Origin and evolution of nuclear envelope proteome – A comparative genomics approach

A Thesis

submitted to the University of Hyderabad for the award of PhD degree

in the Department of Biochemistry, School of Life Sciences

By

Hita Sony Garapati

12LBPH09

Department of Biochemistry

School of Life Sciences

University of Hyderabad

Hyderabad - 500046

Telangana, India

April 2019

University of Hyderabad

Hyderabad - 500 046, India

CERTIFICATE

This is to certify that this thesis entitled **"Origin and evolution of nuclear envelope proteome – A comparative genomics approach"** submitted by **Ms. Hita Sony Garapati** bearing registration number **12LBPH09** in partial fulfillment of the requirements for award of Doctor of Philosophy in the Department of Biochemistry, School of Life Sciences, is a bonafide work carried out by her under my supervision and guidance.

This thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for award of any degree or diploma.

Parts of this thesis have been:

A. Published in the following publications:

1. **Garapati HS, Mishra K: Comparative genomics of nuclear envelope proteins. BMC genomics 2018, 19(1):823 (Chapter 3 & Chapter 4)**

B. Presented in the following conferences:

1. Oral presentation titled "Evolution of Nuclear Envelope Proteome" at the International Congress of Cell Biology (ICCB), held at Hyderabad, India, during 27th-31st January, 2018
2. Awarded prize for the poster presented in the 6th Meeting of Asian Forum of Chromosome and Chromatin Biology held from 3-5th March, 2017
3. Awarded prize for the poster presented in BioQuest 2015, held at University of Hyderabad from 23-24th September, 2015
4. Other conferences attended and presented poster: 9th International Conference on Yeast Biology 2015, International Conference on Genome Architecture and Cell Fate Regulation 2014, 4th Meeting of the Asian Forum of Chromosome and Chromatin Biology 2012.

Further, the student has passed the following courses towards fulfillment of coursework requirement for Ph.D.

Course code	Name	Credits	Pass/Fail
BC 801	Analytical Techniques	4	Pass
BC 802	Research ethics, Data analysis and Biostatistics	3	Pass
BC 803	Lab seminar and Record	5	Pass

Supervisor **Head, Dept. of Biochemistry** **Dean, School of Life Sciences**

University of Hyderabad

Hyderabad - 500 046, India

DECLARATION

I, Hita Sony Garapati, hereby declare that this thesis entitled **"Origin and evolution of nuclear envelope proteome – A comparative genomics approach"** submitted by me under the guidance and supervision of **Professor Krishnaveni Mishra**, is an original and independent research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Date **Signature of the student**

Signature of the Supervisor

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor **Prof. Krishnaveni Mishra** for her continuous support and valuable guidance throughout my research work. Her feedback and suggestions were very helpful in successful completion of the research work and in the writing of this thesis. She has always tried to bring the best out of me and I feel privileged to have worked under her guidance. I am also grateful to her for the financial support that she had extended during the last year of my PhD.

I would also like to express my heartfelt gratitude to **Dr. Rakesh K Mishra**, CCMB for making this PhD possible. I thank him for providing me an opportunity to work as a project assistant in his lab. He introduced me to Bioinformatics and let me work on several challenging projects, during which I learnt a lot. He had always encouraged and motivated me to pursue PhD and he remains the main reason behind this PhD.

I thank **Prof. Mrinal Kanti Bhattacharyya,** Head, Department of Biochemistry, and former Heads, **Prof. N. Siva Kumar** and **Prof. O.H. Shetty** for providing us the best facilities to carry out research.

I thank **Prof. S. Dayananda,** Dean, School of Life Sciences, and former Deans, **Prof K.V.A. Ramaiah**, **Prof. P. Reddanna**, **Prof. Aparna Dutta Gupta** and **Prof. R.P. Sharma** for providing the general facilities of the school.

I would like to thank my Doctoral Committee members **Dr. Rakesh K Mishra** and **Dr. Insaf Ahmed Qureshi** for their critical assessment of my research work and suggestions during the doctoral committee presentations.

I thank the **Bioinformatics Infrastructure Facility** at the School of Life Sciences, University of Hyderabad for the infrastructure.

I am also grateful to **S. B. Chary** and other non-teaching staff of the Department of Biochemistry, School of Life Sciences, for their kind assistance and cooperation.

I thank all my present and the previous lab mates: **Abdul**, **Pradeep**, **Imli**, **Neethu**, **Sangeetha**, **Gurranna**, **Kathir**, **Pallavi**, **Saketh, Preeti** and **Chhaya** for making this PhD experience memorable. Lab outings to mushroom rock and high rock were a lot of fun. I am also very grateful to all of them for being very helpful and supportive at all the times.

I thank **Gurranna** for his help with the cloning and microscopy related work. My special thanks to **Saketh** for all the discussions related to evolution, phylogenetic analysis and all the technical aspects in my work. He always listened to me and helped me gain clarity and understand things better. I would also like to thank **Akhil** for help with preparing some of the supplementary data. I also thank **Srinivas** and our lab attendant **Anil**.

I would like to thank all my friends in CCMB for helping me understand several concepts in biology. I have learnt a lot from all of them during my tenure in CCMB.

I thank the funding bodies: **CSIR**, **UGC**, **DST** and **DBT** for providing all the facilities to carry out the research work. I thank **CSIR** and **UoH** for providing me fellowship during my PhD.

I would like to thank my husband **Hemanth** and my mother-in-law **Vijaya Sree** for understanding me and being very supportive all the time.

Most importantly, I would like to thank my brother **Gopal** and my parents, **Venkata Ratnam** and **Pavani** for everything that they have done for me. I have come this far only because of their unconditional love and support.

Hita

Contents

Contents vi

List of Figures ix

List of Tables x

Chapter 1: Introduction

1.1	Origin of eukaryotic cell	01
1.1.1	Three-domain and eocyte hypothesis	02
1.1.2	Endosymbiotic model and the archaeal host	03
1.2	The Last Eukaryotic Common Ancestor (LECA)	06
1.3	Classification of eukaryotes	07
1.3.1	Opisthokonta	07
1.3.2	Amoebozoa	08
1.3.3	Excavata	08
1.3.4	SAR	09
1.3.5	Archaeplastida	10
1.3.6	Root of the eukaryotic tree	10
1.4	Nucleus	13
1.4.1	Origin of the nucleus	14
1.5	Nuclear envelope	18
1.6	Nuclear envelope composition across eukaryotes	20
1.7	Objectives of the study	23

Chapter 2: Methods

2.1	Bioinformatics methods	24
2.1.1	Identification of NE protein homologs across eukaryotes	24
2.1.2	Prokaryotic homologs of NE proteins	25

2.1.3 Motif analysis 25
2.1.4 Localization data of homologs 26
2.1.5 Phylogenetic analysis 26
2.2 Microscopy methods 26
2.2.1 Live-cell imaging 26
2.2.2 Spheroplast preparation 27
2.2.3 Immunofluorescence 27

Chapter 3: Data set preparation

3.1 Introduction 30
3.2 Results 30
3.3 Conclusions 36

Chapter 4: Nuclear envelope proteins across eukaryotes

4.1 Introduction 37
4.2 Results 38
4.3 Conclusions 56

Chapter 5: Phylogenetic analysis of nuclear envelope proteins

5.1 Introduction 57
5.2 Results 57
5.3 Conclusions 72

Chapter 6: Prokaryotic origins of nuclear envelope proteins

6.1 Introduction 73
6.2 Results 74
6.3 Conclusions 83

Chapter 7: Predicting localization of a protein using PPI data

7.1	Introduction	84
7.2	Results	85
7.3	Conclusions	109

Chapter 8: Discussion

8.1	Core/LECA NE proteins	113
8.2	Fungal specific NE proteins	115
8.3	Origins of nuclear envelope proteins	116
8.4	Caveats of the study	118
8.5	Future prospects	119

Appendix 121

Legends to Additional data (provided in CD) 144

References 146

List of Figures

Figure 1	Three-domain tree and eocyte tree	02
Figure 2	Endosymbiotic theory for the origin of eukaryotes	05
Figure 3	A schematic eukaryotic tree of life	12
Figure 4	Nucleus of eukaryotic cell	13
Figure 5	Models for the origin of nucleus	16
Figure 6	Proteins involved in chromatin organization and NE homeostasis across eukaryotes	41
Figure 7	Domain organization in Ebp2 and Rrs1 proteins	42
Figure 8	Domain organization in Heh2 and Src1 proteins	43
Figure 9	Domain organization in Hmg1 & Hmg2 proteins	45
Figure 10	Gene regulation and transport proteins across eukaryotes	47
Figure 11	Other NE proteins found across eukaryotes	50
Figure 12	Domain organization in the SUN domain proteins	52
Figure 13	Non-linearly conserved proteins	53
Figure 14	LECA nuclear envelope proteome	54
Figure 15	Motifs identified in Saccharomycetes specific proteins	55
Figure 16	Maximum likelihood tree of Ebp2 protein homologs	59
Figure 17	Maximum likelihood tree of Rrs1 protein homologs	61
Figure 18	Maximum likelihood tree of HMG-CoA reductases	63
Figure 19	Maximum likelihood tree of Pct1 protein homologs	65
Figure 20	Maximum likelihood tree of Cse1 protein homologs	67
Figure 21	Maximum likelihood tree of Ntf2 protein homologs	68
Figure 22	Maximum likelihood tree of Trm1 protein homologs	70
Figure 23	Maximum likelihood tree of Slp1 protein homologs	71
Figure 24	Prokaryotic homologs of nuclear envelope proteins	74
Figure 25	Maximum likelihood tree of HMG-CoA reductases across eukaryotes and prokaryotes	77
Figure 26	Maximum likelihood tree of tRNA methyltransferases across eukaryotes and prokaryotes	79
Figure 27	Maximum likelihood tree of Ntf2 protein homologs across eukaryotes and prokaryotes	81

Figure 28	Subcellular localization of selected proteins with predicted localization co-stained with ER marker	106
Figure 29	Subcellular localization of selected proteins with predicted localization co-stained with mitochondrial marker	107
Figure 30	Nuclear organization defects observed in *yhr140w*Δ and *yhl042w*Δ	108
Figure 31	Mitochondrial morphology defects observed in *ydr124w*Δ cells.	109
Figure 32	A box plot showing the number of interactors for verified and uncharacterized ORFs	110
Figure 33	Localization of the homologs of LECA NE proteins in a few model organisms	113

List of Tables

Table 1	Yeast strains used in this study	27
Table 2	Plasmids used in this study	28
Table 3	Primers used in this study	28
Table 4	Nuclear envelope proteome of *Saccharomyces cerevisiae*	31
Table 5	List of organisms used in this study	32
Table 6	Functional classification of yeast NE proteins	39
Table 7	Core and non-linearly conserved NE proteins with prokaryotic domains	82
Table 8	SCL terms considered in the script	86
Table 9	Predicted and the known localization of 100 verified ORFs	88
Table 10	Comparison of the existing web-server predictors with our script	94
Table 11	Localization predicted for 249 ORFs with no SCL data	95
Table 12	Predicted localization and the known localization from high throughput studies for 192 ORFs	99
Table 13	Predicted and experimentally determined SCL for six proteins	105
Table 14	Localization data of the LECA NE protein homologs	112

Chapter 1

Introduction

1.1 Origin of eukaryotic cell

Eukaryotic cells have complex internal organization and are characterized by the presence of a membrane bound nucleus containing the genetic material, an elaborate endomembrane system and organelles such as mitochondria. The prokaryotic cells, bacteria and archaea, lack all of the above features and are a 1000 times smaller compared to the eukaryotic cells. The emergence of eukaryotes from prokaryotes is a major evolutionary milestone on the path to complex life forms, and the sequence of events that led to the appearance of eukaryotes is still not understood clearly.

Lynn Margulis proposed a theory for the origin of eukaryotes as early as 1967 (Sagan 1967), arguing that mitochondria and chloroplasts of the eukaryotes were once prokaryotic cells. This hypothesis was based on the findings that mitochondria and chloroplasts are self-duplicating and have their own DNA and RNA. The origin of eukaryotes was hypothesized to be linked to the increasing amounts of oxygen in the atmosphere (Sagan 1967). According to the theory proposed by Lynn Margulis, evolution of the prokaryotes with photosynthetic abilities, led to increase in the oxygen content in the atmosphere, thus causing a shift from the primitive reducing atmospheric conditions to oxidizing conditions. The increased amount of oxygen in the atmosphere was hypothesized to have led to the evolution of aerobic bacteria (proto-mitochondria). Survival in the presence of oxygen would have necessitated the endosymbiosis between an aerobic prokaryote and a heterotrophic anaerobic prokaryote. Eukaryotic cell is thus proposed to be a result of ancient endosymbiotic events and the organelles such as the mitochondria and chloroplasts have descended from free-living bacteria (Sagan 1967). The endosymbiotic theory gained further support with the development of molecular biology tools and the availability of nucleic acid sequences. Phylogenetic analysis of 16S ribosomal RNA sequences showed that chloroplasts are of cyanobacterial origin and mitochondria originated from α-proteobacteria (Zablen et al. 1975; Bonen and Doolittle 1976; Yang et al. 1985).

However, this hypothesis could not explain the origin of other eukaryotic specific features like the nucleus, cytoskeleton and endomembrane system.

1.1.1 Three-domain and eocyte hypothesis

Life on earth was traditionally classified into two primary kingdoms, namely Eukaryotes and Prokaryotes. Archaea, which were until then classified within bacteria, were first identified to be a separate prokaryotic group in 1977, based on rRNA sequences (Woese and Fox 1977). In 1990, based on the universal phylogenetic tree constructed using the rRNA sequences, Carl Woese argued that the Bacteria, Archaea and Eukaryotes are three independent domains of life that diverged early in evolution and that the eukaryotes and archaea are more closely related to each other and share a common ancestor (Figure 1) (Woese et al. 1990). Phylogenetic analysis aimed at inferring the root of the universal tree using conserved paralogous gene pairs (genes that duplicated before the divergence of bacteria, archaea and eukaryotes) such as the elongation factors (EF-Tu and EF-G) and the α and β subunits of ATPase also showed that the eukaryotes and archaea are closely related to each other and positioned the root of the tree within bacteria (Iwabe et al. 1989). This topology was also found to be consistent in the phylogenetic studies using RNA polymerases II and III (Puhler et al. 1989). However, the three-domain topology of the tree has been challenged.

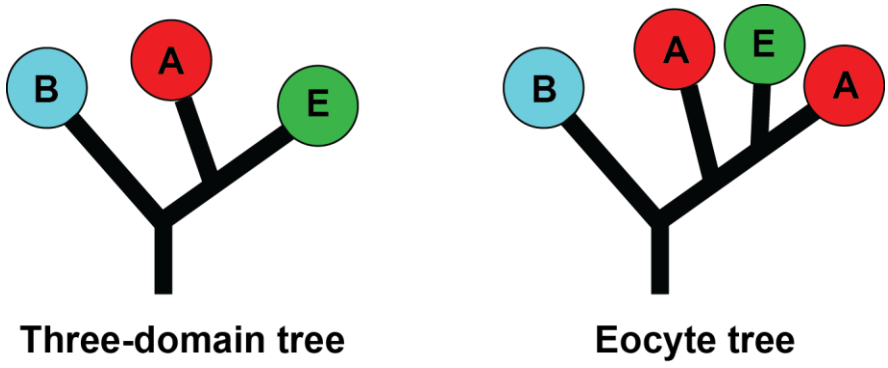

Figure 1: Three-domain tree and eocyte tree. The three-domain tree of life proposed by Carl Woese and the eocyte tree. B: Bacteria; A: Archaea and E: Eukaryotes.

Alternatively, an eocyte tree has been proposed in which the eukaryotes are considered to have evolved from within the archaea, specifically the Crenarchaeota. The analysis of the ribosomal structures of bacteria, archaea and eukaryotes showed that the large and the small ribosomal subunits of the eocytes (or Crenarchaeota) and eukaryotes share many structural features in common (Lake et al. 1984). Phylogenetic analysis of rRNA genes also showed a close relationship between the eukaryotes and the eocytes (Lake 1988). Further, comparison of elongation factor EF-1α sequences across various organisms showed that there is a 11 amino acid insertion that is shared uniquely by the eukaryotes and the eocytes (Crenarchaeota) (Rivera and Lake 1992). These findings suggested that eocytes are the closest realtives of the eukaryotes and share a sister relationship (Figure 1). This led to the eocyte hypothesis or the two-domain hypothesis according to which eukaryotes and archaea are considered as a single lineage.

However, as sequences of more eukaryotic genes became available, it was found that a number of them were closely related to bacteria rather than to archaea. Extensive analysis of whole genome based trees showed that the information processing genes such as those involved in replication, transcription and translation are related to the archaeal homologs, while the operational genes, like those involved in metabolic and biosynthetic roles, are related to the bacterial homologs (Rivera et al. 1998). These studies highlighted the chimeric nature of the eukaryotes and suggested a fusion event between archaea and bacteria in the evolution of eukaryotes.

1.1.2 Endosymbiotic model and the archaeal host

Several endosymbiotic theories have been put forward to explain the origin of eukaryotes. These theories vary with respect to the proposed partners and the sequence of events. The traditional endosymbiotic theories propose that the eukaryotic specific features like the nucleus and an endomembrane system evolved prior to the origin of mitochondria. The amitochondriate eukaryotic lineages such as the microsporidia, parabasalids and diplomonads were considered to have diverged from the protoeukaryote that had evolved eukaryotic specific features but without the mitochondria. The presence of these lineages at the base of the eukaryotic tree very distant from the other eukaryotic lineages supported the notion of early branching of these eukaryotic

lineages (Cavalier-Smith 1987a). However, these theories were ruled out by the discovery of mitochondria related organelles such as the mitosomes (in diplomonads, microsporidia) and hydrogenosomes (parabasalids), which share similar biochemical features with mitochondria (Embley et al. 2003). Additionally, the position of these lineages at the base of the tree was found to be an artifact due to long-branch attraction thus destroying the idea of ancient origin of these eukaryotes and associated theories (Philippe et al. 2000).

The understanding that all extant eukaryotes evolved from a mitochondria-bearing ancestor gave rise to a new set of theories according to which eukaryote specific features evolved after the origin of the mitochondria accompanied by extensive transfer of genes from the bacterial endosymbiont to the host (Embley and Martin 2006). One of the first of this kind is the hydrogen hypothesis, which proposes that the eukaryotic cell evolved from a symbiotic association between two prokaryotes; an autotrophic hydrogen-dependent archaeaon that was the host, and an α-proteobacterial symbiont. This endosymbiosis was driven by the dependence of the host on the hydrogen produced by the α-proteobacterium in the absence of oxygen (Martin and Muller 1998). Thus the widely accepted theories that account for the ancestral presence of mitochondria and the chimeric nature of the eukaryotic genomes involve an archaebacterial host and a α-proteobacterial symbiont. However, the nature of the archaeal host was not clear (Figure 2).

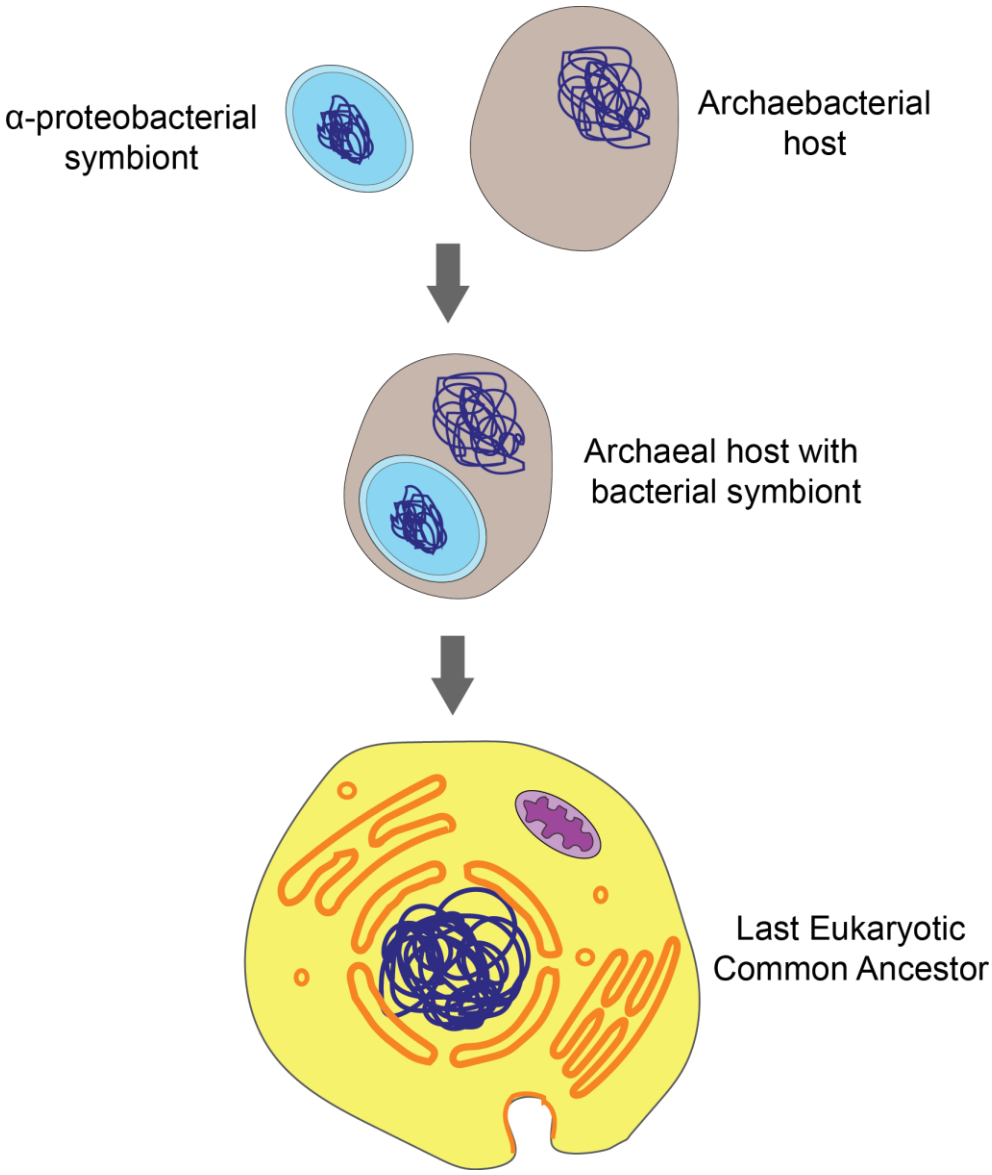

Figure 2: Endosymbiotic theory for the origin of eukaryotes. The α-proteobacterial symbiont (shown in blue) became the mitochondria of eukaryotes (shown in purple). Endomembrane system is shown in orange.

As new organisms belonging to the Archaeal domain were discovered and their genomes sequenced, several eukaryotic signature proteins including actin, tubulin and components of the ubiquitin mediated protein degradation system were identified in the TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota) superphylum of Archaea. This discovery supported the eocyte hypothesis and suggested that the archaeal host was likely to have been

within the TACK superphylum (Williams et al. 2013). Further, the discovery of new phylum of archaea called the Lokiarchaeota, closely related to the TACK superphylum, which contains additional eukaryotic specific proteins such as components of ESCRT complex, proteins that regulate actin dynamics, and expansion of small GTPases strengthened this hypothesis (Spang et al. 2015). The identification of a large number of eukaryotic specific features suggests that the archaeal host was sophisticated and possessed the components that supported the evolution of eukaryotic cellular complexity. However, the issue of the archaeal host of endosymbiosis is far from settled as new archaeal phyla are being continually discovered. The recently identified other close relatives of Lokiarchaeota, which together are now called the Asgard archaea, are seen as a promising source of information to understand the transition from prokaryotes to eukaryotes (Zaremba-Niedzwiedzka et al. 2017).

1.2 The Last Eukaryotic Common Ancestor (LECA)

The striking conservation of the internal organization of the eukaryotic cells such as the presence of membrane bound nucleus containing the DNA, a fully developed endomembrane system, mitochondria or mitochondria derived organelles such as hydrogenosomes or mitosomes across all extant eukaryotes, and the clustering of eukaryotes into a single monophyletic group in phylogenetic analysis, confirms that all the present day eukaryotes evolved from a single ancestor, conveniently termed the Last Eukaryotic Common Ancestor. Additionally, the conservation of several genes such as those involved in signaling (kinases and phosphatases), ubiquitin system, transcriptional regulation, spliceosome, etc., across diverse eukaryotes reinforce the existence of LECA (Koonin 2010).

Several studies so far have contributed to defining the properties of LECA (reviewed in (Koumandou et al. 2013)). Our current understanding is that LECA possessed a sophisticated cytoskeleton with actin, tubulin and the motor proteins; kinesin and dynein. It was capable of endocytosis, exocytosis and phagocytosis as several proteins part of these pathways such as ESCRT machinery, Rab GTPases, SNAREs, tethers and coat proteins are widely conserved. It had a complex and flexible metabolism with several key genes like the AMP-activated kinase, which plays a crucial role in maintaining the cellular energy homeostasis, genes involved in the

synthesis of essential amino acids and the components of cytosolic Fe-S cluster assembly. It possessed a nucleus with the complete set of nuclear pore complex proteins and a full-fledged nucleocytoplasmic transport system. The genome consisted of introns and it had the required splicing machinery.

This suggests that LECA was a complex organism and perhaps as sophisticated as any of the modern day eukaryotes. All the necessary components for the evolution of the diverse eukaryotes were established early in the LECA itself. However, our understanding of LECA so far has been majorly based on the pathways and processes present in animals and fungi. As more genes and pathways from several diverse organisms are analysed, we would most likely find LECA to be much more complex than many of the existing eukaryotes.

1.3 Classification of eukaryotes

The eukaryotic tree of life has undergone several modifications and rearrangements in the past three decades. The traditional eukaryotic classification was based on the level of cellular organization (multicellularity with two or more differentiated tissues) and comprised of four kingdoms namely Animalia, Plantae, Fungi and Protists (Corliss 1984). The availability of genome sequence data and use of molecular phylogenetics led to a better understanding of the relationship between the extant eukaryotes. The current classification of eukaryotes comprises of five supergroups viz., Opisthokonta, Amoebozoa, Excavata, SAR and Archaeplastida. The eukaryotic lineages that belong to each of these supergroups and the major findings that led to this classification are discussed in the following sections.

1.3.1 Opisthokonta

The supergroup Opisthokonta contains Metazoa (animals) and Fungi along with few other lineages with unicellular organisms such as choanoflagellates (e.g. *Monosiga brevicollis*). In 1987, Cavalier-Smith proposed the clade Opisthokonta based on the presence of a single posterior flagellum on flagellated cells such as sperm in animals. Phylogenetic analysis of rRNA

sequences first showed that the metazoan lineage is monophyletic, shares a recent ancestor with the choanoflagellates and that fungi are more closely related to animals than they are to other eukaryotic lineages (Wainright et al. 1993). This finding was further strengthened by the identification of a 12 amino acid insertion that is shared uniquely between fungi and animals in the sequence of elongation factor 1α (Baldauf and Palmer 1993). Till date, several phylogenetic studies based on multiple proteins have consistently recovered the sister relationship between the metazoa and fungi and have supported the existence of this clade (Derelle et al. 2015; Ren et al. 2016). Within the fungi, the relationship within and between the phyla Ascomycota, Basidiomycota, Chytridiomycota and Microsporidia have been well resolved and are shown to be consistent in phylogenomic studies (Wang et al. 2009; Ren et al. 2016).

1.3.2 Amoebozoa

The organisms belonging to this supergroup are characterized by the presence of broad pseudopodia. This supergroup includes lobose testate and naked amoeba (e.g. *Amoeba proteus*), and the mycetozoa or slime molds (the free-living single cell organisms that are capable of forming multicellular structures such as *Dictyostelium discoideum*). Amoebozoa also includes Archamoeba that contains parasitic and amitochondriate organisms such as *Entamoeba histolytica* (Pawlowski and Burki 2009). The monophyly of this supergroup has been established by several single gene studies as well as phylogenomic studies (Smirnov et al. 2005; Brown et al. 2012).

1.3.3 Excavata

This supergroup comprises of diverse organisms that are mainly heterotrophic and a large number of them are parasitic in nature. The name of this supergroup is derived from the presence of an “excavated” feeding groove in some of the organisms (Cavalier-Smith 2002). Many organisms in this supergroup are amitochondriate and contain modified forms of mitochondria such as hydrogenosomes or mitosomes. This supergroup includes kinetoplastids such as *Trypanosoma brucei*, parabasalids such as *Trichomonas vaginalis*, and diplomonads such as

Giardia lamblia. The euglenids, such as *Euglena viridis*, which contain chloroplasts that are derived from green algae through secondary endosymbiosis, are also part of this supergroup. The monophyly of this supergroup has been recently established in phylogenomic studies (Hampl et al. 2009). However, the placement of some of the descendants of this supergroup still remains controversial due to their fast evolving nature.

1.3.4 SAR

The name of this supergroup is an acronym derived from the three main constituent lineages: stramenopiles, alveolates and Rhizaria (Burki et al. 2007). The members of this supergroup do not share any morphological features in common. They are grouped together exclusively based on phylogenomic studies. This is the youngest of all supergroups and has been obtained as a merger of two previously recognized supergroups: Chromalveolata and Rhizaria. Alveolates and stramenopiles were originally part of the supergroup Chromalveolata, which also consisted of the lineages haptophytes and cryptophytes (Reyes-Prieto et al. 2007). The group Chromalveolata was initially proposed to unite organisms with plastids containing chlorophyll *c*, though several lineages included within it do not contain plastids and are not photosynthetic (Cavalier-Smith 1999). The origin of plastids is considered to be through the endosymbiosis of red alga into the common ancestor of chromalveolates and the absence of plastids in some of them is due to secondary loss. Earlier, lack of molecular phylogenetic support made this supergroup controversial. However, as genome sequence data from diverse organisms became available, several phylogenomic studies found Rhizaria clustering with stramenopiles and alveolates, excluding haptophytes and cryptophytes, which still remain as orphan lineages (Burki et al. 2007; Hackett et al. 2007). Stramenopiles consists of diatoms (e.g. *Phaeodactylum tricornutum)* and oomycetes (e.g. *Phytophthora infestans*). Alveolates are characterized by the presence of cortical alveoli and includes apicomplexans such as *Plasmodium falciparum* and ciliates such as *Tetrahymena thermophila*. Rhizaria includes amoeboid protists with filose or reticulose pseudopods. Some lineages in Rhizaria are photosynthetic and include organisms such as *Bigelowiella natans.*

1.3.5 Archaeplastida

This supergroup comprises of three major lineages namely, Glaucophyta, Rhodophyta and Viridiplantae. The organisms belonging to each of these lineages are photosynthetic and contain plastids, the organelle derived from the endosymbiosis of cyanobacteria. Glaucophytes, for example, *Cyanophora paradoxa*, consist of unusual plastids called cyanelles (contain peptidogylycan layer) that are very similar to that of cyanobacteria. Rhodophytes, also called red algae, consist of unicellular as well as multicellular algae such as the seaweed *Chondrus crispus*. The lineage Viridiplantae consists of green algae such as *Chlamydomonas reinhardtii* (unicellular algae), *Volvox carteri* (multicellular algae) and land plants such as mosses, flowering plants etc. Phylogenetic studies based on plastid genes support the monophyly of this group and suggest a single origin for the plastids in the common ancestor of Archaeplastida (Hagopian et al. 2004).

1.3.6 Root of the eukaryotic tree

The position of root of the eukaryotic tree is still heavily debated. Several positions for the root have been proposed based on phylogenomic studies. The supergroups Opisthokonta and Amoebozoa are together called unikonts (organisms with cells having either single or no flagellum), while the Excavata, SAR and Archaeplastida are classified as bikonts (organisms with two flagella) (Stechmann and Cavalier-Smith 2002; Stechmann and Cavalier-Smith 2003). The root of the tree was initially proposed to be lying between the unikonts and the bikonts based on the derived fusion status of two genes; dihydrofolate reductase (DHFR) and thymidylate synthase (TS). While all the bikonts shared a fused DHFR-TS gene, they were found to be separate genes in organisms belonging to Opisthokonta and Amoebozoa supergroups (Stechmann and Cavalier-Smith 2003). In accordance with this, phylogenetic analysis using the genes derived from bacteria, using bacteria as the outgroup, positioned the root between unikonts and the bikonts (Derelle and Lang 2012; Derelle et al. 2015). A recent study using highly conserved nuclear genes supported the root of the eukaryotic tree between unikonts and bikonts and also resolved the fungal phylogeny, which is in accordance with previous studies (Ren et al. 2016). However, a few other studies have suggested different positions for the root. The root of

the eukaryotic tree was inferred to be between opisthokonts and all other eukaryotes based on phylogenetic analysis of multiple genes from diverse taxa (Katz et al. 2012). Phylogenetic analysis of the nuclear encoded proteins of bacterial origin using bacterial sequences as the outgroup showed the root to lie between excavates and all other eukaryotes (He et al. 2014). As genome sequences of more diverse organisms become available, we would be able to better resolve the eukaryotic tree and position of the root with much better consistency. In this study the phylogenetic relationship between the various eukaryotic lineages as shown in Figure 3 (generated based on the phylogenomic studies discussed above) is considered.

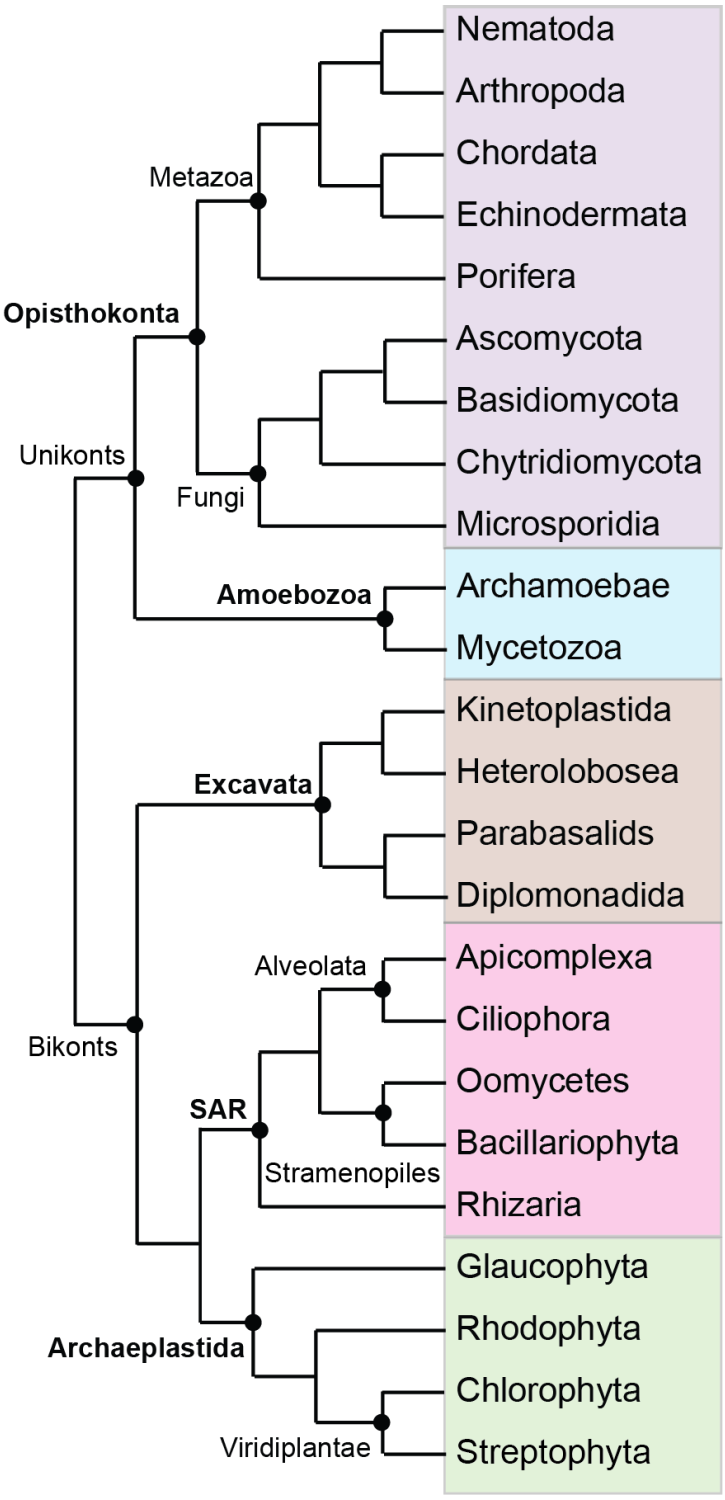

Figure 3: A schematic eukaryotic tree of life. The five eukaryotic supergroups and the major lineages in each of the supergroups are shown. The relationship between the major lineages within the supergroups and between the supergroups is adapted from phylogenomic studies as described in the above sections. The name of the five eukaryotic supergroups is in bold.

