulfing of other cells requires a complex internal cytoskeleton, which interacts with the plasma membrane. This cellular configuration, in the absence of a cell wall, allows phagocytosis to take place. Prokaryotes, whether they are Archaea or Bacteria, do not have a complex internal cytoskeleton and, in general, they do have a cell wall, and therefore they are incapable of phagocytosis. This AB conjecture is complicated further by the fact that a whole set of new cellular structures (i.e., endoplasmic reticulum, spliceosome, etc.) other than the cytoskeleton had to be constructed from prokaryotes that lacked them.

These difficulties with the AB hypothesis led us to consider a second conjecture for the origin of the nucleus. This hypothesis assumes that the nucleus formed from the endosymbiosis of an archaeon and a bacterium in a third cell, which we will call C. One can symbolize this new conjecture as E = A + B + C. We have named this third cell a chronocyte (9). We call this second theory the ABC hypothesis. The simplest prediction of this theory is the

Abbreviations: ESP, eukaryotic signature protein; ESR, eukaryotic signature RNA; ER, endoplasmic reticulum.

[‡]To whom reprint requests should be addressed. E-mail: hhartman@mit.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. List of 91 ESPs associated with cytoplasm and membrane systems

Category	Subcategories (ID)
Cytoskeleton	
Tubulin	α-Tubulin (Tub1; Tub3)
	β-Tubulin (Tub2)
	γ-Tubulin-like protein (Tub4)
Tubulin-associated proteins	Kinesin-related protein (Kip2; Kar3)
	Kinesin-related protein involved in mitosis (Kip3)
	Kinesin heavy chain homolog (Smy1)
	Microtubule-binding protein (Bim1)
	Putative light chain of dynein (Dyn2)
Actin	(Act1)
Actin-related	(Arp1; Arp2; Arp3; Arp4; Arp5; Arp6; Arp7)
Actin-associated	Light chain for myosin (Mic1)
Protein synthesis and breakdown	
Small ribosomal proteins	Ribosomal protein S7 [rp30] (Rps7a; Rps7b)
	Ribosomal protein S21 (Rps21a; Rps21b)
	Ribosomal protein S24 (Rps24a; Rps24b)
	Ribosomal protein S26A (Rps26a; Rps26b)
	Ribosomal protein S27 (Rps27a; Rps27b)
	Ribosomal protein S31 [ubiquitin related] (Rps31)
Large ribosomal proteins	Ribosomal protein L13 (Rpl13b; Rpl13a)
	Ribosomal protein L14 (Rpl14a; Rpl14b)
	Ribosomal protein L18 (Rpl18a; Rpl18b)
	Ribosomal protein L20 (Rpl20b; Rpl20a)
	Ribosomal protein L21 (Rpl21b; Rpl21a)
	Ribosomal protein L24 (Rpl24a; Rpl24b)
	Ribosomal protein L29 (Rpl29)
	Ribosomal protein L33 (Rpl33a)
	Ribosomal protein L35 (Rpl35b; Rpl35a)
	Ribosomal protein L36 (Rpl36a; Rpl36b)
	Ribosomal protein L40 [ubiquitin related] (Rpl40a; Rpl40b)
Translation factors	Translation elongation factor EF-1β (Efb1)
	Translation elongation factor EF-1γ (Tef4; Cam1)
Proteasome-associated	Subunits of proteasome regulatory particle (Rpn1; Rpn8; Rpn10; Rpn11)
Signal peptidase	(Spc3)
Membrane	
Lipid attachments	Geranylgeranyltransferase type II β subunit (Bet2)
	Geranylgeranyltransferase type II α (Bet4)
	Geranylgeranyltransferase type I subunit (Cdc43)
	Farnesyltransferase β subunit (Ram1)
	CAAX farnesyltransferase α subunit (Ram2)
	Farnesyl cysteine-carboxyl methyltransferase (Ste14)
	N-myristoyltransferase (Nmt1)
ER and Golgi	Transport protein particle [TRAPP] component (Bet3)
	HDEL receptor (Erd2)
	Integral membrane proteins (Sac1; Fig4)
	Subunit of coatomer (Sec26)
	Vesicle coat component (Sec24)
Vacuole	Vacuolar protein (Pep8)
	Retromer complex component (Vps35)
	Vacuolar ATPase V_0 domain subunit c (Cup5)
	Vacuolar ATPase V_0 domain c″ (Ppa1; Tfp3)
Endocytosis	Clathrin (Chc1)
	Clathrin-associated proteins (Apm1; Apm2; Apm4; Aps1; Aps2; Aps3; Apl1; Apl2; Apl3; Apl4; Apl5)
	Dynamin (Dnm1; Mgm1; Vps1)

