OPEN ACCESS
Freely Available Online

openaccessPub

PROTEOMICS AND GENOMICS RESEARCH
ISSN NO: 2326-0793

Research Article

DOI : 10.14302/issn.2326-0793.jpgr-12-101

Eukaryotic Signature Proteins

Jian Han[1]* and Lesley J. Collins[2]
[1] Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand
[2] Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Abstract

Eukaryotic Signature Proteins (ESPs) are proteins that delineate the eukaryotes from the archaea and bacteria. They have no recognisable homologues in any prokaryotic genome, but their homologues are present in all main branches of eukaryotes. ESPs are thus likely to have descended from ancient proteins that have existed since the first eukaryotic cell. The last dataset of ESPs was calculated more than a decade ago, thus with advances in technology and the rapid completion of many evolutionary important genomes, this dataset required recalculating. This study recalculated the Giardia lamblia ESP dataset and provides a procedure to calculate signature proteins beginning with any species. The G. lamblia ESP dataset contained a range of proteins including many associated with the membrane, cytoskeleton, nucleus and protein synthesis. ESP datasets have implications on current models of eukaryotic evolution, having high importance in phylogenetic analysis due to ESPs' consistency and conservation in all eukaryotic species.

Corresponding author: Jian Han , Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand, Email: jian272@hotmail.com*

Running title: ESP's of Giardia
Key Words: Eukaryotic signature protein, Giardia, eukaryotes
Received : Jul 30, 2012; **Accepted** : Sep 15, 2012; **Published** : Dec 03, 2012;

Introduction

Eukaryotic Signature Proteins

Eukaryotes are remarkably different from prokaryotes (archaea and bacteria), in terms of cellular structure, genetic content and proteome. Finding a set of proteins which can delineate eukaryotes from prokaryotes can be crucial to understanding the major differences in metabolism between the two groups. Eukaryotic signature proteins (ESPs) are such proteins, since by definition they have no recognisable homologues in prokaryotic genomes, but their homologues are present in all the main branches of eukaryotes. They are involved in most core functions of a eukaryote and provide landmarks to track the origin and evolution of eukaryotic genomes [1].

The approach of searching for signature proteins was first used by Graham *et al.* in searching for archaeal signature proteins [2]. Their study in 1999 found 351 clusters of proteins found only in Euryarchaeota species. Hartman and Fedorov [3] then focused on eukaryotes collecting ESPs by searching yeast protein homologues against three kingdoms of life (archaea, bacteria and eukaryotes). Their analysis procedure began with the *Saccharomyces cerevisiae* genome removing proteins without homologues in *Caenorhabditis elegans, Drosophila melanogaster* and *Arabidopsis thaliana.* After that, proteins that have homologues in any of the 44 bacterial and archaeal species (the only available complete bacterial and archaeal genomes at the time) were removed. Lastly they removed proteins without homologues in *Giardia lamblia* (from here on *Giardia*) [3]. By using this procedure Hartman and Fedorov were left with 347 yeast proteins, and they named this dataset the Eukaryotic signature proteins of *Giardia*. The main point of Hartman's paper was to form a novel

(Continued on page 3)

hypothesis on the formation of eukaryotic cells. From the finding of these 347 ESPs, Hartman argued the presence of proteins without any bacterial and archaeal homologues means that they must have come from a cell of a distinct lineage (termed a “chronocyte”) rather than any symbiotic event between an archaeon and bacteria as previously hypothesised [4].

Subsequently the same Hartman’s research group collected ESPs for the microsporidium *Encephalitozoon cuniculi* [5], the organism with the smallest sequenced eukaryotic genome. The procedure was same as that for *Giardia*’s, except for the last step where they compared with the genome of *E. cuniculi* instead of *Giardia*. They found 401 ESPs for *E. cuniculi*, which consisted of 238 ESPs in common with *Giardia* ESPs. This high level of similarity has indicated that even a minimal eukaryotic cell still preserved most of the ESPs, which agrees with their earlier hypothesis that these ESPs must come from a cell of distinct lineage.