1.4 Nucleus

The nucleus is the characteristic feature of eukaryotic cells. The enormous developmental complexity of eukaryotic cells is ascribed to the presence of a membrane bound nucleus that encloses the genetic material and separates the transcription and translation processes and therefore allows exquisite control of gene expression at various stages. The prominent structural components of the nucleus include the nuclear envelope, nuclear pore complexes, nucleolus and the chromatin. The nuclear envelope is a double lipid bilayer that delineates the nucleus. It consists of an inner nuclear membrane (INM) and an outer nuclear membrane (ONM). The nuclear pore complexes (NPCs) are large multi-protein complexes composed of multiple copies of ~30 nucleoporins. They are embedded into the nuclear envelope at sites where the two bilayers converge and facilitate the free diffusion of small molecules and the bi-directional transport of macromolecules between the nucleoplasm and the cytoplasm (Newport and Forbes 1987). The nucleolus contains the rDNA and is the site for ribosome biogenesis (Figure 4).

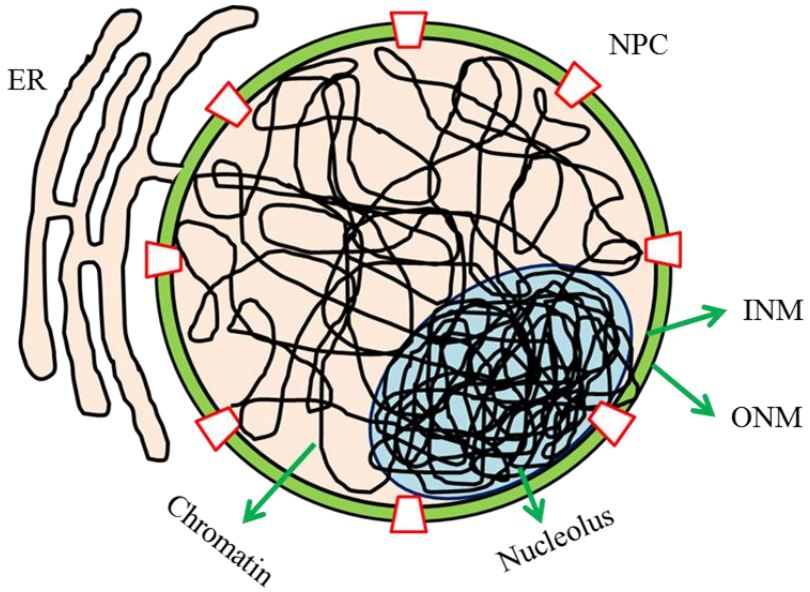

Figure 4: Nucleus of eukaryotic cell. Prominent features of the nucleus; the nuclear envelope (INM & ONM), nuclear pore complex (NPC), chromatin and nucleolus are shown.

1.4.1 Origin of the nucleus

The studies aimed at understanding the origin of eukaryotes could successfully explain the origin of the organelles such as chloroplasts and mitochondria. It is now widely accepted that eukaryotes arose by an endosymbiotic event between an archaeal host and an α-proteobacterial symbiont. However, to date there is no consensus theory to explain the origin of the nucleus. So far, there have been two competing models; the outside-in models and inside-out models. There are two major kinds of outside-in models; the endosymbiotic and the autogenous outside-in models.

The endosymbiotic outside-in models consider the nuclear compartment to be derived from an endosymbiont that was engulfed into the cytoplasm of the host. These models consider eukaryogenesis to be involving three prokaryotic partners; one host, an endosymbiont that became the mitochondria and another that became the nucleus. For example, the endokaryotic model for the origin of nucleus posits the host to be a gram-negative bacteria and that the eocyte archaeon to have formed the nucleus (Lake and Rivera 1994). Alternately, models that involve simultaneous fusion of the symbiont community involving three partners; one for the host, one for nucleus and one for mitochondria have also been proposed (Moreira and Lopez-Garcia 1998; Lopez-Garcia and Moreira 1999). The endosymbiotic model suffered for two main reseasons; first, the absence of a free-living prokaryote that is homologous to the nuclear compartment as it does not possess any metabolic pathway for generating energy for the survival of the cell (Martin 1999). Secondly, phylogenomic analysis suggest the eukaryotic genomes to be a chimera of two genomes; one archaeal and one bacterial and support for a third donor as in the case of endosymbiotic models is lacking (Rochette et al. 2014).

The autogenous outside-in models consider the nuclear compartment to have resulted from the invaginations of the plasma membrane of a prokaryote. According to the model proposed by Cavalier-Smith, the eukaryotes arose from an actinobacterial ancestor. The actinobacteria initially lost its cell wall and developed the ability to phagocytose. The evolution of phagocytosis led to the internalization of the ribosomes that remained attached to the invaginating plasma membrane and gave rise to the rough endoplasmic reticulum and eventually the nuclear envelope (Figure 5) (Cavalier-Smith 1987b; Cavalier-Smith 1988; Cavalier-Smith 2002). The major

drawback of this model is that it considers the proto-eukaryote to have evolved phagocytosis first. The process of phagocytosis requires a full-fledged cytoskeleton, which no prokaryote is known to possess. Also, according to this model the proto-mitochondria is considered to have entered the host-cell via phagocytosis (Cavalier-Smith 2002). The membrane restructuring processes such as phagocytosis is an energy-intensive process and is very unlikely to have occurred without the presence of mitochondria.

The vesicular model, also categorized as an outside-in model, proposes the endomembrane system to have evolved in a host containing the proto-mitochondrial symbiont (Martin 1999). This model posits that the heterotrophic lifestyle of the host led to the transfer of the genes from the symbiont (α-proteobacteria) to the archaeal host. During this process many of the bacterial genes involved in lipid synthesis were transferred to the host genome. This led to the synthesis of bacterial lipids, which accumulated as vesicles in the cytosol of the archaeal host. The fusion of these internal vesicles ultimately resulted in the evolution of an endomembrane system (Figure 5) (Martin 1999). This model accounts for the fact that the eukaryotic lipids are smiliar to that found in bacteria, rather than those found in archaea. However, this model does not explain the mechanism by which the proto-mitochondria entered into the host cell.

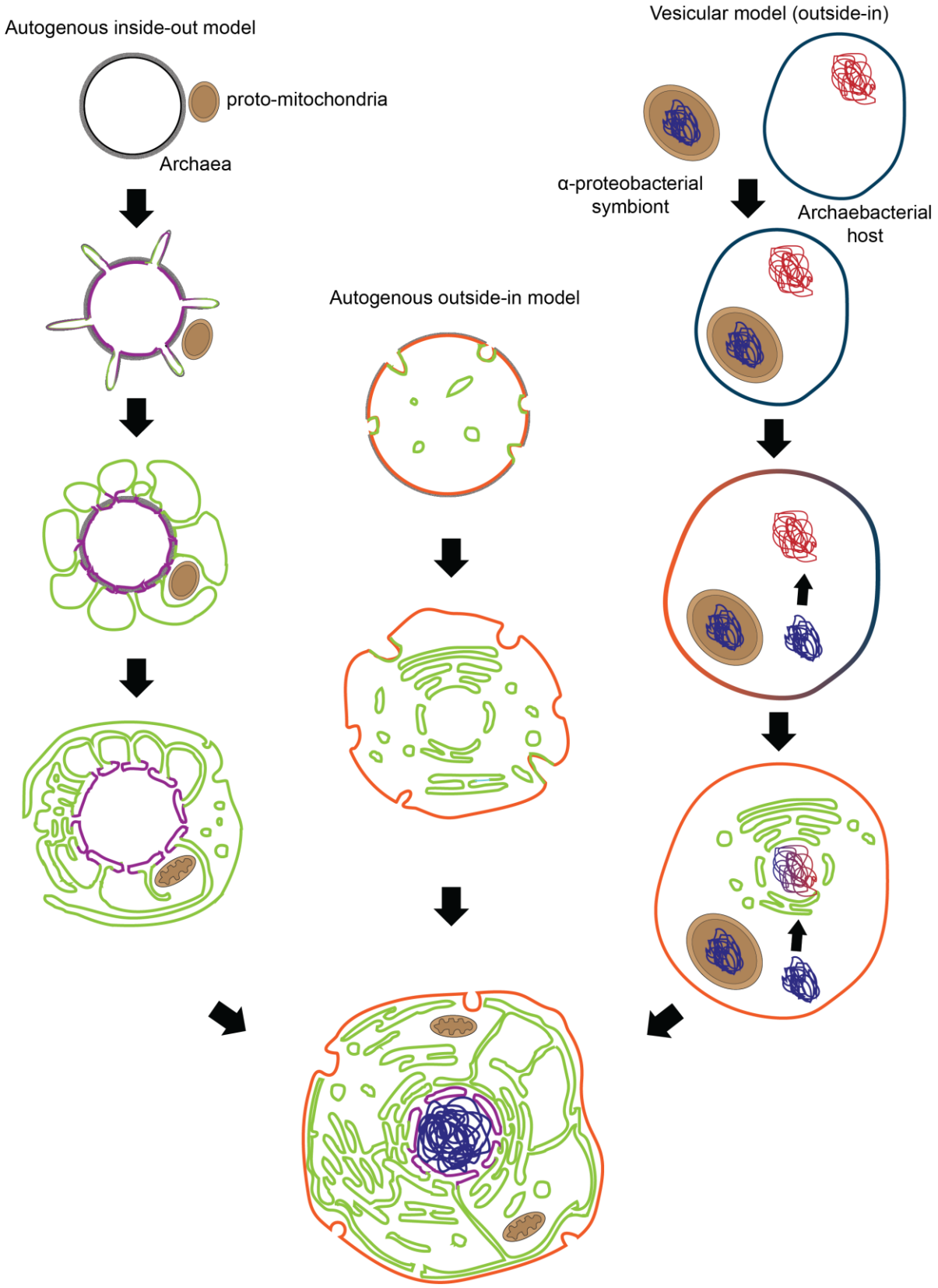
Autogenous inside-out model
proto-mitochondria
Archaea
Autogenous outside-in model
Vesicular model (outside-in)
α-proteobacterial symbiont
Archaebacterial host

Figure 5: Models for the origin of nucleus. The autogenous inside-out (left), autogenous outside-in (centre) and the vesicular outside-in model (right) are shown in the figure. Figure adopted from (Martin and Koonin 2006; Baum and Baum 2014; Baum 2015).

The inside-out model considers a eukaryote to have evolved from an archaeal host and an α-proteobacteria that lived in close association with the host. This model proposes that the archaeal host developed extracellular membrane blebs that eventually expanded and fused to give rise to the plasma membrane (Figure 5). The α-proteobacteria that got trapped between the protrusions gradually entered the cytoplasm and became the mitochondria. According to this model the ancestral prokaryotic cell corresponds to the nuclear compartment and the plasma membrane was formed later (Baum and Baum 2014).

On the other hand, the autogenous inside-out model suggests that the symbiotic relationship between the archaeal host and the proto-mitochondria resulted due to the trapping of the alpha-proteobacteria in the protrusions or the inter-bleb spaces of the archaeal host and does not require any sophisticated energy-intensive membrane manipulation processes. This model considers the eukaryotic endomembranes to have evolved from the archaeal membrane. However, the lipids found in eukaryotic membranes are similar to that found in bacteria and distinct from the archaeal lipids (Lykidis 2007). The lipids in bacteria and eukaryotes are composed of ester-liked fatty acids and make use of a glycerol-3-phosphate backbone, whereas the lipids in archaea utilize a glycerol-1-phosphate backbone and have ether-linked isoprenoids (Koga and Morii 2007). The inside-out model does not explain the need for the shift in the lipid composition of the membranes.

Thus the lack of sufficient support for any of the existing models for the evolution of nucleus leaves the question still unanswered. Studying the origins of genes of nuclear associated complexes such as the nuclear pore proteins, components of the nuclear envelope, nucleolus etc. would be helpful in evaluating the existing hypothesis and also provide new insights into the evolution of nucleus.

1.5 Nuclear envelope

Nuclear envelope, the barrier that separates the nucleoplasm containing the genetic material from the cytoplasm, hosts a number of transmembrane proteins and peripherally associated proteins. Outer nuclear membrane (ONM) is contiguous with the endoplasmic reticulum, is bound by ribosomes and shares a subset of proteins with ER (Gerace and Burke 1988). However, ONM also harbors certain unique set of proteins that interact with the cytoskeletal elements. The composition of inner nuclear membrane (INM) is distinct from ONM and contains proteins that interact with chromosomes and nucleoskeletal components. A network of intermediate filament proteins called lamins lines the INM. Lamin proteins are important structural constituents of the metazoan nuclei and help in maintaining their structural integrity. The interactions between the INM and ONM proteins, which further interact with the chromatin and cytoskeletal elements, respectively, establish nucleocytoplasmic bridges that act as a link between the nucleus and the cytoplasm. These bridges provide structural integrity to the nucleus and also help in proper positioning of the nucleus within the cell (Starr and Fridolfsson 2010). Thus the nuclear envelope, even though acting as a barrier for free movement of large molecules, also serves as a connecting link between the nucleus and cytoplasm. Nuclear envelope also participates actively in various nuclear functions. It plays a key role in the non-random organization of the genome and the multiple interactions of the chromatin with nuclear envelope proteins are crucial for gene regulation, DNA repair and maintaining genome stability. It also serves as an anchor for the centrosome and nucleolus, prominent structures important for cell division and ribosome assembly respectively (Mekhail and Moazed 2010).

DNA inside the nucleus is organized in a non-random fashion. The transcriptionally silent domains such as the telomeres and centromeres are anchored to the nuclear envelope across various organisms (Hochstrasser et al. 1986; Funabiki et al. 1993; Fang and Spector 2005). However, recent studies show the association of also the actively transcribed regions with the nuclear envelope through their interactions with the NPCs (Ahmed et al. 2010). Thus, the nuclear envelope is considered as a platform to organize the genome into transcriptionally active and repressive domains. Chromatin organization is mediated by the interactions between the chromatin or the chromatin associated proteins and the proteins at the INM and NPCs. In budding yeast, the three silent domains, namely, the telomeres, rDNA and the mating type loci,

are associated with the nuclear periphery. Telomeres and the mating type loci are maintained in repressed state by the recruitment of SIR complex proteins, which includes the Sir4 protein (Rusche et al. 2003). The anchoring of the telomeres at the periphery is achieved by the interactions of the Sir4 protein with the INM proteins Esc1 and Mps3 (Taddei et al. 2004; Bupp et al. 2007). Similarly, the rDNA is anchored to the NE by the interaction between the proteins associated with it and the Heh1 and Nur1 proteins at the INM (Mekhail et al. 2008). In Metazoa, lamins together with the INM proteins such as Lap2, Man1 and Emerin mediate the chromatin organization (Liu et al. 2003). The transcriptionally silent domains are found associated with lamins in human and flies (Pickersgill et al. 2006; Guelen et al. 2008). This shows that the inner nuclear membrane proteins are crucial for organizing the silent chromatin domains and this function is conserved from yeast to humans.

The highly repetitive sequences present in the eukaryotic genomes can participate in homologous recombination if not properly regulated and lead to genomic instability. Such regions are anchored to the nuclear periphery and disrupting the connections lead to genomic instability. For example, disrupting the connection between the rDNA (contains 200 repeating units) and the nuclear periphery in yeast leads to aberrant recombination events and affects chromosome stability (Mekhail et al. 2008). Nuclear periphery is also shown to play an important role in DNA repair. Double strand breaks that cannot be repaired by the homologous recombination pathway are recruited to the nuclear pores (Nup84) and are repaired by alternative pathways. The interactions between the Nup84, the SUMO protease (Ulp1) and the SUMO-targeted ubiquitin ligase (STUbL) complex proteins; Slx5 and Slx8, are crucial to enable efficient repair (Palancade et al. 2007; Nagai et al. 2008).

Defects in the nuclear envelope proteins alter the nuclear envelope morphology and functionality, leading to disorders that are collectively called as laminopathies. These disorders are caused due to mutations in the genes coding for lamins and lamin associated proteins such as emerin. Laminopathies include disorders such as progeria (premature aging), muscular dystrophy, and lipodystrophies (Burke and Stewart 2002; Worman et al. 2010). One of the first identified laminopathies was the Emery-Dreifuss muscular dystrophy (EDMD). Patients with EDMD have progressive muscle wastage and weakness in arms and legs, severe changes in posture and cardiomyopathy. The X-linked EDMD disorder was found to be due to mutation in

the gene coding for emerin; an integral inner nuclear membrane protein, leading to the mislocalisation of the emerin from the nuclear periphery (Bione et al. 1994). At the cellular level, the structural integrity of the nuclear envelope is severely affected in patients with EDMD. The nuclear morphology is altered with abnormal nuclear envelope protrusions lacking NPCs and nuclear envelope proteins such as Lap2 (Ognibene et al. 1999).

Though the mechanism of defects in the nuclear envelope proteins leading to diseases is unknown in several cases, these studies highlight the fundamental functions of the nuclear envelope and its importance for health.

1.6 Nuclear envelope composition across eukaryotes

Several studies have been carried out on the nuclear envelope proteins in metazoa and fungi. Apart from the well-studied NE proteins such as lamins, lamin interacting proteins and the SUN domain proteins, several integral membrane proteins have been identified to be part of the nuclear envelope in humans. Initial proteomics study of the nuclear envelope isolated from human cells identified 67 integral membrane proteins (Schirmer et al. 2003). Recent proteomic studies of the nuclear envelope from three different tissues, namely, liver, muscle and blood leukocytes identified over 1000 nuclear envelope transmembrane proteins (NETs). Among the identified transmembrane proteins, only 16% of the NETs were found to be shared in common, indicating that several of the NETs are expressed in a tissue specific manner (Talamas and Capelson 2015). This suggests that the nuclear envelope proteome is highly tissue specific and has a role in tissue specific gene expression patterns. However, a large number of these proteins still remain uncharacterized.

Expansion of nuclear envelope proteins such as lamins and SUN domain proteins has been observed in multicellular organisms. While only one lamin gene is present in most of the invertebrates, three lamin genes are found in humans (Gruenbaum and Foisner 2015). Similarly, the SUN domain proteins have expanded with humans having up to five genes, while most fungi have only a single gene. However, no lamin gene could be identified in any fungi till date. This suggests that the nuclear envelope composition across various eukaryotes is highly variable and

is tailored in an organism specific manner. Knowledge of the conserved nuclear envelope proteins across various eukaryotes would help us infer the LECA nuclear envelope proteome and the fundamental functions of the nuclear envelope, which would further contribute to understanding the origins of the nucleus.

However, to date even partial nuclear envelope composition is known only in few eukaryotes that belong to Metazoa, Fungi and Viridiplantae. Outside the organisms belonging to the Opisthokonta and Archaeplastida supergroups, nuclear envelope composition remains largely unknown. With the rapid advance in sequencing technology, genome sequences of a large number of eukaryotes belonging to all five eukaryotic supergroups are now available. Therefore, comparative genomics approaches could enable us to decipher the composition of the nuclear envelope of the Last Eukaryotic Common Ancestor (LECA).

Comparative studies have shown the presence of lamin like genes outside metazoa (Koreny and Field 2016). For example, NE81 is a nuclear periphery protein in *D. discoideum*, which is structurally and functionally similar to lamins (Kruger et al. 2012). However, the NMCP group of proteins in plants and NUP-1 protein in *T. brucei* are functionally analogous to lamins though they do not share any sequence homology (DuBois et al. 2012; Ciska and Moreno Diaz de la Espina 2013). The chromatin interacting nuclear envelope proteins such as the LEM domain and the SUN domain containing proteins are also shown to be present across various eukaryotes suggesting them to be conserved features of the nuclear envelope across eukaryotes (Mans et al. 2004). Similar studies have shown that the nuclear pore complex proteins are highly conserved across eukaryotes (DeGrasse et al. 2009; Neumann et al. 2010). Extensive comparative genomic studies have also been carried out for components involved in nucleocytoplasmic transport of proteins (karyopherins), RNA export, cell division, and kinetochores, a key component in chromosome segregation in all eukaryotes (Eme et al. 2009; O'Reilly et al. 2011; Serpeloni et al. 2011; van Hooff et al. 2017). These studies have identified the core machinery that is conserved across all supergroups and likely to have been a component of the LECA. Similarly, the nuclear envelope proteins described above and the nuclear pore proteins that are conserved across all eukaryotic supergroups are considered to be traceable to LECA.

However, to date only a small number of the nuclear envelope proteins have been analysed for their presence in LECA and we have no information on the conservation of the overall architecture of the nuclear envelope.

Similar comparative studies using multiple experimentally characterized nuclear envelope proteins in some well-studied eukaryotes would provide a better picture of the conserved components of the nuclear envelope across eukaryotes and hence NE proteome of the LECA. Additionally, tracing the prokaryotic origins of the components of the nuclear envelope and knowledge of the localization of the prokaryotic homologs would help in distinguishing the inside-out and outside-in models for the origin of the nucleus. For example, presence of homologs of the nuclear envelope proteins in archaea, which localize to its plasma membrane, would support the inside-out model. Alternatively, if the nuclear envelope formed by either invaginations of the plasma membrane of a bacterial host or by fusion of vesicles that were generated as a result of transfer of several bacterial genes (involved in lipid biosynthesis and several others) into the host genome, then several of the nuclear envelope proteins would show bacterial origins. Though analysis of several gene families associated with various important cellular processes are required to understand the sequence of events leading to the origin of the nucleus, identifying the components of the LECA nuclear envelope and their prokaryotic counterparts would serve as the first step to understand the origin of nucleus.

1.7 Objectives of the study

Nuclear envelope proteins participate in a number of nuclear functions and are important for health and survival of an organism. In spite of the importance associated with the nuclear envelope, the knowledge of nuclear envelope proteome and their functions are limited to organisms belonging to Opisthokonta and Archaeplastida supergroups. The nuclear envelope proteins conserved across all extant eukaryotes are most likely to have been components of LECA NE proteome, which can provide insights into the origin of the nucleus. The availability of genome sequences of a number of organisms belonging to all the five eukaryotic supergroups has now made it possible to identify the NE proteins across various organisms using comparative genomics.

The aim of this study is to use comparative genomics to identify the conserved nuclear envelope proteins across eukaryotes, which can also be inferred as the nuclear envelope proteome of the Last Eukaryotic Common Ancestor and subsequently identify the prokaryotic origins of the LECA nuclear envelope proteins which would contribute to our understanding of the origin and evolution of the nucleus.

The first objective of this work is to identify the known nuclear envelope proteins of the simple eukaryote *Saccharomyces cerevisiae* from the available experimental data and to look for the homologs of those proteins across organisms belonging to all the five eukaryotic supergroups to identify the NE proteins conserved across all eukaryotes and those that have been gained in a lineage and specific specific manner.

The second objective is to identify the homologs of the core and non-linearly conserved proteins in bacteria and archaea and to perform phylogenetic analysis to comment on the origins of the NE proteins with prokaryotic homologs.

The last objective is to identify additional components of the nuclear envelope as well as other organelles using computational methods. This is achieved by developing a Perl script that can make use of the available protein-protein interaction data and predict the localization of proteins with no data, and validate the localization of some of the proteins using experimental studies.

Chapter 2

Methods

2.1 Bioinformatics methods

2.1.1 Identification of NE protein homologs across eukaryotes

The homologs of the 45 nuclear envelope proteins across the 73 eukaryotic species were identified using HMMER (www.hmmer.org). Unless specified, all analyses were performed using default parameters of the respective software versions mentioned. In order to build the profile HMMs, for each of the NE proteins, homologs in opisthokonts with E-value less than 10^{-10} were first retrieved using online PSI-BLAST (3 rounds of iteration against nr database) with the yeast protein as query (Altschul et al. 1997). The paralogous proteins that arose by gene duplication in *S. cerevisiae* were analyzed together. The retrieved homologs were subjected to multiple sequence alignment using ClustalX version 2.1 (Larkin et al. 2007). The non-conserved regions of the multiple alignments were trimmed off manually using Jalview (version 2.9) (Waterhouse et al. 2009). The conserved region(s) obtained from multiple alignment was then converted into a profile HMM using *hmmbuild* (Finn et al. 2011).

The profile HMM generated was used to search the proteomes of each of the 74 organisms (including *S. cerevisiae*) using *hmmsearch* (version HMMER 3.1b2) with an E-value cut-off of 0.01. The homologs identified using *hmmsearch* were further assessed using reciprocal BLAST searches against the *S. cerevisiae* genome (online BLASTp version 2.7.1 against nr database restricted to *Saccharomyces cerevisiae* S288c sequences) and by looking for the presence of conserved domains using *hmmscan* (version HMMER 3.1b2) with GA cutoffs option against Pfam database (version 28.0). The homologs for which no domains could be detected were further scanned using CD-search at NCBI (Marchler-Bauer et al. 2015).

When multiple homologs sharing the same conserved region/domain were obtained in the *hmmsearch*, the homolog(s) that returned the *S. cerevisiae* query protein as the top-most hit with an E-value less than 10^{-5} in rBLAST were considered. A few proteins do not return the *S.*

cerevisiae protein with significant E-value in rBLAST; possibly due to extensive sequence divergence, however they do contain the conserved region. For such proteins, as only a single hit was obtained, the homolog from *hmmsearch* was directly considered.

2.1.2 Prokaryotic homologs of NE proteins

The profile HMMs generated were used to identify the homologs in the proteomes of archaea and bacteria using *hmmsearch* with an E-value cut-off of 0.01. The homologs obtained by *hmmsearch* were assessed based on the presence of the characteristic domain and by performing a BLASTp analysis against the *Saccharomyces cerevisiae S288c* genome (online BLASTp against nr database restricted to *S. cerevisiae* sequences). The homologs that returned the *S. cerevisiae* query protein as the top-most hit with E-value less than 10^{-5} were further analysed by performing a BLASTp analysis against nr database. In this BLASTp analysis against the nr database, the homologs that returned proteins other than the query were excluded.

The prokaryotic proteins, which do not qualify as homologs in the BLASTp analysis, but share the characteristic domain of the yeast protein, are considered as the "proteins sharing characteristic domain".

2.1.3 Motif analysis

For proteins whose homologs were found only in Saccharomycetes, motif analysis was carried out using MEME (version 4.11.2) by setting the minimum and maximum motif width to 6 and 50 respectively and by allowing one occurrence per sequence. The motifs identified were converted into profile HMMs using *hmmbuild* and searched in the proteomes of the fungi using *hmmsearch* (Bailey and Elkan 1994).

2.1.4 Localization data of homologs

The subcellular localization data for the homologs of the LECA NE proteins in *Mus musculus*, *Homo sapiens* and *Arabidopsis thaliana* were obtained from NCBI. Only ones with experimental evidence of nuclear envelope/nuclear pore/ER membrane localization have been considered.

2.1.5 Phylogenetic analysis

The homologs of the nuclear envelope proteins (eukaryotes alone and eukaryotic homologs with selected prokaryotic homologs) were aligned using MAFFT (version 7.397) using auto option (Katoh et al. 2002). The conserved regions in the alignment were obtained using Gblocks (version 0.91b), by setting the "Minimum Number of Sequences For a Flanking Position" to half of the number of sequences, "Minimum Length of A Block" to 5 and "Allowed Gap Positions" to half, using otherwise default parameters (Castresana 2000). The substitution model for the alignment was determined according to AIC from ProtTest (version 3.0) (Abascal et al. 2005). Maximum likelihood trees were generated using PhyML (version 3.0) with the substitution model determined from ProtTest and the branch robustness was estimated using 100 bootstrap replications (Guindon et al. 2010).

2.2 Microscopy methods

2.2.1 Live-cell imaging

Yeast cells transformed with plasmid containing the gene of interest tagged with GFP were selected on appropriate selection media. For live-cell imaging, cells harvested from an overnight culture (OD~0.6-0.8) were washed and adhered to the glass slide using ConA (0.1 mg/ml). The cells were allowed to settle and stained with DAPI (2 ng/ml) for 10 min in dark. The slide was mounted; cover-slip was placed immediately and sealed with nail paint.

2.2.2 Spheroplast preparation

Cells grown to mid-log phase were harvested and washed with water. Primary fixation was done in 3.7% formaldehyde at 30˚C for 20min. Cells were washed twice and incubated at 30˚C for 10 min in the presence of 0.1M EDTA-KOH (pH 8.0) and 10 mM DTT followed by treatment with zymolyase (250 µg/ml) in YPD-sorbitol. After 45 min spheroplasts were confirmed under the light microscope and were collected by spinning down at 1500 rpm for 10 min. Further they were washed twice with YPD-sorbitol and stored at 4 ˚C.

2.2.3 Immunofluorescence

A clean glass slide was coated with 5µl poly-L-lysine and was allowed to air dry. Spheroplasts were added and allowed to settle. Post-fixation was done in pre-chilled methanol for 5 min followed by 30 sec in pre-chilled acetone. Membrane permeabilisation was done by dipping the slide in a coplin jar containing PBS-T (0.1%) for 20 min at room temperature, followed by blocking in 1% BSA prepared in PBS-T (0.1%). Primary antibody prepared in equal volumes of 1% BSA and PBST was added (anti-myc ~ 1:600, anti-Nsp1 ~ 1:800). The slide was kept at 4˚C overnight in a humid chamber. The slide was washed three times by dipping in coplin jars containing PBS-T (0.1%). Alexa 488 and Cy3 labeled secondary antibodies at 1:1000 dilution were used for myc and Nsp1, respectively and incubated at room temperature for 2 hrs followed by washing. 2 ng/ml of DAPI in PBST was added and the slide was kept in dark for 10 min. Mounting media with/without DAPI was added to slide, coverslip was placed and sealed using nail paint.

Yeast strains used in this study

Name	Description	Source
KRY 1492	BY4741 (*his3Δ 1; leu2Δ 0; met15Δ 0; ura3Δ*)*MAT a*	Euroscarf
KRY 1494	*ESC1-13XMYC::HIS3 MAT a(W303)*	This study

KRY 1752	W303 *TEL VII L::ADE2* dsRED-HDEL *MAT a*	This study
KRY 1586	*yhl042w*::*KAN Mx ESC1 13XMYC his3Δ* 1; *leu2Δ 0*; *met15Δ 0*; *ura3Δ TRP1MAT a*	This study
KRY 1588	*yhr140w*::*KAN Mx Esc1 13XMYC HIS3*; *leu2Δ 0*; *met15Δ 0*; *ura3Δ TRP1*	This study

Table 1: Yeast strains used in this study. The yeast strains used in this study along with their description and source are listed.

Plasmids used in this study

Name	Description	Source
CKM629	YJL218W in pUG23 vector (YJL218W C- GFP TAG)	This study
CKM630	YDR124W in pUG23 vector (YDR124W C- GFP TAG)	This study
CKM631	YHR140W in pUG23 vector (YHR140W C- GFP TAG)	This study
CKM632	YHL042W in pUG23 vector (YHL042W C- GFP TAG)	This study
CKM633	YPL088C in pUG23 vector (YPL088C C- GFP TAG)	This study
CKM634	YPL264C in pUG23 vector (YPL264C C- GFP TAG)	This study

Table 2: Plasmids used in this study. The plasmids used in this study along with their description and source are listed.

Primers used in this study

S.No.	Name	Primer sequence	Description
1	YDR124W FP	cgaGAGCTCTGAGTCCTGGTGTGTC	500 bp upstream from start codon with Sac1 site
2	YDR124W RP	CGGgtcgacAATAAAATCTTTACAATCATCGC	16 bp upstream from stop codon with Sal1 site
3	YJL218W FP	CGGGAGCTCACAGTACAGAGTTCA	500 bp upstream from start codon with Sac1 site

4	YJL218W RP	GGCgtcgacTTTTCGATAGTTTGTAGTTGT	21 bp upstream from stop codon with Sal1 site
5	YPL088W FP	cgaGAGCTCTTGGCATTCATCACCC	436 bp upstream from start codon with Sac1 site
6	YPL088W RP	GGCgtcgacACATCTTTGCCTCTGG	16 bp upstream from stop codon with Sal1 site
7	YHL042W FP	5'CGGGAGCTCTTCTAGCTGCCATG3'	250 bp upstream from start codon with Sac1 site
8	YHL042W RP	GGCgtcgacAATCAATTGCTTACCAGC	18 bp upstream from stop codon with Sal1 site
9	YHR140W FP	CGGGAGCTCGCCATGTAGCATTTA	250 bp upstream from start codon with Sac1 site
10	YHR140W RP	CGGgtcgacATTCTTATCACCTTTCTTTGC	21 bp upstream from stop codon with Sal1 site
11	YPL264C FP	cgaGAGCTCTCATTGAAAGAGATACGA	250 bp upstream from start codon with Sac1 site
12	YPL264C RP	GGCgtcgacATCCTCCAAATCATC	15 bp upstream from stop codon with Sal1 site

Table 3: Primers used in this study. The sequence of the primers used in this study and their description are listed.

Chapter 3

Data set preparation

3.1 Introduction

To identify the conserved nuclear envelope proteins across eukaryotes using sequence homology based searches, we first need to identify the known and experimentally characterized components of the nuclear envelope of one eukaryote. *Saccharomyces cerevisiae* is a simple eukaryote with complete genome sequence and elaborate annotation data. A large number of genes in yeast are characterized and have subcellular localization data available from genome-wide studies. The availability of extensive experimental data thus makes it an ideal organism that can be used as a basis for comparative genomic study to trace the core or the LECA nuclear envelope proteins.

In order to draw conclusions about the conservation status of proteins across eukaryotes, we need to look for their presence in diverse phyla within the eukaryotic supergroups. A protein is considered conserved if its homologs are found in organisms belonging to all five supergroups, though absent from certain organisms, which is considered due to secondary loss. Thus, the choice of organisms and inclusion of multiple organisms from each supergroup is crucial. The relationship between the extant eukaryotes and their classification as shown in Figure 3 is used as the basis to choose organisms.

3.2 Results

3.2.1 Nuclear envelope proteins of *Saccharomyces cerevisiae*

The proteins at the nuclear envelope of *Saccharomyces cerevisiae* were retrieved using a Perl script based on the presence of keywords “nuclear envelope”, “nuclear periphery”, “nuclear membrane” in the description of genes in SGD and the localization data in Yeast GFP fusion localization database (Huh et al. 2003). The retrieved proteins were further analyzed manually

and the nuclear pore complex proteins and spindle pole components were excluded. Finally, 45 NE proteins were considered for the analysis (Table 4).

S. No	Protein	Localisation		S. No	Protein	Localisation
1	Ebp2	Nuclear periphery		24	Ntf2	Nuclear pore
2	Rrs1	Nuclear periphery		25	Thp1	Nuclear pore
3	Mps3	INM		26	Pml39	Nuclear pore
4	Heh2	INM		27	Sec39	NE & ER
5	Src1	INM		28	Pga2	NE & ER
6	Nur1	Nuclear periphery		29	Has1	NE
7	Esc1	Nuclear periphery		30	Ptc7	INM
8	Hmg1	ER & ONM		31	Trm1	INM
9	Hmg2	ER & ONM		32	Jem1	ER & ONM
10	Pct1	ER & ONM		33	Slp1	ER & ONM
11	Nem1	ER & NE		34	Scp160	ER & ONM
12	Spo7	ER & NE		35	Wss1	NE
13	Brr6	ER & NE		36	Trl1	INM
14	Brl1	ER & NE		37	Gtt3	Nuclear periphery
15	Apq12	ER & NE		38	Uip4	ER & NE
16	Ulp1	Nuclear pore		39	Mps2	NE
17	Ssm4	ER & INM		40	Nbp1	INM
18	Rrt12	ER & NE		41	Ypr174c	NE
19	Gas1	Nuclear periphery		42	Nvj1	NE
20	Asi1	INM		43	Prm3	ONM
21	Asi2	INM		44	Cos8	NE
22	Asi3	INM		45	Uip3	NE
23	Cse1	Nuclear pore		-	-	-

Table 4: Nuclear envelope proteome of *Saccharomyces cerevisiae.* The proteins present at the nuclear envelope of *Saccharomyces cerevisiae* are listed along with their specific localization.

