J Mol Evol (2004) 59:695–702
DOI: 10.1007/s00239-003-0085-1

JOURNAL OF MOLECULAR EVOLUTION

What Does the Microsporidian *E. cuniculi* Tell Us About the Origin of the Eukaryotic Cell?

Alexei Fedorov,[1] Hyman Hartman[2]

[1] Department of Medicine, Medical College of Ohio, Toledo OH 43614, USA
[2] Center for Biomedical Engineering NE47-377, MIT, Cambridge, MA 02139, USA

Received: 19 August 2003 / Accepted: 1 July 2004 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. The relationship among the three cellular domains Archaea, Bacteria, and Eukarya has become a central problem in unraveling the tree of life. This relationship can now be studied as the completely sequenced genomes of representatives of these cellular domains become available. We performed a bioinformatic investigation of the *Encephalitozoon cuniculi* proteome. *E. cuniculi* has the smallest sequenced eukaryotic genome, 2.9 megabases coding for 1997 proteins. The proteins of *E. cuniculi* were compared with a previously characterized set of eukaryotic signature proteins (ESPs). ESPs are found in a eukaryotic cell, whether from an animal, a plant, a fungus, or a protozoan, but are not found in the Archaea and the Bacteria. We demonstrated that 85% of the ESPs have significant sequence similarity to proteins in *E. cuniculi*. Hence, *E. cuniculi*, a minimal eukaryotic cell that has removed all inessential proteins, still preserves most of the ESPs that make it a member of the Eukarya. The locations and functions of these ESPs point to the earliest history of eukaryotes.

Key words: Eukaryote — *Giardia* — *Encephalitozoon cuniculi* — Microsporidia — Minimal cell — Nucleus — Endosymbiosis

Correspondence to: Hyman Hartman; *email:* hhartman@mit.edu

Introduction

The microsporidia were once considered to be the deepest-branching eukaryote taxon. Carl Woese and his team found the microsporidia to be very close to the root of the Eukarya based on their ribosomal RNA phylogeny (Vossbrinck et al. 1987). However, this deep divergence was due to a long-branch attraction artifact resulting from rapid evolution of the microsporidia (Gribaldo and Phillippe 2002). The recent analyses of combined protein data resulted in the identification of the microsporidian *Encephalitozoon cuniculi* as a member of the fungal kingdom (Baldauf et al. 2000). In fact, *E. cuniculi* is an intracellular fungal parasite. This lifestyle has caused the shrinking of its genome. *E. cuniculi* is thus a candidate for the "minimal" eukaryotic cell. It has the added virtue to have had its genome fully sequenced.

Let us examine the concept of a minimal cell. The search for the minimal cell began in the cellular domain of the Bacteria with the investigation of the Mycoplasma, a group of small parasitic bacteria (Razin 1997). *Mycoplasma genitalium*, an intracellular parasite, has one of the smallest known bacterial genomes. Because of its small size, it was among the first bacterial genomes to be sequenced. The genome of *M. genitalium* has only 580,000 base pairs (bp), which codes for 468 proteins (Fraser et al. 1995). The characterization of the minimal cell began by comparing the proteins found in *M. genitalium* to those found in *Haemophilus influenzae*, an early sequenced gram-negative bacteria causing ear infections (Mushegian and Koonin 1996). Computer analysis

produced a set of 240 proteins of *M. genitalium* that have orthologs in *H. influenzae*. This set was expanded by the addition of 16 other "gene displacement" proteins. Thus the genome of a minimal cell would code for the 256 proteins that would carry out the bare essential cellular functions such as DNA replication, protein synthesis (translation and transcription), metabolism (glycolysis), and various membrane related functions. It was assumed that the minimal bacterial cell would be a minimal cell for the other cellular domains. However, when Mushegian and Koonin compared these 256 proteins with those from other domains, a lack of correspondence unexpectedly appeared. For example, they found that different sets of proteins were involved in Eukarya and Archaea DNA replication. The authors concluded that the last common ancestor of the three cellular domains "had an RNA genome" and that DNA replication had evolved twice: once in Bacteria and independently in Archaea and Eukarya. (Mushegian and Koonin 1996). Hence, a minimal eukaryotic cell or a minimal archaeal cell would differ from a minimal bacterial cell.

To characterize a "minimal" eukaryotic cell we began our study with *E. cuniculi*. The genome of microsporidian *E. cuniculi* has been sequenced (Katinka et al. 2001). Its length is 2.9 million bp, close to the median size of sequenced prokaryotic genomes (2.6 million bp). *E. cuniculi* has 1997 predicted genes, 44% of which have known functions (Katinka et al. 2001). These data imply that *E. cuniculi* is approaching the size of a minimal eukaryotic cell. To understand the differences between minimal cells of Eukarya and Bacteria, we need to characterize the set of proteins in *E. cuniculi* that is unique to the eukaryotes but absent from other cellular domains. Previously, we collected a set of proteins that were found in all sequenced eukaryotic cells and absent in Bacteria or Archaea (Hartman and Fedorov 2002). We called this set eukaryotic signature proteins (ESPs). Each protein from the ESP set has homologs in all main eukaryotic branches—animals (*Drosophila melanogaster* and *Caenorhabditis elegans*), plants (*Arabidopsis thaliana*), fungi (*Saccharomyces cerevisiae*), and protists (*Giardia lamblia*)—but does not have homologs in the Archaea and Bacteria. The extracellular eukaryotic parasite *G. lamblia* was specifically chosen for the characterization of these ESPs, because it is still considered to be one of the deepest-branching taxon of the eukaryotes (i.e., the diplomonads). The numbers and composition of the proteins from the ESP set of *Giardia* led us to conclude that the origin of eukaryotes involved the formation of the nucleus from prokaryotic endosymbionts in an RNA-based cell (Hartman and Fedorov 2002). Here we compare the ESPs with proteins of a minimal and highly diverged eukaryotic cell of *E. cuniculi*. The surprising result is the overwhelming agreement (85%) between the eukaryote-specific proteins of *E. cuniculi* and the previously characterized set of *Giardia* ESPs. What can the genome of the intracellular microsporidian parasite *E. cuniculi* tell us about the origin of the eukaryotic cell? The answer is a great deal as we investigate a minimal eukaryotic cell, which has eliminated all but the most essential functions.