The unique identifier symbols for the proteins are from *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces) and are shown in parentheses. The 12 ESP proteins that have low sequence homology to prokaryotic proteins are underlined (maximal BLAST score from 50 to 55 bits).

existence of an ESP set of proteins that evolved from the Chronocyte. This theory would imply that there are three cellular domains despite the large infusion of prokaryotic proteins into the eukaryotic cell because of endosymbiosis. The obtained set of ESPs is thus consistent with the ABC hypothesis. We will try to reconstruct the chronocyte by using ESPs as a guide.

Materials and Methods

Protein Sequence Samples. Protein sequences of *D. melanogaster* (14,335 entries), *C. elegans* (17,123 entries), and *S. cerevisiae* (6,271 entries) were downloaded from GenBank (10) release 121 as "*.faa" files. Protein sequences of *A. thaliana* (25,470 entries) were downloaded from the Institute for Genomic Research *A. thaliana* database (www.tigr.org/tdb/e2k1/ath1). The *G. lamblia* protein database was generated on the basis of the *Giardia* single-pass nonassembled nucleotide sequence database (laboratory of M. L. Sogin, Woods Hole, MA). This nucleotide database was downloaded on February 20, 2001 from the Marine Biological Laboratory web site (www.mbl.edu/giardia) and contains 53,325 entries (4.7×10^7 nt) overlapping several times the whole *Giardia* genome. Each entry of the *Giardia* nucleotide database was translated into protein sequences in all six possible reading frames. In the end, the *Giardia* protein database was composed of 319,950 possible protein sequences. The database of prokaryotic proteins (72,998 entries) was obtained from GenBank (release 121) by pooling all available protein sequences from 44 completely sequenced genomes of Bacteria and Archaea.

Protein Comparisons. Protein alignments were obtained by using stand-alone BLAST 2.0 binaries downloaded from the National Center for Biotechnology Information (11). Gapped BLAST was used for comparison of all proteins of a species X with a database of all proteins from a species Y. Analyzing the obtained alignment scores, we divided all proteins of species X into two groups: XY-homologous and XY-unique. An alignment score threshold of 55 bits was used for this division. The 55-bit alignment score corresponds to the *E* value of 10^{-6} for the largest *Giardia* and bacterial protein databases used in our study. The 55-bit threshold is a very reliable threshold, which ensures it unlikely that we would get false-positive results for homologous proteins. Further, the set of XY-homologous proteins of species X was compared with a protein database of species Z and divided into the groups of XYZ-homologous and XYZ-unique proteins with the same alignment score threshold of 55 bits. We performed these steps for the consecutive comparison of proteins from five eukaryotic species and also compared them with all bacterial proteins to obtain the ESP sample—the signature set of eukaryotic proteins without a prokaryotic counterpart. We specifically started the generation of ESP sample from the set of *S. cerevisiae* proteins, because, among the flat FASTA-formatted protein databases, the *S. cerevisiae* is the best annotated one.

All programs were executed on a LINUX-running computer with a dual Pentium III processor. Programs for scanning the BLAST output and grouping the proteins were written in PERL. All protein samples described in this paper are available from our web page (www.mcb.harvard.edu/gilbert/ESP).

Results: Chronocyte Reconstruction

Cytoskeleton. We begin our examination of the ESP set with the cytoplasmic proteins (see Table 1) because it contains the proteins of the cytoskeleton and the proteins involved in endocytosis and phagocytosis.

The ESPs we find in the cytoskeleton are actin, seven actin-related proteins, light chain of myosin, tubulins, kinesins, and the light chain of dynein. Actin and tubulin have structural and very weak sequential similarity to FtsA and FtsZ proteins, respectively, in the Bacteria and the Archaea. Is this structural similarity of actin and FtsA or tubulin and FtsZ caused by these

proteins having diverged from a common ancestor or was this structural similarity caused by the convergence of these proteins from different ancestral proteins? There is at present no good methodology for distinguishing between these two alternatives. There are a number of ESP proteins, e.g., ubiquitin, for which there exists a structural similarity to prokaryotic proteins but no sequential homology. In this paper, we assume that in most cases where this situation arises, there was a common ancestral protein and that it existed in the progenote, a cellular domain that was the ancestor to both the chronocyte and the prokaryotic cells. Therefore, when one finds a protein in eukaryotic cells that is structurally similar but has little or no sequential homology to those found in prokaryotic cells, the best that one can surmise is that these proteins shared a common ancestor.