3.2.2 Choice of organisms

To identify the homologs of the NE proteins across eukaryotes, 73 eukaryotic species belonging to diverse phyla within the five supergroups (Opisthokonta, Amoebozoa, Excavata, SAR and Archaeplastida) with complete genome sequences were chosen (Table 5). Preference was given to organisms that are included in RefSeq database and that are used as models. Relatively, a large number of organisms were chosen from fungi (at least two from each class) to study the in-depth distribution patterns of fungal specific NE proteins. The proteomes of all the organisms considered in this study were downloaded from NCBI except for *Bigelowiella natans* and *Cyanophora paradoxa* which were downloaded from JGI genome portal (Curtis et al. 2012) and the Cyanophora Genome Project hosted on the Rutgers University website respectively (Price et al. 2012).

Table 5: List of organisms used in this study

S. No	Organism	Phylum/Class/Common name	Eukaryotic Supergroup	Proteome Downloaded on
1	*Caenorhabditis elegans*	Nematode	Opisthokonta	29/09/2015
2	*Drosophila melanogaster*	Fruit fly	Opisthokonta	29/09/2015
3	*Anopheles gambiae str. PEST*	Mosquito	Opisthokonta	29/09/2015
4	*Ciona intestinalis*	Tunicate	Opisthokonta	29/09/2015
5	*Danio rerio*	Zebrafish	Opisthokonta	29/09/2015
6	*Takifugu rubripes*	Pufferfish	Opisthokonta	29/09/2015
7	*Anolis carolinensis*	Green anole lizard	Opisthokonta	29/09/2015
8	*Gallus gallus*	Chicken	Opisthokonta	30/09/2015
9	*Ornithorhynchus anatinus*	Platypus	Opisthokonta	29/09/2015
10	*Monodelphis domestica*	Opossum	Opisthokonta	29/09/2015
11	*Canis lupus familiaris*	Dog	Opisthokonta	30/09/2015
12	*Sus scrofa*	Pig	Opisthokonta	30/09/2015
13	*Mus musculus*	Mouse	Opisthokonta	29/09/2015
14	*Pan troglodytes*	Chimpanzee	Opisthokonta	30/09/2015

15	*Homo sapiens*	Human	Opisthokonta	28/09/2015
16	*Strongylocentrotus purpuratus*	Sea urchin	Opisthokonta	29/09/2015
17	*Amphimedon queenslandica*	Sponge	Opisthokonta	29/09/2015
18	*Arthroderma otae CBS 113480*	Eurotiomycetes	Opisthokonta	28/09/2015
19	*Aspergillus nidulans FGSC A4*	Eurotiomycetes	Opisthokonta	28/09/2015
20	*Neosartorya fischeri NRRL 181*	Eurotiomycetes	Opisthokonta	28/09/2015
21	*Leptosphaeria maculans JN3*	Dothideomycetes	Opisthokonta	28/09/2015
22	*Parastagonospora nodorum SN15*	Dothideomycetes	Opisthokonta	28/09/2015
23	*Botrytis cinerea B05.10*	Leotiomycetes	Opisthokonta	28/09/2015
24	*Sclerotinia sclerotiorum 1980 UF-70*	Leotiomycetes	Opisthokonta	28/09/2015
25	*Chaetomium globosum CBS 148.51*	Sordariomycetes	Opisthokonta	28/09/2015
26	*Thielavia terrestris NRRL 8126*	Sordariomycetes	Opisthokonta	28/09/2015
27	*Neurospora crassa OR74A*	Sordariomycetes	Opisthokonta	28/09/2015
28	*Tuber melanosporum Mel28*	Pezizomycetes	Opisthokonta	28/09/2015
29	*Candida glabrata CBS 138*	Saccharomycetes	Opisthokonta	28/09/2015
30	*Zygosaccharomyces rouxii CBS 732*	Saccharomycetes	Opisthokonta	28/09/2015
31	*Kluyveromyces lactis NRRL Y-1140*	Saccharomycetes	Opisthokonta	28/09/2015
32	*Schizosaccharomyces pombe 972h-*	Schizosaccharomycetes	Opisthokonta	28/09/2015
33	*Agaricus bisporus var. bisporus H97*	Agaricomycetes	Opisthokonta	28/09/2015
34	*Schizophyllum commune H4-8*	Agaricomycetes	Opisthokonta	29/09/2015
35	*Trametes versicolor FP-101664 SS1*	Agaricomycetes	Opisthokonta	29/09/2015
36	*Auricularia subglabra TFB-*	Agaricomycetes	Opisthokonta	29/09/2015

	10046 SS5			
37	*Cryptococcus neoformans var. grubii H99*	Tremellomycetes	Opisthokonta	29/09/2015
38	*Ustilago maydis 521*	Ustilaginomycetes	Opisthokonta	29/09/2015
39	*Puccinia graminis f. sp. tritici CRL 75-36-700-3*	Pucciniomycetes	Opisthokonta	29/09/2015
40	*Batrachochytrium dendrobatidis JAM81*	Chytridiomycota	Opisthokonta	29/09/2015
41	*Encephalitozoon intestinalis ATCC 50506*	Microsporidia	Opisthokonta	29/09/2015
42	*Entamoeba histolytica HM-1:IMSS*	Archamoebae	Amoebozoa	16/09/2015
43	*Acytostelium subglobosum LB1*	Mycetozoa	Amoebozoa	28/09/2017
44	*Dictyostelium discoideum AX4*	Mycetozoa	Amoebozoa	14/09/2015
45	*Polysphondylium pallidum PN500*	Mycetozoa	Amoebozoa	28/09/2017
46	*Trypanosoma brucei gambiense DAL972*	Kinetoplastid	Excavata	28/09/2015
47	*Leishmania major strain Friedlin*	Kinetoplastid	Excavata	28/09/2015
48	*Naegleria gruberi strain NEG-M*	Heterolobosean	Excavata	28/09/2015
49	*Trichomonas vaginalis G3*	Parabasalid	Excavata	28/09/2015
50	*Giardia lamblia ATCC 50803*	Diplomonad	Excavata	28/09/2015
51	*Plasmodium falciparum 3D7*	Apicomplexan	SAR	13/04/2017
52	*Theileria parva strain Muguga*	Apicomplexan	SAR	28/09/2015
53	*Babesia bovis T2Bo*	Apicomplexan	SAR	28/09/2017
54	*Toxoplasma gondii ME49*	Apicomplexan	SAR	13/04/2017
55	*Cryptosporidium hominis TU502*	Apicomplexan	SAR	28/09/2015

56	*Paramecium tetraurelia strain d4-2*	Ciliate	SAR	28/09/2017
57	*Tetrahymena thermophila SB210*	Ciliate	SAR	16/09/2015
58	*Phytophthora infestans T30-4*	Oomycete	SAR	16/09/2015
59	*Phaeodactylum tricornutum CCAP 1055/1*	Diatom	SAR	16/09/2015
60	*Thalassiosira pseudonana CCMP1335*	Diatom	SAR	16/09/2015
61	*Bigelowiella natans CCMP2755*	Rhizaria	SAR	02/11/2017
62	*Cyanophora paradoxa*	Glaucophyta	Archaeplastida	27/10/2017
63	*Chondrus crispus*	Red alga	Archaeplastida	03/10/2017
64	*Cyanidioschyzon merolae strain 10D*	Red alga	Archaeplastida	16/10/2017
65	*Chlamydomonas reinhardtii*	Green alga	Archaeplastida	30/09/2015
66	*Volvox carteri f. nagariensis*	Green alga	Archaeplastida	16/10/2017
67	*Physcomitrella patens*	Moss	Archaeplastida	16/10/2017
68	*Marchantia polymorpha subsp. ruderalis*	Liverwort	Archaeplastida	18/10/2017
69	*Amborella trichopoda*	Basal Angiosperm	Archaeplastida	18/10/2017
70	*Oryza sativa Japonica Group*	Rice	Archaeplastida	30/09/2015
71	*Zea mays*	Maize	Archaeplastida	16/10/2017
72	*Arabidopsis thaliana*	Thale cress	Archaeplastida	30/09/2015
73	*Glycine max*	Soybean	Archaeplastida	16/10/2017

Table 5: List of organisms used in this study. The table includes the list of organisms chosen for this study along with the phylum/class/common name, the eukaryotic supergroup to which the organism belongs, and the date on which the proteomes were downloaded.

3.2.3 Prokaryotic proteomes

The non-redundant protein sequences of archaea and eubacteria were downloaded from NCBI RefSeq database (Release 80; release date: Jan 09, 2017, which includes 990 archaeal taxids/organisms and 44270 bacterial taxids/organisms). Additionally, the proteomes of 120 bacterial reference genomes and 24 Asgard archaeal proteomes (submitted in Genbank) were downloaded from NCBI (downloaded on 07 December, 2017 and 29 January, 2019 respectively).

3.3 Conclusions

A total of 45 proteins localizing to INM and ONM are shortlisted. In order to identify homologs, 73 organisms belonging to diverse phyla from all five eukaryotic supergroups are considered. Relatively, a large number of organisms are chosen from fungi to study the in-depth distribution patterns in fungi.

Chapter 4

Nuclear envelope proteins across eukaryotes

4.1 Introduction

The nuclear envelope forms a selective barrier and protects the genome from chemical assaults. Maintaining the nuclear envelope integrity is essential for genome stability and cell survival. The proteins at the nuclear envelope such as the lamins, LEM and SUN domain containing proteins maintain integrity of the nuclear envelope. The LINC complexes that connect the interior of the nucleus with the structural elements of the cytoplasm serve as nucleocytoplasmic bridges, influence the size and shape of the nucleus, maintain uniform spacing between the two nuclear membranes, and play a role in proper positioning of the nucleus within the cell (Starr and Han 2003; Starr and Fridolfsson 2010; Rothballer and Kutay 2013). The LINC complex consists of the SUN domain proteins (Mps3, SUN1-5) at the INM, and the KASH domain proteins (ANC-1, Nesprin1-4) at the ONM. Defects in nuclear envelope proteins lead to disorders called laminopathies. In patients with muscular dystrophy, the nuclear envelope morphology is severely altered with abnormal protrusions.

During cell division, the nuclear envelope undergoes dramatic changes ranging from complete breakdown and reassembly in organisms undergoing open mitosis to morphological changes and expansion in organisms undergoing closed mitosis (Hetzer et al. 2005; Takemoto et al. 2016). The proteins associated with the nuclear envelope regulate remodeling during mitosis. For example, the chromatin interacting proteins such as Lap2, Man1 and emerin mediate NE reassembly in human cells at the end of mitosis (Haraguchi et al. 2008; Anderson et al. 2009), and the localization of the SUN domain protein is tightly coupled to NE dynamics during mitosis in *Arabidopsis thaliana* (Oda and Fukuda 2011). Additionally, the NE proteins mediate the non-random organization of the genome, and the interactions of the chromatin with nuclear envelope proteins are crucial for gene regulation, DNA repair and maintaining genome stability (Mekhail and Moazed 2010). The yeast nuclear envelope is also shown to host ubiquitin ligases, which target the aberrant nuclear proteins for degradation to maintain protein homeostasis and

phosphatases that regulate phospholipid synthesis to maintain membrane homeostasis (Lusk et al. 2007).

In spite of the importance associated with the nuclear envelope proteins, the composition of the nuclear envelope and the studies related to their functions are limited to only few model organisms that belong to the Opisthokonta and Archaeplastida supergroups. The composition of the nuclear envelope and the functions of the NE proteins in organisms outside these supergroups are largely unknown. Thus it is not known if most of these functions are common to all eukaryotes or if they have evolved only in these lineages. However, as genome sequences of organisms belonging to all the five supergroups are now available, comparative genomics can be used to identify the fundamental functions of the nuclear envelope.

Previous comparative studies looked for the presence of a few structural NE proteins and NPCs across eukaryotes (Mans et al. 2004; DeGrasse et al. 2009; Neumann et al. 2010; Koreny and Field 2016). However, the status of many other NE proteins in eukaryotes still remains unknown. Identifying the NE proteins that are shared commonly by the extant eukaryotes would also imply that those proteins are traceable to LECA, which would further help in tracing the origin of the nucleus.

In this chapter, we describe how we identified the homologs of the 45 nuclear envelope proteins of *S. cerevisiae* across 73 eukaryotic lineages (described in Chapter 3). Using comparative sequence analysis, we identified the core and the lineage specific nuclear envelope components.

4.2 Results

4.2.1 Nuclear envelope proteins of *Saccharomyces cerevisiae*

The nuclear envelope proteins of *Saccharomyces cerevisiae* were obtained as described in Chapter 3. The 45 NE proteins were classified into four major functional categories viz., “chromatin organization”, “nuclear envelope homeostasis”, “gene regulation”, and “transport related”. The proteins which did not fit into the above four functional categories or whose function is not known were placed under “others” category (Table 6).

Chromatin organisation	**Nuclear envelope homeostasis**	**Gene regulation**	**Transport related**	**Others**	
Ebp2	Hmg1	Ulp1	Cse1	Has1	Uip4
Rrs1	Hmg2	Ssm4	Ntf2	Ptc7	Mps2
Heh2	Pct1	Rrt12	Thp1	Trm1	Nbp1
Src1	Nem1	Gas1	Pml39	Jem1	Ypr174c
Mps3	Spo7	Asi1	Sec39	Slp1	Nvj1
Nur1	Brr6	Asi3	Pga2	Scp160	Prm3
Esc1	Brl1	Asi2		Wss1	Cos8
	Apq12			Trl1	Uip3
				Gtt3	

Table 6: Functional classification of yeast NE proteins. The nuclear envelope proteins belonging to each of the five functional categories are shown in the table.

4.2.2 Homologs of NE proteins

The homologs of the 45 yeast NE proteins were identified across the 73 shortlisted eukaryotes belonging to the Opisthokonta, Amoebozoa, Excavata, SAR and Archaeplastida supergroups. The NE proteins considered in this study include diverse proteins, which have varying rates of evolution. While, some of the proteins are highly conserved and share significant sequence similarity with the yeast proteins, some are found to be rapidly evolving and have diverged extensively. To maximize the identification of homologs even for rapidly evolving proteins, we used a HMM based method as described in the methods section. Briefly, the homologs were detected as follows: the homologs in closely related organisms were first identified using BLASTp; the identified homologs were used to build profile HMMs; the profile HMMs were then used to identify the homologs in the proteomes of the 73 shortlisted organisms using *hmmsearch*. The presence and absence of the homologs of all the 45 proteins across the 73 shortlisted eukaryotes are mapped (Additional data 1).

Based on the presence of the homologs, the NE proteins are classified as “core proteins”, “non-linearly conserved proteins” or “fungal specific proteins”. The proteins whose homologs were identified in at least one organism across all the five supergroups are termed the “core proteins”. The proteins whose homologs were detected in two or more but not all in five supergroups are termed the “non-linearly conserved” proteins. The proteins for which the homologs were found only in fungi are termed the “fungal specific proteins”. As we started with the *S. cerevisiae* (a fungi) NE proteins as the basis of our analysis, we did find proteins that were specifically present in fungi. However, not finding a homolog in a particular lineage does not necessarily indicate absence, as it is possible that the homolog in the given lineage or species would have diverged extensively and hence could not be detected, resulting in false negatives.

4.2.2.1 Chromatin organization

The proteins at the the inner nuclear membrane play an important role in organizing the chromatin into transcriptionally active and repressive domains, a feature conserved from yeast to humans (Akhtar and Gasser 2007; Mekhail and Moazed 2010). *Saccharomyces cerevisiae* consists of about 7 proteins, namely Ebp2, Rrs1, Mps3, Heh2, Src1, Nur1 and Esc1 at the INM that are involved in chromatin organization. These proteins are involved in clustering of the telomeres and/or tethering the telomeres and rDNA to the nuclear periphery. Among the seven proteins, five of them, namely Ebp2, Rrs1, Mps3, Heh2 and Src1 were found to be part of the core proteome (Figure 6). The homologs of these five proteins were identified across all supergroups, although they are found to have varying degrees of conservation.

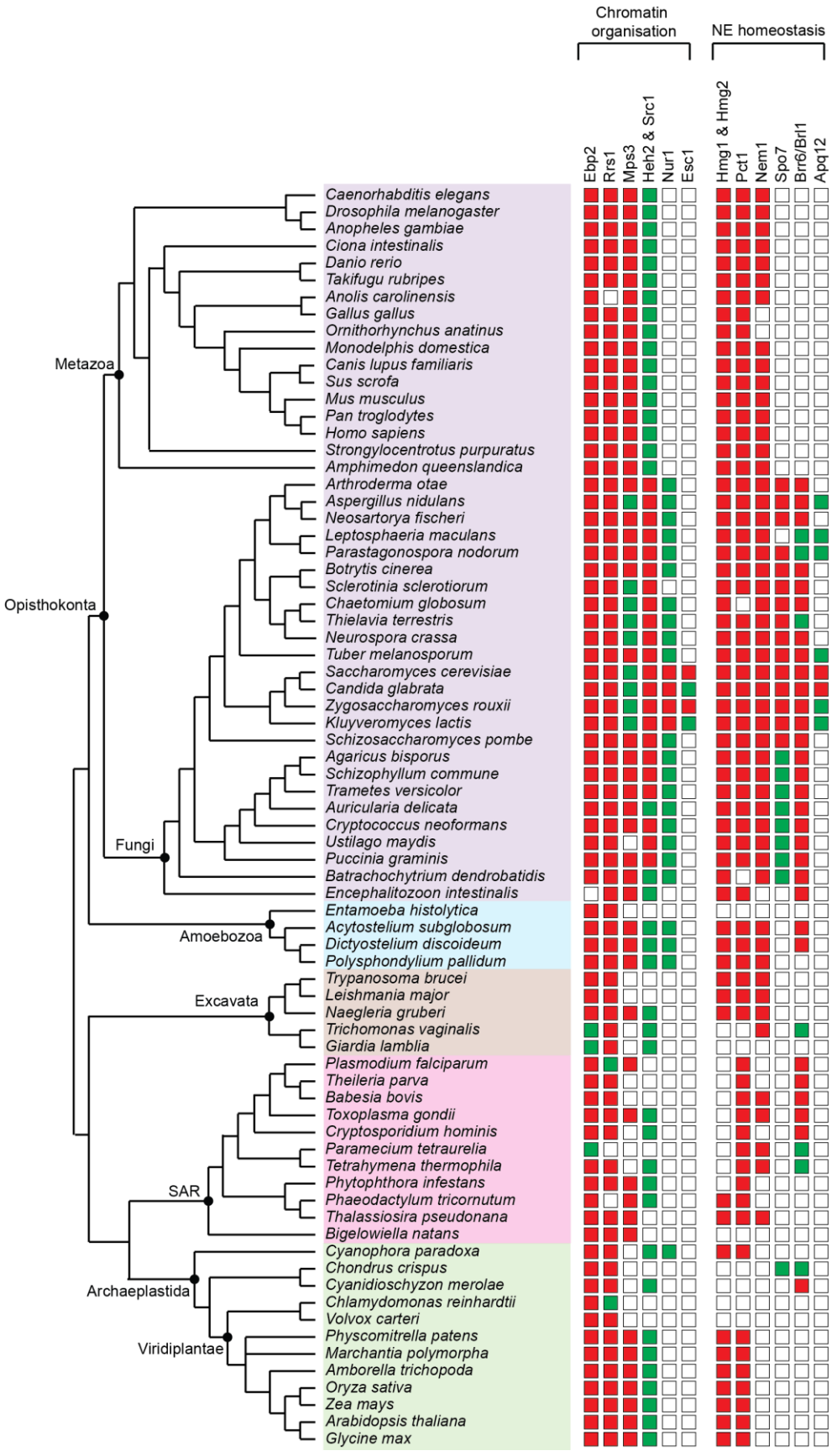
Chromatin organisation
NE homeostasis
Ebp2
Rrs1
Mps3
Heh2 & Src1
Nur1
Esc1
Hmg1 & Hmg2
Pct1
Nem1
Spo7
Brr6/Brl1
Apq12
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Ciona intestinalis
Danio rerio
Takifugu rubripes
Anolis carolinensis
Gallus gallus
Ornithorhynchus anatinus
Monodelphis domestica
Canis lupus familiaris
Sus scrofa
Mus musculus
Pan troglodytes
Homo sapiens
Strongylocentrotus purpuratus
Amphimedon queenslandica
Arthroderma otae
Aspergillus nidulans
Neosartorya fischeri
Leptosphaeria maculans
Parastagonospora nodorum
Botrytis cinerea
Sclerotinia sclerotiorum
Chaetomium globosum
Thielavia terrestris
Neurospora crassa
Tuber melanosporum
Saccharomyces cerevisiae
Candida glabrata
Zygosaccharomyces rouxii
Kluyveromyces lactis
Schizosaccharomyces pombe
Agaricus bisporus
Schizophyllum commune
Trametes versicolor
Auricularia delicata
Cryptococcus neoformans
Ustilago maydis
Puccinia graminis
Batrachochytrium dendrobatidis
Encephalitozoon intestinalis
Entamoeba histolytica
Acytostelium subglobosum
Dictyostelium discoideum
Polysphondylium pallidum
Trypanosoma brucei
Leishmania major
Naegleria gruberi
Trichomonas vaginalis
Giardia lamblia
Plasmodium falciparum
Theileria parva
Babesia bovis
Toxoplasma gondii
Cryptosporidium hominis
Paramecium tetraurelia
Tetrahymena thermophila
Phytophthora infestans
Phaeodactylum tricornutum
Thalassiosira pseudonana
Bigelowiella natans
Cyanophora paradoxa
Chondrus crispus
Cyanidioschyzon merolae
Chlamydomonas reinhardtii
Volvox carteri
Physcomitrella patens
Marchantia polymorpha
Amborella trichopoda
Oryza sativa
Zea mays
Arabidopsis thaliana
Glycine max
Metazoa
Opisthokonta
Fungi
Amoebozoa
Excavata
SAR
Archaeplastida
Viridiplantae

Figure 6: Proteins involved in chromatin organization and NE homeostasis across eukaryotes. The presence/absence and the degree of conservation of homologs identified for chromatin organization and NE homeostasis proteins are shown. Red filled squares represent the homologs validated using rBLAST with significant E-value (less than 10^{-5}). The green filled squares represent the homologs that can be found only using *hmmsearch* and share conserved region/domain. The supergroups Opisthokonta, Ameobozoa, Excavata, SAR and Archaeplastida are shaded in purple, blue, brown, pink and green filled rectangles, respectively.

The homologs of the proteins Ebp2 and Rrs1 were identified across almost all the organisms considered. The homologs of these proteins are well conserved and share significant sequence similarity (Figure 7). In case of the C-terminal SUN domain protein Mps3, homologs could not be detected in a few bikonts. Mps3 homologs in Saccharomycetes have diverged significantly from the rest of the eukaryotes (Figure 6). In rBLAST analysis, most of the homologs identified return the *S. pombe* SUN domain protein with significant E-value but not the Mps3 of *S. cerevisiae*.

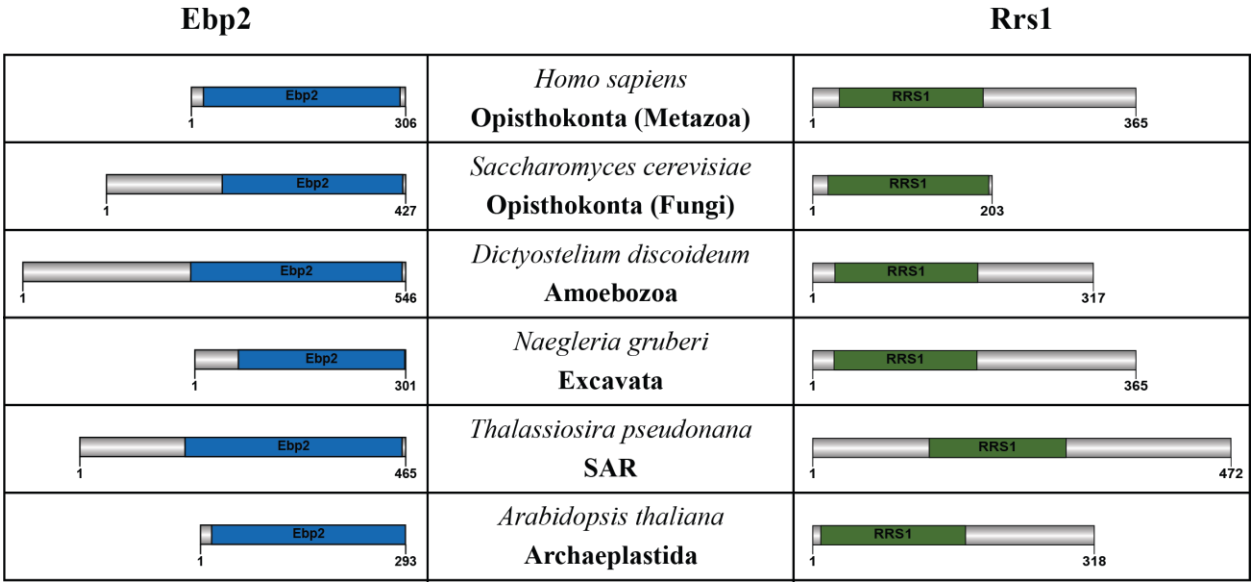

Figure 7: Domain organization in Ebp2 and Rrs1 proteins. The domain architectures of the homologs of Ebp2 and Rrs1 proteins are shown in representative organisms from each eukaryotic supergroup. Ebp2 domain is shown in blue and the RRS1 domain is shown in green. All maps are drawn to scale.

Heh2 and Src1 are paralogous proteins in yeast. These proteins have a HeH domain at the N-terminal and MSC domain at the C-terminal end. Homologs of these proteins, with significant sequence similarity were identified only in fungi (Figure 6). However, we did find proteins with the MSC domain across all supergroups in our *hmmsearch* analysis. The homologs with an N-terminal HeH/LEM domain in combination with MSC domain were found only in opisthokonts and in one excavate, *Naegleria gruberi* (Figure 8). The HeH domain present in fungi and the LEM domain found in metazoa share a similar helix-extension-helix fold (Brachner and Foisner 2011).

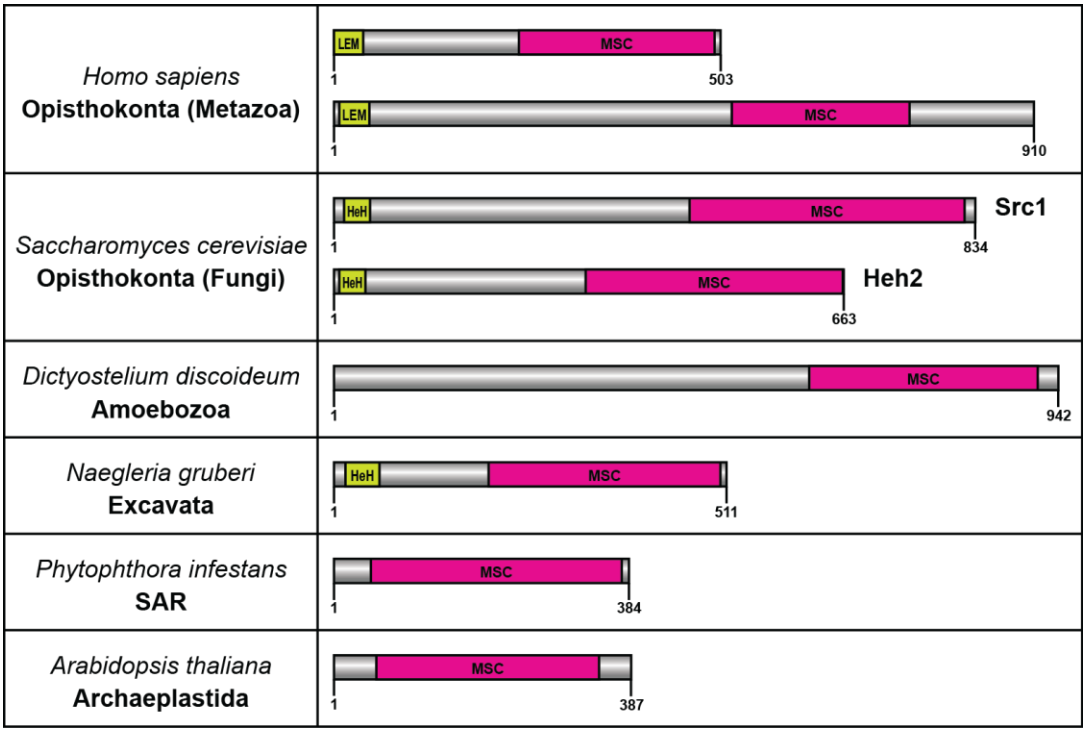

Figure 8: Domain organization in Heh2 and Src1 proteins. The domain architectures of the homologs of Heh2 and Src1 proteins are shown in representative organisms from each eukaryotic supergroup. MSC domain is shown in pink and the N-terminal HeH/LEM domain is shown in green. All maps are drawn to scale.

Among the chromatin interacting proteins, Nur1 and Esc1 were found to be non-linearly conserved and fungal specific respectively. Nur1 protein homologs were found in fungi, mycetozoa and in *Cyanophora paradoxa*, a glaucophyte that belongs to the Archaeplastida supergroup. The homologs identified in Saccharomycetes are found to share significant sequence

similarity, while the rest share the characteristic domain (Figure 6). Esc1 is a lineage specific protein as its homologs could be identified only in Saccharomycetes. The homologs identified share similarity only w.r.t a small motif.

4.2.2.2 Nuclear Envelope homeostasis

The dynamics of the nuclear envelope and its shape are tightly linked to the genes that regulate lipid synthesis and maintain lipid homeostasis. We find 8 proteins at the ONM-ER network of yeast that are involved in maintaining the nuclear envelope homeostasis. Several of these are transmembrane proteins and include the paralogous proteins Hmg1 & Hmg2 that are involved in sterol biosynthesis (Basson et al. 1986); Pct1, Nem1 and Spo7 proteins that regulate the phospholipid biosynthesis (Siniossoglou 2009) and the proteins Brr6, Brl1 and Apq12 that maintain lipid homeostasis (Hodge et al. 2010). Of these, the proteins Hmg1, Hmg2, Pct1, Brr6 & Brl1 are identified to be part of the core proteome (Figure 6). The homologs of the HMG-CoA reductase proteins (Hmg1 & Hmg2) are found across all the opisthokonts considered in this study and are highly conserved. However, they could not be detected in a number of bikonts including alvelolates (SAR), parabasalids and diplomonads (Excavata), red algae and green algae (Archaeplastida) (Figure 6). A lineage specific gain of domains has been found in case of HMG-CoA reductases. Homologs identified in all organisms contain the HMG-CoA_red domain, while an additional Sterol_sensing domain is found only in opisthokonts. This suggests the gain of Sterol_sensing domain in the common ancestor of opisthokonts. Further, the homologs in fungi also have a HPIH domain at the N-terminal in addition to the Sterol_sensing and HMG-CoA_red domains (Figure 9).

The homologs of the Pct1 protein, involved in phospholipid biosynthesis were found across all supergroups. Interestingly, we find that the homologs of Brr6/Brl1 are restricted to only a few organisms across all five supergroups. They were found only in fungi in Opisthokonta, slime molds in Amoebozoa, parabasalids in Excavata, alveolates in SAR and rhodophytes in Archaeplastida. This suggests secondary loss in large subsets of organisms across supergroups.

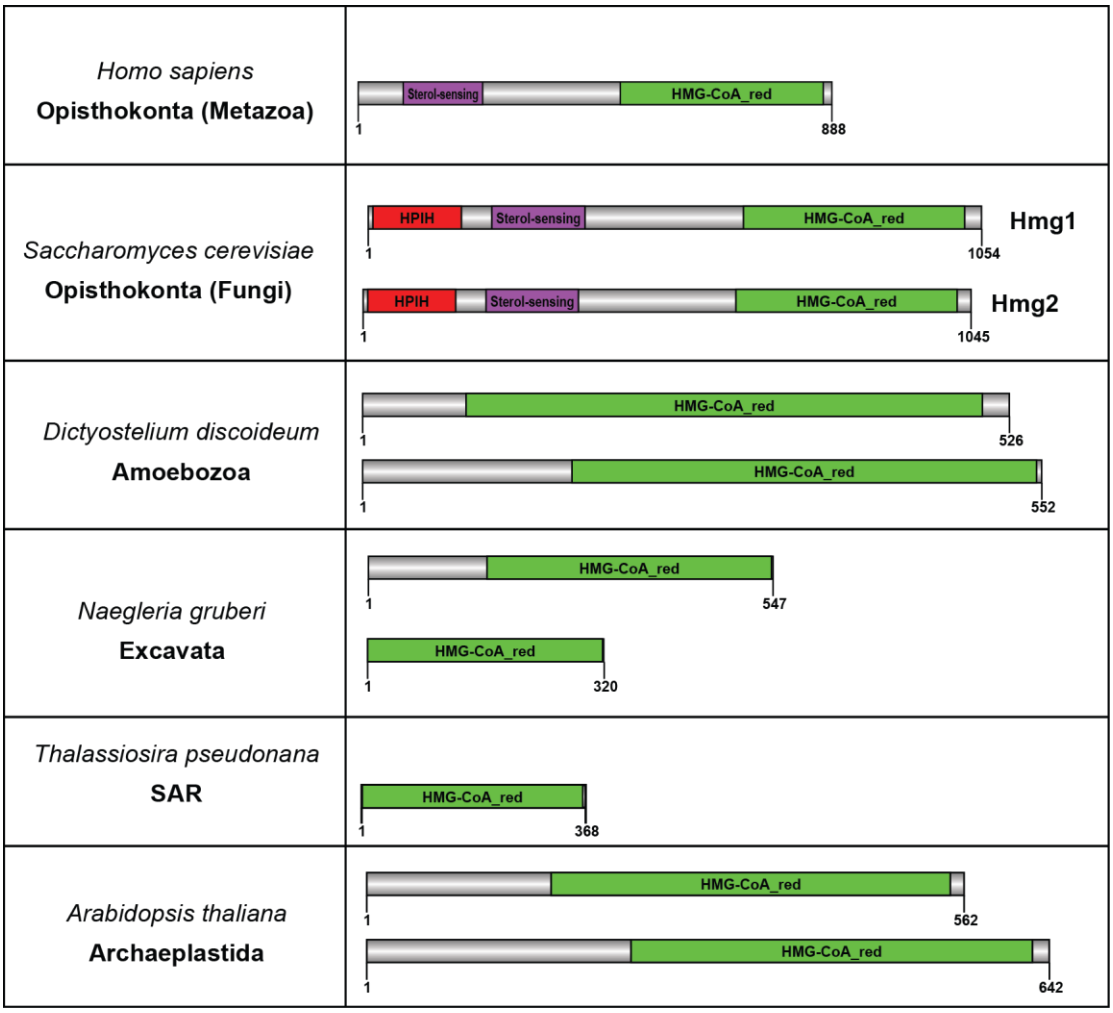

Figure 9: Domain organization in Hmg1 & Hmg2 proteins. Homologs of Hmg1 and Hmg2 proteins are shown in representative organisms. The domains found in each of the homologs are shown in different colors. HMG-CoA_red (green), sterol-sensing (purple), HPIH (red). The homolog in *Thalassiosira pseudonana* with HMG-CoA_red domain is partial. All maps are drawn to scale.

Among the NE homeostasis proteins, Nem1 and Spo7 are non-linearly conserved. The homologs of Nem1 protein are found in four of the five supergroups, while no homolog could be detected in any organism belonging to Archaeplastida. Spo7 homologs were identified only in fungi and in one red alga. However, previous studies have shown the presence of a Spo7 ortholog in mammals, which could be identified using the *S. pombe* Spo7, but not *S. cerevisiae* Spo7 (Han et

al. 2012). The Apq12 protein, which functions along with Brr6 and Brl1 in maintaining lipid homeostasis is present only in ascomycetes and is categorized as fungal specific.

4.2.2.3 Gene regulation

The inner nuclear membrane also hosts proteins that contribute to the spatial and temporal regulation of gene expression. This regulation is achieved by the post-translational modification of the transcription activators/repressors that are targeted to the nuclear envelope. The proteins at the nuclear envelope of yeast that fall into this functional class include Ulp1 (SUMO protease), Ssm4, Asi1 & Asi3 (Ubiquitin ligases), Rrt12 (peptidase) and Gas1 (1,3-beta-glucanosyltransferase). All these proteins play a role in the regulation of gene expression (Deng and Hochstrasser 2006; Zargari et al. 2007; Hontz et al. 2009; Texari et al. 2013; Eustice and Pillus 2014). The homologs of three proteins, namely, Ulp1, Ssm4 and Rrt12 are found across all supergroups and are considered as core proteins (Figure 10). Ulp1 in yeast and its homolog in human are associated with the NPC. The Ssm4 protein, which is found at the INM and ER in yeast, is found at the ER in human. Asi1 and Asi3 are categorized as non-linearly conserved proteins as no homolog could be detected in amoeba. The fungal homologs and the Apicomplexan *B. bovis,* share significant sequence similarity with Asi1/Asi3; however, most of the others return Asi1/Asi3 as top-most hits in rBLAST but with an E-value higher than 10^{-5} but less than 10^{-2}. Asi2 protein, which works together with Asi1 & Asi3 proteins, is present only in Saccharomycetes. The homologs of the Gas1 protein were found in fungi, *P. tricornutum* and in *Zea mays*. Remarkably, the homolog in *Zea mays* shares the same domain architecture and significant sequence similarity with yeast Gas1.