Materials and Methods

Protein sequence databases of *S. cerevisiae*, *D. melanogaster*, *A. thaliana*, *G. lamblia*, and 44 bacteria and archaea were downloaded as previously described (Hartman and Fedorov 2002). Protein sequences of *E. cuniculi* (1997 entries) were downloaded from GenBank (Benson et al. 1999).

ESPs of Giardia. In our previous paper (Hartman and Fedorov 2002) we performed consecutive BLAST2.0 alignments of *S. cerevisiae* proteins with proteins of *D. melanogaster*, *C. elegans*, *A. thaliana*, and *G. lamblia* and those of 44 bacteria and archaea species. We used a blast score of 55 bits. This score was based on our consultation with experts in bioinformatics, as the *Giardia* database was in contigs only and was not assembled or annotated. As a result, 347 yeast proteins that have significant sequence similarity to proteins of all studied eukaryotes, but not to any bacteria or archaea (identity threshold of 55 blast score bits), were selected. We call this set ESPs of *G. lamblia*, since *Giardia* represents the most divergent eukaryotic proteins in the studied group.

ESPs of E. cuniculi. The comparison of *G. lamblia* and *E. cuniculi* proteins began with collecting ESPs of *E. cuniculi* which followed our previous approach for gathering ESPs of *Giardia*. We started from 6271 *S. cerevisiae* proteins and compared them with proteins of *D. melanogaster*, *C. elegans*, *A. thaliana*, *E. cuniculi*, and 44 sequenced bacteria and archaea using BLAST 2.0 alignment program (Altschul et al. 1997). This resulted in 401 *S. cerevisiae* sequences with significant similarity to all sequenced eukaryotes but not to any bacteria or archaea. The same similarity threshold of 55 blast score bits (used in our previous paper on *Giardia* ESPs [Hartman and Fedorov 2002]) has been employed in the present study which approximately corresponds to a P-value of 10^{-6}. This threshold is sufficiently stringent and, thus, allows us to assume that matched proteins share a common origin. We were also obliged to use the same 55 blast score as we were about to compare the *E. cuniculi* ESP results with those of *G. lambia*. We then compared the 347 ESPs of *Giardia* with the 401 ESPs of *E. cuniculi*. It was found that 238 proteins are common to the ESPs of *G. lamblia* and *E. cuniculi*, while 109 ESPs are unique to *G. lamblia* and 163 are unique to ESP of *E. cuniculi*. These unique sets were studied further using PSI-BLAST programs.

PSI-BLAST. For each unique protein from the ESP set of *G. lamblia* that does not match an *E. cuniculi* ESP, we found the best-matched protein from *D. melanogaster* and *A. thaliana* proteome using the BLAST 2.0 program. These three protein sequences were aligned with each other by CLUSTALW 1.8 (Higgins and Sharp 1988) and the multiple alignments obtained was used as input for the PSI-BLAST program (Altschul et al. 1997) in the search of the *E. cuniculi* protein database. The results of the round 2 PSI-BLAST output were analyzed automatically by our PERL program (prog_PSI_401_2). When a protein from the *G. lamblia* ESP set had a PSI-BLAST alignment score with the *E. cuniculi*

Table 1. Comparison of ESP sets of *G. lamblia* and *E. cuniculi*

Protein set	BLAST 2.0		PSI-BLAST (round 2)	
	No. common (≥55 score bits)	No. unique (< 55 score bits)	No. weak-homolog (> 88 score bits)	unique (< 40 score bits)
347 ESPs, *G. lamblia*	238	109	57	52
401 ESPs, *E. cuniculi*	238	163		

database higher than 88 bits, (in 98% of the cases it was > 100 bits), the protein was called a "putative homolog" and is shown in green in Fig. 1. Otherwise, the protein was called "unique" and is shown in red on in Fig. 1. At round 2 of PSI-BLAST all "unique" proteins had a psi-blast score of < 50 bits. We did not perform the reciprocal procedure for the PSI-BLAST comparison of unique *E. cuniculi* ESPs with the *Giardia* database because the *G. lamblia* database consists of multiple short nucleotide sequences translated in six possible reading frames. This would lead to false negative results. All computational procedures were performed automatically by the PERL script prog_PSI_401_2. The program and all described protein sets are available at our web page, www.mco.edu/medicine/fedorov/E_cuniculi.

Protein Groups. All 401 proteins were compared with each other by BLAST 2.0 binaries (Altschul et al. 1997). Next we performed the simplest grouping procedure: (1) two proteins were considered similar and put in the same group if they had a similarity score > 55 bits, and (2) groups were pooled together if any member of one had sequence similarity (≥55 bits) to any protein of another group. This procedure yielded 214 different protein groups.

Results

Since *Giardia* proteins (open reading frames) are still not assembled or annotated, we cannot compare eukaryote-specific proteins of *G. lamblia* with those of *E. cuniculi* directly. In our previous paper, 914 proteins of *S. cerevisiae* which have sequence similarity to *D. melanogaster*, *C. elegans*, and *A. thaliana* (blast score, 55 bits) and no sequence similarity to Bacteria and Archaea were compared with the fragmented *Giardia* database. That comparison resulted in a set of 347 proteins comprised of 180 unrelated protein groups. We called these proteins ESPs of *G. lamblia*.

Here we used the same set of 914 eukaryote-specific proteins of *S. cerevisiae* and aligned it with the entire set of 1997 *E. cuniculi* proteins. This analysis of 401 budding yeast proteins showed significant sequence similarity to the microsporidian proteome (similarity level, > 55 blast score bits). We called the 401 proteins that were obtained *E. cuniculi* ESPs. The 401 ESPs of *E. cuniculi* were compared to the 347 ESPs of *G. lamblia*. The main results of this comparison are summarized in Table 1 and the entire comparison is presented in Fig. 1. Two hundred thirty-eight proteins are common in the two sets, while 109 are unique to *G. lamblia* ESPs and 163 to ESPs of *E. cuniculi* (see Table 1). However, when the 109 unique ESP *G. lamblia* proteins were compared with the entire set of *E. cuniculi* proteins using PSI-BLAST alignment, 57 of these proteins showed weak, yet significant, similarity to the microsporidian (shown in green in Fig. 1). The other 52 proteins remained specific to *Giardia* (shown in red in Fig. 1). Reciprocal psi-blast of *E. cuniculi* unique ESPs was not performed for reasons outlined under Materials and Methods.