In our study, an updated ESP dataset was recalculated using the wealth of genomes now available. The method is similar to that of Hartman’s [3], but instead of *S. cerevisiae* as the starting point, we began with *Giardia* [6]. The resulting set of proteins we consider more likely to represent ancestral forms, because *Giardia* is a basal eukaryote which diverged during the early days of eukaryotic evolution [7], and thus a representative of an early (perhaps the earliest) eukaryotic lineage [8]. *Giardia* has also undergone genomic reduction [8] as is common in parasites and eukaryotes with small genomes such as yeasts. Because the original Hartman dataset originated from a different eukaryote (*S. cerevisiae*), we expect that each ESP dataset will be slightly different depending on which eukaryote was used as the starting point. For this reason, as well as the recalculated *Giardia* ESP dataset, we present a protocol to calculate ESPs with different eukaryotes as the starting point so that datasets can be compared to understand of ancestral protein evolution.

An ESP dataset is a set of functionality important proteins, since all eukaryotes maintain them, and it is also a set of evolutionarily important proteins, because they are not present in any prokaryotes. Thus, ESP datasets, once they are recalculated to include as many different eukaryotes as possible, may hold the key to unravel the many different theories of how eukaryotes and prokaryotes evolved. In addition, because of their universal presence, we find that they are potentially good candidates for performing phylogenetic analysis (publication in preparation).

Materials and Methods

Selection of Species for Analysis

Ideally, the more species involved in the eukaryote-wide search, the more robust the ESP dataset. However, the time the analysis takes also increases as the number of species increase. Therefore, the number of species used has to be compromised to an extent depending on computing resources. Additionally ESP results can be biased to some extent due to species selection. To minimise this effect selected species should cover as wide a range of organisms as possible. For our ESP datasets species which would best represent major branches of each of the three domains were chosen for analysis, including 28 bacteria, 12 archaea, and 17 eukaryotes. Ideally, we would prefer to omit all parasites with a potentially reduced genome, but this is not possible until more completely sequenced genomes are available, and *Giardia*, as a parasite, was included in the analysis to represent the eukaryotic supergroup Excavata. A list of all species used in the study can be found in the supplementary material.

ESP Calculations

The Basic Local Alignment Search Tool (BLAST) [9] was used for comparing homologous proteins between species. In our comparisons, BLAST hits with a bit-score ≥ 55 were considered as “homologues”. This cut-off is same as Hartman’s[3] although we did test to see if other parameters would be more suited (data not shown). ESP datasets were then calculated under the following procedure: The analysis began with all annotated proteins of the chosen starting organism *Giardia lamblia* Assemblage A, downloaded from GiardiaDB version 1.3 (www.giardiadb.org). The *Giardia* genome size is ~12 megabases (Mb) containing ~5000 protein coding genes [10]. *Giardia* proteins that had homologues in any of the 28 bacterial and 12 archaeal species were then discarded; then proteins that did not have homologues in any of the 17 eukaryotic species were removed. The remaining proteins are termed ESPs. Data was managed using MySQL and Perl scripts to facilitate loading and updating of genomic information.

This simple protocol was very effective in calculating ESPs beginning with any eukaryote, or alternatively to include new and updated genome or proteome information. The Perl scripts written for this procedure are available from the authors upon request.

Results

The *Giardia lamblia* ESP dataset

Our ESP dataset contained 274 *Giardia lamblia* eukaryotic signature proteins (ESPs) these comprising of 267 distinctive proteins. Although the *Giardia* genome is annotated, a number of proteins are still to this day,

(Continued on page 5)

Table 1. Protein functions of *Giardia* ESPs. The 274 ESPs were divided into many categories according to their putative functions. Distinct ESP numbers (i.e. counting multiple copies as one) are shown in brackets.