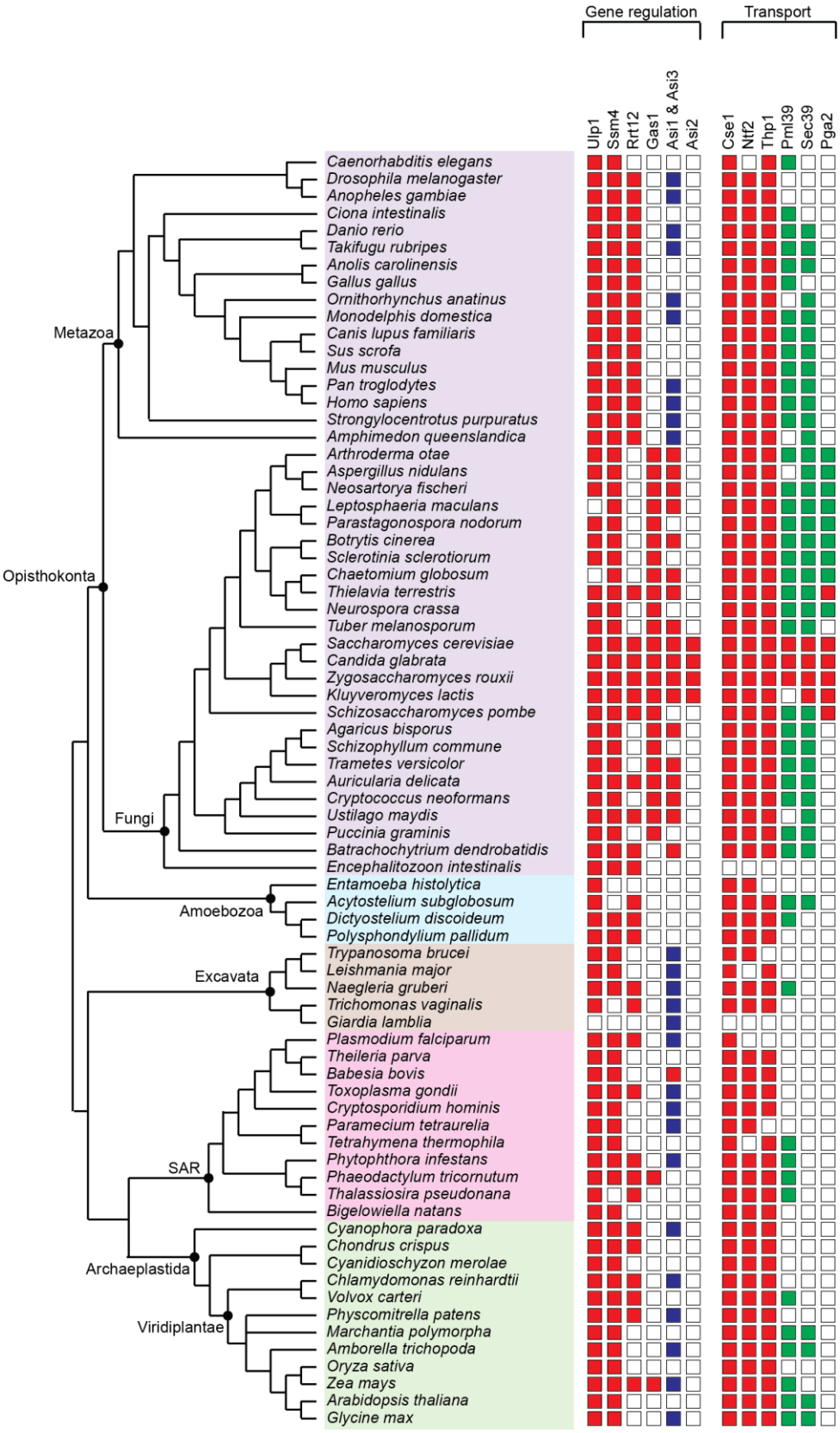
Gene regulation
Transport
Ulp1
Ssm4
Rrt12
Gas1
Asi1 & Asi3
Asi2
Cse1
Ntf2
Thp1
Pml39
Sec39
Pga2
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Ciona intestinalis
Danio rerio
Takifugu rubripes
Anolis carolinensis
Gallus gallus
Ornithorhynchus anatinus
Monodelphis domestica
Canis lupus familiaris
Sus scrofa
Mus musculus
Pan troglodytes
Homo sapiens
Strongylocentrotus purpuratus
Amphimedon queenslandica
Arthroderma otae
Aspergillus nidulans
Neosartorya fischeri
Leptosphaeria maculans
Parastagonospora nodorum
Botrytis cinerea
Sclerotinia sclerotiorum
Chaetomium globosum
Thielavia terrestris
Neurospora crassa
Tuber melanosporum
Saccharomyces cerevisiae
Candida glabrata
Zygosaccharomyces rouxii
Kluyveromyces lactis
Schizosaccharomyces pombe
Agaricus bisporus
Schizophyllum commune
Trametes versicolor
Auricularia delicata
Cryptococcus neoformans
Ustilago maydis
Puccinia graminis
Batrachochytrium dendrobatidis
Encephalitozoon intestinalis
Entamoeba histolytica
Acytostelium subglobosum
Dictyostelium discoideum
Polysphondylium pallidum
Trypanosoma brucei
Leishmania major
Naegleria gruberi
Trichomonas vaginalis
Giardia lamblia
Plasmodium falciparum
Theileria parva
Babesia bovis
Toxoplasma gondii
Cryptosporidium hominis
Paramecium tetraurelia
Tetrahymena thermophila
Phytophthora infestans
Phaeodactylum tricornutum
Thalassiosira pseudonana
Bigelowiella natans
Cyanophora paradoxa
Chondrus crispus
Cyanidioschyzon merolae
Chlamydomonas reinhardtii
Volvox carteri
Physcomitrella patens
Marchantia polymorpha
Amborella trichopoda
Oryza sativa
Zea mays
Arabidopsis thaliana
Glycine max
Metazoa
Opisthokonta
Fungi
Amoebozoa
Excavata
SAR
Archaeplastida
Viridiplantae

Figure 10: Gene regulation and transport proteins across eukaryotes. The presence/absence and the degree of conservation of homologs identified for proteins involved in gene regulation and transport are shown. Red filled squares represent the homologs validated using rBLAST with significant E-value (less than 10^{-5}). Blue filled squares represent the homologs validated using rBLAST but with E-value higher than 10^{-5} but less than 10^{-2}. The green filled squares represent the homologs that can be found only using *hmmsearch* and share conserved region/domain. The supergroups Opisthokonta, Ameobozoa, Excavata, SAR and Archaeplastida are shaded in purple, blue, brown, pink and green filled rectangles, respectively.

4.2.2.4 Transport

The nuclear pore proteins embedded into the nuclear envelope mediate the nucleocytoplasmic transport of macromolecules and are conserved across eukaryotes. Though pore complex proteins were excluded from our analysis, we did consider a few pore-associated proteins namely, Cse1 and Ntf2, that are involved in nucleocytoplasmic transport of the proteins (Corbett and Silver 1996; Hood and Silver 1998); Thp1, involved in mRNA export (Fischer et al. 2002) and Pml39, involved in retaining unspliced mRNAs inside the nucleus (Palancade et al. 2005). Additionally, two proteins, Sec39 and Pga2, involved in vesicle-mediated transport and protein processing/trafficking, respectively, localized to nuclear membrane/ER (Yu et al. 2006; Rogers et al. 2014). The proteins Cse1, Ntf2, Thp1 and Pml39 are identified to be part of the core proteome in this functional class. The homologs of Cse1 and Ntf2 identified across eukaryotes show significant sequence similarity with the yeast protein, while the Pml39 homologs outside Saccharomycetes share only the characteristic domain and cannot be detected using BLASTp (Figure 10). The Thp1 homologs in Saccharomycetes have diverged significantly from the rest of the eukaryotes. The Thp1 homologs identified in *hmmsearch* analysis showed significant sequence similarity with the *A. delicata* Thp1, but not the Thp1 of *S. cerevisiae*. Among the transport related proteins, Sec39 and Pga2 are categorized as non-linearly conserved and fungal specific respectively. The homologs of Sec39 were detected only in Opisthokonts, Amoebozoa and Viridiplantae, while the Pga2 homologs were found only in ascomycetes.

4.2.2.5 Other NE proteins

Apart from the proteins belonging to the above-mentioned functional categories, we find proteins with diverse functions and some with yet unknown functions associated with the nuclear envelope. While some of these proteins are part of the core protein group, a large number of these are found to be fungal specific (Figure 11). The core proteins include the helicase, Has1; phosphatase Ptc7; tRNA methyltransferase Trm1; DnaJ chaperone Jem1 and the mid-SUN domain protein Slp1. Has1 and Trm1 are highly conserved overall at the sequence level, while Jem1 homology is limited to the DnaJ domain. The RNA binding protein Scp160, metalloprotease Wss1 and the tRNA ligase Trl1, are classified as non-linearly conserved proteins. Interestingly, a large number of proteins in this category namely, Gtt3, Uip4, Mps2, Nbp1, Ypr174c, Nvj1, Prm3, Cos8 and Uip3 were found only in ascomycetes.

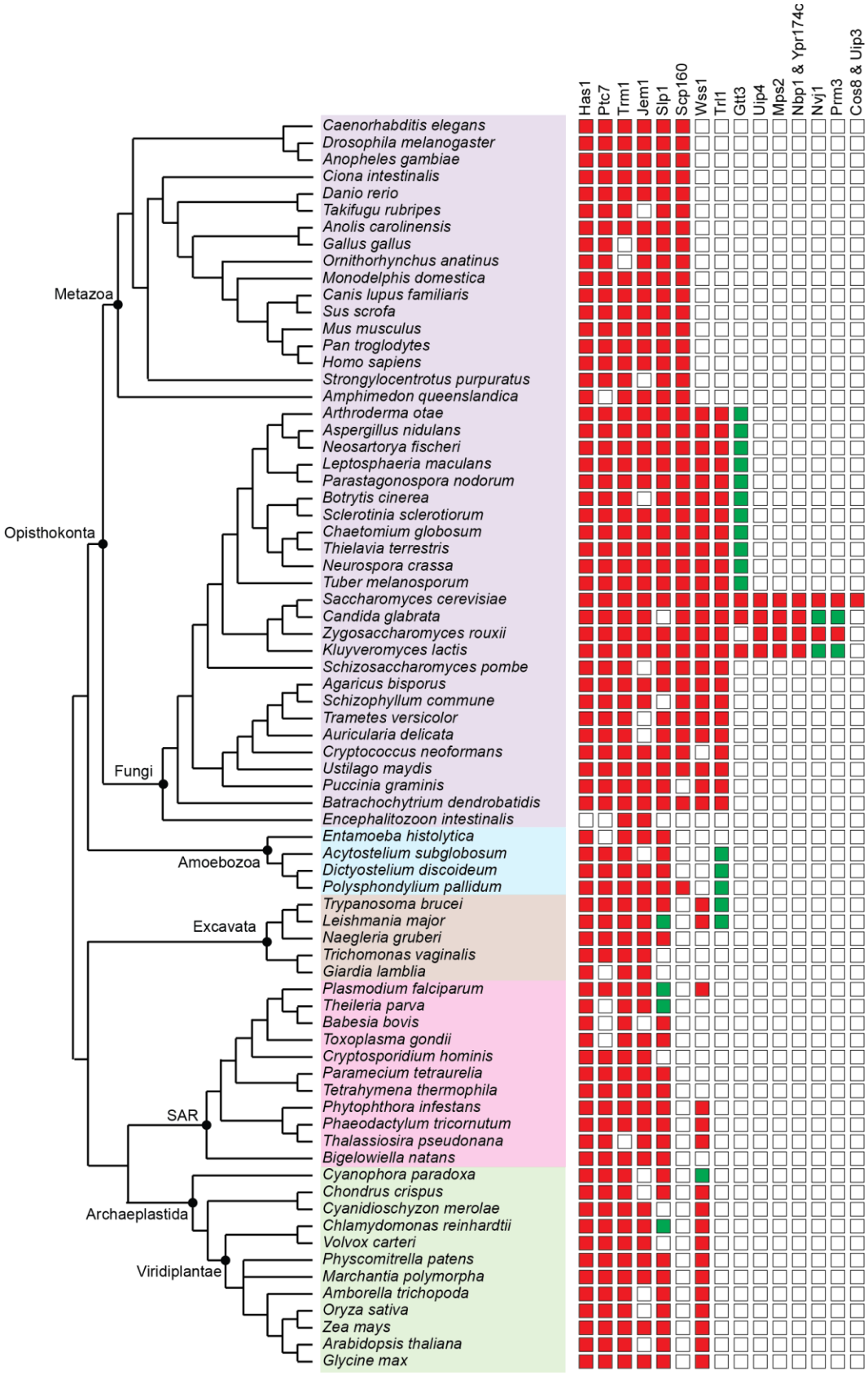
Has1
Ptc7
Trm1
Jem1
Slp1
Scp160
Wss1
Trl1
Gtt3
Uip4
Mps2
Nbp1 & Ypr174c
Nvj1
Prm3
Cos8 & Uip3
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Ciona intestinalis
Danio rerio
Takifugu rubripes
Anolis carolinensis
Gallus gallus
Ornithorhynchus anatinus
Monodelphis domestica
Canis lupus familiaris
Sus scrofa
Mus musculus
Pan troglodytes
Homo sapiens
Strongylocentrotus purpuratus
Amphimedon queenslandica
Arthroderma otae
Aspergillus nidulans
Neosartorya fischeri
Leptosphaeria maculans
Parastagonospora nodorum
Botrytis cinerea
Sclerotinia sclerotiorum
Chaetomium globosum
Thielavia terrestris
Neurospora crassa
Tuber melanosporum
Saccharomyces cerevisiae
Candida glabrata
Zygosaccharomyces rouxii
Kluyveromyces lactis
Schizosaccharomyces pombe
Agaricus bisporus
Schizophyllum commune
Trametes versicolor
Auricularia delicata
Cryptococcus neoformans
Ustilago maydis
Puccinia graminis
Batrachochytrium dendrobatidis
Encephalitozoon intestinalis
Entamoeba histolytica
Acytostelium subglobosum
Dictyostelium discoideum
Polysphondylium pallidum
Trypanosoma brucei
Leishmania major
Naegleria gruberi
Trichomonas vaginalis
Giardia lamblia
Plasmodium falciparum
Theileria parva
Babesia bovis
Toxoplasma gondii
Cryptosporidium hominis
Paramecium tetraurelia
Tetrahymena thermophila
Phytophthora infestans
Phaeodactylum tricornutum
Thalassiosira pseudonana
Bigelowiella natans
Cyanophora paradoxa
Chondrus crispus
Cyanidioschyzon merolae
Chlamydomonas reinhardtii
Volvox carteri
Physcomitrella patens
Marchantia polymorpha
Amborella trichopoda
Oryza sativa
Zea mays
Arabidopsis thaliana
Glycine max
Metazoa
Opisthokonta
Fungi
Amoebozoa
Excavata
SAR
Archaeplastida
Viridiplantae

Figure 11: Other NE proteins found across eukaryotes. The presence/absence and the degree of conservation of homologs identified for proteins categorized under "Others" are shown. Red filled squares represent the homologs validated using rBLAST with significant E-value (less than 10^{-5}). The green filled squares represent the homologs that can be found only using *hmmsearch* and share conserved region/domain. The supergroups Opisthokonta, Ameobozoa, Excavata, SAR and Archaeplastida are shaded in purple, blue, brown, pink and green filled rectangles, respectively.

4.2.3 Core nuclear envelope proteins

Of the 45 NE proteins analyzed, 22 of them are found in at least one organism in each of the eukaryotic supergroups. These 22 proteins constitute the core nuclear envelope proteome that was probably part of the LECA. Of note, a significant number of proteins that are involved in chromatin organization, NE homeostasis, gene regulation and transport are part of the core proteome. The proteins that are involved in chromatin organization and nuclear envelope homeostasis are also important for maintaining the nuclear architecture in yeast (Wright et al. 1988; Hodge et al. 2010; Horigome et al. 2011; Rothballer and Kutay 2013; Schreiner et al. 2015). Interestingly, we find that two SUN domain proteins viz. the C-terminal (Mps3) and mid-SUN (Slp1) domain proteins are part of the core proteome (Figure 12). The mid-SUN domain proteins have expanded in plants and contribute to maintenance of nuclear morphology in *A. thaliana* (Graumann et al. 2014). Thus, the origin and evolution of the two SUN domain families, appears to predate LECA. This suggests that LECA possessed a sophisticated nuclear envelope proteome that mediated various critical nuclear functions. The conservation of NE proteins that anchor chromatin and enzymes that modulate transcription factors suggests that the function of NE as a key architectural component in gene regulation is ancient and potentially existed in LECA. Similarly the presence of the ubiquitin and sumoylation components at the NE suggests an evolutionarily conserved mechanism to maintain nuclear protein homeostasis.

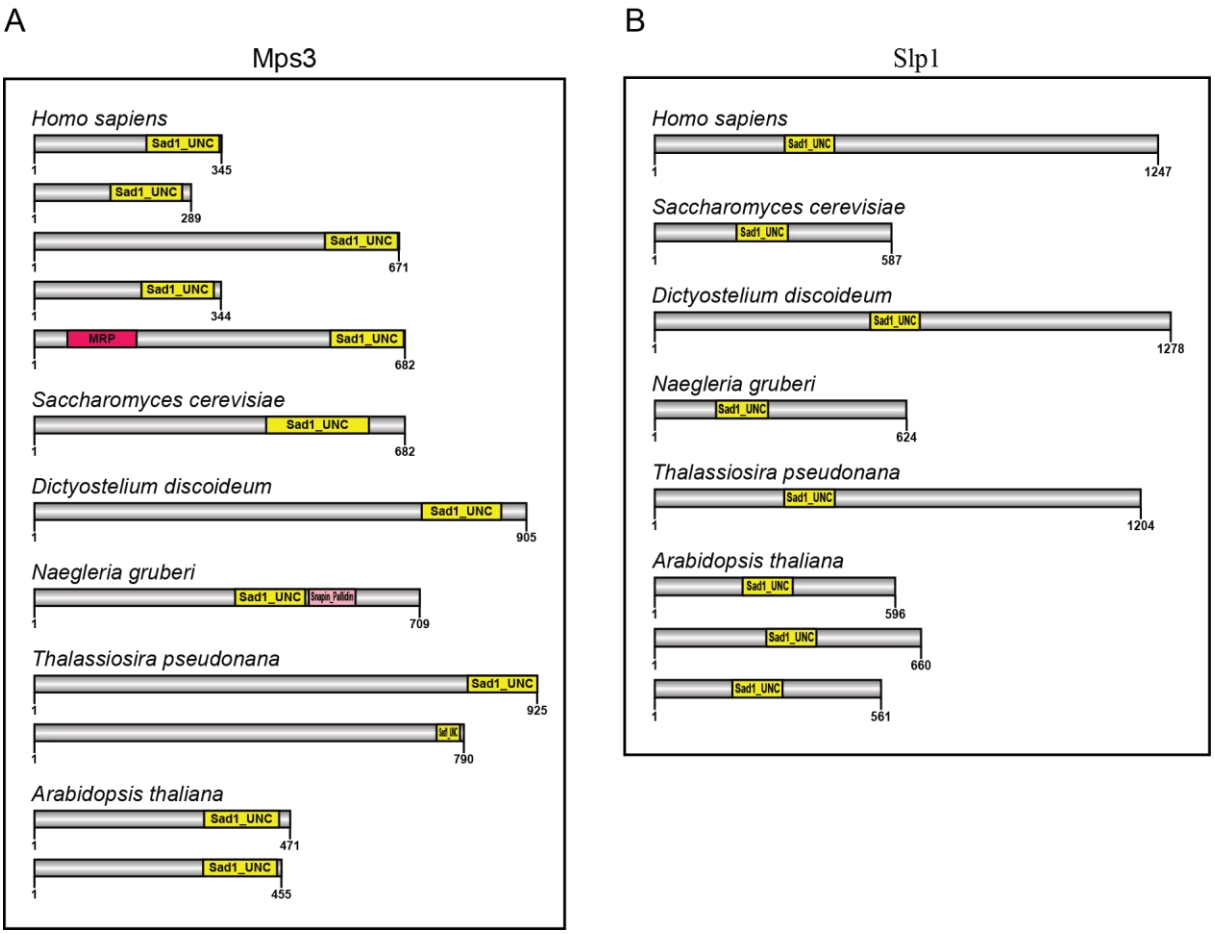

Figure 12: Domain organization in the SUN domain proteins. The domain architectures of the homologs of SUN domain containing proteins A) Mps3 (C-terminal SUN) and B) Slp1 (mid-SUN) are shown in representative organisms. Sad1_UNC domain is shown in yellow. The additional MRP domain found in Mps3 homolog of human is shown in dark pink.

The homologs of 10 proteins could not be obtained across all eukaryotic supergroups (Figure 13). These are termed the non-linearly conserved proteins.

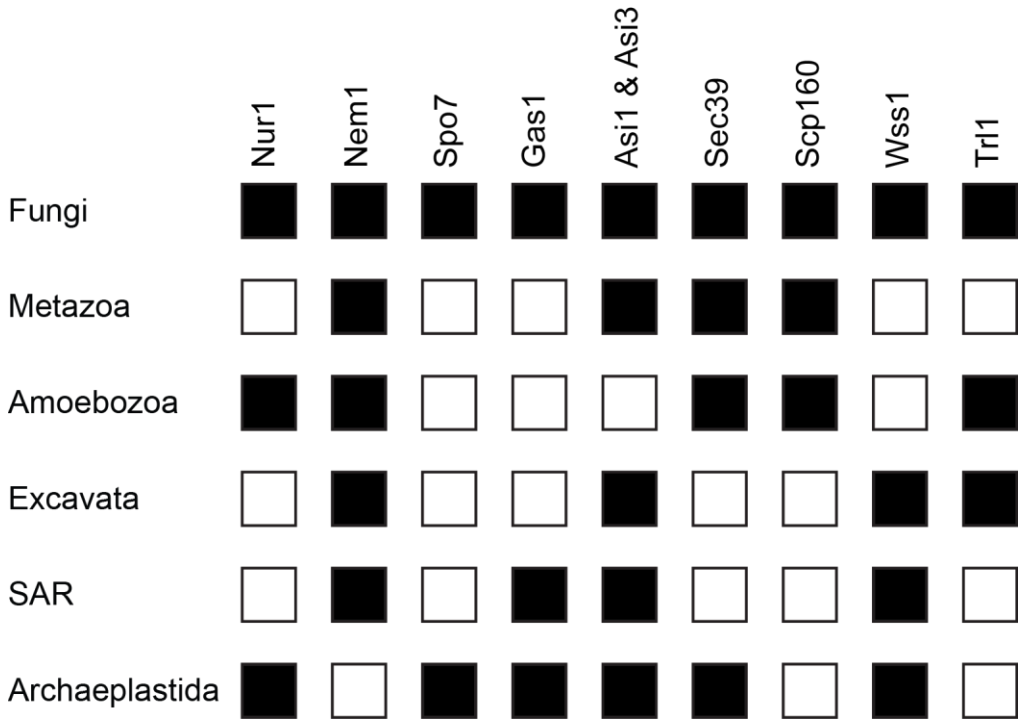

Figure 13: Non-linearly conserved proteins. The occurrence of non-linearly conserved proteins across supergroups is depicted. The black filled rectangles represent the presence of the protein in the respective supergroup. Fungi and Metazoa, belonging to Opisthokonta, are shown separately.

Two of these proteins, Nem1 and Wss1 are found in four of the supergroups and could not be detected only in Archaeplastida and Ameobozoa supergroups, respectively. We speculate that these two proteins were probably present in LECA and were lost in some lineages later. Among the other non-linearly conserved proteins, Gas1 homologs are predominantly found in fungi and are found only in *P. tricornutum* and *Z. mays* outside fungi. The presence of these homologs outside fungi is possibly due to an HGT event. Similarly, Nur1, which is present in fungi and Amoeba, which are unikonts, is found only in one organism in bikonts viz. *C. paradoxa*. Thus, 24 proteins that are present in all or at least four supergroups probably were constituents of the LECA nuclear envelope proteome (Figure 14).

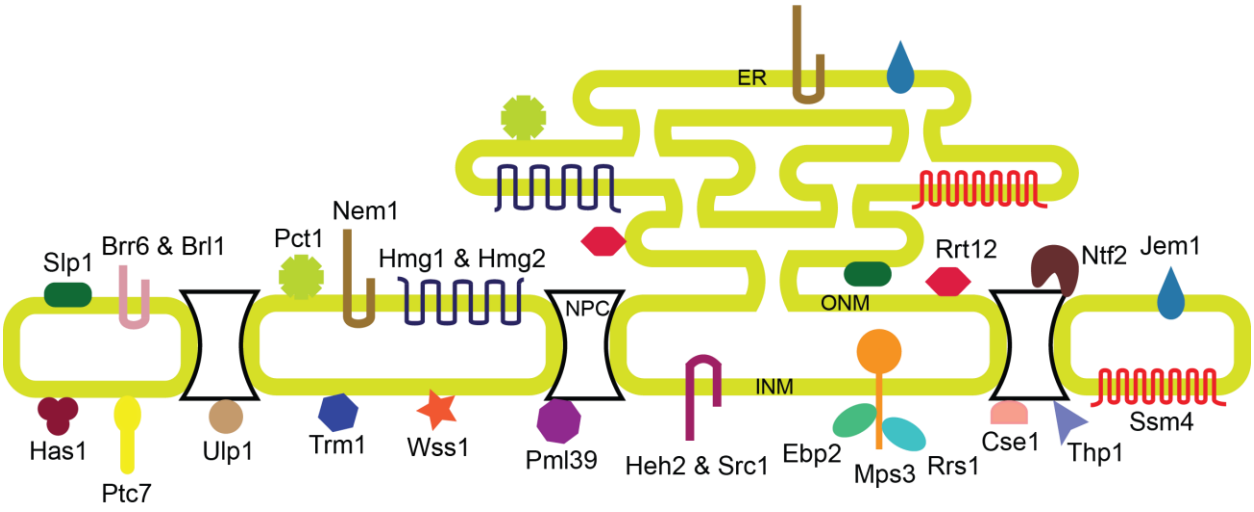

Figure 14: LECA nuclear envelope proteome. A pictorial representation of the LECA nuclear envelope proteome. The core nuclear envelope proteins identified and the non-linearly conserved proteins present in at least 4 eukaryotic supergroups are represented in different shapes and colors based on their available localization data in *S. cerevisiae*. The proteins present at ER-ONM network are shown both at the ER and ONM.

4.2.4 Fungal specific NE proteins

Out of the 45 proteins used for query, 13 are found only in the fungal kingdom, the ascomycetes, and among them 10 are found only in Saccharomycetes. Among the Saccharomycetes specific proteins, a majority is rapidly evolving and the homologs share very low sequence similarity amongst them. We further analysed these sequences to see if there were any conserved motifs, and identified short conserved motifs in three of them. One motif each was identified in the N-terminal region of Esc1 protein, C-terminal of Nvj1 and Prm3 proteins (Figure 15). These motifs were used to mine homologs in other fungi; however, no additional homolog could be identified outside Saccharomycetes.

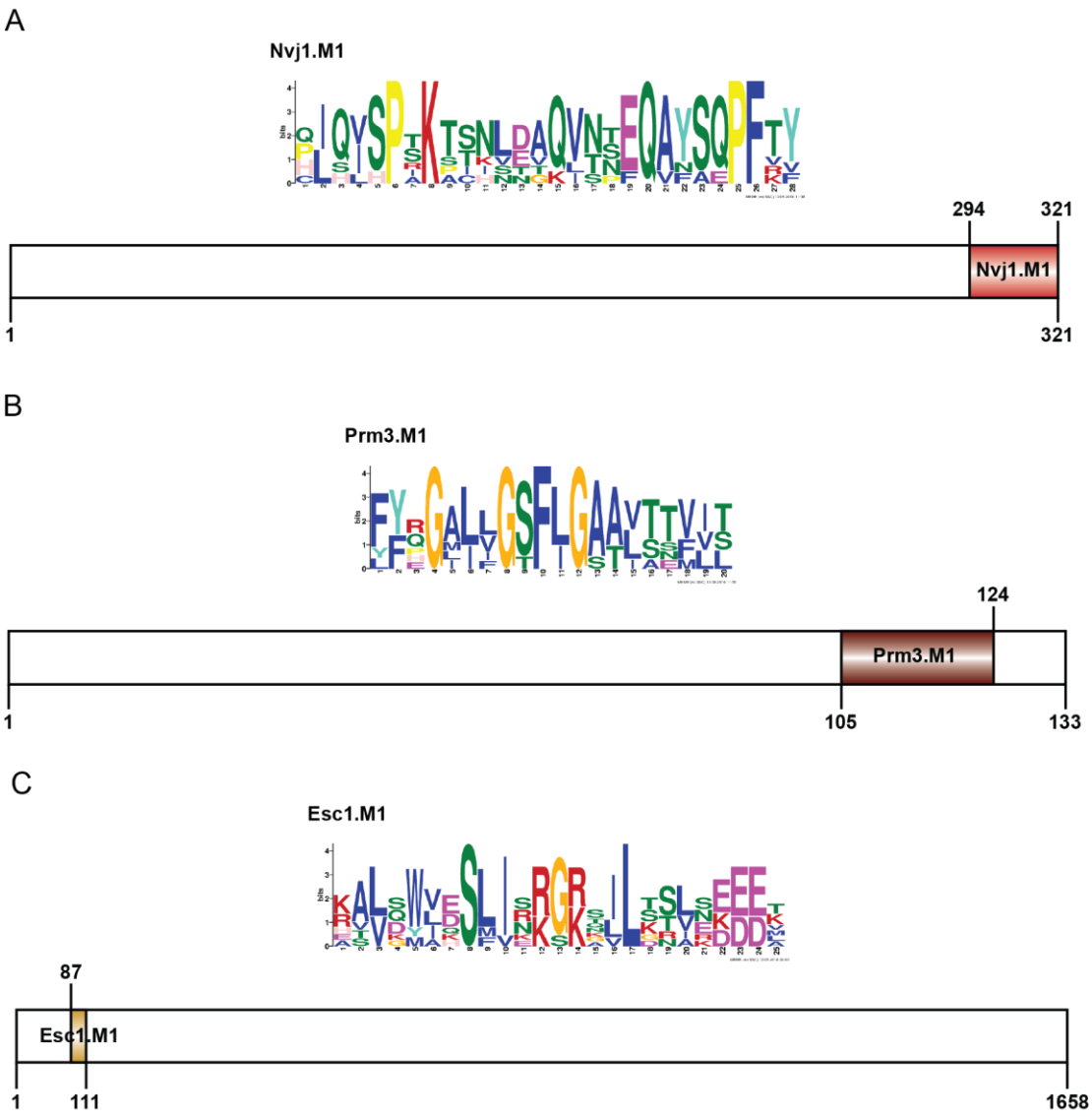

Figure 15: Motifs identified in Saccharomycetes specific proteins. The sequence logos of the motifs identified in A) Nvj1, B) Prm3, and C) Esc1 are shown. Shown below is the scaled map of the *S. cerevisiae* protein indicating the position of the motif.

The identified motifs coincide with regions experimentally tested for function in the three proteins. Nvj1 forms nucleus-vacuole junctions through its interaction with Vac8 and promotes piecemeal microautophagy of the nucleus (Roberts et al. 2003). The motif identified overlaps with the region that was earlier shown to be sufficient and necessary for interaction with Vac8

(Pan et al. 2000) suggesting that this function is likely conserved across Saccharomycetes. Prm3 protein plays an important role in nuclear fusion event, which is the final step in yeast mating pathway. The motif identified is part of the region that was shown to be important for stability, localization and function of this protein (Shen et al. 2009). While most of these genes are non-essential for the survival of yeast, only Nbp1, which is required for the insertion of spindle pole body into the nuclear membrane, is essential. The presence of around 20% of NE proteome unique to Saccharomycetes is an indication of the fast evolving nature of the nuclear envelope proteome.

4.3 Conclusions

NPCs and a subset of NE proteins have been shown to be present in LECA. However, to date a comprehensive analysis of the NE proteome across a wide range of organisms has not been done. Using a comparative genomics approach we identified the core nuclear envelope proteins that are present across all eukaryotic supergroups, the non-linearly conserved and the fungal specific NE proteins. A significant number of proteins involved in chromatin organization, nuclear envelope homeostasis, gene regulation and transport are found to be part of the core proteome, suggesting that they are conserved NE functions that were present in LECA. As more experimental data from diverse organisms becomes available, this study along with other similar studies will help in understanding the origin and the evolution of the nucleus.

Chapter 5

Phylogenetic analysis of nuclear envelope proteins

5.1 Introduction

The evolutionary relationship between species can be inferred by performing phylogenetic analysis of homologous sequences. Understanding the evolutionary relationship between the extant eukaryotes is useful for inferring the 1) genes or functions that are present in the ancestral eukaryote as well as those that have been gained in specific lineages, 2) shared similarities and differences between the organisms, 3) origin of characteristics related to the development or the physiology of the organisms, etc. Previous studies using several conserved proteins have shown the phylogenetic relationship between the supergroups and within the supergroups (discussed in the Introduction). However, the position of some of the lineages still remains controversial. From the comparative analysis of nuclear envelope proteins, we identified 22 proteins that are conserved across all eukaryotic supergroups. Using the conserved proteins, we performed maximum likelihood analysis to determine the evolutionary relationship between the organisms chosen in this study and compared it with the reference tree shown in Figure 3. For constructing ML trees, we used proteins whose homologs are present in a large number of organisms considered and which have not undergone extensive duplication.

5.2 Results

Among the proteins identified to part of the core proteome, the proteins Ebp2, Rrs1, Hmg1, Hmg2, Pct1, Cse1, Ntf2, Trm1, and Slp1 are present in a number of organisms, show very high conservation and have not duplicated extensively. Maximum likelihood trees were constructed with the above proteins.

5.2.1 Ebp2

The nuclear envelope proteins Ebp2 and Rrs1 are involved in ribosome biogenesis and organization of chromatin. The protein Ebp2 is highly conserved and is found to have terminal duplications in few organisms (*S. purpuratus*, *P. graminis*, *P. tetraurelia*, *E. histolytica*, *G. max*, *Z. mays* and *T. vaginalis*) (Figure 16). The ML tree generated using the homologs of this protein shows a split between the unikonts (opisthokonts and amoebozoa) and the bikonts (excavates, SAR and archaeplastidans). Monophyly could be recovered for the major lineages such as Fungi, Metazoa and Viridiplantae. Within SAR, the clades ciliates and stramenopiles could be recovered. Among excavates, the kinetoplastids and the heterolobosean *N. gruberi* are obtained as sister groups in accordance with the reference tree. However, none of the supergroups could be recovered as monophyletic groups. Within the unikonts, monophyly of amoebozoa could not be recovered as *E. histolytica* was found within the bikonts. Fungi are found closely related to the metazoa and mycetozoa. The sister relationship between stramenopiles and alveolates was not recovered. The two rhodophytes do not cluster together and are found with the alveolates. This is probably due to the extensive diversification of *C. merolae* homolog (Figure 16).