The ESP sets have some redundancy because of recent evolutionarily duplication of a substantial part of the *S. cerevisiae* genome. In order to take this redundancy into account, we divided the ESPs into unique protein groups. The similarity threshold of a blast score equaling 55 bits was used for grouping procedure. As a result, 401 ESPs of *E. cuniculi* were divided into 214 unique groups, while 347 ESPs of *G. lamblia* were divided into 180 unique groups. Comparison of these groups demonstrated that 108 of the 180 *Giardia* ESP groups are common to *E. cuniculi* ESPs (55-bit threshold of BLAST 2.0). When "putative homologs" revealed by psi-blast were taking into account, the number of common groups increased to 142.

Discussion

There are three groups of proteins in the ESPs of *E. cuniculi*: (1) those that can be matched to the ESPs of *G. lamblia* by means of a blast score of 55 bits, (2) those that can be matched to the ESPs of *G. lamblia* by means of psi-blast, and (3) those that have no sequence similarity to the ESPs of *Giardia*. We discuss the relevance of these three groups to the structures of the eukaryotic cell.

The Plasma Membrane and the Cytoskeleton. As we discussed in our previous paper, one of the deepest distinctions between prokaryotes and eukaryotes is found in the plasma membrane (Hartman and Fedorov 2002). There is a lack of clathrin and associated proteins in the ESPs of *E. cuniculi* (see Fig. 1). Since *E. cuniculi* is an intracellular parasite, it possibly has a diminished need for clathrin-based endocytosis. This conjecture points out a difference between being an intracellular parasite (*E. cuniculi*) and being an extracellular parasite (*Giardia*).

Giardia	E. cuniculi
PLASMA MEMBRANE AND ENDOCYTOSIS	
Clathrin (Chc1)	----
clathrin-associated proteins (Apl2; Apm1; Aps1; Aps2; Aps3; Apl1; Apl3; Apl4; Apl5; Apm2; Apm4;)	clathrin-associated proteins (Apl2; Apl6; Aps1; Aps2; Aps3; Apl1; Apm4; Apl4; Apm1)
dynamin (Dnm1; Mgm1; Vps1)	dynamin (Dnm1; Mgm1; Vps1)
CYTOSKELETON	
Tubulin	
alpha-tubulin (Tub1; Tub3)	alpha-tubulin (Tub1; Tub3)
beta-tubulin (Tub2)	beta-tubulin (Tub2)
gamma tubulin-like protein (Tub4)	gamma tubulin-like protein (Tub4)
Tubulin-assosiated proteins	
kinesin-related protein (Kip2; Kar3)	kinesin-related protein (Kip2; Kar3)
kinesin-related protein involved in mitosis (Kip3)	kinesin-related protein involved in mitosis (Kip3)
kinesin heavy chain homolog (Smy1)	kinesin heavy chain homolog (Smy1)
microtubule-binding protein (Bim1)	microtubule-binding protein (Bim1)
putative light chain of dynein (Dyn2)	putative light chain of dynein (Dyn2)
Actin	
(Act1)	(Act1)
Actin-related proteins	
(Arp1; Arp2; Arp3; Arp4; Arp5; Arp6; Arp7)	(Arp1; Arp2; Arp3; Arp4; Arp5; Arp6; Arp7)
Actin-assosiated proteins	
light chain for myosin (Mlc1)	light chain for myosin (Mlc1)
----	myosin (Myo2; Myo3; Myo4; Myo5)
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 S24 (Rps24a; Rps24b)
ribosomal protein S26A (Rps26a; Rps26b)	ribosomal protein S26A (Rps26a; Rps26b)
ribosomal protein S27 (Rps27a; Rps27b)	ribosomal protein S27 (Rps27a; Rps27b)
ribosomal protein S31 [ubiquitin related] (Rps31)	ribosomal protein S31 [ubiquitin related] (Rps31)
Large ribosomal proteins	
----	ribosomal protein L6 (Rpl6b; Rpl6a)
ribosomal protein L13 (Rpl13b; Rpl13a)	ribosomal protein L13 (Rpl13b; Rpl13a)
ribosomal protein L14 (Rpl14a; Rpl14b)	----
ribosomal protein L18 (Rpl18a; Rpl18b)	ribosomal protein L18 (Rpl18a; Rpl18b)
ribosomal protein L20 (Rpl20b; Rpl20a)	ribosomal protein L20 (Rpl20b; Rpl20a)
ribosomal protein L21 (Rpl21b; Rpl21a)	ribosomal protein L21 (Rpl21b; Rpl21a)
ribosomal protein L24 (Rpl24a; Rpl24b)	ribosomal protein L24 (Rpl24a; Rpl24b)
ribosomal protein L29 (Rpl29)	----
ribosomal protein L33 (Rpl33a)	ribosomal protein L33 (Rpl33a)
ribosomal protein L35 (Rpl35b; Rpl35a)	ribosomal protein L35 (Rpl35b; Rpl35a)
ribosomal protein L36 (Rpl36a; Rpl36b)	ribosomal protein L36 (Rpl36a; Rpl36b)
ribosomal protein L40 [ubiquitin related] (Rpl40a; Rpl40b)	ribosomal protein L40 [ubiquitin related] (Rpl40a; Rpl40b)
Translation factors	
translation elongation factor EF-1beta (Efb1)	----
translation elongation factor EF-1gamma (Cam1; Tef4)	----
Proteasome associated proteins	
subunits of proteasome regulatory particle (Rpn1; Rpn8; Rpn11; Rpn10;)	subunits of proteasome regulatory particle (Rpn1; Rpn8; Rpn11; Rpn2; Rpn3; Rpn5; Rpn6; Rpn7; Rpn10)
Signal peptidase	
(Spc3)	(Spc3)
MEMBRANE	
ER and Golgi	
transport protein particle TRAPP comp. (Bet3)	transport protein particle TRAPP comp. (Bet3)
HDEL receptor (Erd2)	HDEL receptor (Erd2)
integral membrane proteins (Sac1; Fig4)	integral membrane proteins (Sac1; Fig4)
----	SNARE docking (Sec1; Sly1; Vps45)
subunit of coatomer (Sec26)	subunit of coatomer (Sec26)
vesicle coat component (Sec24)	vesicle coat component (Sec24)
miscellaneous (Gpi8)	miscellaneous (Gpi8; Hrd1; Rer1; Yip1; Sso1; Sso2; Sec21; Syg1)