Protein category	Sub category	# of proteins (distinctive copies)	
Cytoskeleton	actins	4	37(34)
	microtubule related	1	
	tubulins	8(5)	
	kinesins	24	
Membrane	cell adhesion	2	34
	clathrin related	11	
	endocytosis	1	
	ER and Golgi	9	
	lipid attachments	4	
	vacuole	7	
Nucleus	DNA polymerase	1	45(41)
	histones	11(7)	
	histone-associated	4	
	LIM related	4	
	ribonucleoproteins	2	
	RNA enzymes	9	
	topoisomerase	1	
	transcriptional factors	5	
	transcriptional transactivators	2	
	Zinc fingers	6	
Protein synthesis and breakdown	ribosome biogenesis proteins	4	17
	large ribosomal proteins	4	
	small ribosomal proteins	3	
	proteasome associated	2	
	translation factors	4	
Signalling system	14-3-3 protein	1	97
	calmodulins	5	
	cell cycle related	9	
	GTP-binding proteins	20	
	kinases and phosphatases	35	
	Phosphatidylinositol proteins	7	
	ubiquitins	2	
	ubiquitin conjugation enzymes	15	
	ubiquitin proteases	5	
Others	others	33	33
Hypothetical proteins	hypothetical proteins	10	10

designated as hypothetical proteins. The 274 ESPs were divided into seven groups according to predicted conserved function based on their description and homology to *S. cerevisiae* proteins. The seven protein groups are:

1. Proteins related to the plasma membrane and endocytosis (34 proteins).

2. Proteins associated with the cytoskeleton (39 proteins).

3. Proteins are involved in the signalling system (97 proteins).

4. Proteins in the nucleus (45 proteins).

5. Proteins involved with protein synthesis and breakdown (15 proteins).

6. Proteins with unknown function (34 proteins).

7. Hypothetical proteins (10 proteins).

Table 1 lists all ESPs by these categories. Some ESPs have multiple gene copies, and thus numbers of distinctive ESPs (i.e. not including the repeated ones) are also included in brackets. The identifiers of all ESPs in these groups and a FASTA file of all ESPs are given in the supplementary material.

The *Giardia* ESP dataset contains some protein families where multiple proteins are descended from a common ancestor, and hence each protein in these families has a high sequence similarity to the others. An example from our ESP dataset is the histone family, consisting of H2A, H2B, H3 and H4, but not H1 because H1 is not found in *Giardia*. Another example is the tubulin family where the alpha, beta, gamma, delta and epsilon tubulin are all found as ESPs. By starting with *Giardia* proteins which have only a few protein families we did not have to do any manual curation to resolve different family members. However, if a different eukaryote was used as the starting point then there may be more issues in resolving individual members of protein families.

There were 39 ESPs designated to the cytoskeleton, including a number of actins (proteins that make microfilaments and thin filaments), tubulins (proteins that make microtubule), kinesins (protein motors) and a microtubule-binding protein. The cytoskeleton is thought to be a eukaryotic cellular signature structure (CSS) that defines eukaryotes, but recently a prokaryotic cytoskeleton has been identified [11,12]. It has also been reported that the eukaryotic actin and tubulin genes are weakly homologous to FtsA and FtsZ, both of which are part of the bacterial cell division machinery [13]. The 3-dimensional structure of FtsA and FtsZ are remarkably similar to that of actin and tubulin, respectively, but their primary structures (i.e. sequence) have little similarity [14,15]. The actin and tubulin families remain in the ESP dataset since the similarity to prokaryotic proteins do not reach the thresholdevolution.

The majority of membrane ESPs appear to be involved in the transportation of macromolecules, including a large number of clathrin (involved in forming coated vesicles), endoplasmic reticulum (ER) and Golgi apparatus related proteins, vacuolar proteins, proteins involved in attachment, and one protein involved in endocytosis. It is hoped that closer study of these ESPs will aid in understanding if the eukaryotic cell arose by engulfing other cells, (i.e. that progenitors of these ancient proteins might once have functioned to enable the proposed "raptor" cell [1] or chronocyte [3] to engulf ancestral bacterial and archaeal cells).