A recent case was investigated by R. F. Doolittle (12), who has considered the evolution of the eukaryotic cytoskeletal proteins actin and tubulin and their prokaryotic counterparts FtsA and FtsZ. He determined that the rate by which actin and tubulin varied in the eukaryotic cells was very slow (10% change per billion years). The calculated time when the bacterial FtsA and FtsZ proteins began to diverge from their possible eukaryotic actin and tubulin homologs, according to R. F. Doolittle, is greater than the age of the earth (4.5 billion years). On the other hand, the divergence times of other noncytoskeletal proteins such as metabolic proteins were about two billion years ago. This finding presents a paradox. The solution to this paradox was, according to R. F. Doolittle, "to have [an] RNA-based 'urkaryote' that was capable of making cytoskeletal proteins. . . The rate of sequence change in an RNA-based system would have been enormously greater than occurs in DNA-based systems" (12). This solution to the paradox implies that if the actin and FtsA (tubulin and FtsZ) did have a common ancestor, then it was not to be found in either in Archaea or Bacteria, but in some hypothetical RNA-based "urkaryote," or perhaps in the progenote of Woese. This solution implicates a third cell in the evolution of the eukaryotic cell and hence is a variant of the ABC hypothesis. This solution to the paradox of the relation of actin and tubulin to their prokaryotic counterparts will also apply to other proteins in our set of ESPs such as ubiquitin, histones, and GTP-binding proteins. Histones may be an exception, as the histone fold is found only in one branch of the Archaea and not in Bacteria.

The Plasma Membrane. One of the deepest distinctions between the prokaryotes and the eukaryotes is to be found in the membrane. The prokaryotes have perfected the use of a proton gradient across a membrane to synthesize ATP. The eukaryote has perfected the interaction of cytoskeletal proteins with its membrane, modulated by calcium ions. This distinction could be the major driving force for endosymbiosis. The eukaryote could engulf a prokaryotic cell and it would benefit from the ability of the prokaryotic cell to generate ATP from a proton gradient; this was the case in the formation of the chloroplast from a cyanobacterium and the formation of a mitochondrion from a purple bacterium, and may have been the case in the early stages of the formation of the nucleus as well.

The ability of the original host cell to engulf prokaryotic cells resides in an interaction between its cytoskeleton and the proteins associated with the plasma membrane that are involved in endocytosis and phagocytosis. The ESPs that are particularly involved in endocytosis include clathrin, clathrin-related proteins, and dynamin (Table 1).

The ability of this array of proteins to engulf a prokaryotic cell is coordinated by a signaling system that involves the calcium ion. The ESPs involved in controlling the concentration and the effects of the calcium ion are calmodulin, phosphatidylinositol kinases, and phosphatases (Table 2).

This finding would imply that the proteins involved in phagocytosis, endocytosis, and the calcium ion control system (modulated by calmodulin and phosphatidylinositol) were inherited by the eukaryotic cell from the chronocyte.

Endoplasmic Reticulum (ER) and Protein Synthesis. The use of calcium ion as a major internal signal in the eukaryotic cell involves not only the cytoskeleton and the plasma membrane but also the ER, a membrane system that lies between the nucleus and the plasma membrane. There are three domains in the ER: the outer nuclear membrane, the smooth ER, and the rough ER. Phospholipids, cholesterol, and steroids are synthesized in the smooth ER. The rough ER is so designated because it is a membrane decorated by ribosomes, where proteins are synthesized, folded, and packaged for transport to the Golgi apparatus. This relationship between the smooth and rough ER is indicative of an evolutionary relationship between lipid biosynthesis and protein biosynthesis. A large number of our ESPs are implicated in this system, especially the ribosomes (Table 1) and the GTP-binding proteins (Table 2).