Giardia	E. cuniculi
Vacuole	
vacuolar protein (Pep8)	----
retromer complex component (Vps35)	----
vacuolar ATPase V0 domain subunit c (Cup5)	vacuolar ATPase V0 domain subunit c (Cup5)
vacuolar ATPase V0 domain c" (Ppa1; Tfp3)	vacuolar ATPase V0 domain c" (Ppa1; Tfp3)
----	Vacuolar membrane protein (Vps13)
----	vacuolar ATPase V1 domain su F (Vma7)
----	vacuolar ATPase V1 domain su H (Vma13)
SIGNALING CASCADE	
Calmodulin	
(Cmd1)	(Cmd1)
Ca-binding protein (Cdc31)	Ca-binding protein (Cdc31)
Phoshatidylinositol	
phosphatidylinositol kinases (Vps34; Pik1; Stt4; Mss4; Tel1; Tor2; Tor1; Mec1)	phosphatidylinositol kinases (Vps34; Pik1; Stt4; Mss4; Tel1; Tor2; Tor1; Mec1; Tra1)
phosphatidylinositol phosphatases (Inp51; Inp52; Inp53)	phosphatidylinositol phosphatases (Inp51; Inp52; Inp53)
Ubiquitin	
ubiquitin (Ubi4)	ubiquitin (Ubi4)
ubiquitin-like protein (Smt3)	ubiquitin-like protein (Smt3)
ubiquitin-like protein (Rub1)	ubiquitin-like protein (Rub1)
Ubiquitin conjugation enzymes	
(Cdc34; Ubc1; Ubc4; Ubc5; Ubc6; Ubc8; Ubc9; Ubc11; Ubc12; Ubc13; Pex4; Qri8; Rad6)	(Cdc34; Ubc1; Ubc4; Ubc5; Ubc6; Ubc8; Ubc9; Ubc11; Ubc12; Ubc13; Pex4; Qri8; Rad6)
Ubiquitin proteases	
(Ubp5; Ubp14; Ubp8; Ubp15; Ubp12; Doa4; Ubp6; Ubp10)	(Ubp5; Ubp8; Ubp12; Ubp14, Ubp15; Doa4; Ubp7; Ubp9; Ubp11; Ubp13; Ubp6; Ubp10)
----	ubiquitin ligases (Hul5)
----	miscellaneous (Ufd2; Ufd4; Ulp1; Apc11)
GTP-binding proteins	
Ras (Ras1; Ras2; Rsr1; Rsg1; Tem1)	Ras (Ras1; Ras2; Rsr1; Rsg1; Tem1)
Rho (Rho1; Rho2; Rho3; Rho4; Rho5; Cdc42; Rdi1)	Rho (Rho1; Rho2; Rho3; Rho4; Rho5; Cdc42)
Arf (Arf1; Arf2; Arf3; Sar1; Arl1)	Arf (Arf1; Arf2; Arf3; Sar1; Arl1)
Arf gap (Age2; Gcs1; Age1)	Arf gap (Age2; Gcs1; Age1)
Ran (Gsp1; Gsp2; Yrb1)	Ran (Gsp1; Gsp2; Yrb1)
Rab (Ypt1; Ypt6; Ypt7; Ypt10; Ypt31; Ypt32; Ypt52; Ypt53; Vps21)	Rab (Ypt1; Ypt6; Ypt7; Ypt10; Ypt31; Ypt32; Ypt52; Ypt53; Vps21)
Rab gap (Mdr1; Msb3)	Rab gap (Mdr1; Msb3)
GTP-binding related (Cin4; Ypt11; Arl3)	GTP-binding related (Cin4; Ypt11; Arl3; Rna1; Srm1; Gyp1; Gyp7)
Cyclin	
B-type cyclin (Clb1; Clb2; Clb3; Clb4; Clb5; Clb6)	B-type cyclin (Clb1; Clb2; Clb3; Clb4; Clb5; Clb6)
cell cycle checkpoint protein (Bub3)	cell cycle checkpoint protein (Bub3)
cyclin-dependent kinase-activating kinase (Cak1)	cyclin-dependent kinase-activating kinase (Cak1)
Kinases and phosphateses	
serine/threonine protein kinase (Cdc7; Sky1; Ykl171w; Vps15; Iks1)	serine/threonine protein kinase (Cdc7; Sky1 Ykl171w; Vps15; Iks1)
involved in cell cycle (Cdc50)	involved in cell cycle (Cdc50)
subunit of the Cdc28 protein kinase (Cks1)	----
LAMMER Protein Kinases (Kns1)	LAMMER Protein Kinases (Kns1)
β subunit of casein kinase II (Ckb1; Ckb2)	β subunit of casein kinase II (Ckb1; Ckb2)
dual-specificity tyrosine phosphatases (Pps1; Yvh1; Tep1; Cdc14)	----
protein phosphatase regulatory subunits (Cdc55; Cnb1; Sds22; Rts1; Tpd3)	protein phosphatase regulatory subunits (Cdc55; Cnb1; Sds22; Rts1; Tpd3)
protein phosphatase type 2C (Ptc1; Ptc2; Ptc3; Ptc4)	----
myotubularin dual specif phosphatase (Yjr110w)	----
14-3-3 proteins	
(Bmh1; Bmh2)	(Bmh1; Bmh2)

A

Fig. 1. Continued.