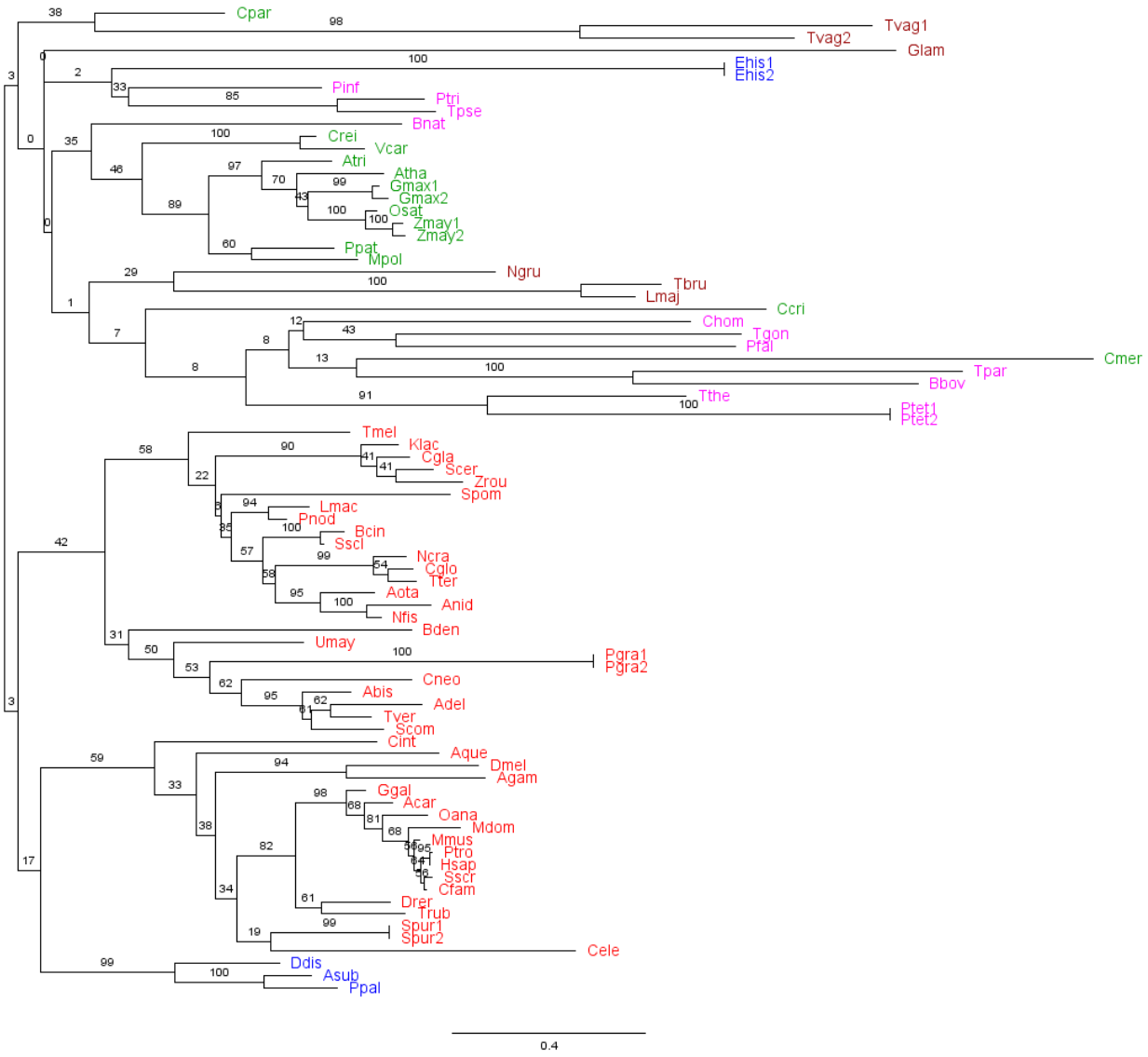

Figure 16: Maximum likelihood tree of Ebp2 protein homologs: Phylogenetic tree is constructed using the homologs of Ebp2 identified across eukaryotes. The 157 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.1. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.2 Rrs1

The maximum likelihood tree of Rrs1 homologs recovered the monophyly of metazoa and the sister relationship between fungi and metazoa (Figure 17). Within fungi, the relationship between the major phyla viz., Ascomycota, Basidiomycota and Chytridiomycota could be recovered. Within Amoebozoa, the clade mycetozoa was recovered and among the bikonts monophyly of the lineages Viridiplantae and Alveolata was observed. However, the fast-evolving microsporidian, *E. intestinalis,* was not found to cluster within fungi and was found with the excavate *G. lamblia*. The sister relationship between the mycetozoa and the archameoba *E. histolytica* was not observed. Similar to Ebp2, the two rhodophytes did not cluster together and were not found close to Viridiplantae.

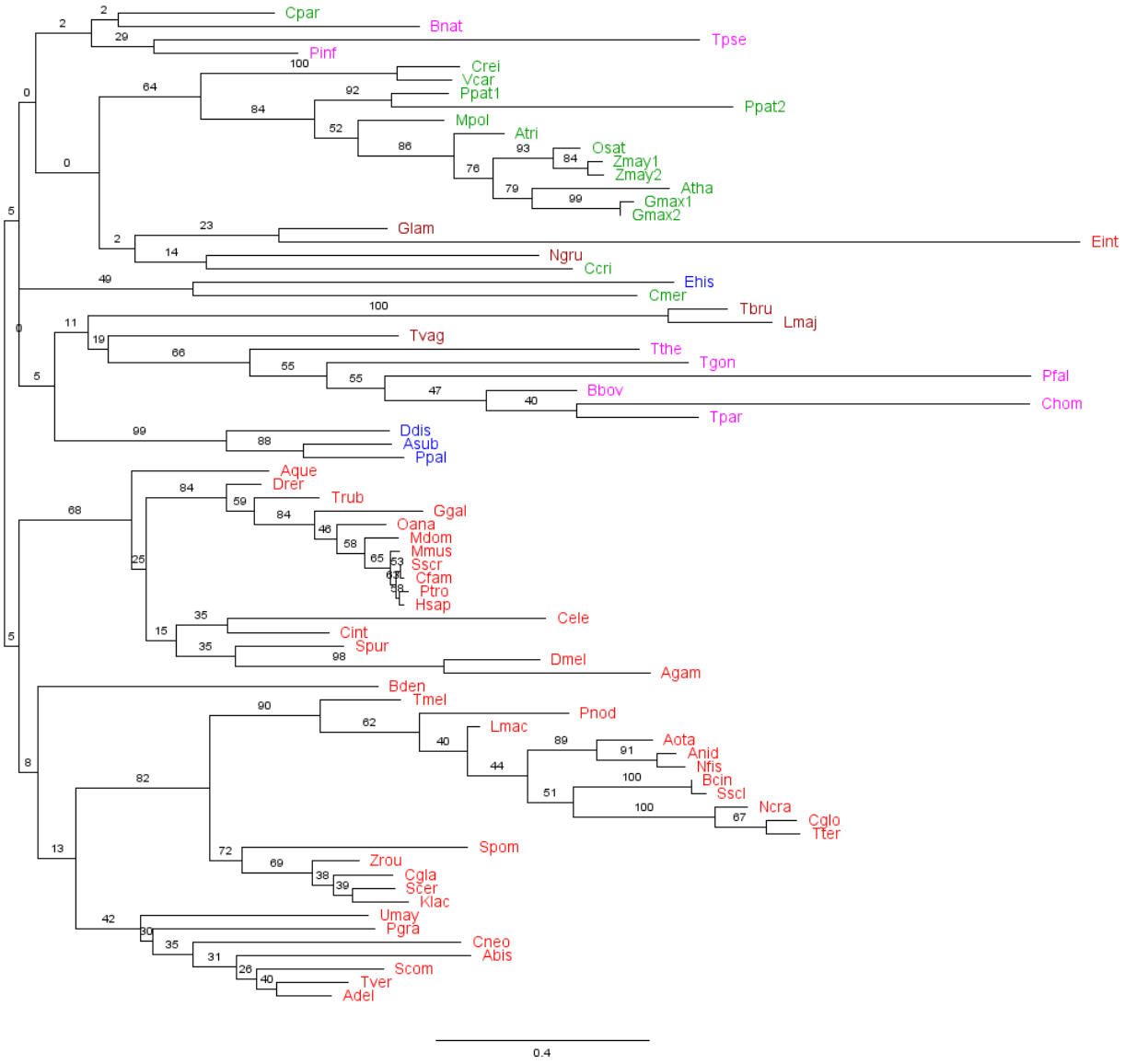

Figure 17: Maximum likelihood tree of Rrs1 protein homologs: Phylogenetic tree is constructed using the homologs of Rrs1 identified across eukaryotes. The 126 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.2. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.3 HMG-CoA reductases

The core proteins Hmg1, Hmg2 are involved in maintaining NE homeostasis. Hmg1 and Hmg2 are paralogous genes that arose by a duplication event in *S. cerevisiae*. Both the paralogs were considered for analysis. The homologs identified are found to have siginificant sequence similarity with Hmg1/Hmg2 and duplication is observed in Viridiplantae, and in the ancestor of mycetozoa (Figure 18). The maximum likelihood tree of HMG-CoA reductases recovered the monophyly of the fungi and metazoan lineages and the supergroup Opisthokonta. The glaucophyte *C. paradoxa* was found as sister group to Viridiplantae, and the monophyly of the supergroup Archaeplastida could be recovered. However, rhodophyta are not present as no homolog of HMG-CoA reductase could be identified in the red algae. Though part of the core proteome, the homologs of HMG-CoA reductases are absent from a number of organisms in SAR. The homolog found in the microsporidian *E. intestinalis* is found to have diverged extensively from the rest.

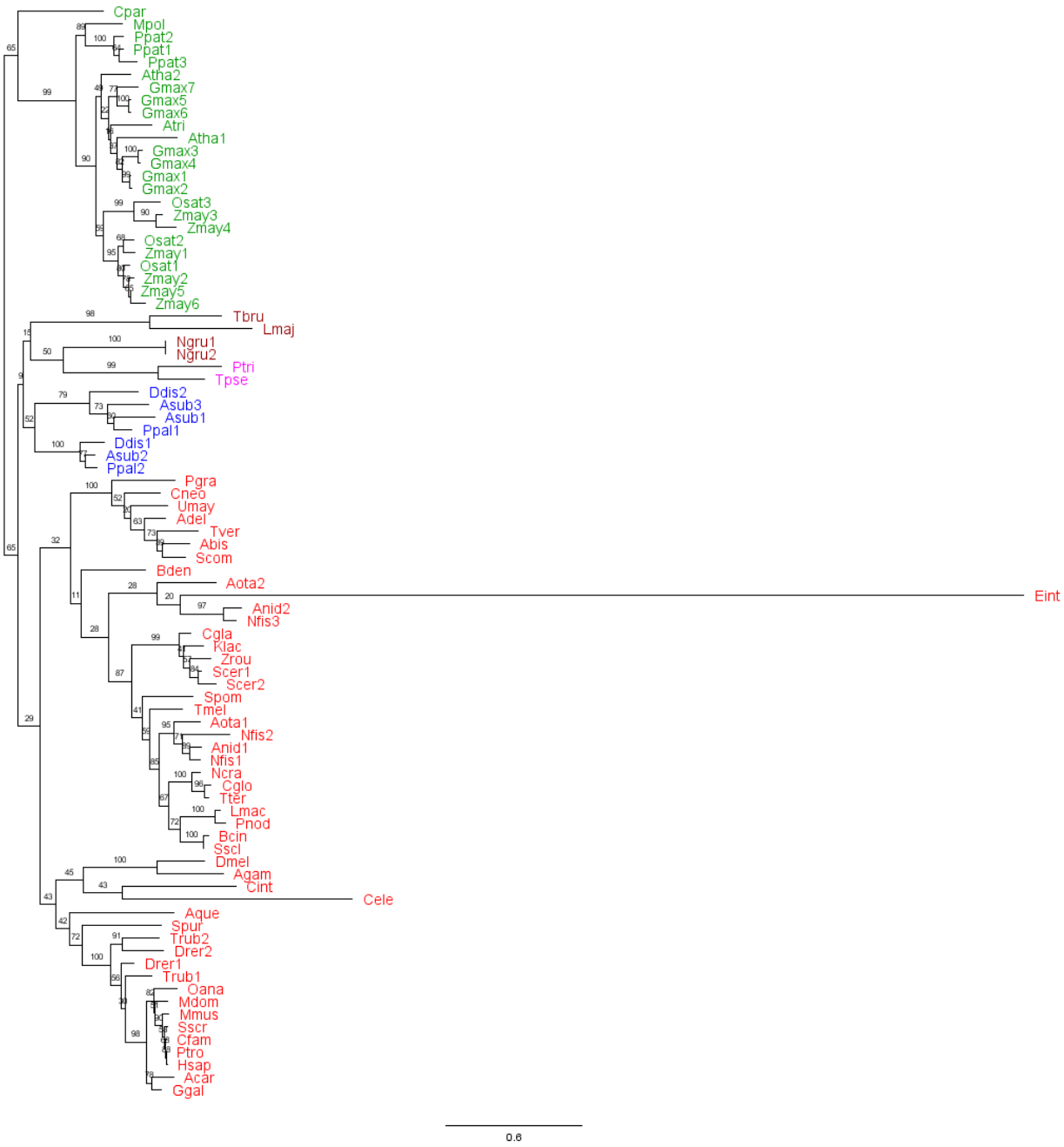

Figure 18: Maximum likelihood tree of HMG-CoA reductases: Phylogenetic tree is constructed using the homologs of Hmg1 & Hmg2 identified across eukaryotes. The 360 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.3. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.4 Pct1

The Pct1 gene that is involved in phosphatidylcholine synthesis is also part of the core proteome. This gene has undergone duplication in the ancestor of vertebrates. Terminal duplications are observed in *C. elegans, D. melanogaster* and *S. purpuratus* belonging to metazoa; *P. infestans* and *P. tetraurelia* belonging to SAR. The maximum likelihood tree with homologs of Pct1 protein recovered monophyly of metazoa and Streptophyta (Figure 19). The monophyly of fungi could not be recovered. The homologs in basidiomycetes have diverged extensively and the homologs in alveolates are found to share close similarity with the ascomycetes. The Pct1 homologs in kinetoplastids and the diatom have very long branches indicating that they evolved rapidly.

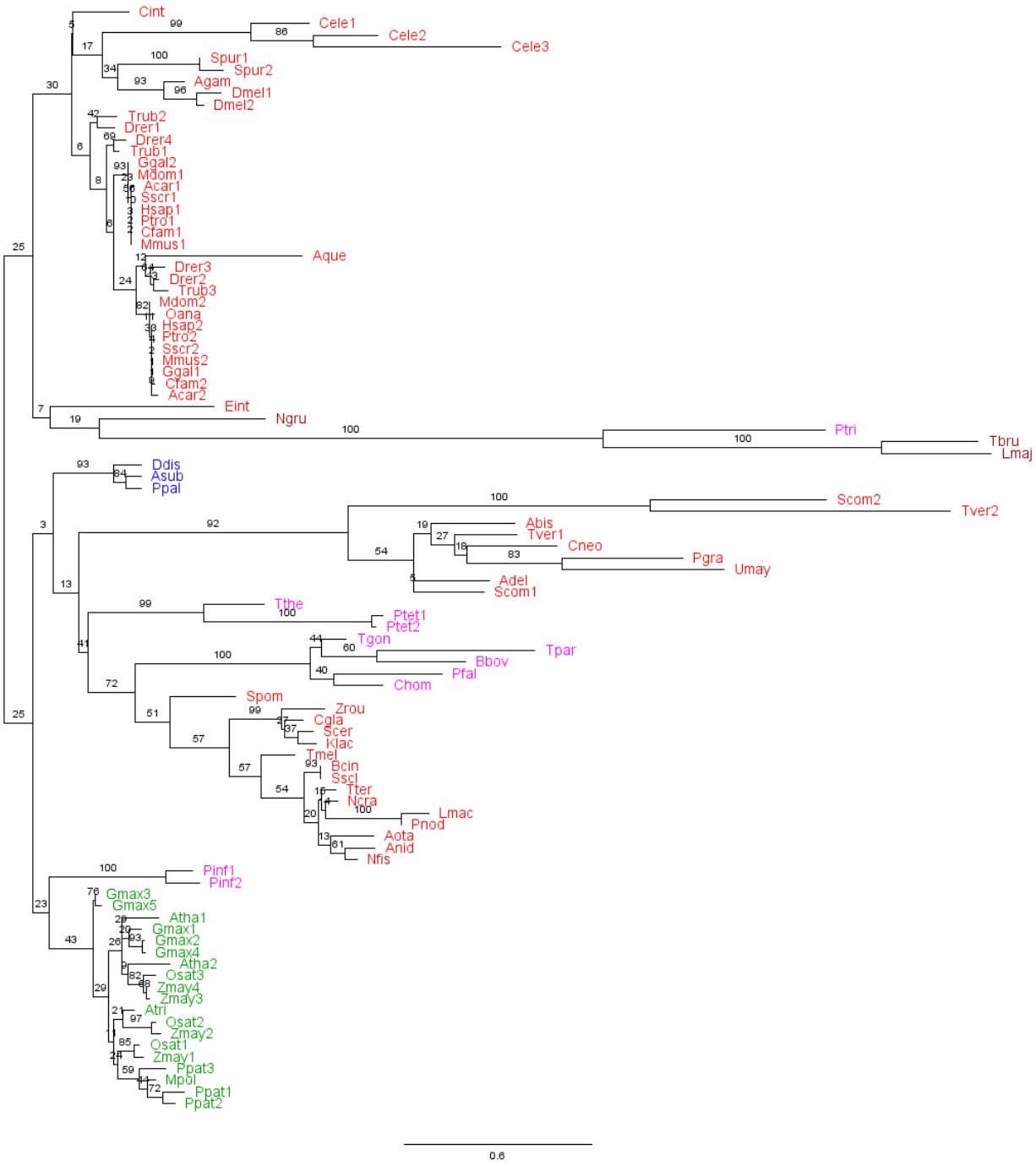

Figure 19: Maximum likelihood tree of Pct1 protein homologs: Phylogenetic tree is constructed using the homologs of Pct1 identified across eukaryotes. The 147 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.4. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.5 Cse1

The protein Cse1 involved in nucleocytoplasmic transport is part of the core proteome. The homologs of this protein are found in almost all the eukaryotes and share significant homology. The ML tree generated with this protein recovered the lineages fungi and Viridiplantae as monophyletic groups and the sister relationship between fungi and metazoa (Figure 20). Within fungi, the relationship between the organisms considered was recovered in accordance with the reference tree, except for *S. pombe*. However, none of the supergroups could be recovered. The metazoan homologs cluster together except for the homolog in *C. elegans*. The ciliates, the diplomonad, *G. lamblia* and the archameoba, *E. histolytica* have long branches and are found to cluster together probably due to long-branch attraction. Stramenopiles are found as sister group to the clade containing glaucophyte and Viridiplantae. Within Archaeaplastida, the two red algae cluster together but are not found to cluster with glaucophytes and Viridiplantae. Mycetozoa are closely associated with excavates (kinetoplastids and *N. gruberi*).

Figure 20: Maximum likelihood tree of Cse1 protein homologs: Phylogenetic tree is constructed using the homologs of Cse1 identified across eukaryotes. The 514 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.5. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.6 Ntf2

The homologs of Ntf2 protein are found in most of the organisms considered in the study and are found to have duplicated in a number of organisms belonging to metazoa and Viridiplantae. The ML tree of Ntf2 protein did not recover monophyly of any of the eukaryotic supergroups or the monophyly of any of the major eukaryotic lineages (Figure 21).

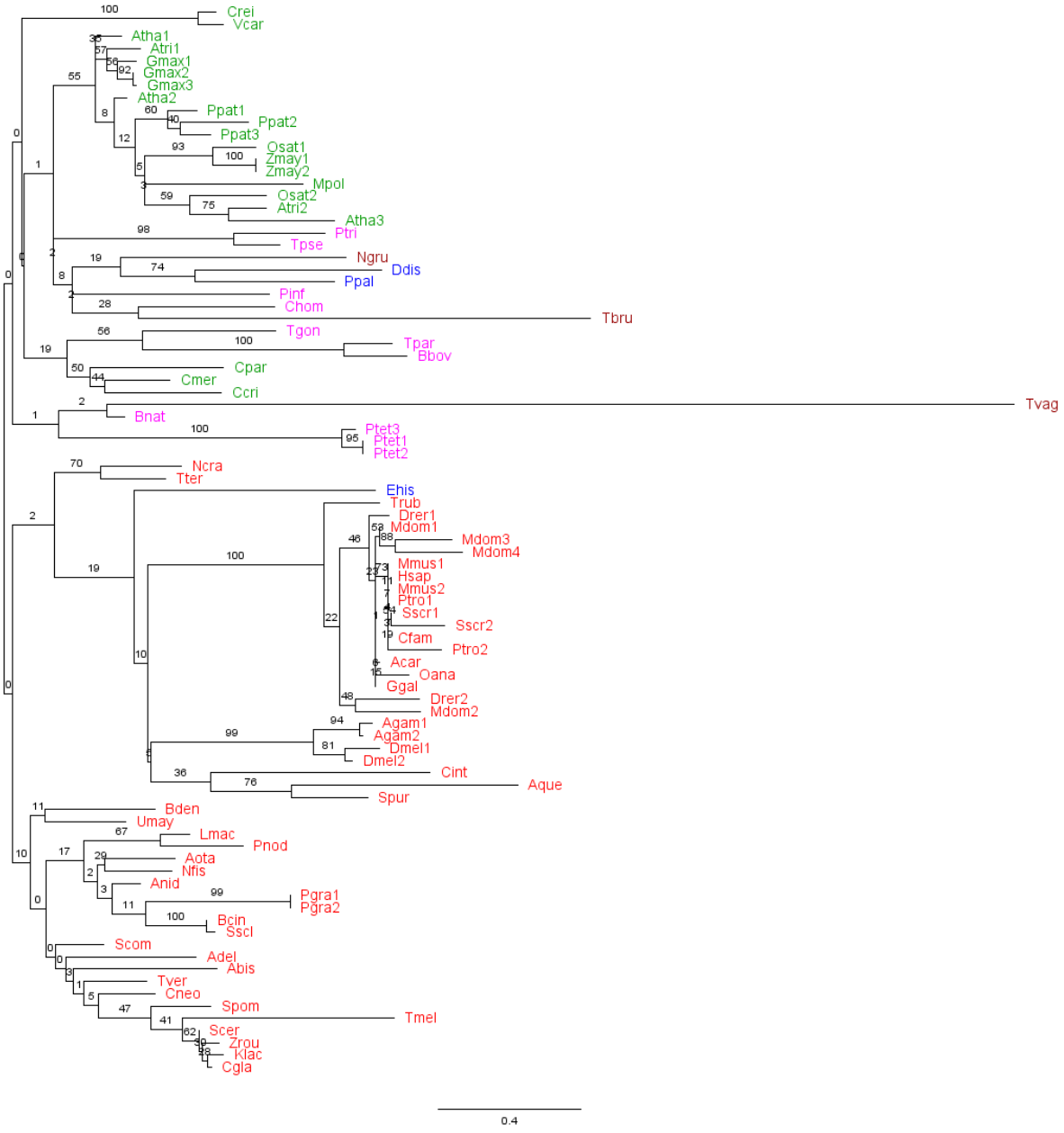

Figure 21: Maximum likelihood tree of Ntf2 protein homologs: Phylogenetic tree is constructed using the homologs of Ntf2 identified across eukaryotes. The 92 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.6. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.7 Trm1

The homologs of tRNA methyl transferase gene are found in all eukaryotic supergroups. More than one homolog of this protein is found in organisms belonging to Archaeplastida supergroup. Some of the homologs of Trm1, found in all organisms across Archaeplastida, though they returned the Trm1 protein with significant E-value in rBLAST, are found to have diverged extensively from the rest of the homologs. In the ML tree constructed with the homologs of Trm1 protein, we find that these homologs of Trm1 found in Archaeplastida and in *B. natans* cluster separately from the rest of the homologs (Figure 22).

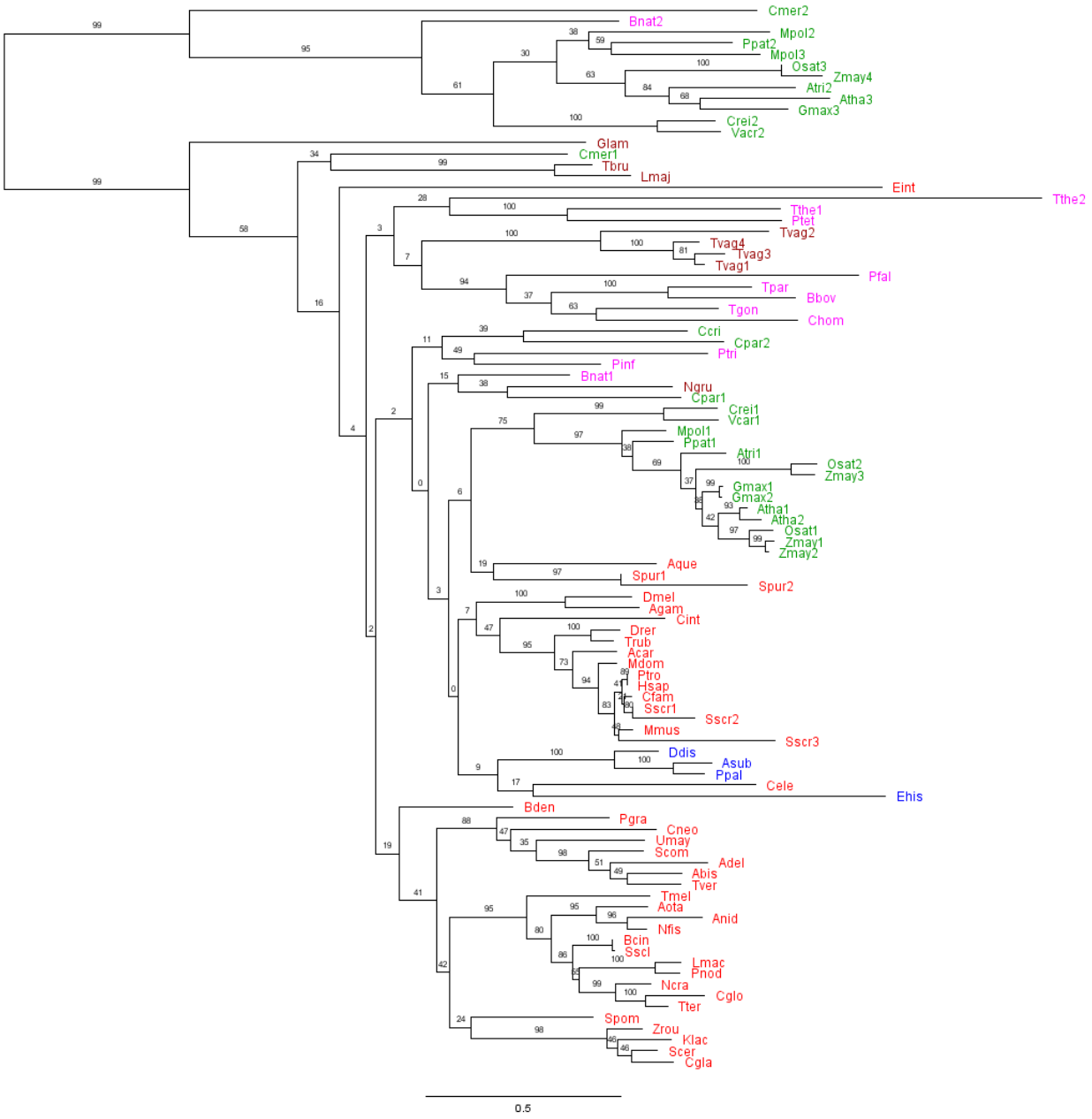

Figure 22: Maximum likelihood tree of Trm1 protein homologs: Phylogenetic tree is constructed using the homologs of Trm1 identified across eukaryotes. The 255 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.7. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.2.8 Slp1

The mid-SUN domain protein Slp1 is found across all eukaryotes and is found to have duplicated in Viridiplantae. The ML tree constructed with the homologs of this protein recovered the split between unikonts and bikonts. Monophyly could be recovered for fungi and Viridiplantae. Amoebozoa are found as sister group to metazoa and the Archamoeba *E. histolytica* is found within the metazoa. The three major lineages of the SAR viz., stramenopiles, alveolates and Rhizaria are found to cluster together. However, the kinetoplastids that belong to Excavata are also found closely associated with ciliates (Figure 23).

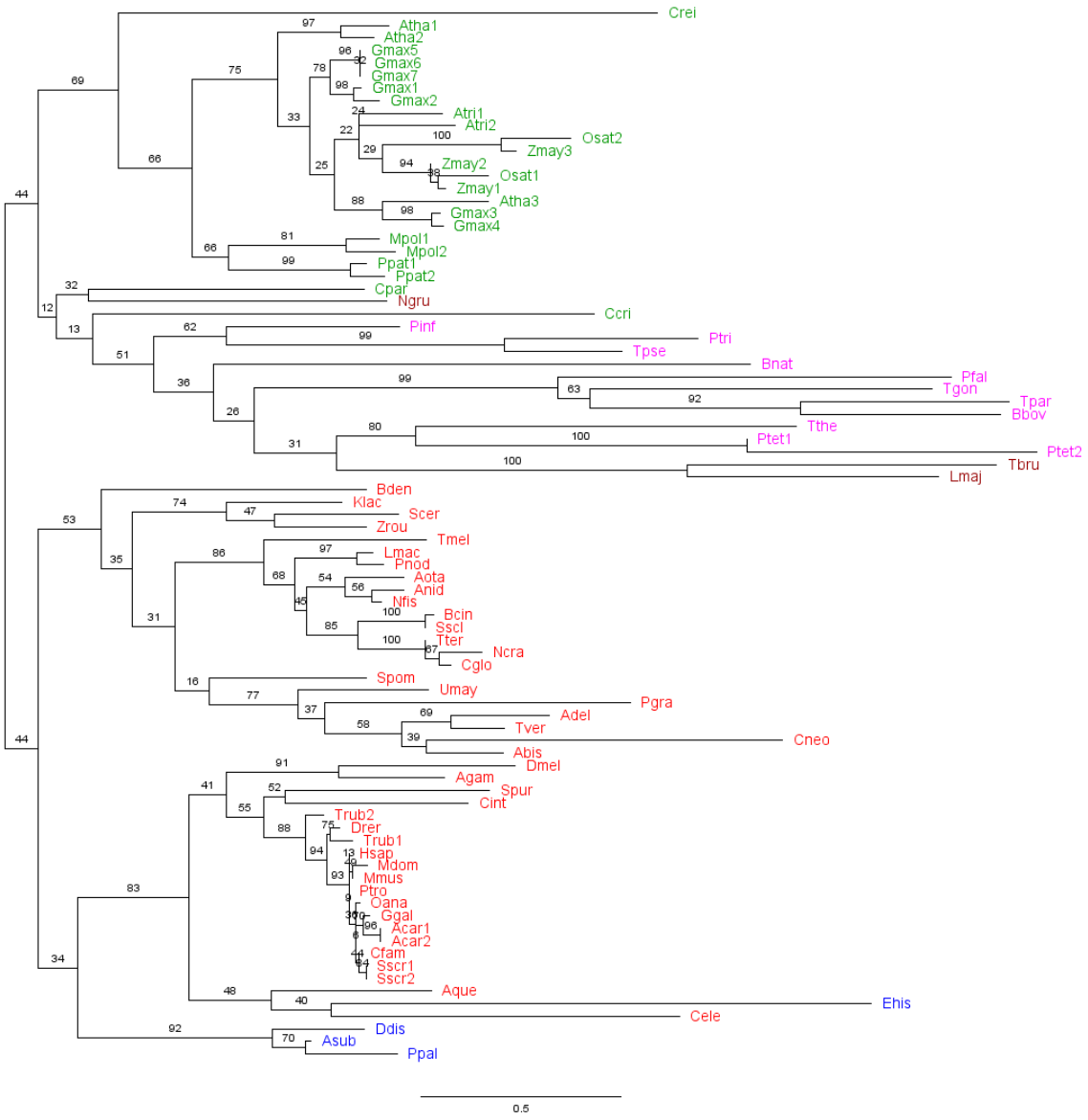

Figure 23: Maximum likelihood tree of Slp1 protein homologs: Phylogenetic tree is constructed using the homologs of Slp1 identified across eukaryotes. The 132 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.8. The five eukaryotic supergroups are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green).

5.3 Conclusions

Phylogenetic analysis has been carried out for a subset of the nuclear envelope proteins identified across all eukaryotic supergroups. The maximum likelihood trees generated recovered monophyly of the major lineages such as those of Fungi, Metazoa and Viridiplantae. The close relationship between fungi and metazoa was found in the ML trees constructed from Rrs1, Hmg1 & Hmg2, Cse1 and Ntf2 proteins. All the organisms belonging to Excavata are never found together possibly due to their fast evolving nature. Within Amoebozoa, the mycetozoa are found as a monophyletic group, but the Archamoeba, *E. histolytica* is never found as a sister group to mycetozoa. Similarly, in case of SAR supergroup, though, the organisms belonging to alveolates and stramenopiles cluster together, they are not often found as sister groups, except for in the ML tree of Slp1 protein. The lack of correlation between the reference tree and the ML trees generated with each of these proteins is probably due to the fast-evolving nature of the organisms such as *G. lamblia*, *E. intestinalis* included in the study. In previous studies, removal of some problematic taxa such as the ciliates, microsporidia, foraminifera etc. have shown to improve the support for some clades (Yoon et al. 2008). Additionally, horizontal gene transfer events can also lead to artifacts in the phylogenetic trees. A recent study showed that a number of genes in the Entamoeba are gained by HGT and it has gained genes from the Parabasalid *T. vaginalis* as well (Grant and Katz 2014). Phylogenetic analysis by concatenating several conserved genes and by sampling multiple taxa from each of the supergroups will probably be able to resolve the eukaryotic tree with better support and consistency.

Chapter 6

Prokaryotic origins of nuclear envelope proteins

6.1 Introduction

Understanding the origins and evolution of several nuclear associated systems would provide insights into the origin of the nucleus. In this context, extensive comparative studies have been carried out for nuclear pore complex proteins, nucleolar proteins and few structural nuclear envelope proteins. The composition and the overall architecture of nuclear pore complexes are conserved across all eukaryotes (DeGrasse et al. 2009; Neumann et al. 2010). The nuclear pore complex proteins, though not sharing significant similarity at the sequence level, are found to have conserved their secondary structures (DeGrasse et al. 2009). Thus LECA is considered to have possessed fully functional NPCs. Interestingly, it has been found that the scaffold Nups contain β-propeller and/or α-solenoid folds, which are also found in vesicle coating complex proteins (Devos et al. 2004). Both the scaffold Nups and the vesicle coating complexes are important for bending the membrane and stabilizing their curvature. No sequence homolog of nuclear pore complex proteins could be identified in either bacteria or archaea so far. Intriguingly, proteins sharing similar architecture to that of the scaffold NPCs are found in PVC superphylum bacteria (Santarella-Mellwig et al. 2010). Similarly, comparative studies with the structural nuclear envelope proteins have shown their presence in LECA and failed to identify any prokaryotic homologs. However, the HeH domain found in the yeast INM proteins Heh2 & Src1 is shown to be of bacterial origin and the SUN domain is shown to be related to the discoidin domain (F5_F8_type_C), which is widespread in bacteria (Mans et al. 2004). In contrast, the nucleolar components such as small-subunit (SSU) rRNA processing proteins are shown to be of archaeal origin (Feng et al. 2013). Earlier studies also showed that the information processing genes such as those involved in replication, transcription etc. are of archaeal origin, while the operational genes such as those involved in metabolic pathways are of bacterial origin (Thiergart et al. 2012).

In order to gain a better understanding of origins of the components of the nuclear membrane, we looked for the homologs of the core and non-linearly conserved nuclear envelope proteins (identified in Chapter 4) in bacteria and archaea.

6.2 Results

6.2.1 Prokaryotic Homologs of core and non-linearly conserved proteins

The comparative sequence analysis of 45 nuclear envelope proteins of *Saccharomyces cerevisiae* identified 22 proteins to be part of the core proteome and 10 non-linearly conserved proteins (Chapter 4). In order to trace the origins of the nuclear envelope proteins, we looked for the homologs of the 22 core and 10 non-linearly conserved NE proteins in bacteria and archaea as described in the Methods section. Of the 32 NE proteins, homologs of 4 of them were identified in prokaryotes, while for 13 of them, proteins sharing the characteristic domains could be identified. The remaining 15 NE proteins appear to be eukaryotic innovations (Figure 24). The NE proteins with prokaryotic homologs include the Hmg1, Hmg2, Trm1 and Ntf2. The homologs of HMG-CoA reductase, tRNA methyl transferase proteins are found in both archaea and bacteria with significant E-value, while those of Ntf2 are found only in Streptomycetaceae bacteria.