ESPs associated with the nucleus included histones, RNA associated enzymes and proteins from the DNA replicating machinery. Histones, as mentioned before, are responsible for packing DNA into chromatin structures. Prokaryotes and archaea do not have complicated DNA packaging systems, thus histones H2A, H2B, H3 and H4 are expected to be ESPs. H1, the linker of chromatin, is an exception because it is less conserved and is absent in some eukaryotes such as *Giardia*. Although archaeal (euryarchaeotes) genomes also contain ancient histone homologues [16], the similarity is more at a structural level rather than the sequence level. RNA-based ESPs include enzymes involved in RNA editing, which has been proposed as an ancient mechanism [17]. The presence of ESPs could suggest RNA editing in eukaryotes existed since the divergence of eukaryotes. Finally, as expected since the DNA replication process in eukaryotes is much different from that of the prokaryotes, with different polymerases and different transcription factors (including some proteins annotated as "zinc finger proteins") utilised, DNA replication proteins were well represented in the ESP dataset.

Several ESPs were involved in protein synthesis as expected since this mechanism in eukaryotes is well known to be different from the prokaryotic mechanisms. The eukaryotic 80S ribosome is also different to the prokaryotic 70S ribosome. Several ribosomal proteins, translational factors fulfilled the criteria of ESP indicating clearly that although *Giardia* has a smaller ribosome than most other eukaryotes, it is eukaryotic rather than prokaryotic. There are also two proteasome-related ESPs indicating perhaps that the protein degradation mechanism is universal to all eukaryotes.

Signalling-system ESPs contain many kinases and phosphatases. These are enzymes performing a variety

(Continued on page 6)

of functions by adding and removing phosphate groups to a molecule (such as proteins or ATP). Phosphatidylinositol kinases and phosphatases are involved in cellular functions such as cell growth, proliferation, differentiation, motility, survival and intracellular trafficking. GTP-binding proteins are prominent (they function as "molecular switches"), and give more sophisticated regulation of enzymes, ion channels, transporters, controlling numerous cell activity such as transcription, motility, contractility, and secretion [18]. Ubiquitin-related proteins are very abundant and are involved in directing protein degradation. Five calmodulins were found as ESPs, indicating that regulation by means of calcium-binding is a distinct mechanism in eukaryotes.

There are nine proteins associated directly with the cell cycle, such as cyclins and cyclin-dependent kinases (CDKs). Cyclins and cyclin-dependent kinases form complexes which upon activation will dictate which phase the cell will go through. This is a unique scenario in eukaryotes because prokaryotes do not possess nuclei, and their cell division is relatively simple. Interestingly a protein annotated "notchless" was found as an ESP. The homology between *Giardia* notchless and other eukaryotic notchless is very high (e.g. a bit-score of 296 to the *D. melanogaster* notchless protein), which suggested high confidence in the annotation. Notchless is a regulator of the notch pathway, which plays a central role in the control of cell fate decisions in a wide variety of cell lineages during invertebrate and vertebrate development [19]. It is unknown why homologues of the gene for this protein would be present in *Giardia*, *Phytophthora infestans* and *Dictyostelium discoideum*, the three single celled eukaryotes used in this study.

ESPs which fell into the category of "others" are those which have not been annotated very well in any species. Their annotations typically only have suggestion to their sequence or predicted 3D structure, for example the "Glycine-rich protein" or "WD-40 repeat protein". Some proteins also have suggested functions such as "ATPase". Lastly, ESPs in the "hypothetical protein" category all have "Hypothetical proteins" as their annotation and were not able to be resolved further. It is hoped that future proteomic and genomic studies will increase the specificity of the annotation of these proteins and enable us to categorise them more precisely.

Comparison with Hartman's Dataset

Hartman and Fedorov initially calculated *Giardia* 347 ESPs in 2002 using *S. cerevisiae* as the starting eukaryote. Comparisons between the ESP dataset obtained in this study and Hartman's dataset showed that from the new dataset of 274 ESPs, 203 proteins had homologues in Hartman's dataset, and 71 did not. A reverse BLAST search was performed (i.e. the use of Hartman's ESPs dataset as the input searching against the new ESP dataset), and out of the 347 Hartman's ESPs, 237 had homologues in our ESP dataset, and 110 did not.