Giardia	E. cuniculi
Lipid attachements	
geranylgeranyltransferase type II β (Bet2)	geranylgeranyltransferase type II β (Bet2)
geranylgeranyltransferase type IIα (Bet4)	geranylgeranyltransferase type IIα (Bet4)
geranylgeranyltransferase type I (Cdc43)	geranylgeranyltransferase type I (Cdc43)
farnesyltransferase beta subunit (Ram1)	farnesyltransferase beta subunit (Ram1)
CAAX farnesyltransferase α (Ram2)	CAAX farnesyltransferase α (Ram2)
farnesyl cysteine-carboxyl methyltransferase (Ste14)	farnesyl cysteine-carboxyl methyltransferase (Ste14)
N-myristoyl transferase (Nmt1)	N-myristoyl transferase (Nmt1)
Rab geranyltransferase regulatory (Mrs6)	Rab geranyltransferase regulatory (Mrs6)
NUCLEUS	
Histones	
histone H2A (Hta1; Hta2)	histone H2A (Hta1; Hta2)
histone H2B (Htb1; Htb2)	histone H2B (Htb1; Htb2)
histone H3 (Hht2; Hht1)	histone H3 (Hht2; Hht1)
histone H4 (Hhf1; Hhf2)	histone H4 (Hhf1; Hhf2)
Histone-associated proteins	
histone acetyltransferase (Gcn5; Hat2)	histone acetyltransferase (Gcn5; Hat2)
Chromatin binding proteins (Cse4)	Chromatin binding proteins (Cse4; Esp1; Rsc4; Pho23; Yng1)
Topoisomerase I	
(Trf5; Trf4)	(Trf5; Trf4)
Transcriptional factors	
(Hap3; Set2; Sps18; Ssl1; Gts1; Htz1; Mob1; Mob2; Sip2)	(Hap3; Set2; Sps18; Ssl1; Hap2; Hap5; Fcp1; Bdf1; Bdf2; Spt5; Spt7; Taf60; Tfa1;Taf19; Toa1; Asf1; Fhl1; Tsm1; Cdc39; Bas1; Ada2; Bur6; Arg80; Ccr4; Sin3; Azf1; Pzf1; Fzf1; Crz1; Gts1; Mob2)
Zinc fingers	
(Mot2; Cth1; Sas2; Glo3; Tis11; Ybr267w)	(Mot2; Cth1; Sas2; Glo3; Tis11; Ssu72; Sfb2; Ybr267w)
DNA-repair enzymes	
----	(Pol12; Pob3; Hys2; Soh1; Mus81; Rfa1; Ixr1; Mre11)
RNA-polymerase	
subunits of RNA polymerases (Rpc19 Rpb8)	subunits of RNA polymerases (Rpc19)
Spliceosomal proteins	
(Smd3; Prp8; Prp9; Lsm2)	(Smd3; Prp8; Prp9; Smd2; Smx3; Cef1; Hsh49; Cus1; Cbc2; Luc7; Lsm2)
RNA enzymes	
RNA exonuclease (Rex3)	RNA exonuclease (Rex3; Rex4)
ribonuclease H (Rnh70)	ribonuclease H (Rnh70)
mRNA guanylyltransferase [capping] (Ceg1)	mRNA guanylyltransferase [capping] (Ceg1)
RNA (guanine-7-)methyltransferase (Abd1)	RNA (guanine-7-)methyltransferase (Abd1)
poly(A) polymerase (Pap1)	poly(A) polymerase (Pap1)
----	miscellaneous (Rna14; Cft2)
Nucleolus	
nucleolar protein (Ebp2)	nucleolar protein (Nop13)
small nucleolar RNP proteins (Gar1)	small nucleolar RNP proteins (Gar1)
protein required for biogenesis of the 60S ribosomal subunit (Brx1)	protein required for biogenesis of the 60S ribosomal subunit (Brx1)
Nuclear Pore and Transport	
nuclear pore protein (Gle2; Nsp1; Ntf2)	nuclear pore protein (Gle2; Nsp1)
karyopherin alpha (Srp1)	karyopherin alpha (Srp1; Kap95)
putative nuclear protein (Mak16)	putative nuclear protein (Mak16)
----	nuclear pore complex (Nup100; Nup116)
RNA binding and export	
----	(Mip6; Lhp1; Gbp2; Pes4)