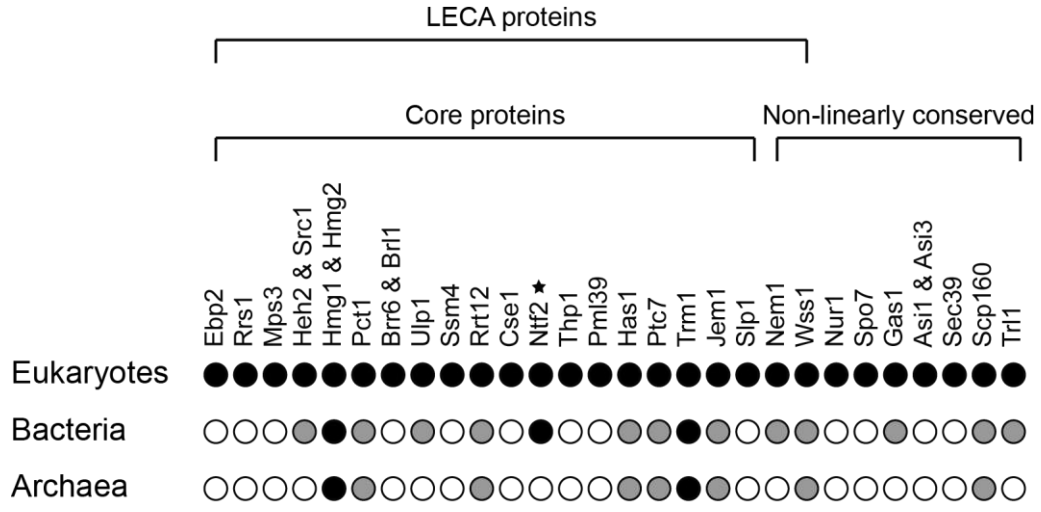

Figure 24: Prokaryotic homologs of nuclear envelope proteins. The core and non-linearly conserved nuclear envelope proteins whose homologs or characteristic domains are found in bacteria and/or archaea are shown. The black filled circles indicate the presence of a homolog and the grey filled circles represent the presence of the characteristic domain. The asterisk above the Ntf2 is used to represent the presence of the homologs only in a particular family of bacteria.

6.2.2 HMG-CoA reductases

S. cerevisiae contains two paralogous HMG-CoA reductases, Hmg1 and Hmg2. HMG-CoA reductases catalyse the conversion of HMG-CoA to mevalonate in the sterol biosynthesis pathway (Brown and Goldstein 1980). The homologs of the yeast HMG-CoA reductase proteins Hmg1 & Hmg2 are found across all eukaryotic supergroups. However, within the SAR and Excavata supergroups, though we retrieved hits in the hmmsearch analysis, a homolog sharing significant sequence similarity (E-value less than 10^{-5}) to Hmg1/Hmg2 could not be identified in *P. tetraurelia* (Alveolates), *P. infestans* (Stramenopiles) within SAR, *T. vaginalis* (Parabasalids) within Excavata, amongst others. The domain analysis of the hits found in these organisms showed that they contained a class II HMG-CoA reductase domain, while the homologs found in all other eukaryotes contained the class I HMG-CoA reductase domain.

The class I enzymes of mammals function in the conversion of HMG-CoA to mevalonate (Istvan and Deisenhofer 2000), whereas the class II enzyme of the bacteria, *Pseudomonas mevalonii* acts as a catabolic enzyme and converts mevalonate to HMG-CoA (Jordan-Starck and Rodwell 1989). However, both the class I and II enzymes are capable of catalyzing the conversion of HMG-CoA to mevalonate and the reverse reaction (Sherban et al. 1985). At the sequence and structural level, the class I and II enzymes are distinguished by the absence of a *cis*-loop in the catalytic domain of the class II enzymes. Additionally, the class II enzymes also lack the membrane anchor domain present in the class I proteins and are soluble in nature (Friesen and Rodwell 2004).

Earlier studies showed the presence of a class II HMG-CoA reductase in *G. lamblia* (Boucher and Doolittle 2000). Using the *P. infestans* sequence as the query in BLASTp analysis, we further identified the presence of class II proteins in several eukaryotes (Additional data 2). The

class II enzymes were found to be present in a number of anaerobic organisms such as *Trichomonas vaginalis* and *Tritrichomonas foetus* (parabasalids), *Giardia lamblia* (diplomonad), *Neocallimastix californiae* and *Piromyces finnis* (chytridiomycetes). All these organisms have reduced forms of mitochondria such as hydrogenosomes or mitosomes. The chytridiomycete *B. dendrobatidis* is found to have both class I and class II enzymes.

In accordance with the previous studies, we find both class I and II HMG-CoA reductases in both archaea and bacteria (Friesen and Rodwell 2004). Archaea have predominantly class I proteins, while the class II proteins were limited to Archaeoglobi and DHVE2 group of archaea. Intriguingly, all the Asgard archaea considered in the study were found to have class II proteins. In bacteria, both class I and II proteins were found across all major phyla (Additional data 2). In order to understand the phylogenetic relationship between the eukaryotic and prokaryotic homologs, we constructed maximum likelihood trees with all the eukaryotic homologs identified earlier (including the class II proteins found in hmmsearch) and selected homologs in archaea and bacteria (including the class II proteins). We find that the eukaryotic class I proteins branch as a sister group to the archaeal and bacterial class I proteins (Figure 25). All the class I proteins in eukaryotes cluster as a monophyletic group, except for the second homolog found in *A. queenslandica*. It is found to cluster with the class I bacterial proteins. The class II HMG-CoA reductases branch separately and contain the class II proteins of bacteria, archaea and some of the eukaryotes that belong to fungi, SAR and excavates. The presence of class II HMG-CoA reductases in eukaryotes belonging to three different supergroups suggests that it was possibly present in LECA and lost from several eukaryotes later. However, in the ML tree, the class II proteins of eukaryotes are not obtained as a monophyletic group, which rules out the possibility of a common origin for the class II enzymes in eukaryotes.

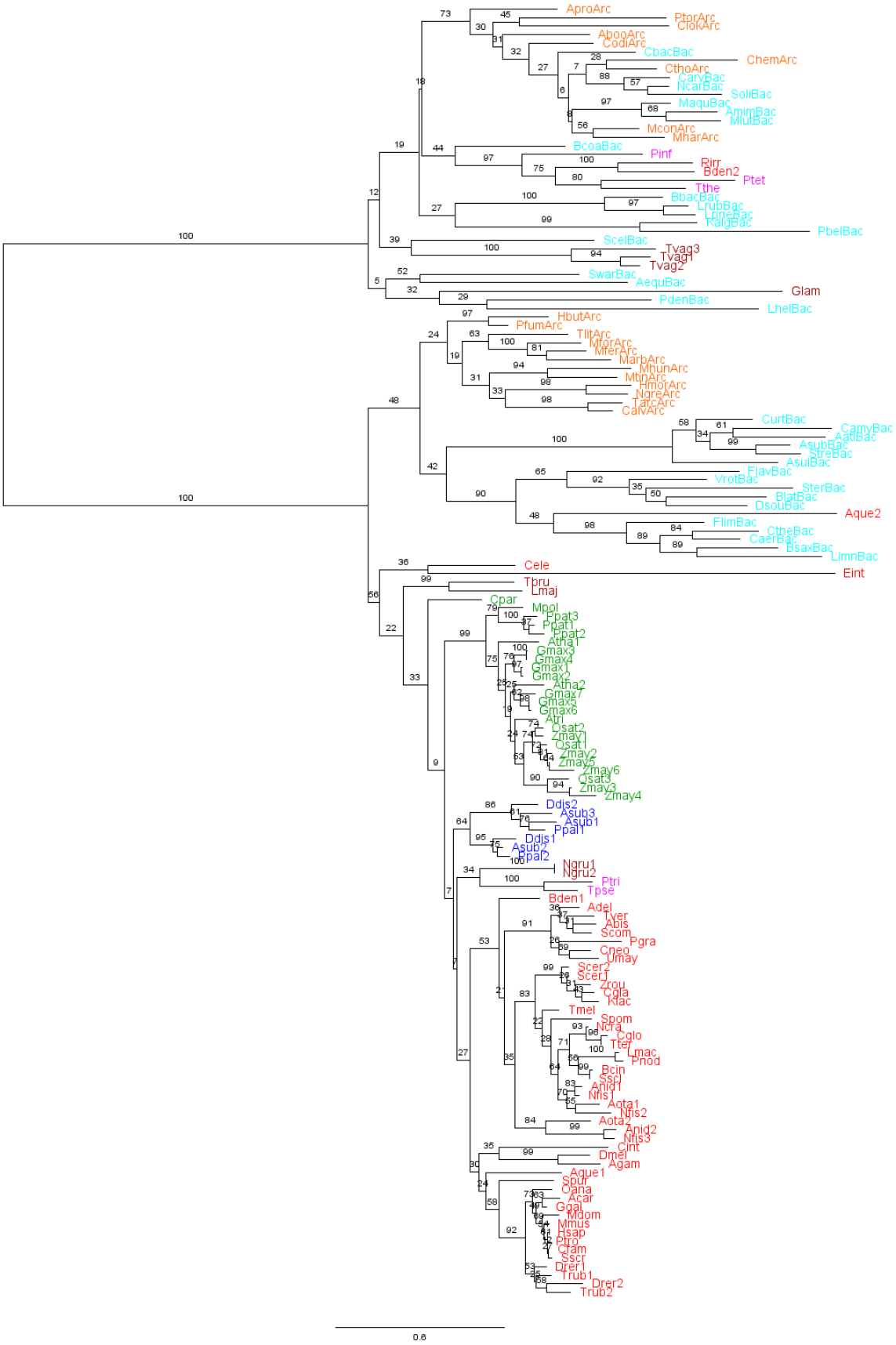
AproArc
PtorArc
ClorArc
AbooArc
CodiArc
CbacBac
ChemArc
CthoArc
CarvBac
NcarBac
SoliBac
MaquBac
AmimBac
MlutBac
MconArc
MharArc
BcoaBac
Pinf
Rirr
Bden2
Ptet
Tthe
BbacBac
LrubBac
LpneBac
RalgBac
PbelBac
ScelBac
Tvag3
Tvag1
Tvag2
SwarBac
AequBac
Glam
PdenBac
LhelBac
HbutArc
PfumArc
TlitArc
MtorArc
MrerArc
MarbArc
MhunArc
MthnArc
HmorArc
NgreArc
TacrArc
CalvArc
CurtBac
CamyBac
AatiBac
AsubBac
StreBac
AsuiBac
FlavBac
VrotBac
SterBac
BlatBac
DsouBac
Aque2
FlimBac
CtheBac
CaerBac
BsaxBac
LimnBac
Cele
Eint
Tpru
Lmaj
Cpar
Mpol
Ppat3
Ppat1
Ppat2
Atha1
Gmax3
Gmax4
Gmax1
Gmax2
Atha2
Gmax7
Gmax5
Gmax6
Atri
Csat2
Zmay1
Csat1
Zmay2
Zmay5
Zmay6
Csat3
Zmay3
Zmay4
Ddis2
Asub3
Asub1
Ppal1
Ddis1
Asub2
Ppal2
Ngru1
Ngru2
Ptri
Tpse
Bden1
Adel
Tver
Abis
Scom
Pgra
Cneo
Umay
Scer2
Scer1
Zrou
Cgla
Klac
Tmel
Spom
Ncra
Cglo
Tter
Lmac
Pnod
Bcin
Ssci
Anid1
Nfis1
Aota1
Nfis2
Aota2
Anid2
Nfis3
Cint
Dmel
Agam
Aque1
Spur
Cana
Acar
Ggal
Mdom
Mmus
Hsap
Ptro
Cfam
Sscr
Drer1
Trub1
Drer2
Trub2
0.6

Figure 25: Maximum likelihood tree of HMG-CoA reductases across eukaryotes and prokaryotes. The phylogenetic tree is constructed using both class I and class II proteins of eukaryotes, archaea and bacteria and includes 96 eukaryotic, 21 archaeal and 34 bacterial sequences. The 271 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.9. The five eukaryotic supergroups, bacteria and archaea are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green), Archaea (orange) and Bacteria (cyan).

6.2.3 tRNA methyl transferase

The Trm1 gene in yeast encodes for a tRNA methyltransferase which produces the N2,N2-dimethylguanosine modified base in tRNAs (Ellis et al. 1986). The homologs of this gene are found across all eukaryotes. In Archaeplastida supergroup, apart from the actual Trm1 homolog, we find an additional homolog across all organisms that shares less sequence similarity with the Trm1 protein. In prokaryotes, the homologs of Trm1 were found in almost all phyla of archaea, while in bacteria they are found only in Cyanobacteria and in few organisms belonging to the Aquificae and Armatimonadetes phyla (Additional data 3).

Phylogenetic analysis was performed using all the homologs of Trm1 gene obtained in eukaryotes and selected homologs from archaea and bacteria. In the ML tree, archaea are obtained as a sister group to eukaryotes suggesting vertical descent of Trm1 gene from archaea (Figure 26). The Trm1 homolog found in *A. aeolicus* bacteria belonging to the phyla Aquificae is found within the archaeal clade suggesting gain of Trm1 gene into Aquificae by horizontal gene transfer. Interestingly, the additional homolog found in plants and in *B. natans* with less similarity to Trm1, branch as sister group to cyanobacteria. This suggests that the Trm1 gene was gained from archaea by vertical descent into eukaryotes, while a significantly diverged variant of this gene was gained from cyanobacteria into the common ancestor of Archaeplastida (Figure 26). However, the additional homolog found in *C. merolae* is found to branch separately from the rest of plant homologs. This is probably due to extensive sequence divergence in the *C. merolae* homolog.


AbacBac
Cmer2
AcaspBac
SynspBac
LrosBac
HalspBac
SmajBac
CwatBac
MyxspBac
Mpol2
Ppat2
Mpol3
Atri2
Atha3
Gmax3
Zmay4
Osat3
Bnat2
Crei2
Vacr2
TgamArc
ClagArc
AbooArc
MlumArc
TherArc
PoccArc
CkorArc
AsulArc
MmetArc
HalspArc
HvolArc
NmagArc
SsolArc
AaeoBac
TpenArc
CdiArc
CthoArc
CheiArc
NitspArc
MvilArc
MforArc
ClokArc
Glam
Eint
Tvag2
Tvag4
Tvag1
Tvag3
Tthe2
Tthe1
Ptet
Spom
Cneo
Abis
Tver
Adel
Scom
Umay
Pgra
Zrou
Klac
Scer
Cgla
Tmel
Aota
Anid
Nfis
Bcin
Ssci
Lmac
Pnod
Ncra
Cglo
Tter
Bden
Ehis
Ptri
Pfal
Tpar
Bbov
Chom
Tgon
Cpar2
Ccri
Cmer1
Tbru
Lmaj
Pinf
Spur1
Spur2
Aque
Crei1
Vcar1
Mpol1
Ppat1
Atri1
Osat2
Zmay3
Gmax1
Gmax2
Atha1
Atha2
Osat1
Zmay1
Zmay2
Dmel
Agam
Cint
Drer
Trub
Acar
Mdom
Sscr3
Mmus
Cfam
Hsap
Ptro
Sscr1
Sscr2
Cele
Ddis
Asub
Ppal
Bnat1
Cpar1
Ngru
0.4

Figure 26: Maximum likelihood tree of tRNA methyltransferases across eukaryotes and prokaryotes. The phylogenetic tree is constructed using 96 eukaryotic, 21 archaeal and 9 bacterial sequences. The 207 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.10. The five eukaryotic supergroups, bacteria and archaea are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green), Archaea (orange) and Bacteria (cyan).

6.2.4 Nucleocytoplasmic transport factor

The Ntf2 gene involved in the nucleocytoplasmic transport of proteins is part of the core LECA proteome. We find homologs of this protein in bacteria specifically in those belonging to Streptomycetaceae family. No homolog could be identified in any archaea. The presence of the homologs only in a specific family of bacteria suggests a horizontal gene transfer event from eukaryotes to bacteria. In the maximum likelihood tree constructed, the Ntf2 homologs of bacteria cluster within the eukaryotes. However, they are not recovered as a monophyletic group (Figure 27). The Ntf2 homologs in Streptomycetaceae family cluster into two groups, one clusters with the Arthropods; *D. melanogaster* and *A. gambiae* while, the other group is found to branch with the excavates (Figure 27). The two groups of Ntf2 homologs in bacteria do not find each other in the BLASTp analysis against the NCBI nr database, suggesting they differ significantly at the sequence level. In accordance with the ML tree, the Kita1Bac and Kita2Bac Ntf2 sequences return the *D. melanogaster* Ntf2 sequence as the top-hit following the Ntf2 homologs of this group of bacteria. The Ntf2 homologs in the other group of bacteria are found to return the fungal Ntf2 homologs as the top hits in the BLASTp analysis. This suggests the possibility of two independent HGT events into bacteria followed by sequence divergence.

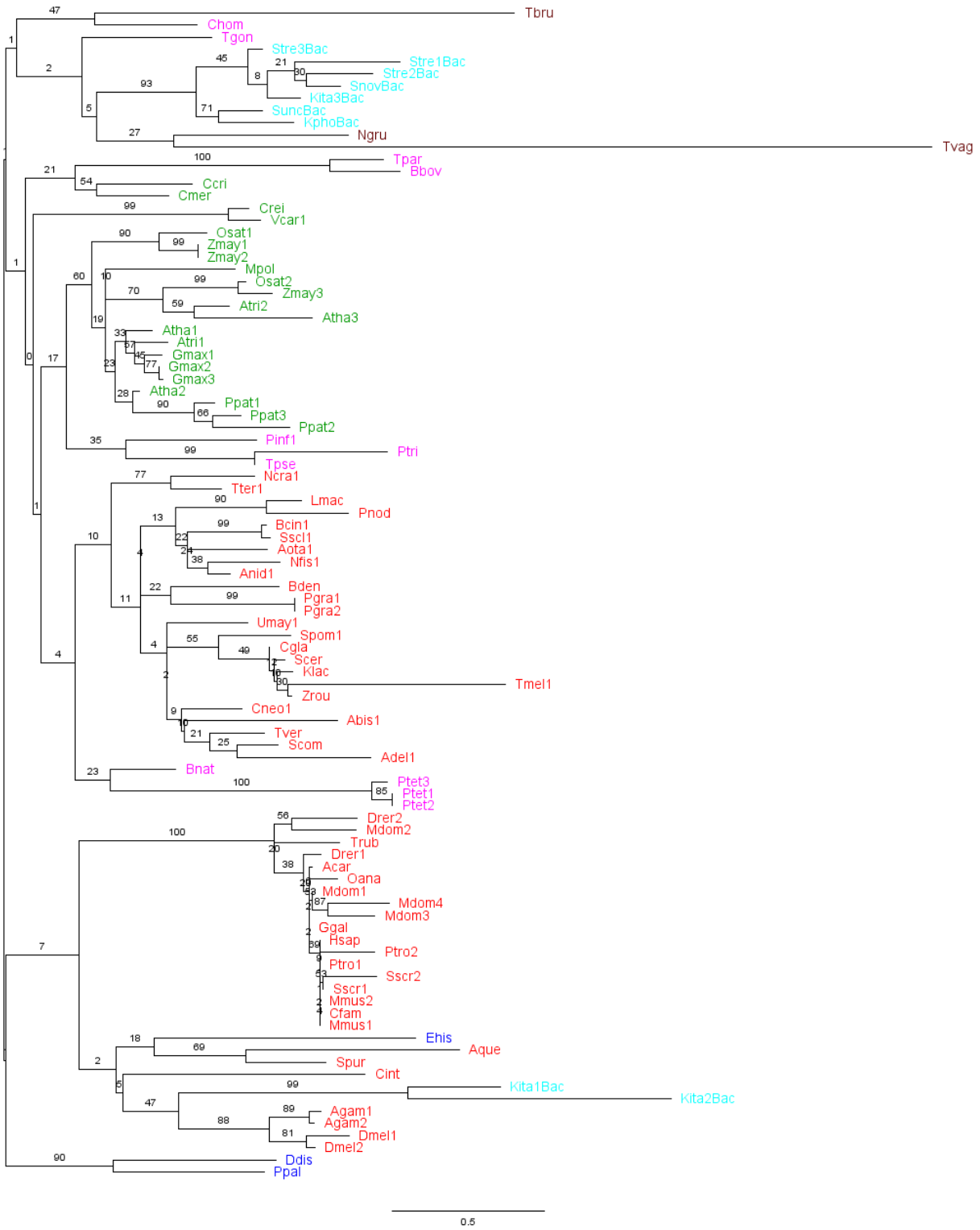

Figure 27: Maximum likelihood tree of Ntf2 protein homologs across eukaryotes and prokaryotes. The phylogenetic tree is constructed using 87 eukaryotic, and 9 bacterial sequences. The 94 conserved positions in the alignment were used for phylogenetic analysis. Bootstrap values are shown at the nodes. The scale bar represents the

average number of substitutions per site. The full name of the organisms and the protein accession numbers are provided in Appendix table A.11. The five eukaryotic supergroups and bacteria are color coded as follows: Opisthokonta (red), Amoebozoa (blue), Excavata (brown), SAR (pink), Archaeplastida (green), and Bacteria (cyan).

6.2.5 NE proteins with prokaryotic domains

Among the 32 core and non-linearly conserved NE proteins, 13 of them are found to contain characteristic domains that are of prokaryotic origins (Table 7). The domains found in the proteins Pct1, Has1, Ptc7, Jem1, Rrt12, Wss1 and Scp160 are found in both archaea and bacteria suggesting that they are ancestral domains, while those found in Heh2, Src1, Ulp1, Nem1, Gas1 and Trl1 are found only in bacteria.

S. No	Protein	Domains in yeast protein	Domains in Bacteria	Domains in Archaea
01	Heh2 & Src1	HeH & MSC	HeH	-
02	Pct1	CTP_transf_like	CTP_transf_like	CTP_transf_like
03	Ulp1	Peptidase_C48	Peptidase_C48	
04	Rrt12	Peptidase_S8	Peptidase_S8	Peptidase_S8
05	Has1	DEAD & Helicase_C & DUF4217	DEAD & Helicase_C	DEAD & Helicase_C
06	Ptc7	SpoIIE	SpoIIE	SpoIIE
07	Jem1	DnaJ	DnaJ	DnaJ
08	Nem1	NIF	NIF	
09	Wss1	WLM	WLM	WLM
10	Gas1	Glyco_hydro_72 & X8	Glyco_hydro_72	
11	Scp160	KH_1	KH_1	KH_1
12	Trl1	RNA_lig_T4_1 & tRNA_lig_Kinase & tRNA_lig_CPD	RNA_lig_T4_1	

Table 7: Core and non-linearly conserved NE proteins with prokaryotic domains. The characteristic domain present in the yeast protein and the respective domains found in bacteria and archaea are shown.

The proteins Heh2 and Src1 contain an N-terminal HeH domain and a C-terminal MSC domain. No protein could be found in archaea with either MSC or HeH domain, except for a single partial protein containing a HeH domain in only one Lokiarchaea. However, we find a number of proteins with HeH domain in bacteria. In accordance with the previous studies we obtained a

Rho termination protein and a number of uncharacterized proteins containing either a HeH or Rho_N or SAP domain all of which share a helix-extension-helix fold (Aravind and Koonin 2001). Interestingly, for the first time, we could identify 50S ribosomal proteins in bacterial species with a HeH domain. In gram-positive bacteria, ribosomal protein L20 contains a HeH domain towards the C-terminal end, while in gram-negative bacteria, the L21 ribosomal proteins have either a Rho_N domain or a helix-hairpin-helix domain at the C-terminal. The presence of nucleic acid binding domains such as helix-extension-helix (HeH and Rho_N) and helix-hairpin-helix domains in ribosomal proteins suggests additional regulatory roles of these proteins.

The Peptidase_C48 domain found in the SUMO protease Ulp1 is found in Chlamydiae and Proteobacteria phyla of bacteria. The NIF domain present in the phosphatase Nem1 involved in phospholipid synthesis is found in several bacterial phyla. Glyco_hydro_72 domain found in Gas1 protein and the RNA_lig_T4_1 domain found in the tRNA ligase Trl1 are found in bacteria but not archaea. Though we could not identify the Sad1_UNC domain of Mps3 in either bacteria or archaea, we did identify the related F5_F8_type_C domain proteins in bacteria in accordance with the previous studies.

6.3 Conclusions

A number of nuclear envelope proteins are found to have evolved in the LECA. A significant number of them are found to have evolved from the domains contributed by bacteria and archaea, with a higher number of domains being traceable to bacteria. Only, 4 of the NE proteins have homologs, which share significant sequence similarity with the eukaryotic proteins. We find the class II HMG-CoA reductases in several eukaryotes belonging to Fungi, Excavata and SAR. A number of anaerobic eukaryotes are found to contain class II enzymes. Additionally, the Asgard archaea are found to contain class II enzymes, but not class I. The tRNA methyl transferases are found to have descended from Archaea with an additional homolog of these proteins being contributed from bacteria. In this study, we also find the case of horizontal gene transfer from eukaryotes to bacteria, as observed in the case of Ntf2 protein.

Chapter 7

Predicting localization of a protein using PPI data

7.1 Introduction

Cells are complex biological machines in which a number of intricately connected pathways and processes operate within and between the various subcellular compartments. Proteins play a central role in carrying out the various biological processes. The presence of a protein in the correct subcellular compartment is essential to carry out its functions. It is important to know the function of all the proteins in an organism to understand the physiology of the cells. Knowledge of the localization of a protein can provide hints about its function.

The conventional functional studies to determine the protein subcellular location are resource and time-intensive, thus failing to keep up with the pace with which new sequence data is being generated. The genome of *Saccharomyces cerevisiae* was sequenced over two decades ago; however, there still remain around 700 genes for which we have no functional analysis and over 1000 genes for which no localisation data is available. Computational methods serve as useful tools in predicting the potential localization of proteins, thus bridging the time gap that exists between identifying the entire set of proteins in an organism to knowing their subcellular locations and hence their function.

Analysis of the protein-protein interactions (PPI) in yeast showed that 76% of the interactions are between proteins occurring in the same subcellular compartment (Schwikowski et al. 2000). Similarly, analysis of the human PPI network showed that a significant number of interactions are between proteins residing in the same subcellular compartment (Gandhi et al. 2006). This suggests that protein-protein interaction data can be effectively used to predict the subcellular location (SCL) of proteins. This hypothesis was further strengthened by the development of algorithms that used PPI data to predict the SCL of proteins. For example, Scott et al. developed PSLT2 that predicts the localisation of yeast proteins by combining the PPI data and the knowledge of motifs and domains in the protein sequence (Scott et al. 2005). Lee et al.

developed the DC-kNN classification algorithm, which integrates features of individual protein such as the amino acid composition, presence of signal motifs, functional annotations and network dependent properties like the features and the localization of the interacting proteins to predict localization (Lee et al. 2008). For each subcellular localization, the best combination of features of the individual proteins and their interactors is extracted. This information is then used to predict the localization of proteins. Jiang and Wu developed ensemble classifier to predict multiplex subcellular localization of proteins in yeast. It was built by combining four graph-based semi-supervised learning algorithms developed earlier (Majority, χ^2-score, GenMultiCut and Functional Flow) for predicting the function of a protein from PPI data (Jiang and Wu 2012).

Currently available tools to predict the subcellular localization of proteins make use of features such as the amino acid composition, sorting signals and motifs, GO terms, phylogeny information etc (Horton et al. 2007; Huang et al. 2008; Blum et al. 2009; Chou and Shen 2010a). Though, several algorithms that make use of PPI data have been developed to predict the SCL of proteins, none of them are available as online tools. Yeast-PLoc has been developed to predict localization of yeast proteins using both sequence and PPI information with high accuracy and was earlier available as a user-friendly web-server, but no longer maintained (Hu et al. 2012). In this study, we have developed a Perl script to predict the SCL of proteins for which no localization data is available in *Saccharomyces cerevisiae* using PPI data. We predicted the SCL of 249 yeast proteins annotated as "cellular component unknown" and another 192 proteins for which cellular component data are available from only high throughput studies. We further validated a few of those predictions using experimental studies.

7.2 Results

To identify the subcellular localization of proteins with unknown cellular component in yeast, we made use of the extensive protein-protein interaction data available for *Saccharomyces cerevisiae*. The rationale we used was that if a large number of interactors of the protein are in a particular subcellular location, the protein is likely to be physically present in the same location. Based on this logic, we developed a Perl script to predict the SCL of yeast proteins that have more than 4 unique physical interactors. Briefly, the script works as follows: first, the unique

physical interactors of the protein of interest are obtained from the physical interaction data file downloaded from BioGRID database (version 3.4.164, downloaded on August 25th, 2018) (Stark et al. 2006). For the proteins with more than 4 unique physical interactors, the cellular component data with the GO evidence codes IDA and HDA for each interactor is obtained from the gene2go file downloaded from NCBI (downloaded on August 24th, 2018). The parent SCL terms and their child terms (that are similar to parent) are provided as input (Table 8). The cellular component terms associated with each of the interactors are matched with the child terms and if a match is found they are converted into a parent SCL term. This was done to make the cellular component terms uniform across the various interactors.

S. No	Parent terms (SCL)	Child terms
1	Mitochondria	mitochon
2	Nucleus	nucleus|nucleoplasm|chromatin|chromosome
3	Nuclear periphery	nuclear pore|nuclear envelope|nuclear membrane|nuclear periphery
4	Nucleolus	nucleolus|nucleolar
5	Cytosol	cytoplasm|cytosol
6	Endoplasmic reticulum	endoplasmic| ER
7	Golgi	golgi
8	Vacuole	vacuole|vacuolar
9	Plasma membrane	plasma|cell periphery
10	Spindle	spindle
11	Endosome	endosome|endosomal
12	Peroxisome	peroxisome|peroxisomal
13	Bud neck	bud neck

Table 8: SCL terms considered in the script. The parent terms or the subcellular locations that are considered in the script and the corresponding child terms that are used to identify them are shown in the table.

The number of interactors associated with each of the SCL is then counted and the percentage of the interactors that localize to a particular compartment out of the total interactors with SCL data is calculated. An interactor associated with more than one SCL is counted under all the SCL categories associated. The SCL of the protein of interest is then predicted based on the

percentage of interactors associated with each SCL. The different possible SCLs are ranked based on the percentage of interactors. The SCL with the highest percentage of interactors is considered as the most probable SCL of the protein. However, in order to be able to document potential multiple localizations of a protein, the percentage of interactors for the first, second and third predicted SCLs are compared. If the SCL with the highest percentage of interactors differs from the second highest SCL by more than 10%, then the one with highest percentage of interactors is only reported as the predicted SCL. But, if the difference between the first and the second SCL percentage of interactors is less than 10% then both are reported as the predicted SCL. Similarly, the second and third SCL are also compared and reported based on the 10% cut-off. The cut-off was chosen on a trial basis. Three different values; 10%, 20% and 30% were tried. The 10% cut-off minimized the false positives and hence was considered (the Perl script used and the required database files are provided as Additional data 4).

7.2.1 Validating the script

In order to test our hypothesis and the script, we first used our script to predict the SCL of 100 proteins for which the SCL is available from various experimental studies. We consider a prediction to match exactly if all the predicted SCL(s) are part of the known localization of the protein. We consider a partial match if one of the predicted SCL matches but one or more of the predicted SCL terms do not match with the known localization. Similarly, a prediction is considered to be a mismatch, if none of the terms in the predicted SCL match the known localization(s) of the protein. Based on these criteria we evaluated our predictions (Table 9) and find that our script is able to predict the SCL of 62 proteins as exact match, 31 as partial match and 7 as mismatch. Thus our script is able to predict the localization of 93 out of 100 proteins. Additionally, the predictions with our script were also compared with the localization data in CYCLoPs database (Koh et al. 2015). For each protein the localization(s) for which the LOC-score in CYCLoPs database was equal to or above the respective cutoffs mentioned in Chong et al. (Chong et al. 2015) were assigned to it. Of the 100 proteins, 82 had localization data in CYCLoPs database and among the 82 proteins we found an exact match for 37 proteins, partial match for 28 proteins and mismatch for 17 proteins (Table 9).