There are 17 ribosomal proteins in our ESPs. These proteins are found in all sequenced eukaryotic cells and in particular in *Giardia*. There are 72 ribosomal proteins in *Giardia* (13), of which 55 ribosomal proteins have homology to either bacterial or archaeal ribosomal proteins. Because many of the remaining 17 ribosomal proteins are small, we lowered the criterion for homology to 50 bits, and then 6 ribosomal proteins were homologous to ribosomal proteins of the Archaea. There are 11 proteins that are not homologous to prokaryotic ribosomal proteins at this lower criterion for homology. The eukaryotic ribosomal proteins therefore have evolved from a mixture of prokaryotic ribosomal proteins and a small set of ribosomal proteins that came from the host cell or chronocyte.

Among the ESPs, there are two ESP ribosomal proteins that are ubiquitin-fusion proteins. This relationship suggests that ubiquitin may have been involved with protein synthesis before it became involved with protein degradation. Ubiquitin and ubiquitin ligases and proteases are prominent in our ESPs. They are involved in the eukaryotic protein degradation system. The proteins to be degraded are linked to ubiquitin by ubiquitin ligases and then handed over to the proteasome for degradation. The eukaryotic proteasome is made up of 14 different proteins (seven α and seven β), whereas in the Archaea there are only 2 different proteins (one α and one β). The proteasome may have originated in the Archaea and diversified in the eukaryotic cell, as there is a significant sequence homology of the α and β proteins of the Archaea to the 7 α and 7 β proteins of the eukaryotic cell. *Giardia*, however, has the full complexity of the eukaryotic proteasome (14). We do not find any proteasome proteins in our ESPs; however, we do find a number of proteasome-associated proteins in our ESPs. They are found in the 19S proteasome regulatory particle. We hypothesize that the eukaryotic protein degradation system is a chimeric structure with the ubiquitin, ubiquitin ligases, ubiquitin proteinases, and the 19S regulatory proteasome particle coming from the host cell (chronocyte) and the original proteasome most likely came from an archaeal endosymbiont.

A general theme in both the eukaryotic protein synthesis and degradation pathways is that part of the machinery came from the prokaryotic endosymbiont and part came from the host cell or chronocyte. This is the case with respect to the eukaryotic translation elongation factors (EF)-1α, -β, and -γ. The yeast elongation factor EF-1α has 33% sequence homology to the bacterial EF-Tu (15), and hence we do not see it in our ESPs. The prokaryotic elongation factor EF-Tu and its eukaryotic homolog EF-1α are both GTP-binding proteins. The prokaryotic EF-Ts and the eukaryotic EF-1β and EF-1γ act as GDP-exchanging proteins. The elongation factors EF-1β and EF-1γ are found in our ESPs and most probably came from the chronocyte.

What happened to the host's EF-Tu? We conjecture that it was the precursor of the largest family of proteins in our ESPs-

Table 2. List of 108 ESPs associated with signaling systems

Category	Subcategories (ID)
Calmodulin	(Cmd1) Ca-binding protein (Cdc31)
Phosphatidylinositol	Phosphatidylinositol kinases (Vps34; Pik1; Stt4; Mss4; Tel1; Tor2; Tor1; Mec1) Phosphatidylinositol phosphatases (Inp51; Inp52; Inp53)
Ubiquitin	Ubiquitin (Ubi4) Ubiquitin-like protein (Smt3) Ubiquitin-like protein (Rub1)
Ubiquitin conjugation enzymes	(Cdc34; Ubc1; Ubc4; Ubc5; Ubc6; Ubc8; Ubc9; Ubc11; Ubc12; Ubc13; Pex4; Qri8; Rad6)
Ubiquitin protease	(Ubp5; Ubp6; Ubp8; Ubp10; Ubp12; Ubp14; Ubp15; Doa4)
GTP-binding proteins	Ras (Ras1; Ras2; Rsr1; Rsg1; Tem1) Rho (Rho1; Rho2; Rho3; Rho4; Rho5; Cdc42; Rdi1) Arf (Arf1; Arf2; Arf3; Sar1; Arl1) Arf gap (Age1; Age2; Gcs1) Ran (Gsp1; Gsp2; Yrb1) Rab (Ypt1; Ypt6; Ypt7; Ypt10; Ypt31; Ypt32; Ypt52; Ypt53; Vps21) Rab gap (Mdr1; Msb3) GTP-binding-related (Cin4; Ypt11; Arl3)
Cyclin	B-type cyclin (Clb1; Clb2; Clb3; Clb4; Clb5; Clb6) Cell cycle checkpoint protein (Bub3) Cyclin-dependent kinase-activating kinase (Cak1)
Kinases and phosphatases	Serine/threonine protein kinase (Cdc7; Sky1; Iks1; Ykl171w; Vps15) Involved in cell cycle (Cdc50) Subunit of the Cdc28 protein kinase (Cks1) LAMMER protein kinases (Kns1) β subunit of casein kinase II (CKII) (Ckb1; Ckb2) Dual-specificity protein tyrosine phosphatases (Pps1; Yvh1; Tep1; Cdc14) Protein phosphatase regulatory subunits (Tpd3; Cdc55; Cnb1; Rts1; Sds22) Protein phosphatase type 2C (Ptc1; Ptc2; Ptc3; Ptc4) Myotubularin dual-specificity phosphatase (Yjr110w)
14–3–3 proteins	(Bmh1; Bmh2)