Giardia	E. cuniculi
OTHERS	
Enzymes	
riboflavin kinase (Fmn1)	----
FAD synthetase (Fad1)	----
protein carboxyl methylase (Ycr047c)	protein carboxyl methylase (Ycr047c)
N-terminal acetyltransferase (Nat3)	N-terminal acetyltransferase (Nat3)
acetyltransferase (SAS gene family) (Esa1)	acetyltransferase (SAS gene family) (Esa1)
glucosamine-phosphate N-acetyltransferase (Gna1)	glucosamine-phosphate N-acetyltransferase (Gna1)
UDP Glucose pyrophosphorylase (Yhl012w)	UDP Glucose pyrophosphorylase (Yhl012w)
phosphoryltransferase (Gpi13)	phosphoryltransferase (Gpi13)
----	Similar to NADH dehydrogenase (Yif1)
----	Phosphoacetylglucosamine mutase (Pcm1)
----	Desaturase/hydroxylase enzyme (Scs7)
----	N-AGPinositol de-N-acetylase (Gpi12)
----	Polyphasphate synthetase (Vtc2, Vtc3, Vtc4)
----	Mannosyltransferase (Pmt2; Pmt3; Pmt5)
----	Lipid phosphate phosphatese (Dpp1; Lpp1)
----	Phosphorylcholine transferase (Muq1; Pct1)
----	Sterol-ester synthetase (Are1; Are2)
Clusters of unknown proteins	
(Ydr126w; Erf2; Ydr459c; Ynl326c; Yol003c)	(Ydr126w; Erf2; Ydr459c; Ynl326c; Yol003c)
(Psr2; Ypl063w; Psr1; Nem1)	(Psr2; Ypl063w; Psr1; Nem1)
(Ygl014w; Mpt5; Yll013c)	(Ygl014w; Mpt5; Yll013c)
(Tom1; Rsp5; Hul4)	(Tom1; Rsp5; Hul4)
(Yil088c; Ybl089w; Ynl101w)	(Yil088c; Ybl089w; Ynl101w)
(Ypl249c; Msb4)	(Ypl249c; Msb4)
(Msi1; Rsa2)	(Msi1; Rsa2)
(Imp4; Rpf1)	(Imp4; Rpf1)
(Mrd1; Ynl110c)	(Mrd1; Ynl110c)
(Gdi1)	(Gdi1)
(Ydl060w; Bms1)	(Ydl060w; Bms1)
(Yfr021w; Ypl100w; Ygr223c)	(Yfr021w, Ypl100w; Ygr223c)
(Ykl121w; Ymr102c)	(Ykl121w; Ymr102c)
(Ssf1; Ssf2)	(Ssf1; Ssf2)
(Vps24; Fit1; Ykl002w)	----
(Ylr328w; Ygr010w)	----
----	(Ymr285c; Yol042w; Yor353c)
----	(Yml072c; Ynl087w; Yor086c)
Unique unknown proteins	
Sas3; Ydl216c; Sfb3; Las21; Ynr048w; Rlp7; Yhr122w; Ylr409c; Nip7; Ypr031w; Hrt1; Nmd3; Yer082c; Ykl099c; Pri2; Yth1; Yer126c; Nud1; Ypl247c; Ypl236c; Plp1; Ufd1; Ygr145w; Crm1; Ptk1; Ykt6; Ydr266c; Ydr339c; Ydr411c; Ydr083w; Yil005w; Yhr186c; Ydr365c; Abp140; Bph1; Ymr068w; Yjl109c; Sgt1; Vip1; Ybr228w; Yol093w; Enp1; Yil113w; Ykl100c; Ygl047w; Ykl056c; Pac10; Ent3; Hym1; Yor289w	Sas3; Ydl216c; Sfb3; Las21; Ynr048w; Rlp7; Yhr122w; Ylr409c; Nip7; Ypr031w; Hrt1; Nmd3; Yer082c; Ykl099c; Pri2; Yth1; Yer126c; Nud1; Ypl247c; Ypl236c; Plp1; Ufd1; Ygr145w; Crm1; Ptk1; Ykt6; Ydr266c; Ydr339c; Ydr411c; Ydr083w; Yil005w; Yhr186c; Ydr365c; Abp140; Bph1; Ymr068w; Yjl109c; Sgt1;
----	Ama1; Caf40; Cdc20; Cdc33; Dcp2; Dst1; Erv1; Fkh2; Frq1; Grd19; Hut1; Ybl071w-a; Lem3; Lsm1; Per1; Sac6; Smp2; Yip2; Yaf9; Ybr042c; Ycr063w; Ycs4; Ydr018c; Ydr128w; Ydr255c;Ygl131c; Ygr054w; Yjr070c;Yju2; Ykl059c; Ykl146w; Ylr002c; Ylr323c; Ymr073c; Ymr093w;Ymr288w; Yor091w; Yor175c;Ypl110c; Ypr037c; Tos8; Yng2; Ypr133c; Zap1; Jsn1

B

Fig. 1. Comparison of proteins from the 347-ESP set of *G. lamblia* and 401-ESP set of *E. cuniculi*. The unique identifier symbols for the proteins are from the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces) and are shown in parentheses in different colors. The **black** color of a protein identifier from the ESP set of *G. lamblia* shows that this protein is also present among 401 ESPs of *E. cuniculi* or has a ''putative homolog'' in the *E. cuniculi* proteome revealed by PSI-BLAST. The **black** color of a protein identifier from the ESP set of *E. cuniculi* means that this protein is also present among the 347-ESP set of *G. lamblia*. The **green** color of a protein identifier shows that this protein is missing from the 401-ESP set of *E. cuniculi*, however, its ''putative homolog,'' detected by psi-blast, is present in the *E. cuniculi* proteome. The **red** color of a protein identifier from the 347-ESP set of *G. lamblia* shows that this protein does not have any sequential similarity to *E. cuniculi* proteins. The **red** color of a protein identifier from the 401-ESP set of *E. cuniculi* means that this protein does not show sequence similarity above 55 bits (BLAST 2.0 score) while screening the *G. lamblia* contig database. The sum of all black and red identifiers for *G. lamblia* is equal to 347, and that for *E. cuniculi* is equal to 401.

The ESPs most closely associated with the plasma membrane are cytoskeletal proteins, such as actin and associated proteins. *G. lamblia* and *E. cuniculi* have a full complement of actin and actin-related proteins (see Fig. 1). The three-dimensional structure of actin showed unexpected structural similarities to hexokinase and HSP 70 (Kabsch and Holmes 1995). These three proteins have a common nucleotide-binding motif, as they all bind ATP in the presence of calcium or magnesium ions. The "actin fold" was also found in the bacterial proteins FtsA and MreB (Kabsch and Holmes 1995). What is the evolutionary relationship between actin and bacterial proteins that have the "actin fold"? The role of MreB filaments in the cell shape of bacteria has suggested that the actin cytoskeleton evolved from MreB filaments (van den Ent et al. 2001). However, a careful sequence comparison of MreB and actin demonstrated that MreB and actin most likely had a common ancestor in the very distant past but that MreB was not an evolutionary precursor of actin (Doolittle and York 2002). Presumably, actin and all the related nucleotide-binding proteins evolved from a common ancestor which bound a nucleotide. This gene may have undergone gene doubling and various insertions, eventually evolving into MreB and, independently, actin proteins (Kabsch and Holmes 1995). Finally, in all such cases there is an alternative explanation of an independent origin of proteins with very little or without sequence similarity due to convergent evolution. Both *G. lamblia* and *E. cuniculi* have a full set of actin-related proteins (abbreviated *Arp* in Fig. 1). *Arp1* is involved in spindle alignment. *Arp2* and *Arp3* are involved in actin polymerization. *Arp4*, *Arp5*, and *Arp7* are involved in chromatin remodeling. *Arp6* is localized to heterochromatin (Goodson and Hawse 2002).

Both *G. lamblia* and *E. cuniculi* have tubulins among their ESPs. The three-dimensional structure of tubulin showed structural similarities to glyceraldehyde–3-phosphate dehydrogenase and bacterial protein FtsZ (Nogales et al. 1998). These three types of proteins have a common nucleotide-binding motif which is similar to the Rossmann fold, distinct from the actin fold (Doolittle 1995). As in the case with actin, we consider the existence of a deep ancestor to tubulin and FtsZ rather than bacterial FtsZ as the possible precursor to tubulin.