Table 9: Predicted and the known localization of 100 verified ORFs

S.No.	Gene	Predicted localisation	SGD cellular component	CYCLoPs localization
1	YLR319C	bud neck; cytosol	cellular bud neck (IDA) cellular bud tip (IDA) spindle pole body (IDA) incipient cellular bud site (IDA)	Bud; Bud site; Cortical patches
2	YMR032W	bud neck; cytosol	cellular bud neck (IDA) cellular bud neck septin ring (IDA)	Cytoplasm
3	YMR124W	bud neck; cytosol; nucleus	cytoplasm (HDA) cellular bud neck (IPI)	Bud; Bud site; Cortical patches
4	YNL233W	nucleus; cytosol; bud neck	incipient cellular bud site (IDA) cellular bud neck (IDA) cellular bud neck septin collar (IDA)	Bud; Cytoplasm
5	YNL166C	bud neck; cytosol	cellular bud neck septin ring (IDA) cellular bud neck (IDA)	Cytoplasm
6	YNR049C	cytosol; plasma membrane; bud neck	cellular bud tip (IDA) cellular bud neck (IDA) prospore membrane (IDA) plasma membrane (IDA) cytosol (HDA) nucleus (HDA)	Cytoplasm
7	YOR188W	cytosol; bud neck	cytosol (HDA) mitochondrion (HDA) plasma membrane (HDA) cellular bud neck (HDA) cellular bud tip (HDA)	Cytoplasm
8	YPL242C	bud neck; cytosol	cellular bud neck contractile ring (IDA) cytoplasm (HDA)	Cytoplasm
9	YPR055W	cytosol; bud neck	cellular bud neck (IDA) cellular bud tip (IDA) cytoplasm (HDA) exocyst (IDA) prospore membrane (HDA)	Cortical patches; Bud; Bud site
10	YBR102C	cytosol; bud neck; nucleus	incipient cellular bud site (IDA) exocyst (IDA) cellular bud neck (IDA) cellular bud tip (IDA) prospore membrane (HDA)	Bud, Bud site, Cortical Patches
11	YDR037W	cytosol	cytoplasm (IDA)	No data
12	YIR004W	cytosol	cytosol (IDA) endoplasmic reticulum (HDA) cellular bud (HDA) cell periphery (HDA)	ER
13	YKR059W	cytosol	cytoplasmic stress granule (IDA)	Cytoplasm
14	YLR262C	cytosol	golgi apparatus (IDA) cytosol (IDA)	Cytoplasm
15	YNL007C	cytosol	nucleus (IDA) cytosolic small ribosomal subunit (IDA)	Cytoplasm
16	YOL133W	cytosol; nucleus	nucleus (IDA) cytoplasm (IDA)	No data

17	YOR042W	cytosol	cytoplasm (HDA)	Cytoplasm
18	YMR215W	endoplasmic reticulum	plasma membrane (IDA) fungal-type cell wall (IDA) fungal-type vacuole (HDA)	Cytoplasm; Endoplasmic reticulum
19	YAL023C	endoplasmic reticulum	endoplasmic reticulum (HDA)	Cytoplasm, Endoplasmic reticulum
20	YJL192C	endoplasmic reticulum	endoplasmic reticulum (IDA)	Endoplasmic reticulum
21	YKL154W	endoplasmic reticulum	integral component of endoplasmic reticulum membrane (IDA)	Endoplasmic reticulum
22	YLR088W	endoplasmic reticulum	endoplasmic reticulum (HDA)	Cytoplasm; Endoplasmic reticulum
23	YLR372W	endoplasmic reticulum	endoplasmic reticulum (IDA)	Endoplasmic reticulum
24	YOL003C	endoplasmic reticulum	endoplasmic reticulum (IDA)	No data
25	YNL101W	endoplasmic reticulum	fungal-type vacuole (IDA) vacuole-mitochondrion membrane contact site (IDA)	No data
26	YPL274W	endoplasmic reticulum	plasma membrane (IDA) endoplasmic reticulum (HDA)	Cytoplasm
27	YBR159W	endoplasmic reticulum	endoplasmic reticulum membrane (IDA)	Endoplasmic reticulum
28	YMR218C	cytosol	early endosome (IDA) trans-Golgi network (IDA)	Endosome; Golgi
29	YOL018C	golgi; cytosol; vacuole	trans-Golgi network (IDA) integral component of endosome membrane (IDA)	No data
30	YOR357C	golgi; cytosol	cytosol (IDA) endosome (IDA) fungal-type vacuole membrane (HDA)	Vacuole
31	YDR189W	cytosol; golgi; vacuole	COPII-coated ER to Golgi transport vesicle (IDA) endoplasmic reticulum (IDA) golgi membrane (IDA)	Endoplasmic reticulum; Golgi
32	YEL036C	cytosol; endoplasmic reticulum; golgi	Golgi cis cisterna (IDA)	Mitochondria
33	YGL198W	cytosol; golgi	golgi apparatus (IDA)	No data
34	YJL166W	mitochondria	mitochondrial respiratory chain complex III (IDA) mitochondrion (HDA)	Mitochondria
35	YAL010C	mitochondria	mitochondrial sorting and assembly machinery complex (IPI) mitochondrial outer membrane (HDA)	No data
36	YJL063C	mitochondria	mitochondrial large ribosomal subunit (IDA) mitochondrion (HDA)	Mitochondria
37	YJL062W-A	mitochondria	integral component of mitochondrial inner membrane (IDA) mitochondrion (HDA)	Mitochondria
38	YJL003W	mitochondria	mitochondrial inner membrane (IDA) mitochondrion (HDA)	No data
39	YJR045C	mitochondria	mitochondrial inner membrane	No data

			(IDA) mitochondrial nucleoid (IDA) mitochondrion (HDA)	
40	YJR101W	mitochondria	mitochondrial small ribosomal subunit (IDA) mitochondrion (HDA)	Mitochondria
41	YKL148C	mitochondria	mitochondrial respiratory chain complex II, succinate dehydrogenase complex (ubiquinone) (IDA) mitochondrion (IDA)	No data
42	YKL134C	mitochondria	mitochondrial matrix (IDA) mitochondrion (HDA)	Mitochondria
43	YLR439W	mitochondria	mitochondrial large ribosomal subunit (IDA) mitochondrion (HDA)	Mitochondria
44	YJL061W	nuclear periphery; cytosol; nucleus	nucleus (IDA) nuclear pore (IDA) cytosol (IDA) nuclear pore cytoplasmic filaments (IDA)	Nuclear periphery
45	YJL041W	cytosol; nuclear periphery	nuclear pore (IDA) nuclear pore central transport channel (IDA) nuclear pore nuclear basket (IDA)	Nuclear periphery; Vacuole
46	YJL039C	nuclear periphery; cytosol	nuclear pore (IDA) nuclear pore inner ring (IDA) nucleus (HDA)	Nuclear periphery
47	YJR042W	nuclear periphery	nuclear pore (IDA) nuclear pore outer ring (IDA)	Nuclear periphery
48	YKL057C	nuclear periphery	nuclear pore (IDA) nuclear pore outer ring (IDA)	No data
49	YKR082W	nuclear periphery; cytosol	nuclear pore (IDA) nucleus (IDA) cytosol (IDA) nuclear pore inner ring (IDA)	Nuclear periphery
50	YLR018C	nuclear periphery; endoplasmic reticulum; cytosol	nuclear pore (IDA) integral component of nuclear outer membrane (IDA) nuclear pore transmembrane ring (IDA)	Nuclear periphery
51	YML031W	nuclear periphery; cytosol	nuclear pore (IDA) spindle pole body (IDA) nuclear pore transmembrane ring (IDA)	Nuclear periphery
52	YMR129W	nuclear periphery	nuclear pore (IDA) nuclear envelope lumen (IDA) nuclear pore transmembrane ring (IDA) mitochondrion (HDA) cell periphery (HDA)	Vacuole
53	YBL079W	nuclear periphery	nuclear pore (IDA) nuclear pore inner ring (IDA)	Nuclear periphery; Vacuole
54	YJL010C	nucleolus	nucleolus (IDA) 90S preribosome (IDA)	Nucleolus

			nucleus (HDA)	
55	YJR002W	nucleolus	small-subunit processome (IDA) nucleolus (HDA) nucleus (HDA)	No data
56	YLR129W	nucleolus	small-subunit processome (IDA) nucleolus (IDA)	Nucleus
57	YMR093W	nucleolus	nucleolus (IDA) small-subunit processome (IDA) rDNA heterochromatin (IDA)	Nucleolus; Nucleus
58	YDR324C	nucleolus	nucleolus (IDA) small-subunit processome (IDA)	Nucleolus; Nucleus
59	YOR287C	nucleolus	nucleolus (IDA) 90S preribosome (IDA)	No data
60	YPL126W	nucleolus	nucleolus (IDA) small-subunit processome (IDA) rDNA heterochromatin (IDA)	Nucleolus; Nucleus
61	YHR066W	nucleolus	nucleolus (IDA) preribosome, large subunit precursor (IDA)	Nucleolus; Nucleus
62	YNL022C	cytosol	nucleolus (IDA) nucleus (IDA)	Nucleus
63	YFL008W	nucleus	nuclear mitotic cohesin complex (IDA) nucleus (HDA)	Nucleus
64	YNR010W	nucleus; cytosol	nucleus (IDA) cytosol (IDA)	Nucleus
65	YNR023W	nucleus	nucleus (IDA) cytosol (IDA)	Nucleus
66	YEL009C	nucleus; cytosol	nucleus (IDA)	Nucleus
67	YNR033W	cytosol; nucleus	cytoplasm (HDA)	Cytoplasm
68	YPL138C	nucleus; cytosol	nucleus (IDA) cytosol (IDA)	Nucleus
69	YPL137C	cytosol; nucleus	cytoplasm (HDA) mitochondria (HDA) endoplasmic reticulum (HDA)	Cell periphery; Cytoplasm
70	YJL185C	cytosol; peroxisome	peroxisome (IDA)	No data
71	YKL197C	peroxisome	peroxisomal membrane (IDA) cytosol (HDA)	Mitochondria; Peroxisome
72	YMR026C	peroxisome	integral component of peroxisomal membrane (IDA) peroxisomal importomer complex (IDA)	Mitochondria; Peroxisome
73	YNL329C	peroxisome	peroxisome (IDA) cytosol (IDA)	Mitochondria; Peroxisome
74	YNL214W	peroxisome	peroxisomal membrane (IDA) peroxisomal importomer complex (IDA)	Mitochondria; Peroxisome
75	YOL044W	peroxisome; cytosol	integral component of peroxisomal membrane (IDA)	Cytoplasm
76	YDL065C	peroxisome; cytosol	peroxisomal membrane (IDA) endoplasmic reticulum (IDA) cytosol (IDA)	Cytoplasm
77	YDR142C	cytosol; peroxisome	peroxisome (IDA) cytosol (IDA)	No data

78	YPL147W	cytosol	integral component of peroxisomal membrane (IDA) peroxisomal importomer complex (IDA)	No data
79	YDR244W	peroxisome	peroxisome (IDA) cytosol (IDA) peroxisomal importomer complex (IDA)	Bud; Bud site; Peroxisome
80	YDR265W	peroxisome; cytosol	peroxisomal membrane (IDA) peroxisomal importomer complex (IDA)	Mitochondria; Peroxisome
81	YLR332W	plasma membrane	integral component of plasma membrane (IDA) fungal-type vacuole (HDA)	Cell periphery
82	YNL142W	vacuole; plasma membrane; endoplasmic reticulum	plasma membrane (IDA)	Cell periphery
83	YOR153W	plasma membrane; cytosol; endoplasmic reticulum	plasma membrane (IDA) mitochondrion (HDA) cell periphery (HDA)	Cell periphery
84	YPL221W	plasma membrane	endoplasmic reticulum (IDA) fungal-type vacuole (HDA) cellular bud neck (HDA)	Bud; Cell periphery; ER
85	YPL058C	plasma membrane	plasma membrane (IDA) cell periphery (HDA)	Cell periphery
86	YPR156C	plasma membrane; cytosol; endoplasmic reticulum	plasma membrane (IDA) cell periphery (HDA)	Cell periphery; Endoplasmic reticulum
87	YGR055W	plasma membrane; cytosol	plasma membrane (IDA) cell periphery (HDA) fungal-type vacuole (HDA)	Nucleus; Vacuole
88	YNL188W	spindle	half bridge of spindle pole body (IDA) endoplasmic reticulum (HDA)	No data
89	YPL255W	spindle; nucleus	central plaque of spindle pole body (IDA)	Spindle pole
90	YPL124W	nucleus; spindle; cytosol	central plaque of spindle pole body (IDA)	Spindle pole
91	YBL031W	nucleus; cytosol	spindle (IDA) cellular bud neck (IDA) nuclear microtubule (IDA)	Cytoplasm; Spindle pole
92	YJL178C	vacuole; cytosol; golgi; mitochondria	vacuolar membrane (IDA) trans-Golgi network (IDA) mitochondrion (IDA) phagophore assembly site (IDA)	Vacuole
93	YKR007W	vacuole	fungal-type vacuole membrane (IDA) late endosome membrane (IDA) cytosol (HDA)	Vacuole
94	YLR090W	vacuole	integral component of mitochondrial outer membrane (IDA) mitochondrion (HDA) nucleus (HDA)	Cytoplasm
95	YLR148W	vacuole; cytosol	extrinsic component of vacuolar	Endosome

			membrane (IDA) fungal-type vacuole membrane (IDA)	
96	YLR211C	cytosol	phagophore assembly site (IDA) vacuolar membrane (IDA) cytoplasm (HDA)	Cytoplasm
97	YEL013W	vacuole	fungal-type vacuole membrane (IDA) nucleus-vacuole junction (IDA) cytosol (HDA)	Vacuole
98	YLR396C	vacuole	fungal-type vacuole membrane (IDA) cytosol (IDA)	Endosome
99	YMR197C	vacuole; cytosol	integral component of Golgi membrane (IDA) fungal-type vacuole membrane (HDA)	No data
100	YMR231W	vacuole	fungal-type vacuole membrane (IDA)	Endosome

Table 9: Predicted and the known localization of 100 verified ORFs. A comparison of the predicted and the experimentally determined localization data in SGD (column 4) and in CYCLoPs database (column 5) for 100 ORFs is shown. The data in column 4 and 5 are color-coded based on the comparison with the predicted localizations in column 3. The cells filled in yellow represent an exact match of predicted with the available localization, green filled cells represent partial match and red filled cells represent mismatch.

To compare the performance of our script with the other existing softwares, we randomly chose a few proteins from among the 100 ORFs for which localization data is available from experimental studies and predicted their localization using two existing web-server predictors WoLF PSORT (Horton et al. 2007) and Euk-mPLoc 2.0 (Chou and Shen 2010b). The prediction accuracy was found to be very low for WoLF PSORT as an exact match of the predicted and known localization was found only for 4 out of 12 proteins. Euk-mPLoc 2.0 could predict the localization of 7 proteins accurately. However, it fails to predict the subnuclear localization of the proteins. Our script based on PPI data performed better than the two available predictors (Table 10). Our script could predict the SCL of proteins part of the nuclear pore, nucleolus and spindle pole body accurately.

Protein	WoLF PSORT	Euk-mPLoc 2.0	Our Script	Actual localisation
YLR319C	Nucleus; Cytosol & Nucleus	Nucleus	Bud neck; cytosol	Cellular bud neck; spindle pole body
YDR037W	Nucleus; Cytosol	Cytoplasm	Cytosol	Cytoplasm
YJL192C	Plasma membrane; ER	ER	ER	ER
YGL198W	Plasma membrane	ER	Cytosol; Golgi	Golgi apparatus
YJL003W	Mitochondria	Mitochondria	Mitochondria	Mitochondrial inner membrane
YJR042W	Nucleus; Cytosol & Nucleus; Cytosol; Cytosol & Mitochondria; Mitochondria; Peroxisomes	Nucleus	Nuclear periphery	Nuclear pore
YDR324C	Nucleus; Mitochondria; Cytosol & Nucleus	Nucleus	Nucleolus	Nucleolus
YFL008W	Nucleus	Nucleus	Nucleus	Nucleus
YPL138C	Nucleus	Nucleus	Nucleus; Cytosol	Nucleus; Cytosol
YNL214W	Nucleus; Cytosol & Nucleus; Mitochondria; Cytosol	Peroxisome	Peroxisome	Peroxisomal membrane
YPL058C	Plasma membrane	Cell membrane	Plasma membrane	Plasma membrane
YNL188W	Nucleus; Cytosol & Nucleus	Nucleus	Spindle	Half bridge of spindle pole body

Table 10: Comparison of the existing web-server predictors with our script. The localization predicted using WoLF PSORT, Euk-mPLOC 2.0 and our script along with the localization of these proteins determined using experimental studies are shown in the table. Yellow filled cells show an exact match, red filled cells show a mismatch. Partial matches are not considered.

7.2.2 Predicting the SCL of proteins with no data

After validating our hypothesis and the script, we went on to predict the localization of the proteins for which the cellular component is unknown and also for those that have SCL data only from high throughput studies (proteins with GO Cellular Component annotations with only HDA evidence code). Based on our criteria of having more than 4 interactors, we could predict the localization of 249 proteins for which no SCL data is available (Table 11). Among the proteins with single predicted SCL, we find that 176 of them are in cytosol followed by 21 in nucleus, 3 are predicted to localize to mitochondria and 1 each to bud neck and endoplasmic reticulum. We checked if localization data for these 249 proteins was available in CYCLoPs database. We did find localization data for 41 proteins and found our predictions to match for 35 proteins (Table 11). For the ORFs with high throughput data we compared our predictions with the reported SCL. Of the 192, for 104 of them we find a correspondence, while for 88 we did not (Table 12). Among the 88 proteins for which the predicted localization did not match the known localization, 78% of them are predicted to localize to cytosol, while the SCL reported from high-throughput studies was found to be of a specific organelle within the cytosol (eg. endoplasmic reticulum or mitochondria or vacuole). This is possibly due to the absence of extensive interaction data and cellular component annotations for the available interactors.

Table 11: Localization predicted for 249 ORFs with no SCL data

S. No	ORF	Predicted localisation	S. No	ORF	Predicted localisation
1	YBR285W	Bud neck	**126**	YFR043C	Cytosol
2	YBR255C-A	Mitochondria	**127**	YFR055W	Cytosol
3	YDR124W	Mitochondria	**128**	YGL050W	Cytosol
4	YPR121W	Mitochondria	**129**	YGL121C	Cytosol
5	YPL264C	Endoplasmic reticulum	**130**	YGL196W	Cytosol
6	YBL086C	Nucleus	**131**	YGL242C	Cytosol
7	YBR063C	Nucleus	**132**	YGL248W	Cytosol
8	YBR296C-A	Nucleus	**133**	YGR109C	Cytosol
9	YDL186W	Nucleus	**134**	YGR146C	Cytosol
10	YIL177C	Nucleus	**135**	YHL010C	Cytosol
11	YLL054C	Nucleus	**136**	YHR015W	Cytosol

12	YLR125W	Nucleus	**137**	YHR025W	Cytosol
13	YBL043W	Nucleus	**138**	YHR029C	Cytosol
14	YDL214C	Nucleus	**139**	YHR044C	Cytosol
15	YDL246C	Nucleus	**140**	YHR138C	Cytosol
16	YDR403W	Nucleus	**141**	YIL009W	Cytosol
17	YER180C	Nucleus	**142**	YJL057C	Cytosol
18	YHR209W	Nucleus	**143**	YJL105W	Cytosol
19	YKL090W	Nucleus	**144**	YJR083C	Cytosol
20	YKL161C	Nucleus	**145**	YJR142W	Cytosol
21	YNR059W	Nucleus	**146**	YKL093W	Cytosol
22	YNR064C	Nucleus	**147**	YKL098W	Cytosol
23	YNR069C	Nucleus	**148**	YKL104C	Cytosol
24	YOL063C	Nucleus	**149**	YKL198C	Cytosol
25	YOR134W	Nucleus	**150**	YKL218C	Cytosol
26	YPL171C	Nucleus	**151**	YKR017C	Cytosol
27	YBL081W	Cytosol	**152**	YKR056W	Cytosol
28	YBR053C	Cytosol	**153**	YKR058W	Cytosol
29	YBR225W	Cytosol	**154**	YKR098C	Cytosol
30	YBR284W	Cytosol	**155**	YLL057C	Cytosol
31	YCR024C-B	Cytosol	**156**	YLL063C	Cytosol
32	YCR102C	Cytosol	**157**	YLR070C	Cytosol
33	YDL109C	Cytosol	**158**	YLR128W	Cytosol
34	YDL177C	Cytosol	**159**	YLR137W	Cytosol
35	YDR210W	Cytosol	**160**	YLR149C	Cytosol
36	YDR249C	Cytosol	**161**	YLR219W	Cytosol
37	YDR286C	Cytosol	**162**	YLR243W	Cytosol
38	YDR336W	Cytosol	**163**	YLR306W	Cytosol
39	YEL023C	Cytosol	**164**	YLR359W	Cytosol
40	YEL077C	Cytosol	**165**	YLR405W	Cytosol
41	YER158C	Cytosol	**166**	YML050W	Cytosol
42	YGR035C	Cytosol	**167**	YML068W	Cytosol
43	YGR079W	Cytosol	**168**	YML083C	Cytosol
44	YGR127W	Cytosol	**169**	YMR081C	Cytosol
45	YGR153W	Cytosol	**170**	YMR087W	Cytosol
46	YGR240C-A	Cytosol	**171**	YMR154C	Cytosol
47	YHR210C	Cytosol	**172**	YMR191W	Cytosol

48	YIL151C	Cytosol	**173**	YMR210W	Cytosol
49	YIL152W	Cytosol	**174**	YMR217W	Cytosol
50	YIR016W	Cytosol	**175**	YMR318C	Cytosol
51	YJL049W	Cytosol	**176**	YMR322C	Cytosol
52	YJL107C	Cytosol	**177**	YNL024C-A	Cytosol
53	YJL181W	Cytosol	**178**	YNL278W	Cytosol
54	YJL206C	Cytosol	**179**	YNL307C	Cytosol
55	YJL218W	Cytosol	**180**	YNL333W	Cytosol
56	YJR011C	Cytosol	**181**	YNL334C	Cytosol
57	YJR012C	Cytosol	**182**	YOL064C	Cytosol
58	YJR107W	Cytosol	**183**	YOL128C	Cytosol
59	YKL033W-A	Cytosol	**184**	YOR021C	Cytosol
60	YKL068W-A	Cytosol	**185**	YOR032C	Cytosol
61	YLR154C-G	Cytosol	**186**	YOR126C	Cytosol
62	YLR407W	Cytosol	**187**	YOR137C	Cytosol
63	YLR446W	Cytosol	**188**	YOR155C	Cytosol
64	YLR460C	Cytosol	**189**	YOR202W	Cytosol
65	YML002W	Cytosol	**190**	YOR280C	Cytosol
66	YML020W	Cytosol	**191**	YOR349W	Cytosol
67	YMR102C	Cytosol	**192**	YPL003W	Cytosol
68	YMR130W	Cytosol	**193**	YPL095C	Cytosol
69	YMR181C	Cytosol	**194**	YPL113C	Cytosol
70	YMR209C	Cytosol	**195**	YPL241C	Cytosol
71	YMR262W	Cytosol	**196**	YPL250C	Cytosol
72	YMR265C	Cytosol	**197**	YPL281C	Cytosol
73	YMR317W	Cytosol	**198**	YPR106W	Cytosol
74	YNL165W	Cytosol	**199**	YCR099C	Cytosol
75	YNL193W	Cytosol	**200**	YDR338C	Cytosol
76	YNL295W	Cytosol	**201**	YFL034W	Cytosol
77	YOL014W	Cytosol	**202**	YPL257W	Cytosol
78	YOL036W	Cytosol	**203**	YLR030W	Plasma membrane;Nucleus
79	YOR097C	Cytosol	**204**	YJR115W	Bud neck;Cytosol
80	YOR111W	Cytosol	**205**	YGR067C	Cytosol;Nucleus
81	YPL034W	Cytosol	**206**	YHR022C	Cytosol;Nucleus
82	YPL039W	Cytosol	**207**	YJR061W	Cytosol;Nucleus
83	YPL056C	Cytosol	**208**	YKL121W	Cytosol;Nucleus

84	YPL077C	Cytosol	**209**	YLR031W	Cytosol;Nucleus
85	YPL216W	Cytosol	**210**	YPL088W	Cytosol;Nucleus
86	YPR013C	Cytosol	**211**	YPR117W	Cytosol;Nucleus
87	YPR015C	Cytosol	**212**	YBL049W	Cytosol;Nucleus
88	YPR084W	Cytosol	**213**	YBR157C	Cytosol;Nucleus
89	YAL034C	Cytosol	**214**	YDR436W	Cytosol;Nucleus
90	YBL036C	Cytosol	**215**	YDR466W	Cytosol;Nucleus
91	YBL066C	Cytosol	**216**	YDR475C	Cytosol;Nucleus
92	YBL067C	Cytosol	**217**	YER136W	Cytosol;Mitochondria
93	YBR022W	Cytosol	**218**	YGL158W	Cytosol;Nucleus
94	YBR153W	Cytosol	**219**	YHR185C	Cytosol;Nucleus
95	YBR213W	Cytosol	**220**	YJL165C	Cytosol;Nucleus
96	YBR256C	Cytosol	**221**	YJR108W	Cytosol;Nucleus
97	YBR259W	Cytosol	**222**	YKL050C	Cytosol;Nucleus
98	YBR276C	Cytosol	**223**	YML118W	Cytosol;Nucleus
99	YCL036W	Cytosol	**224**	YNL092W	Cytosol;Nucleus
100	YCR015C	Cytosol	**225**	YPL258C	Cytosol;Nucleus; plasma membrane
101	YCR026C	Cytosol	**226**	YCR100C	Cytosol; Nucleus
102	YCR073C	Cytosol	**227**	YHR078W	Nucleus; Cytosol
103	YCR076C	Cytosol	**228**	YEL076C	Nucleus;Cytosol
104	YCR091W	Cytosol	**229**	YGL117W	Nucleus;Cytosol
105	YDL037C	Cytosol	**230**	YIL002W-A	Nucleus;Spindle;Cytosol
106	YDL073W	Cytosol	**231**	YLR456W	Nucleus;Cytosol
107	YDL176W	Cytosol	**232**	YNL050C	Nucleus;Bud neck
108	YDL237W	Cytosol	**233**	YPR089W	Nucleus; Cytosol
109	YDR183W	Cytosol	**234**	YAR018C	Nucleus; Cytosol
110	YDR242W	Cytosol	**235**	YCR105W	Nucleus; Cytosol
111	YDR277C	Cytosol	**236**	YCR106W	Nucleus; Cytosol
112	YDR287W	Cytosol	**237**	YDL079C	Nucleus; Cytosol
113	YDR428C	Cytosol	**238**	YER007W	Nucleus; Cytosol
114	YDR435C	Cytosol	**239**	YJR159W	Nucleus; Cytosol
115	YDR512C	Cytosol	**240**	YKL072W	Nucleus; Cytosol
116	YDR541C	Cytosol	**241**	YLL033W	Nucleus; Cytosol
117	YEL029C	Cytosol	**242**	YMR041C	Nucleus; Cytosol
118	YEL041W	Cytosol	**243**	YPL023C	Nucleus; Cytosol

119	YEL072W	Cytosol	**244**	YBL009W	Cytosol; Nucleus; Nucleolus
120	YER010C	Cytosol	**245**	YGL081W	Cytosol; vacuole; Mitochondria; Nucleus
121	YER051W	Cytosol	**246**	YGL021W	Nucleus; Cytosol; vacuole; Mitochondria
122	YER055C	Cytosol	**247**	YOR393W	Endoplasmic reticulum; Golgi; Mitochondria; Bud neck; Cytosol; Nucleus
123	YER130C	Cytosol	**248**	YDR506C	Cytosol; Endoplasmic reticulum; Mitochondria
124	YFL059W	Cytosol	**249**	YGL146C	Cytosol; Endoplasmic reticulum
125	YFR007W	Cytosol			

Table 11: Localization predicted for 249 ORFs with no SCL data. The localization predicted with our script for the 249 ORFs with unknown cellular component in *Saccharomyces cerevisiae* is shown. The colored rows indicate the proteins for which localization data is available in CYCLoPs database. The rows filled in yellow represent an exact match of predicted and the available localization, green filled rows represent partial match and red filled rows represent mismatch.

Table 12: Predicted localization and the known localization from high throughput studies for 192 ORFs

S. No	ORF	Predicted localisation	Cellular component (as in SGD)
1	YBL029W	Cytosol	nucleus; cytoplasm
2	YBR090C	Cytosol	nucleus; cytoplasm
3	YBR138C	Cytosol	cytoplasm
4	YBR242W	Cytosol	nucleus; cytoplasm
5	YCL002C	Cytosol	nucleus; cytoplasm
6	YCR043C	Cytosol	Golgi apparatus; cytosol; mating projection tip
7	YCR095C	Cytosol	cytoplasm
8	YDL085C-A	Nucleus	nucleus; cytoplasm
9	YDL086W	Cytosol	cytoplasm; mitochondrion
10	YDL233W	Nucleus; Cytosol	nucleus; cytoplasm
11	YDR020C	Cytosol	nucleus; cytoplasm
12	YDR111C	Cytosol	nucleus; cytoplasm

13	YDR222W	Cytosol	cytoplasm
14	YDR248C	Cytosol	fungal-type vacuole; cytoplasm
15	YDR307W	Endoplasmic reticulum	endoplasmic reticulum
16	YDR391C	Cytosol	nucleus; cytoplasm
17	YDR520C	Cytosol	nucleus; cytoplasm
18	YEL025C	Cytosol	nucleus; cytoplasm
19	YER079W	Nucleus	nucleus; cytoplasm
20	YER156C	Cytosol	nucleus; cytoplasm
21	YFL040W	Vacuole;plasma membrane	fungal-type vacuole; prospore membrane
22	YFR006W	Cytosol	cytoplasm
23	YGL036W	Cytosol	cytoplasm
24	YGL101W	Cytosol	nucleus; cytoplasm
25	YGL185C	Cytosol	cytoplasm
26	YGR017W	Cytosol	nucleus; cytoplasm
27	YGR111W	Cytosol	nucleus; cytoplasm
28	YGR126W	Nucleus	nucleus; cytoplasm
29	YGR161C	Cytosol; Nucleus	nucleus; cytoplasm
30	YGR210C	Cytosol	cytoplasm
31	YGR237C	Cytosol	cytoplasm
32	YHL029C	Cytosol	cytoplasm
33	YHR097C	Cytosol	nucleus; cytoplasm; nuclear periphery
34	YHR131C	Cytosol	cytoplasm
35	YHR140W	Endoplasmic reticulum	cytoplasm; endoplasmic reticulum
36	YHR159W	Cytosol	cytoplasm
37	YHR202W	Cytosol	fungal-type vacuole; cytosol
38	YIL077C	Mitochondria	mitochondria
39	YIL092W	Cytosol	nucleus; cytoplasm
40	YIL161W	Cytosol	cytoplasm
41	YIR014W	Endoplasmic reticulum	endoplasmic reticulum; fungal-type vacuole
42	YIR035C	Cytosol	cytoplasm
43	YJL016W	Cytosol	cytoplasm
44	YJL055W	Cytosol	cytoplasm; nucleus
45	YJL070C	Cytosol	cytoplasm; mitochondrion
46	YJR015W	Endoplasmic reticulum	cytoplasm; endoplasmic reticulum
47	YJR149W	Cytosol	cytoplasm
48	YKL023W	Cytosol	cytoplasm

49	YKL075C	Cytosol	cytoplasm
50	YKR045C	Cytosol	cytoplasm
51	YLR177W	Cytosol	cytoplasm
52	YLR287C	Cytosol	cytoplasm
53	YLR326W	Plasma membrane	cell periphery
54	YLR345W	Cytosol	cytoplasm
55	YML096W	Cytosol	cytoplasm
56	YML119W	Cytosol	cytoplasm
57	YMR090W	Cytosol	cytoplasm
58	YMR196W	Cytosol	cytoplasm
59	YMR221C	vacuole	fungal-type vacuole membrane; mitochondrion
60	YMR253C	Cytosol	cytoplasm
61	YNL010W	Cytosol	nucleus; cytoplasm
62	YNL024C	Cytosol	cytoplasm
63	YNL300W	Cytosol	cytosol; fungal-type cell wall
64	YNR014W	Cytosol	cytoplasm
65	YNR029C	Cytosol	cytoplasm
66	YOR062C	Cytosol	nucleus; cytoplasm
67	YOR238W	Cytosol	cytoplasm
68	YOR289W	Cytosol	nucleus; cytoplasm
69	YOR385W	Cytosol	cytoplasm
70	YPL067C	Cytosol	cytoplasm
71	YPL108W	Cytosol	cytoplasm
72	YPL199C	Cytosol	cytoplasm; cell periphery
73	YPL245W	Cytosol	nucleus; cytoplasm
74	YPL247C	Cytosol	nucleus; cytoplasm
75	YPR010C-A	Cytosol	cytosol
76	YAL061W	Cytosol	nucleus; cytoplasm
77	YCR051W	Cytosol	nucleus; cytoplasm
78	YCR090C	Cytosol	nucleus; cytoplasm
79	YGL082W	Cytosol	nucleus; cytoplasm; plasma membrane; cell periphery
80	YKR023W	Cytosol	cytoplasm; mitochondrion
81	YKR075C	Cytosol	nucleus; cytoplasm
82	YLR419W	Cytosol	cytoplasm; mitochondrion
83	YML053C	Cytosol	nucleus; cytoplasm

84	YML079W	Cytosol	nucleus; cytoplasm
85	YML082W	Cytosol	nucleus; cytoplasm
86	YMR122W-A	Cytosol	cytoplasm; endoplasmic reticulum
87	YPL071C	Cytosol	nucleus; cytoplasm
88	YAR028W	vacuole; ER; Cytosol	fungal-type vacuole membrane
89	YBL095W	Mitochondria; Cytosol	mitochondrion; mitochondrial inner membrane
90	YCR016W	Cytosol; Nucleus	nucleus; nucleolus
91	YDR119W	Cytosol; vacuole; Mitochondria	fungal-type vacuole membrane
92	YDR476C	Cytosol; Plasma membrane; ER	endoplasmic reticulum
93	YGR117C	Cytosol; Nucleus	cytoplasm
94	YGR168C	Peroxisome; Cytosol	peroxisome
95	YHR033W	Cytosol; Nucleus	cytoplasm
96	YHR048W	Cytosol; vacuole; plasma membrane	cell periphery
97	YIL001W	Cytosol; Nucleus	cytoplasm
98	YJL161W	Mitochondria; Cytosol	mitochondrion
99	YJR154W	Cytosol; Nucleus	cytoplasm
100	YKL077W	ER; Cytosol; Golgi	endoplasmic reticulum; fungal-type vacuole
101	YLR278C	Cytosol; Nucleus	nucleus
102	YMR310C	Cytosol; Nucleus	nucleus
103	YOR093C	Cytosol; Nucleus	cytosol; cellular bud neck
104	YOR296W	Cytosol; Nucleus	cytoplasm
105	YBL010C	Cytosol	clathrin-coated vesicle
106	YBL059W	Cytosol	mitochondria
107	YBR047W	Cytosol	mitochondrion
108	YBR241C	Endoplasmic reticulum	fungal-type vacuole membrane
109	YBR287W	Cytosol	endoplasmic reticulum
110	YCL021W-A	Nucleus	fungal-type vacuole
111	YCR007C	Cytosol	fungal-type vacuole; cell periphery
112	YCR061W	vacuole; Mitochondria; ER; Plasma membrane	cytoplasm
113	YCR087C-A	Nucleus	nucleolus
114	YCR101C	Cytosol	fungal-type vacuole
115	YDL027C	Cytosol	mitochondrion; endoplasmic reticulum

116	YDL121C	Cytosol	endoplasmic reticulum
117	YDL157C	Cytosol	mitochondrion
118	YDL180W	Cytosol	fungal-type vacuole membrane
119	YDL199C	Cytosol	fungal-type vacuole membrane
120	YDL206W	Cytosol	fungal-type vacuole membrane
121	YDL211C	Cytosol	fungal-type vacuole
122	YDL241W	Cytosol	endoplasmic reticulum
123	YDR056C	Cytosol	endoplasmic reticulum
124	YDR061W	Cytosol	mitochondrion
125	YDR090C	Cytosol	plasma membrane
126	YDR262W	Cytosol	fungal-type vacuole
127	YDR371W	Nucleus; Nucleolus	cytoplasm
128	YDR387C	Cytosol	fungal-type vacuole membrane
129	YDR415C	Cytosol	fungal-type vacuole
130	YDR524C-B	Cytosol	endoplasmic reticulum
131	YER076C	Cytosol	mitochondrion; endoplasmic reticulum
132	YER077C	Cytosol; Nucleus	mitochondrion
133	YER182W	Cytosol	mitochondrion
134	YFL054C	Nucleus; Cytosol	cell periphery
135	YFR045W	Cytosol	mitochondria
136	YGL085W	Cytosol	mitochondrion
137	YGR016W	Cytosol	endoplasmic reticulum
138	YGR021W	Cytosol	mitochondrion
139	YGR026W	Cytosol	endoplasmic reticulum; cell periphery
140	YGR052W	Cytosol	mitochondrion
141	YGR125W	ER; Cytosol	fungal-type vacuole
142	YHL008C	Cytosol	fungal-type vacuole
143	YHL017W	Cytosol	clathrin-coated vesicle
144	YHL018W	Cytosol	mitochondrion
145	YHL026C	Cytosol	cell periphery
146	YHL042W	Endoplasmic reticulum	fungal-type vacuole
147	YHR045W	Cytosol; Nucleus	endoplasmic reticulum
148	YIL055C	Nucleus	mitochondrion
149	YIL067C	Cytosol	fungal-type vacuole
150	YIL127C	Cytosol	nucleolus
151	YJL132W	Cytosol	fungal-type vacuole

152	YJL147C	Cytosol	mitochondrion
153	YJR003C	Cytosol; Nucleus	mitochondrion
154	YJR039W	Cytosol; Nucleus	mitochondrion
155	YKL063C	Cytosol	Golgi apparatus
156	YKL071W	Nucleus	cytoplasm
157	YKL133C	Cytosol	mitochondrion
158	YKR051W	Cytosol	endoplasmic reticulum
159	YKR070W	Cytosol	mitochondrion
160	YLR001C	Cytosol	fungal-type vacuole membrane; mitochondrion
161	YLR046C	Cytosol	fungal-type vacuole
162	YLR050C	Cytosol	endoplasmic reticulum
163	YLR104W	Cytosol	fungal-type vacuole
164	YLR173W	Cytosol	fungal-type vacuole membrane
165	YLR283W	Cytosol	mitochondrion; endoplasmic reticulum
166	YLR426W	Cytosol	mitochondrion; endoplasmic reticulum
167	YLR454W	Cytosol	mitochondrion
168	YML037C	Cytosol	clathrin-coated vesicle
169	YMR144W	Cytosol	nucleus
170	YMR155W	Cytosol	fungal-type vacuole
171	YNL058C	Cytosol	fungal-type vacuole; endoplasmic reticulum
172	YNL144C	Cytosol	mitochondrion
173	YNL146W	Cytosol	endoplasmic reticulum
174	YNL168C	Cytosol	mitochondrion
175	YNL181W	Cytosol	nuclear envelope; endoplasmic reticulum
176	YNL217W	Cytosol	fungal-type vacuole
177	YNR021W	Cytosol	endoplasmic reticulum
178	YNR040W	Cytosol	mitochondrion
179	YNR065C	Nucleus; Cytosol	endoplasmic reticulum
180	YNR066C	Cytosol	fungal-type vacuole; membrane
181	YOR271C	Cytosol	mitochondrion
182	YOR292C	Cytosol	fungal-type vacuole
183	YPL107W	Cytosol	mitochondria
184	YPL109C	Cytosol; Nucleus	mitochondria
185	YPL168W	Nucleus	mitochondrion
186	YPL222W	Nucleus; Cytosol	mitochondrion
187	YPR003C	Cytosol	endoplasmic reticulum

188	YPR063C	Cytosol	endoplasmic reticulum
189	YPR071W	Cytosol	endoplasmic reticulum; cell periphery
190	YPR109W	Cytosol	endoplasmic reticulum
191	YPR114W	Cytosol	endoplasmic reticulum
192	YPR159C-A	Endoplasmic reticulum	cytosol

Table 12: Predicted localization and the known localization from high throughput studies for 192 ORFs. A comparison of the predicted and known localization for 192 ORFs is shown. The rows filled in yellow represent an exact match of predicted and the available localization, green filled rows represent partial match and red filled rows represent mismatch.

7.2.3 Experimental validation of predicted SCL of selected proteins

To verify if the predictions are correct, we experimentally tested the localization of six proteins chosen arbitrarily. Of the six proteins, three of them are predicted to localize to endoplasmic reticulum (Yhr140w, Yhl042w and Ypl264c), one to mitochondria (Ydr124w) and two to cytosol (Ypl088w and Yjl218w) (Table 13). While, Ypl264c, Ydr124w, Ypl088w and Yjl218w do not have any prior SCL data, Yhl042w is shown to localize to vacuole (Yofe et al. 2016) and Yhr140w is shown to localize to endoplasmic reticulum and cytoplasm in high throughput studies (Tkach et al. 2012).