The unique identifier symbols for the proteins are from *Saccharomyces* Genome Database and are shown in parentheses. The 10 ESP proteins that have low sequence homology to prokaryotic proteins are underlined (maximal BLAST score from 50 to 55 bits).

the GTP-binding proteins (Table 2). This family of GTP-binding proteins include the subfamilies labeled Ras, Rho, Rab, Arf, and Ran (16). Members of all of these subfamilies serve as biological switches. Before the GTP-binding proteins Ras, Rho, Rab, and Arf can act as switches, they must be localized to the various cell membranes by a posttranslational modification with a lipid (farensyl, geranylgeranyl, and myristyl groups).

Ras is localized to the plasma membrane. Ras in its GTP form activates a cascade of serine/threonine kinases. Rho GTP-binding proteins are involved in the organization of the actin cytoskeleton. Rho proteins are also involved in phagocytosis and endocytosis. Rab GTP binding proteins are localized to the ER and the Golgi apparatus and are involved in vesicle transport. Arf is localized to the Golgi apparatus, where it is involved in the budding of vesicles from the Golgi apparatus.

We assume that these proteins—Ras, Rho, Rab, and Arf—have evolved from the membrane-protein-synthesizing machinery and cytoskeleton of the chronocyte. They are now localized on the cytoskeleton and membranes (the plasma, ER, and Golgi) of the eukaryotic cell.

The GTP-binding protein Ran does not have a lipid attachment and is thus not localized to a membrane. In fact, its main function is in the transport of molecules into the nucleus from the cytoplasm and the transport of molecules from the nucleus into the cytoplasm. Thus, Ran is a system that came into existence when the nucleus became an endosymbiont. There has recently been a phylogenetic comparison of prokaryotic and eukaryotic GTP-binding proteins. There was a clear separation of the prokaryotic families from the eukaryotic (Rab, Ran, Ras, and Rho) GTP-binding proteins (17).

The Nucleus. The ESPs found in the nucleus are dominated by proteins involved in the synthesis, processing and transport of the RNAs of the nucleus into the cytoplasm. These nuclear ESPs are the transcription factors, zinc fingers, proteins associated with the RNA polymerases, spliceosomal proteins, a poly(A) polymerase, an mRNA capping enzyme, and the nuclear pore proteins (Table 3). There are also nucleolar proteins associated with the synthesis and transport of ribosomal RNA. The formation of mRNA in the eukaryotic cells frequently involves the splicing-out of introns from a larger precursor RNA. According to the exon–early theory, the origin of splicing mechanisms is assumed to have arisen in an RNA-based cell. Because RNA replication is far more error-prone than DNA replication, splicing may have originated as an error correction mechanism (18). Among the nuclear ESPs there are four histones: H2A, H2B, H3, and H4. However, the eukaryotic histones share the same three-dimensional structure with the archaeal histone-like proteins of the Euryarchaeota (methanogens, etc.) (19, 20). Unlike actin, tubulin, ubiquitin, and the GTP-binding proteins, whose three-dimensional homologs are found throughout the Archaea and Bacteria, the histone fold is found only in the Euryarchaeota and not in the Crenarchaeota or the Bacteria. The simplest explanation at the present time for the evolution of histones is that a histone-like protein came in with an ancient archaeal endosymbiont and subsequently evolved into the full eukaryotic complement of histones.