In close association with cytoskeleton composed of actin and tubulin, there are the motor proteins dynein, myosin, and kinesin. This association allows the eukaryotic cell to move, phagocytize, and divide. The control of this structure is modulated through interaction with calcium ions. Both *E. cuniculi* and *G. lamblia* have kinesins in their ESPs (Fig. 1). *E. cuniculi* has four myosins among its ESPs, whereas none exists in *G. lamblia*. It appears that myosin is a later addition to the evolving eukaryotic cell, as fungi branched much later in the phylogenetic tree than the diplomonads. The absence of dynein in the ESPs of both *E. cuniculi* and *G. lamblia* is due to its relationship to AAA proteins which are widely dispersed among the bacteria (Vale 2000). The motor proteins myosin and kinesin have a common structural motif with the G proteins (Kull et al. 1998). The similarity of these proteins is due to a common nucleotide-binding motif, which differs from that of actin or tubulin.

The fact that many ESPs have structural similarity to the bacterial proteome is not surprising because, likely, all cells share a common origin. However, the lack of sequential similarity of the discussed conservative structural proteins with their prokaryotic counterparts gave us a foundation to conjecture that the cytoskeleton including the motor proteins of the eukaryotic cell evolved out of a set of nucleotide-binding proteins from a primitive RNA-based cell and not out of a bacterial or an archaeal cell (Hartman and Fedorov 2002).

The Endoplasmic Reticulum and Protein Synthesis. The cytoskeleton of the eukaryotic cell is closely associated with the endoplasmic reticulum (ER). *E. cuniculi* has a significantly larger number of ER and Golgi ESPs than found in *G. lamblia* (16 and 6 proteins, respectively). This difference implies that the ER and Golgi of the parasitic fungus *E. cuniculi* are more complex than those of *G. lamblia*. The evolutionary study of the ER and Golgi complex is still in its infancy (Beznoussenko and Mironov 2002). We need a much larger sample of ER and Golgi proteins from single-celled protozoans, which are not parasitic, to infer the origin and evolution of these membranes.

Prominent among the ESPs of *E. cuniculi* and *G. lamblia* are ubiquitin, ubiquitin ligases, and proteases (Fig. 1). Some proteins of Archaea and Bacteria have a 3D structural fold similar to that of ubiquitin and are involved in the biosynthesis of sulfur-containing coenzymes (Wang et al. 2001). These prokaryotic proteins and ubiquitin might have diverged from a common ancestor, yet they have evolved independently in prokaryotes and eukaryotes. We found several proteasome-associated proteins from the 19S proteasome regulatory particle in the ESP sets of both *G. lamblia* and *E. cuniculi* (Fig. 1). There are more proteins associated with the proteasome regulatory particle of *E. cuniculi* than that of *G. lamblia* (eight and four proteins, respectively). This might point to an evolution from the simpler regulatory system of *G. lamblia* to that of the fungi. We hypothesize that the eukaryotic protein degradation complex composed of the ubiquitin, ubiquitin ligases, ubiquitin proteinases, and 19S regulatory proteasome

particle did not originate from prokaryotes. At the same time, the ancient proteasome likely came from an archaeal endosymbiont (Bouzat et al. 2000).

The Nucleus. Using the relatively strong threshold for protein similarity of 55 blast score bits ($\sim 10^{-6}$ e-values), we revealed four histones (H2A, H2B, H3, and H4) in the nuclear ESPs of *G. lamblia* and only two histones (H3 and H4) in *E. cuniculi*. However, if we search for sequence similarity using psi-blast, as opposed to blast, then we pick up highly diverged histones H2A and H2B of *E. cuniculi*. The eukaryotic histones share the same 3D structure with the archaeal histone-like proteins of the Euryarchaeota (methanogens, etc.) (Arents and Moudrianakis 1995). Unlike actin, tubulin, ubiquitin, and the GTP-binding proteins whose 3D counterparts are found throughout the Archaea and Bacteria, the histone fold is only found in the Euryarchaeota, and not in the Crenarchaeota or the Bacteria. Presently, the simplest explanation for the evolution of histones is that a histone-like protein came from an ancient archaeal cell and subsequently evolved into the full eukaryotic complement of histones. As for other proteins connected to the DNA structure, there are two topoisomerase I ESPs (Trf4 and Trf5) found in both *E. cuniculi* and *G. lamblia*. There are eight *E. cuniculi* ESPs involved with DNA repair that are not found in *G. lamblia*.

The ESPs found in the nucleus are dominated by proteins involved in the synthesis, processing, and transport of RNAs out of the nucleus into the cytoplasm. There are two ESPs (Rpc19 and Rpb8) associated with the RNA polymerases of *G. lamblia*. We could only detect by psi-blast one diverged ESP (Rpc19) in *E. cuniculi*. The ESPs representing transcription factors are much more diverse in *E. cuniculi* (29 proteins) than in *G. lamblia* (9 proteins). They share only four transcription factors in common. These deep distinctions show that transcriptional factors play a major role in the evolution of eukaryotes. However, when we compare groups of zinc finger ESPs between these two cells, we find more than 70% identity. The groups of nucleolar proteins ESPs associated with the synthesis and transport of ribosomal RNA are very similar in both *E. cuniculi* and *G. lamblia*. There are more ESP spliceosomal proteins in *E. cuniculi* (10) than in *G. lamblia* (4). The ESPs involved in the transport proteins and the nuclear pore proteins are found in equal abundance in *E. cuniculi* and *G. lamblia*.

The Cell Cycle and Cellular Coordination. The regulators of the eukaryotic cell cycle (cyclins, serine/threonine kinases, and ubiquitin proteins) are present among ESPs (Fig. 1). When we compare the cyclin ESPs of *E. cuniculi* with those of *G. lamblia*, we see complete agreement between these two cells. However, when we compare the ESPs representing kinases and phosphatases of these two groups, we get a different picture. There is a significantly higher number of kinases and phosphatases in *G. lamblia* than in *E. cuniculi* (24 versus 10 proteins). This must be due to the difference in intracellular versus extracellular lifestyles.