Overall we can conclude that the datasets are in fact very similar, mainly because the principle was the same in both sets, i.e. to find proteins conserved across eukaryotes and not found in prokaryotes. The slight variation seen between the datasets can be attributed to the way the ESPs are calculated, and the addition of protein information from newly sequenced genomes. Hartman started with the *S. cerevisiae* proteome because the *Giardia* genome was hardly annotated at that time, and performed BLAST searches with the yeast proteins against the only 44 prokaryotes' genomes available at the time, then only four eukaryotes and lastly *Giardia*. We used a more straightforward approach and started BLAST searches directly with *Giardia* proteins, giving the benefit of obtaining a set of ESPs in most likely their ancient form.

Conclusion

An ESP dataset for *Giardia* has been recalculated taking advantage of recent genome sequencing and increased availability of proteomic information. We feel the new dataset is more precise than that of Hartman's. However, it is very difficult to obtain an exact list of ESPs due to factors such as distant, unrecognisable homologues. In addition, there may still be a few false positives due to the lack of any completed genomes from some branches of prokaryotes and eukaryotes (namely species from the bacterial phylum Caldiserica or species from eukaryotic supergroup Rhizaria). It should be made clear that ESPs are not a complete set of ancestral eukaryotic proteins [20,21] because some eukaryotes may have lost some ancestral proteins or ancestral proteins may share homology with those present in prokaryotes (if only in domain structure). One example is Dicer which is considered an ancestral eukaryotic enzyme involved in RNAi [22], but some lineages such as yeast have secondary loss of this protein [23]. A future method that is able to capture these proteins such as Dicer could be useful to obtain more than just ESPs as signatures of eukaryotes, and expand our datasets to ancestral eukaryotic proteins.

The protocol described here for ESP calculation is flexible and permits future recalculations to include newly sequenced genomes as well as updated genomic information. Increased computer power means that more species can now be more readily included. Perl scripts and command line scripts have already been

(Continued on page 7)

prepared to make the process of ESP re-calculation straightforward, and one only has to run these scripts to have an updated dataset. If ESPs are re-calculated in the future, more organisms can be included in the calculations especially when more complete genomes become available. The continuing increase in computing power will thus permit even more robust ESP calculations.

ESP datasets are expected to be useful for phylogenetic analysis. Previously, molecular based phylogenetic studies between distantly related species (such as between different supergroups of eukaryotes) was done by using 18S rRNAs, or based on a single gene when these happened to be sequenced. This approach can tend to give misleading results if the gene in question has undergone rates of change different from what is 'typical' for that species, or if there if there has been more than one change per site. A wider variety and larger quantities of molecular data is needed to accurately build the eukaryotic trees [24]. ESPs are a set of proteins conserved in all eukaryotes, so this potentially makes them ideal candidates for phylogenetic studies, by permitting a large number of proteins to be used to build phylogenetic relationships with no missing data. We would also expect the ancient proteins to have a slow and constant evolutionary rate (to keep them conserved), which also make them ideal for studying phylogenetic relationships between distant organisms. Hence the possibility of the use of ESPs in phylogenetic studies is currently under investigation and a manuscript discussing this topic is in preparation.

Dataset availability

A list of ESPs has been included in the supplementary material. The dataset will also be made available to www.giardiadb.org in the near future.

Acknowledgments

This work was funded by Health Research Council of New Zealand (HRC) - Emerging Researcher Grant 07/168. Eukaryotic Signature Proteins - Guides to Modern Eukaryotic Parasites.

Conflict of Interest

The authors declare that there is no conflict of interest surrounding this dataset and associated manuscript.

Affiliations

Abbreviations

ESP, Eukaryotic Signature Protein.

BLAST, Basic Local Alignment Search Tool.

CSS, Cellular Signature Structure.