Table 3. List of 47 ESPs associated with the nucleus

Category	Subcategories (ID)
DNA-associated proteins	
Histones	Histone H2A (Hta1; Hta2) Histone H2B (Htb1; Htb2) Histone H3 (Hht2; Hht1) Histone H4 (Hhf1; Hhf2)
Histone-associated	Histone acetyltransferase (Gcn5; Hat2) (Cse4)
Topoisomerase I	(Trf5; Trf4)
Transcriptional factors	(Mob2; Mob1; Hap3; Sip2; Set2; Sps18; Ssl1; Gts1; Htz1)
Zinc fingers	(Ybr267w; Mot2; Cth1; Sas2; Glo3; Tis11)
RNA-associated proteins	
RNA polymerase	Subunits of RNA polymerases (Rpc19; Rpb8)
Splicesome	Core snRNP protein (Smd3) RNA splicing factor (Prp9) SnRNA-associated protein of the Sm class (Lsm2) U5 snRNP and spliceosome component (Prp8)
RNA enzymes	RNA exonuclease (Rex3) Ribonuclease H (Rnh70) mRNA guanylyltransferase [mRNA capping] (Ceg1) RNA (guanine-7-)methyltransferase [capping] (Abd1) Poly(A) polymerase (Pap1)
Nucleolus	Nucleolar protein (Ebp2) Small nucleolar RNP proteins (Gar1) Protein required for biogenesis of the 60S ribosomal subunit (Brx1)
Nuclear pore and transport	Nuclear pore protein (Nsp1; Ntf2; Glc2) Karyopherin α (Srp1) Putative nuclear protein (Mak16)

The unique identifier symbols for the proteins are from *Saccharomyces* Genome Database and are shown in parentheses. Two ESP proteins that have low sequence homology to prokaryotic proteins are underlined (maximal BLAST score from 50 to 55 bits).

The Cell Cycle. The formation of the nucleus created a problem for the primitive eukaryotic cell: how to coordinate the division of the cytoplasm with that of the nucleus. The problem was solved by packaging the DNA in the nucleus into a small number of chromosomes. These chromosomes would double in each cell cycle. The cytoskeleton (actin filaments and microtubules) that had played such an important role in phagocytosis was now enlisted in the separation of the duplicated chromosomes to the daughter cells. This complex molecular ballet, called mitosis, was coordinated by the cyclins, a group of proteins that in turn activated the serine/threonine kinases. The cyclins oscillate because of their synthesis early in the cell cycle and their hydrolysis in the later phases of the cell cycle mediated by ubiquitin. This oscillation appears to be the master cycle in the eukaryotic cell cycle. The regulators of the eukaryotic cell cycle (cyclins, serine/threonine kinases, and ubiquitin proteins) are present among ESPs (Tables 2 and 3). However, the wide distribution of serine/threonine kinases among the Bacteria and the Archaea has led Leonard *et al.* (21) to postulate the existence of these proteins in the common eukaryote and prokaryote ancestor. This observation would imply that perhaps some of the serine/threonine kinases came in with the bacterial and archaeal endosymbionts and some were already present in the host cell. The great differences in the cell cycles of prokaryotic cells and eukaryotic cells could be explained if the cyclins, some of the serine/threonine kinases (found in our ESPs), and the cytoskeleton came from the host cell or chronocyte. This scenario would be a novel perspective on the evolution of the cell cycle, as it implies that the cell cycle did not simply evolve from the prokaryotic cell cycle (22).

Discussion

The eukaryotic cell is not a simple fusion of an archaeon and bacterium. This statement is borne out by the existence of 347 ESPs. This finding agrees with the predictions of the ABC hypothesis. The 254 proteins that have an assigned function are intimately related to the structure and function of the eukaryotic cell. They are the components of the cytoskeleton, inner membranes, RNA-modification machinery, and the major elements of intracellular control systems such as ubiquitin, inositol phosphates, cyclins, and the GTP-binding proteins.

Thus, the ESPs with an assigned function are able to give us a partial picture of the chronocyte. It had a plasma membrane and a cytoskeleton. The coordination of this membrane–cytoskeleton system by means of calcium ions allowed the chronocyte to phagocytize archaea and bacteria. There was a complex inner membrane system where proteins were synthesized and broken down which eventually evolved into the ER, the GTP-binding proteins, ubiquitin, and the 11 ESP ribosomal proteins. This result is not a complete picture of the chronocyte, as we still cannot account for the functions of 93 ESPs (Table 4).