The GTP-binding proteins *ras* (plasma membrane), *rho* (cytoskeleton), *rab* (endoplasmic reticulum), *arf* (Golgi), and *ran* (nucleus–cytoplasm) are very prominent among the ESPs of *E. cuniculi* and *Giardia*. We hypothesize that these proteins *ras*, *rho*, *rab*, and *arf* have evolved from the membrane-protein synthesizing machinery and the cytoskeleton of the host cell. They are now localized on the cytoskeleton and membranes (the plasma, endoplasmic reticulum, and Golgi) of the eukaryotic cell (Hartman and Fedorov 2002).

Enzymes. There are 10 enzymes found in *E. cuniculi* that are not found in the ESPs of *G. lamblia*. This may be due to the fact that *E. cuniculi*, as a fungus, belongs to the crown of the eukaryotic tree (Katinka et al. 2001), while *G. lamblia*, as a diplomonad, branched earlier. Finally, there are 108 and 91 of ESPs of *E. cuniculi* and *G. lamblia*, respectively, that have no assigned function. Among them, 57 proteins were found in both *E. cuniculi* and *G. lamblia*.

Conclusions

We found 401 ESPs in *E. cuniculi*. These ESPs represent 214 unique protein groups. Comparison of the ESPs of *G. lamblia* with the proteins of *E. cuniculi* demonstrated that 85% of the 347 ESPs of *G. lamblia* have sequence similarity to ESPs found in *E. cuniculi*. Proteins from the ESP sets fall into two main categories: (1) proteins related to the observable structures in the cytoplasm of the eukaryotic cell such as the plasma membrane (clathrin), the cytoskeleton (actin and arps, tubulin, and associated kinesins), the endoplasmic reticulum, the Golgi, and the nucleus (histones, etc.), and (2) proteins involved in the coordination of the eukaryotic cell such as GTP-binding proteins (i.e., *ras*, *rho*, *rab*, *arf*, and *ran*), calmodulin, ubiquitin, cyclin, serine–threonine kinases, and phosphatases, 14–3–3 proteins, and enzymes modulating PIP (phospatidyl inositol phosphates). These cellular structures and their defining proteins are unique to the eukaryotic cell and so are the control proteins.

A significant number of ESP sets prominent in the cytoplasm, such as actin, tubulin, kinesins, ubiquitin, and GTP-binding proteins, all have counterparts in

prokaryotes with similar 3D folds but no significant sequential similarity. These proteins are also found among the most conserved proteins in the eukaryotic cell (Copley et al. 1999). The best that can be inferred from these facts is that prokaryotic proteins and their ESP counterparts had a common ancestor, and that ancestor was a more primitive RNA-based cell (Doolittle and York 2002; Nogales et al. 1998).

The large number and diversity of ESPs point to a very ancient origin of the eukaryotic cell. These proteins are congruent with the recent hypothesis which places eukaryotes at the root of the universal tree of life (Gribaldo and Phillippe 2002; Penny and Poole 1999). The data presented here are also consistent with our previous hypothesis that the eukaryotic cell had an RNA-based cell as one of its ancestors (chronocyte) and that the nucleus was formed by the engulfing of prokaryotic cells that became the nuclear endosymbiont (Hartman and Fedorov 2002).

Acknowledgments. This research was supported in part by NSF Grandt DBI-0205512. Support for this work was provided by the Medical College of Ohio Foundation. We would also like to thank Ms. Lisa Johnston for her excellent secretarial assistance.

References

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

Arents G, Moudrianakis EN (1995) The histone fold: A ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proc Natl Acad Sci USA 92:11170–11174

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–976

Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, Wheeler DL (1999) GenBank. Nucleic Acids Res 27:12–17

Beznoussenko GV, Mironov AA (2002) Models of intracellular transport and evolution of the Golgi complex. Anat Rec 268:226–238

Bouzat JL, McNeil LK, Robertson HM, Solter LF, Nixon JE, Beever JE, Gaskins HR, Olsen G, Subramaniam S, Sogin ML, Lewin HA (2000) Phylogenomic analysis of the alpha proteasome gene family from early-diverging eukaryotes. J Mol Evol 51:532–543

Copley RR, Schultz J, Ponting CP, Bork P (1999) Protein families in multicellular organisms. Curr Opin Struct Biol 9:408–415

Doolittle RF (1995) The origins and evolution of eukaryotic proteins. Philos Trans R Soc Lond B Biol Sci 349:235–240

Doolittle RF, York AL (2002) Bacterial actins? An evolutionary perspective. Bioessays 24:293–296

Fraser CM, Gocayne JD, White O, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

Goodson HV, Hawse WF (2002) Molecular evolution of the actin family. J Cell Sci 115:2619–2622

Gribaldo S, Phillippe H (2002) Ancient phylogenetic relationships. Theor Population Biol 61:391–408

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

Higgins DG, Sharp PM (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. Gene 73:237–244

Kabsch W, Holmes KC (1995) The actin fold. FASEB J 9:167–174

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:450–453

Kull FJ, Vale RD, Fletterick RJ (1998) The case for a common ancestor: Kinesin and myosin motor proteins and G proteins. J Muscle Res Cell Motil 19:877–886

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268–10273

Nogales E, Downing KH, Amos LA, Lowe J (1998) Tubulin and FtsZ form a distinct family of GTPases. Nat Struct Biol 5:451–458

Penny D, Poole A (1999) The nature of the last universal common ancestor. Curr Opin Genet Dev 9:672–677

Razin S (1997) The minimal cellular genome of mycoplasma. Indian J Biochem Biophys 34:124–130

Vale RD (2000) AAA proteins: Lords of the ring. J Cell Biol 150:F13–F19

Van den Ent F, Amos L, Lowe J (2001) Prokaryotic origin of the actin cytoskeleton. Nature 413:39–44

Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR (1987) Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326:411–414

Wang C, Xi J, Begley TP, Nicholson LK (2001) Solution structure of ThiS and implications for the evolutionary roots of ubiquitin. Nat Struct Biol 8:47–51