ER, Endoplasmic Reticulum.

Supplemental data

List of supplementary files

S1. List of 274 *Giardia* ESPs.

S2. FASTA file of 274 *Giardia* ESPs.

S3. List of selected species used as comparison.

References

1. Kurland CG, Collins LJ, Penny D (2006) Genomics and the irreducible nature of eukaryote cells. Science 312: 1011-1014.

2. Graham DE, Overbeek R, Olsen GJ, Woese CR (2000) An archaeal genomic signature. Proceedings of the National Academy of Sciences of the United States of America 97: 3304-3308.

3. Hartman H, Fedorov A (2002) The origin of the eukaryotic cell: A genomic investigation. Proceedings of the National Academy of Sciences of the United States of America 99: 1420-1425.

4. Lake JA, Rivera MC (1994) Was the nucleus the 1st endosymbiont. Proceedings of the National Academy of Sciences of the United States of America 91: 2880-2881.

5. Fedorov A, Hartman H (2004) What does the microsporidian E-cuniculi tell us about the origin of the eukaryotic cell? Journal of Molecular Evolution 59: 695-702.

6. Brown DM, Upcroft JA, Edwards MR, Upcroft P (1998) Anaerobic bacterial metabolism in the ancient eukaryote Giardia duodenalis. International Journal for Parasitology 28: 149-164.

7. Vanacova S, Liston DR, Tachezy J, Johnson PJ (2003) Molecular biology of the amitochondriate parasites, Giardia intestinalis, Entamoeba histolytica and Trichomonas vaginalis. International Journal for Parasitology 33: 235-255.

8. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, et al. (2007) Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. Science 317: 1921-1926.

9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

10. Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, et al. (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. Nucleic Acids Research 37: 526-530.

(Continued on page 8)

11. Shih Y-L, Rothfield L (2006) The bacterial cytoskeleton. Microbiology and Molecular Biology Reviews 70: 729-754.

12. Watters C (2006) The bacterial cytoskeleton. CBE life sciences education 5: 306-310.

13. Jimenez M, Martos A, Vicente M, Rivas G (2011) Reconstitution and Organization of Escherichia coli Proto-ring Elements (FtsZ and FtsA) inside Giant Unilamellar Vesicles Obtained from Bacterial Inner Membranes. Journal of Biological Chemistry 286: 11236 -11241.

14. van den Ent F, Lowe J (2000) Crystal structure of the cell division protein FtsA from Thermotoga maritima. Embo Journal 19: 5300-5307.

15. Desai A, Mitchison TJ (1998) Tubulin and FtsZ structures: functional and therapeutic implications. Bioessays 20: 523-527.

16. Spitalny P, Thomm M (2008) A polymerase III-like reinitiation mechanism is operating in regulation of histone expression in archaea. Molecular Microbiology 67: 958-970.

17. Collins LJ, Chen XS (2009) Ancestral RNA The RNA biology of the eukaryotic ancestor. Rna Biology 6: 495-502.

18. Neves SR, Ram PT, Iyengar R (2002) G protein pathways. Science 296: 1636-1639.

19. Royet J, Bouwmeester T, Cohen SM (1998) Notchless encodes a novel WD40-repeat-containing protein that modulates Notch signaling activity. Embo Journal 17: 7351-7360.

20. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, et al. (2010) The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. Cell 140: 631-642.

21. Koonin EV (2010) The Incredible Expanding Ancestor of Eukaryotes. Cell 140: 606-608.

22. Chen XW, Collins LJ, Biggs PJ, Penny D (2009) High Throughput Genome-Wide Survey of Small RNAs from the Parasitic Protists Giardia intestinalis and Trichomonas vaginalis. Genome Biology and Evolution 1: 165-175.

23. Drinnenberg IA, Fink GR, Bartel DP (2011) Compatibility with Killer Explains the Rise of RNAi-Deficient Fungi. Science 333: 1592-1592.

24. Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, et al. (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proceedings of the National Academy of Sciences of the United States of America 106: 3859-3864.