We have also found that if we began our search for ESPs with *Schizosaccharomyces pombe* instead of *S. cervesiae*, we got back a majority of the ESPs as analyzed above but also got some new ESPs (results are posted at www.mcb.harvard.edu/gilbert/ESP). If, in our search for ESPs, we excluded *Drosophila*, we found an RNA-directed RNA polymerase. This enzyme is involved in the replication of RNAi, an RNA involved in posttranscriptional silencing. Because this enzyme is not found in Bacteria and Archaea, it suggests that the chronocyte was an RNA-based cell. This finding is also consistent with the differences found between the proteins of *S. cervesiae* and those of *Sch. pombe*. When the proteins of *S. cervesiae* are compared with the proteins of *Sch. pombe*, there is evidence that 300 proteins have been lost by *S. cervesiae*, including many components of the spliceosome, signalosome, and the posttranscriptional gene-silencing systems (23). Thus, many cellular genomes are necessary for the reconstruction of the chronocyte, as some proteins can be missing in an individual genome. To get a full reconstruction of the chronocyte, we need more eukaryotic sequenced genomes, that are well annotated.

The results from *Sch. pombe* point to the genome of the chronocyte as being based on RNA. It was hypothesized that the chronocyte was an RNA-based cell (9). Thus far we have focused on the ESPs of the eukaryotic cell. A full picture of the chronocyte would be completed by a set of eukaryotic signature RNAs (ESRs). There are the ribosomal and tRNAs to be considered, and to complete this set of ESRs we would require a complete catalog of all of the noncoding RNAs found in the Bacteria, Archaea, and the Eukarya. We already have some members of the noncoding ESRs, such as spliceosomal RNAs and a host of small RNAs of unknown function, which are found in the eukaryotic cell and not in Bacteria or Archaea. The non-coding RNAs in the eukaryotic cell is an area of cellular research that bears close attention in the near future.

The hypothesis that the nucleus was a prokaryotic endosymbiont in an RNA-based host cell (chronocyte) can explain why transcription occurs in the nucleus and translation occurs in the cytoplasm. The separation between transcription and translation would be the result of the communication setup between the endosymbiont, a DNA-based cell, and the host cell, which was an RNA-based cell. mRNA made in the endosymbiont would be transported and translated in the "cytoplasm" of the chronocyte. There are other important processes found in the eukaryotic cell,

Table 4. List of eight ESP enzymes and 93 ESPs with unknown functions

Category	Subcategories (ID)
Enzymes	Riboflavin kinase (Fmn1) FAD synthetase (Fad1) Protein carboxyl methylase (Ycr047c) N-terminal acetyltransferase (Nat3) Acetyltransferase in the SAS gene family (Esa1) Glucosamine-phosphate *N*-acetyltransferase (Gna1) UDP–glucose pyrophosphorylase (Yhl012w) Phosphoryltransferase (Gpi13)
Clusters of unknown proteins	(Ydr126w; Erf2; Ydr459c; Ynl326c; Yol003c) (Psr2; Ypl063w; Nem1; Psr1) (Ygl014w; Mpt5; Yll013c) (Tom1; Rsp5; Hul4) (Yil088c; Ybl089w; Ynl101w) (Vps24; Fti1; Ykl002w) (Yfr021w; Ypl100w; Ygr223c) (Ylr328w; Ygr010w) (Ypl249c; Msb4) (Msi1; Rsa2) (Imp4; Rpf1) (Mrd1; Ynl110c) (Ykl121w; Ymr102c) (Ssf1; Ssf2) (Gdi1; Mrs6) (Ydl060w; Bms1)
Unique unknown proteins	Sas3, Ydl216c, Sfb3, Las21, Ynr048w, Hym1, Abp140, Rlp7, Yor289w, Yhr122w, Hrt1, Nmd3, Yol093w, Yhr186c, Yer082c, Yer126c, Nud1, Ypl247c, Yil113w, Ypl236c, Ylr409c, Nip7, Vip1, Yil005w, Ybr228w, Enp1, Bph1, Ymr068w, Yjl109c, Ypr031w, Yth1, Ent3, Ptk1, Ykt6, Ydr083w, Ykl100c, Ykl099c, Ygl047w, Ykl056c, Pri2, Plp1, Ufd1, Pac10, Ygr145w, Crm1, Sgt1, Ydr266c, Gpi8, Ydr339c, Ydr365c, Ydr411c

The unique identifier symbols for the proteins are from *Saccharomyces* Genome Database and are shown in parentheses. The 11 ESP proteins that have low sequence homology to prokaryotic proteins are underlined (maximal BLAST score from 50 to 55 bits).

such as reverse transcription, splicing, etc., that have evolved out of the cellular processes of the chronocyte and were not brought into the eukaryotic cell by the prokaryotic symbionts.

Conclusions

We agree with Horiike *et al*. (1) that the nucleus is an endosymbiont with inputs from Bacteria and Archaea. We disagree that the host cell came from the Bacteria. The host cell or chronocyte was not a prokaryotic cell but one that had a cytoskeleton composed of actin and tubulin and an extensive membrane system.

The chronocyte donated to the resulting eukaryotic cell, its cytoskeleton, ER, Golgi apparatus, and major intracellular control systems, such as calmodulin, ubiquitin, inositol phosphates, cyclin, and the GTP-binding proteins.

Finally, the full characterization of the Chronocyte will come from an understanding of the ESRs.

We thank C. Woese, T. Smith (Boston University), W. Gilbert (Harvard University), and C. Burge (MIT) for valuable discussions. We also thank M. L. Sogin and his group at the Marine Biological Laboratory (Woods Hole, MA) for their assistance in accessing the *Giardia* database. We are grateful to W. Gilbert for allowing us to use the computational facilities at Harvard, which made this study possible.

1. Horiike, T., Hamada, K., Kanaya, S. & Shinozawa, T. (2001) *Nat. Cell Biol.* **3,** 210–214.
2. Doolittle, W. F. & Logsdon, J. M., Jr. (1998) *Curr. Biol.* **8,** R209–211.
3. Graham, D. E., Overbeek, R., Olsen, G. J. & Woese, C. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3304–3308.
4. Gillin, F. D., Reiner, D. S. & McCaffery, J. M. (1996) *Annu. Rev. Microbiol.* **50,** 679–705.
5. Mereschowsky, C. (1910) *Biol. Zentralbl.* **30,** 278–367.
6. Gupta, R. S. & Golding, G. B. (1996) *Trends Biochem. Sci.* **21,** 166–171.
7. Lake, J. A. & Rivera, M. C. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2880–2881.
8. Gupta, R. S. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 1435–1491.
9. Hartman, H. (1984) *Speculations Sci. Technol.* **7,** 77–81.
10. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A. & Wheeler, D. L. (1999) *Nucleic Acids Res.* **27,** 12–17.
11. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acid Res.* **25,** 3389–3402.
12. Doolittle, R. F. (1995) *Philos. Trans. R. Soc. London B* **349,** 235–240.
13. Shirakura, T., Maki, Y., Yoshida, H., Arisue, N., Wada, A., Sanchez, L. B., Nakamura, F., Muller, M. & Hashimoto, T. (2001) *Mol. Biochem. Parasitol.* **112,** 153–156.
14. Bouzat, J. L., McNeil, L. K., Robertson, H. M., Solter, L. F., Nixon, J. E., Beever, J. E., Gaskins, H. R., Olsen, G., Subramaniam, S., Sogin, M. L. & Lewin, H. A. (2000) *J. Mol. Evol.* **51,** 532–543.
15. Negrutskii, B. S. & El'skaya, A. V. (1998) *Prog. Nucleic Acid Res. Mol. Biol.* **60,** 47–78.
16. Takai, Y., Sasaki, T. & Matozaki, T. (2001) *Physiol. Rev.* **81,** 153–208.
17. Mittenhuber, H. (2001) *J. Mol. Microbiol. Biotechnol.* **3,** 21–35.
18. Reanney, D. C. (1984) *J. Theor. Biol.* **110,** 315–321.
19. Arents, G. & Moudrianakis, E. N. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11170–11174.
20. Sandman, K. & Reeve, J. N. (2000) *Arch. Microbiol.* **173,** 165–169.
21. Leonard, C. J., Aravind, L. & Koonin, E. V. (1998) *Genome Res.* **8,** 1038–1047.
22. Nasmyth, K. (1995) *Philos. Trans. R. Soc. London B* **349,** 271–281.
23. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11319–11324.