Protein	**Predicted SCL**	**Verified SCL**
YHL140W	ER	ER
YHL042W	ER	ER
YPL264C	ER	ER
YDR124W	Mitochondria	Mitochondria
YPL088W	Cytosol and Nucleus	Cytosol
YJL218W	Cytosol	Mitochondria

Table 13: Predicted and experimentally determined SCL for six proteins. Comparison between the SCL predicted using our script and the experimentally determined localization for six proteins chosen arbitrarily.

In order to test the localization of the six proteins, we cloned the genes of interest into a vector containing GFP protein and expressed it as a fusion protein. We then checked the localization of

the fusion-proteins using live-cell imaging by co-localising with either dsRed-HDEL (ER marker) or Mito Tracker Red (mitochondrial dye) and DAPI (nuclear stain). As shown in Figure 28, we find that Yhr140w, Yhl042w and Ypl264c co-localize with the ER marker dsRed-HDEL, suggesting that they are localizing to ER in accordance with the prediction. The protein Ypl088w that was predicted to localize to cytoplasm and nucleus, was found to localize to cytosol.

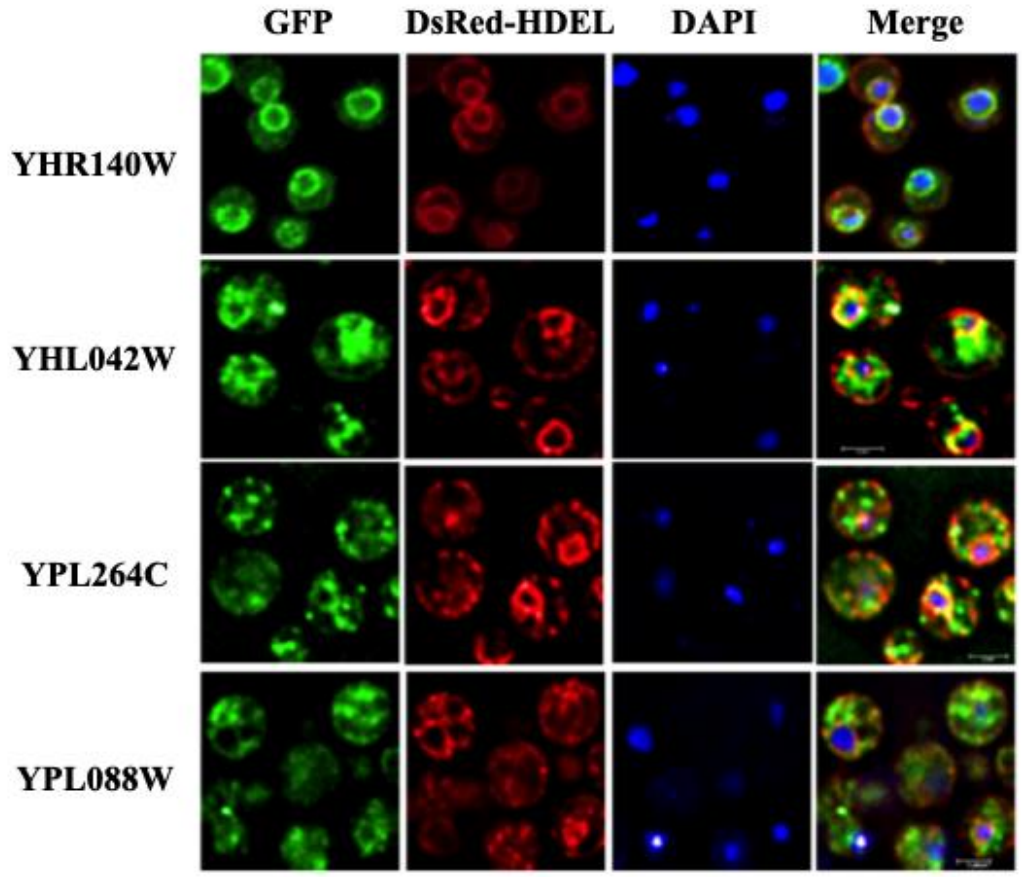

Figure 28: Subcellular localization of selected proteins with predicted localization co-stained with ER marker. Live-cell imaging of the strains transformed with the vector containing the protein of interest fused with GFP. DsRed-HDEL is used as the marker for ER and DAPI is used to mark the nucleus.

The proteins Ydr124w and Yjl218w that were predicted to localize to mitochondria and cytosol, respectively are found to co-localize with the Mito Tracker Red (Figure 29). This suggests that they are localizing to mitochondria.

Thus, of the 6 proteins for which the localization is tested, we find a correspondence between the prediction and the experimental result for five of them, while for one we did not. As mentioned earlier, this is possibly due to the limited interaction data and lack of precise annotation data for the interactors.

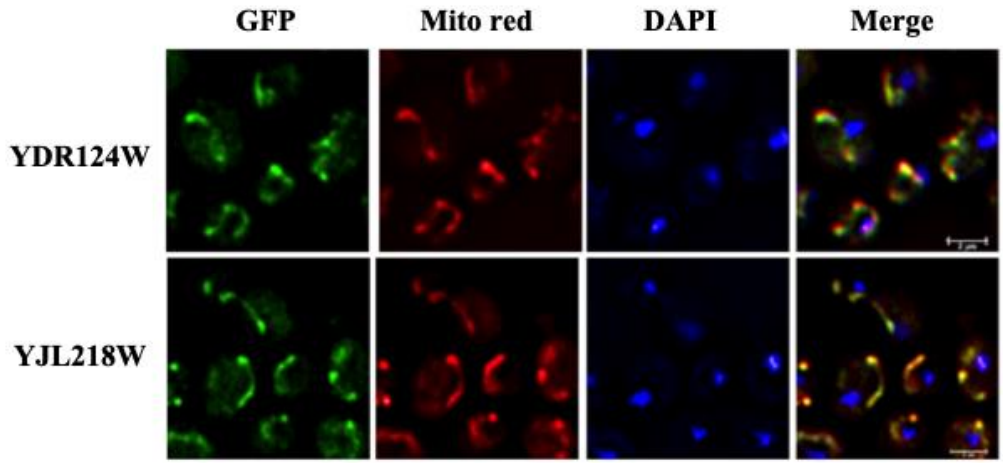

Figure 29: Subcellular localization of selected proteins with predicted localization co-stained with mitochondrial marker. Live-cell imaging of the strains transformed with the vector containing the protein of interest fused with GFP. Mito Red is used as the marker for mitochondria and DAPI is used to mark the nucleus.

To get some insight into the function of these proteins, the strains deleted for the three ER proteins were screened for ER, nuclear envelope and pore complex defects and the strains deleted for the two mitochondrial proteins were screened for mitochondrial morphology defects. In case of the ER proteins, immunofluorescence was performed to screen for nuclear organization defects. The deletion strains of each of the proteins were crossed with the strain in which an inner nuclear membrane protein, Esc1 is endogenously tagged with 13Myc. Immunofluorescence was then performed as described in methods. Antibodies against Myc and Nsp1 were used to stain the nuclear envelope and nuclear pore complex respectively. Wild type cells have smooth round nuclei with evenly distributed nuclear pore complexes. *ypl264cΔ* strain showed no defects in either nuclear envelope or pore complex and looked just like wild type. *yhr140wΔ* showed discontinuity in the nuclear envelope as well as clustering of nuclear pores, while in *yhl042wΔ* the nuclear envelope appeared distorted and was not spherical (Figure 30).

The deletion strains of the above 3 ER proteins were also looked for ER morphology defects and two of them (Yhr140w and Yhl042w) were found to have defects. This suggests that these two genes contribute to the maintenance of ER and nuclear morphology.

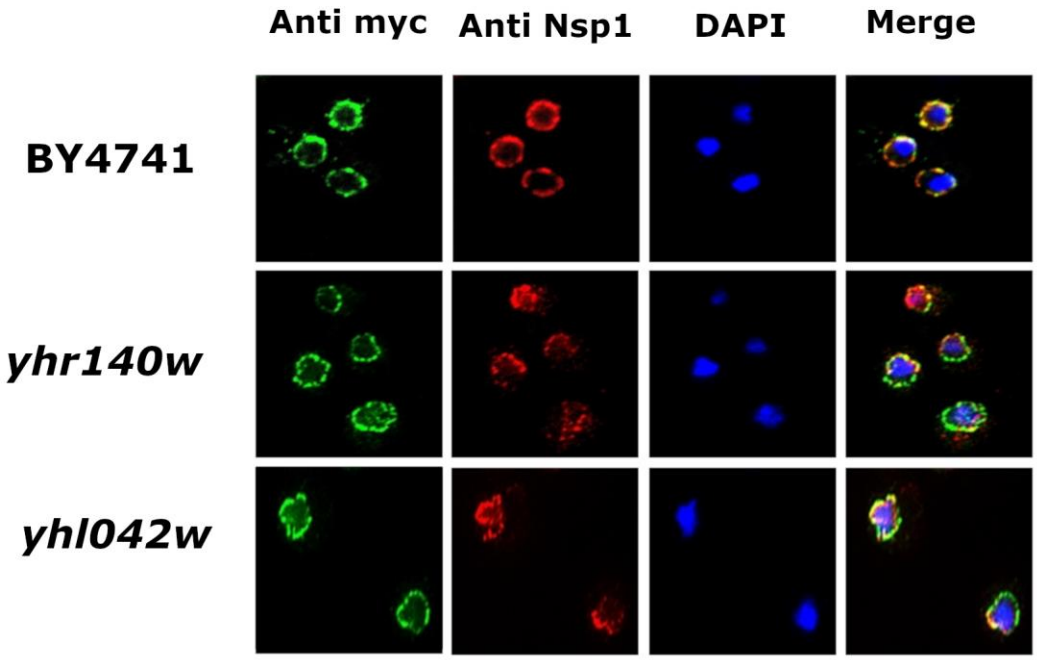

Figure 30: Nuclear organization defects observed in *yhr140w*Δ and *yhl042w*Δ. Immunofluorescence of the strains deleted for Yhr140w and Yhl042w. The inner nuclear membrane protein, Esc1-Myc (green) and nuclear pore protein, Nsp1 (red) were detected using anti-Myc and anti-Nsp1 antibodies.

Similarly, the strains lacking the two mitochondria localizing proteins were checked for mitochondrial morphology defects, if any. Live-cell imaging was performed with the deletion strains of each of these proteins that were transformed with a mitochondrial marker mito-GFP. In wild type cells mitochondria appear as a tubular network, while severe fragmentation of mitochondria is observed in *ydr124w*Δ cells (Figure 31). No defects were observed in the *yjl218w*Δ cells. This shows that Ydr124w has a role in the maintenance of mitochondrial morphology.

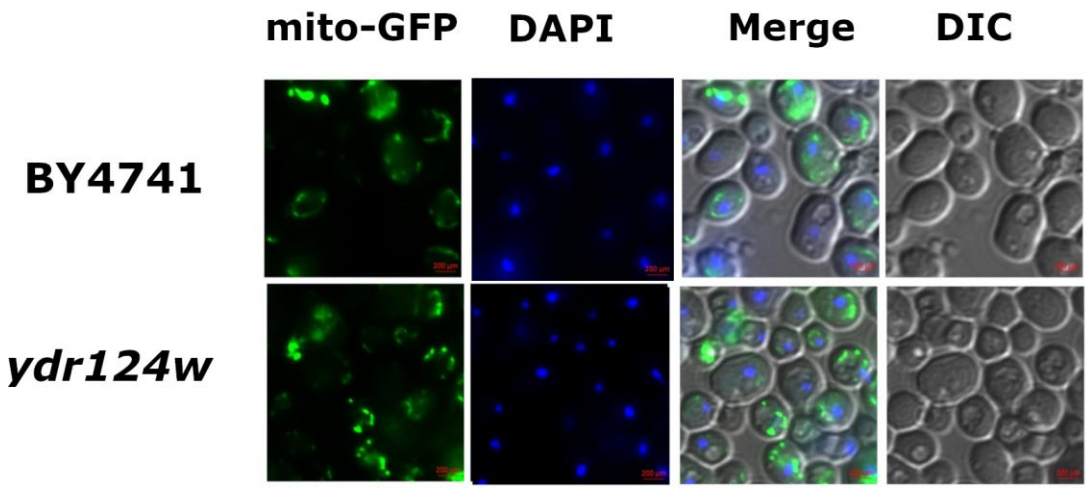

Figure 31: Mitochondrial morphology defects observed in *ydr124w*Δ cells. Live-cell imaging of the WT and the strain deleted for Ydr124w. Mito-GFP is used as marker for mitochondria and DAPI marks the nucleus.

7.3 Conclusions

The subcellular location of a protein can be predicted based on its interacting partners. A Perl script was developed to predict the SCL of proteins. The script could predict with high accuracy the SCL of proteins for which localization data is available. For 93 proteins out of 100, the prediction matched that of the known localization. Further, the script was used to predict the SCL of proteins with no localization data. Experimental validation was carried out for few proteins and the predictions were found to be correct for 5 out of 6 proteins. The protein, for which the experimental results did not match the prediction, was predicted to localize to cytosol, while its actual localization was found to be to a specific organelle within the cytosol (mitochondria). Similarly, in case of the uncharacterized ORFs for which high-throughput data is available we find quite a number of mismatches (88/192). Of these 88, around 70 of them are predicted to localize to cytosol, while the high-throughput data shows its localization to a specific organelle in the cytosol such as ER/Mitochondria/Vacuole.

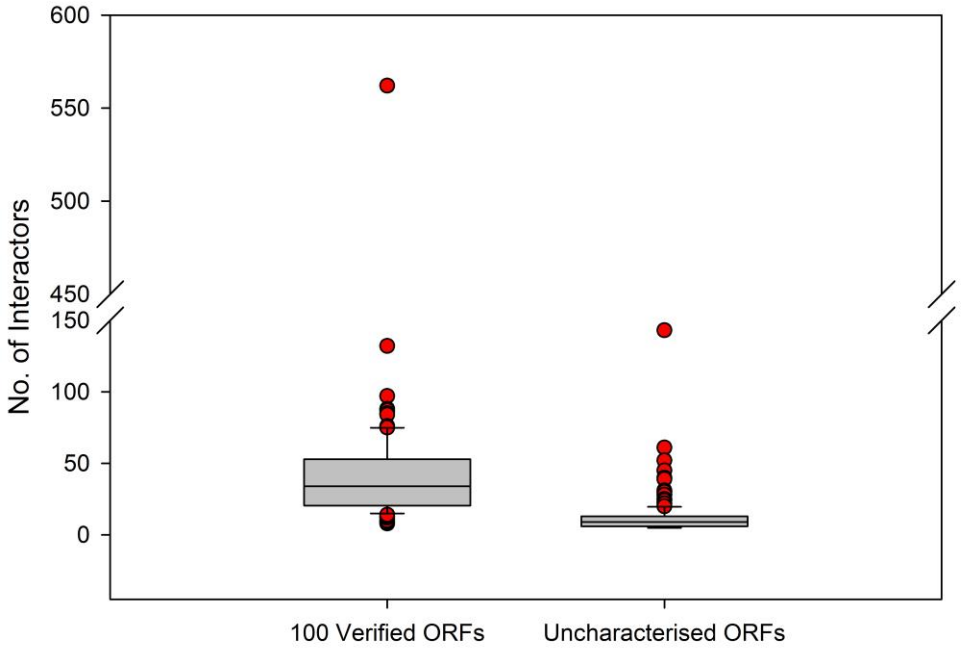

Figure 32: A box plot showing the number of interactors for verified and uncharacterized ORFs. The number of interactors for the 100 verified ORFs (prediction accuracy 93/100) and the 192 uncharacterized ORFs with high-throughput data (prediction accuracy 104/192) are shown in the box plot.

The inability of the script to predict the specific localization in these cases is possibly due to the availability of only a limited number of interactors for the uncharacterized ORFs. We compared the number of interactors for the 100 verified ORFs and the 192 uncharacterised ORFs. We find that the verified ORFs for which the prediction accuracy is high have higher number of interactors (mean= 44; median = 34) compared to the uncharacterized ORFs (mean= 11.9; median = 9) (Figure 32). Thus the accuracy of the predictions with our script increases with increase in the interaction data. The performance of the script depends on the amount of interaction data available and the associated annotation data.

Chapter 8

Discussion

Nuclear envelope forms a protective barrier around the genome and separates it from the rest of the cellular components. The NE also partakes in essential functions like maintenance of nuclear architecture, chromatin organization, control of transcription and DNA repair. The proteins associated with or integrated in the nuclear envelope mediate these functions. Defects in the nuclear envelope proteins are shown to affect the nuclear morphology and lead to disorders such as progeria, muscular dystrophy etc. Thus, the nuclear envelope proteins play an important role in the health and survival of the organism. However, it is still unknown if these functions are conserved across all extant eukaryotes or if they have specifically evolved in certain lineages. Knowledge of the nuclear envelope proteins across various eukaryotes will help us infer the conserved NE components and hence the composition of the LECA nuclear envelope. However, the composition of the nuclear envelope is known only in very few model organisms.

In this work, using the experimentally known NE proteins of a model organism, we have taken a sequence comparison approach to unravel the fundamental components of the nuclear envelope. In contrast to previous studies, which focused only on a small subset of NE proteins, in this study we considered all known nuclear envelope proteins of *S. cerevisiae* that participate in various nuclear functions.

Our comparative genomic study of nuclear envelope proteins in eukaryotic supergroups has identified a set of 24 proteins of which 22 are present in all supergroups and 2 in four of the supergroups. Of the 24, 11 localize to NE/ER in either human/mouse/Arabidopsis (Table 14; Figure 33; Additional data 5). We speculate that these were likely components of the early ancestor of eukaryotes, the LECA and perhaps carry out similar functions in all organisms including LECA. This comprehensive analysis serves as a starting point to understand the composition and complexity of the ancestral nuclear envelope.

Yeast		Human		Mouse		Arabidopsis	
Protein	**Loc**	**Protein**	**Loc**	**Protein**	**Loc**	**Protein**	**Loc**
Nem1	NE & ER	**Dullard**	NE, ER				
Ulp1	Nuclear pore	**Senp2**	Nuclear pore				
Hmg1 & Hmg2	ONM & ER	**Hmgcr**	ER			**Hmgr1**	ER
Ssm4	INM & ER	**March6**	ER				
Pct1	ONM & ER			**Pcyt1a**	ER		
Ntf2	Nuclear pore					**Ntf2A**	NE
						Ntf2b	NE
Heh2 & Src1	INM	**Man1**	INM	**Man1**	NE		
		Lemd2	INM	**Lem2**	NE		
Mps3	INM	**Sun2**	NE	**Sun2**	NE	**Sun1**	NE
		Sun1	NE	**Sun1**	NE	**Sun2**	NE
Slp1	ONM & ER					**Sun4**	NE, ER
						Sun3	NE, ER

Table 14: Localization data of the LECA NE protein homologs. The LECA proteins whose homologs in either human or mouse or Arabidopsis are known to localize to NE or ER or nuclear pore are shown.

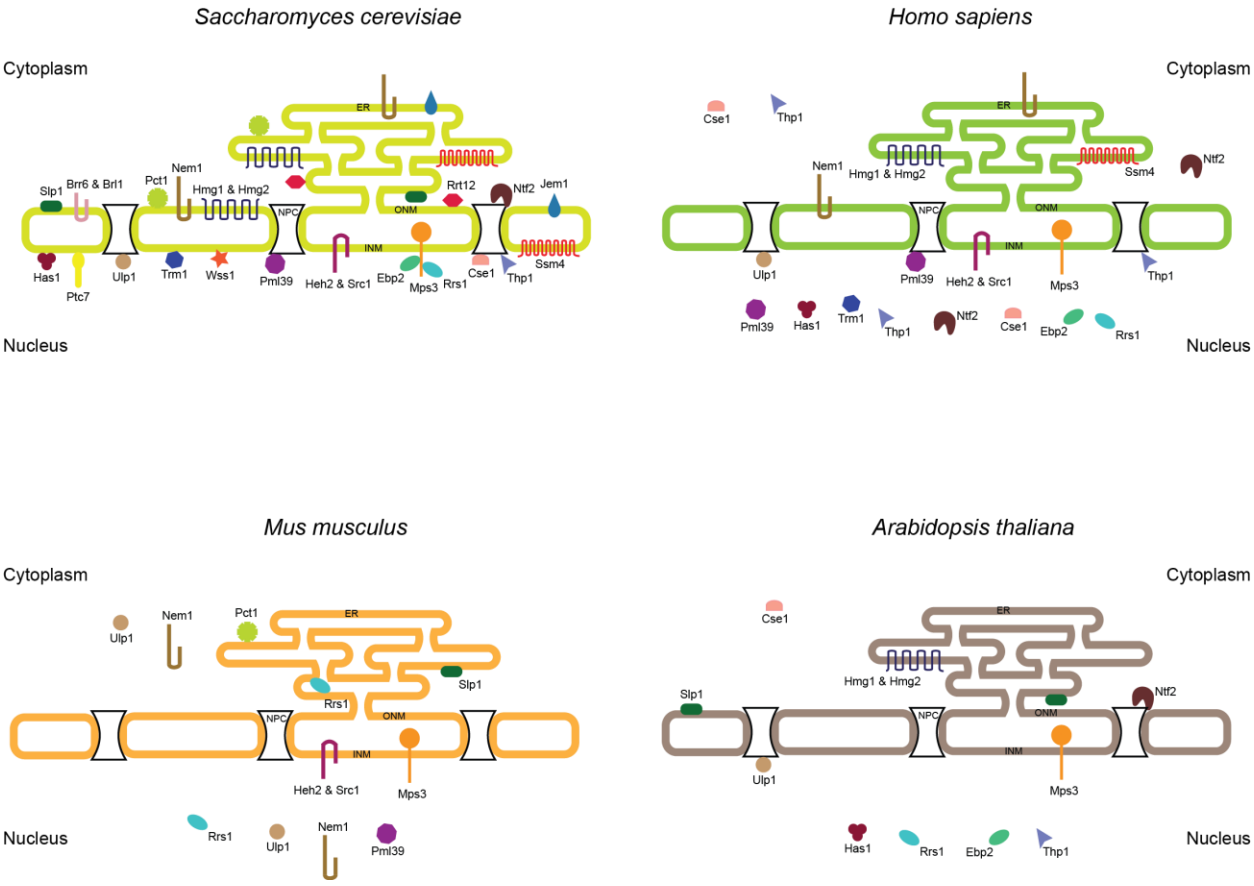

Figure 33: Localization of the homologs of LECA NE proteins in a few model organisms. The figure shows the *Saccharomyces cerevisiae* NE proteins identified to be part of the LECA proteome and the localization of the homologs of these proteins in *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*. Only the proteins whose localization is detrmined from experimental studies and which localize to nucleus/INM/ONM/nuclear pore/ER/cytoplasm are shown.

8.1 Core/LECA NE proteins

In *S. cerevisiae*, loss of the chromatin interacting proteins of the core proteome, leads to nuclear morphology defects. The function of some of these proteins is conserved across eukaryotic supergroups. For example, the C-terminal SUN domain protein in *S. cerevisiae* is required for SPB duplication and insertion into the nuclear envelope (Jaspersen et al. 2006), while the ortholog identified in the evolutionarily distant amoeba, *D. discoideum* maintains the connection between centrosome and nuclear envelope through its interaction with chromatin (Xiong et al.

2008). In addition, the C-terminal SUN domain proteins in yeast, animals and plants are known to tether telomeres to the nuclear periphery during meiosis (Rothballer and Kutay 2013; Varas et al. 2015). The paralogous proteins Heh2 and Src1, containing the HeH and MSC domains tether telomeres and rDNA to the nuclear periphery in yeast. The orthologs of these proteins in *S. pombe* and in human are critical for maintaining nuclear envelope morphology through their interactions with chromatin (Ulbert et al. 2006; Schreiner et al. 2015). Positioning of chromosomes in the nucleus, which in turn influences gene expression, is regulated through interaction with these nuclear envelope proteins. This suggests an early evolution of the chromatin-NE interaction and consequent gene regulation mechanisms.

Two proteins, Ebp2 and Rrs1 involved in ribosome biogenesis and telomere clustering in *Saccharomyces cerevisiae* (Horigome et al. 2011) are present in almost all the organisms considered in this study. The human Rrs1 ortholog also contributes to proper separation of the chromosomes during mitosis in addition to regulating the ribosome synthesis (Gambe et al. 2009). This suggests a conserved function of the Rrs1 protein in chromatin interaction in opisthokonts. As more functional data from eukaryotes become available we would know if the chromatin interaction in ribosome biogenesis proteins is ancient or a feature evolved only in opisthokonts.

Another important class of proteins conserved across supergroups is the SUMO proteases and ubiquitin ligases. The SUMO protease Ulp1, is associated with nuclear pores in *S. cerevisiae* where it desumoylates, among others, specific transcription activators and repressors and regulates the transcription of genes in an NPC dependent manner (Texari et al. 2013). One of the Ulp1 orthologs in *Arabidopsis thaliana* Esd4, also identified as a part of the core proteome in this study, localizes to the nuclear periphery and the mutants have low levels of a transcription factor which acts as repressor for flowering (Murtas et al. 2003; Hermkes et al. 2011). Similarly, the Ubiquitin ligase Ssm4, present at the INM in yeast, degrades the transcription factor matα2 that represses a-specific genes in α cells (Chen et al. 1993; Swanson et al. 2001; Deng and Hochstrasser 2006). We find orthologs of Ssm4 across all eukaryotes and the ortholog in *Arabidopsis*, Sud1 is found to regulate HMG-CoA reductase activity (Doblas et al. 2013). However, the mechanism of this regulation is still unknown. Together, these data indicate that the SUMO and ubiquitin mediated protein homeostasis is a conserved function associated with

the nuclear envelope. Since this is found in both unikonts and bikonts, we speculate that this property evolved in the ancient nuclear envelope.

A significant number of proteins that are involved in lipid biosynthesis are part of the LECA nuclear envelope proteome. The nuclear envelope expansion during cell division (Hetzer et al. 2005; Takemoto et al. 2016) requires additional nuclear membrane synthesis that is regulated by the proteins associated with the ER-ONM network. The Nem1-Spo7 phosphatase complex in yeast dephosphorylates the PA phosphatase, Lipin/Pah1, that mediates the conversion of phosphatidic acid (PA) to diacylglycerol (DAG) and thus restrict membrane growth. On the other hand, the phosphorylation of Pah1 allows the nuclear membrane growth (Siniossoglou 2009). The human Nem1 ortholog, Dullard, is an NE protein and ectopic expression in yeast rescues the NE defects of *nem1Δ* cells (Siniossoglou et al. 1998; Kim et al. 2007). Recent studies demonstrated the Pah1 and Nem1 mediated regulation of lipid droplet number in the ciliate *Tetrahymena thermophila* (Pillai et al. 2017). This suggests the presence of a conserved mechanism for regulating lipid biosynthesis and membrane homeostasis across eukaryotes. Pct1 gene involved in phosphatidylcholine synthesis and HMG-CoA reductase involved in sterol biosynthesis are found in organisms across all supergroups. In yeast, over-production of Hmg1 leads to karmellae formation (Wright et al. 1988). Similarly, the deletion of HMG-CoA reductase in Arabidopsis leads to altered ER morphology around the nucleus (Ferrero et al. 2015). In *S. cerevisiae*, the proteins Brr6, Brl1 and Apq12 are integral membrane proteins and form a complex. The mutants of these proteins are found to have altered lipid composition in membranes along with defects in nuclear envelope morphology and NPC biogenesis (Hodge et al. 2010; Zhang et al. 2018). Although we could not detect Archaeplastida homologues for Nem1 or Pct1 and Hmg1/2 in algae and the Brr6/Brl1 are less widely distributed among members of the supergroups, association of proteins regulating membrane biosynthesis is a widely conserved feature of nuclear envelopes of most eukaryotes.

8.2 Fungal specific NE proteins

Another important finding from this study is the identification of 13 proteins specific to ascomycetes, potentially appearing after the Ascomycota-Basidiomycota split. Of the 13 specific

to ascomycetes, 10 are restricted to Saccharomycetes. As most of these proteins are found to be rapidly evolving, it is possible that the homolog in organisms outside ascomycetes have diverged to an extent that they cannot be identified by sequence based searches. Nevertheless, the presence of around 20% of NE proteome unique to Saccharomycetes is an indication of the fast evolving nature of the nuclear envelope proteome. An early proteomic study revealed that there are over 60 nuclear envelope proteins in animals (Schirmer et al. 2003; Wilkie et al. 2011) suggesting that the nuclear envelope proteome has undergone tremendous expansion. These data hint at the potential for multiple NE proteins specific to each lineage to have evolved.

8.3 Origins of nuclear envelope proteins

The evolutionary origin of the nucleus is not yet understood. The existing theories for the evolution of nucleus suffer from lack of enough evidence. Models propose either an archaeal or bacterial origin of nuclear envelope (Martin 1999; Baum and Baum 2014). Tracing the origin of the structural components of the nucleus could be one means of providing insights into its origin. In this context, identifying the prokaryotic origins of the LECA NE proteome identified in this study would enhance our knowledge of the evolution of nucleus. We thus searched for the homologs of the core and non-linearly conserved NE proteins in bacteria and archaea. Of the 32 NE proteins, homologs of only 3 proteins viz., Hmg1, Hmg2 and Trm1 were found in both archaea and bacteria, and the homologs of one protein Ntf2 were found only in bacteria. No homolog could be identified for the remaining 28 proteins and are considered to be eukaryotic specific. However, 13 of the eukaryotic specific proteins are found to contain prokaryotic domains. While, 7 of these proteins contain domains that are present in both archaea and bacteria, 6 of them contain domains that are present exclusively in bacteria. These findings suggest that most of the nuclear envelope proteins evolved in the LECA either as eukaryotic specific innovations, or by using prokaryotic domains or a combination of both. Interestingly, the nuclear envelope proteins although contain bacteria specific domains we could not find domain(s) exclusively present in archaea.

The nuclear pore complex proteins that are thought to have coevolved with the NE to allow transport of molecules between the cytoplasm and nucleus, are also found to lack prokaryotic

homologs and are considered as eukaryotic innovations. However, proteins sharing similar folds to that of the nuclear pore proteins are found in bacteria belonging to the PVC superphylum and are found to localize close to the vesicle membranes (Santarella-Mellwig et al. 2010), while no such proteins could be found in archaea. In this study, we find several domains present in eukaryotic NE proteins to be of bacterial origin. For example, the HeH domain found in chromatin interacting proteins is found in bacteria but not in archaea. The presence of nucleic acid binding domain in multiple ribosomal proteins suggests additional regulatory roles of these proteins. HeH domains have been reported in iridoviruses and phages of firmicute bacteria and are proposed to be involved in organization of viral DNA in the capsid (de Souza et al. 2010). This suggests that the proteins that are a part of the nuclear membrane predominantly evolved by recruiting the bacterial domains. It is likely that the nucleus evolved by vesicular biogenesis model, which involves the transfer of bacterial genes (involved in lipid biosynthesis and several others) to the archaeal host followed by the synthesis of vesicles that fused and gave rise to the endomembrane system (Martin 1999).

The homologs of HMG-CoA reductases and tRNA methyl transferases are the only two universally conserved eukaryotic proteins found in both archaea and bacteria. In this study, we find that the class I and class II HMG-CoA reductases are present across all three domains of life, suggesting ancient origins for the two classes of enzymes. Interestingly, we find that the Asgard archaea, which are now considered as the closest relatives of eukaryotes (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017) are found to contain class II proteins but not class I. However, most of the eukaryotes contain the class I proteins. If we accept the idea that the host for the endosymbiotic event that led to the evolution of eukaryotes lies within the Asgard superphylum, the eukaryotic class I HMG-CoA reductases are then probably of bacterial origin.

The tRNA methyl transferases are found in a large number of archaea, while in bacteria it is found only in Cyanobacteria, Aquificae (specifically only in the family Aquificaceae) and Armatimonadetes. The presence of the Trm1 homologs in only one family of Aquificae, which are found within the archaeal clade in the ML tree, suggests an HGT event from archaea to Aquificaceae. As Aquificaceae are found in extreme environments like hot springs and hydrothermal vents where they cohabit with archaea, this is a strong possibility. The clustering of additional homolog of Trm1 in Archaeplastida with the Cyanobacteria suggests an independent

gain of this homolog into plants probably corresponding to the endosymbiotic event leading to the evolution of chloroplasts. Consistently, the additional Trm1 homolog in Arabidopsis that is found to cluster with cyanobacteria is localized to the plastid. Previous phylogenomic studies have also pointed to the episodic influx of bacterial genes into the eukaryotes, which corresponds to the endosymbiotic origins of mitochondria and chloroplasts (Ku et al. 2015). The archaeal origin of Trm1 in eukaryotes is consistent with the informational genes having been acquired from archaea and the chimeric nature of the eukaryotic genome.

Interestingly, we find the homologs of Ntf2 protein only in Streptomycetaceae bacteria. In the ML tree, the Ntf2 homologs cluster with the eukaryotic homologs suggesting horizontal gene transfer from eukaryotes to bacteria. The fact that the Ntf2 homologs in bacteria do not form a monophyletic group in the maximum likelihood tree suggests a possibility of multiple HGT events from eukaryotes to bacteria. Few cases of HGT from eukaryotes to bacteria have also been reported earlier. For example, some of the *Wolbachia* strains that infect several arthropod species are found to share salivary gland surface proteins found in mosquitoes (Klasson et al. 2009; Woolfit et al. 2009). Similarly, *Legionella pneumophila* are shown to contain several proteins of eukaryotic origins (Lurie-Weinberger et al. 2010). In this study, we also find a homolog of Ptc7 protein in a bacterium, *Candidatus Cardinium*, an endosymbiont of the arthropod *Bemisia tabaci*. This is probably a recent HGT event. Thus, horizontal gene transfer events from eukaryotes to bacteria are found to be more prevalent than thought earlier.

8.4 Caveats of the study

This study presents a comprehensive picture of the ancient nuclear envelope proteome and identifies the prokaryotic origins of the components. However, there are some limitations. One, many eukaryotes and especially yeasts, have undergone reductive evolution, and therefore, many NE proteins, originally part of LECA NE, may have been lost in yeast but present in other organisms. These would not be identified in this study. Second, there are limitations of sequence-based methods for capturing homologs. Though careful analysis with stringent cut-offs was performed, the identified homologs may still contain some false positives (a protein which is not a homolog) and/or false negatives (failure to detect a homolog). As many proteins included in the

analysis are rapidly evolving, there is a high chance for false negatives being present. For example, while no homolog for Spo7 could be identified in metazoa in this study using the *S. cerevisiae* protein sequence, homology searches using the *S. pombe* protein sequence did find a Spo7 ortholog (Han et al. 2012). Similarly, a more significantly diverged counterpart of Wss1, Spartan, was identified in mammals recently (Stingele et al. 2015). The failure to detect homologs because of sequence divergence in such cases would falsely implicate gene loss and may also lead to under-representation of genuinely conserved proteins. Using multiple experimental datasets for NE to start this search would be more comprehensive; however, this sort of data is not available currently. Despite these caveats, this study serves as a first step towards reconstructing the LECA NE proteome and identifying their prokaryotic homologs. With further experimental evidence of NE proteins from diverse organisms we would be able to build a complete picture of this key evolutionary innovation.

8.5 Future prospects

Identifying the complete nuclear envelope proteome of LECA would be a key contribution towards tracing the origins of the nucleus and understanding eukaryogenesis better. In this study, using the *S. cerevisiae* nuclear envelope proteome as the basis, we identified the NE proteome of LECA. We find that proteins belonging to various functional classes such as those involved in maintaining membrane and protein homeostasis are present in LECA. However, this study identifies only the minimal nuclear envelope proteome, as there would be several proteins that are present in other eukaryotes but lost in *S. cerevisiae*. Thus, similar studies using the experimentally determined nuclear envelope proteins in various diverse eukaryotes, as and when available, will be useful in reconstructing the complete NE proteome of LECA.

In this work, we have used a sequence-based approach. Studies using structure-based approach have identified more potential homologs (Devos et al. 2004; Santarella-Mellwig et al. 2010). Therefore, a similar approach for the proteins identified as part of nuclear envelope in yeasts and other organisms would provide a wider picture of the conserved nuclear envelope proteome as well as their evolutionary origins.

The nuclear envelope proteome seems to be one of the classes of proteins that have expanded with increasing developmental complexity. Similarly, in our analysis we discovered about 20% of yeast NE proteins that do not seem to have any homologs outside of the fungal kingdom. For those proteins that are fungal specific, the sequences beyond the conserved short motif are not conserved. Together these data indicate that there are several NE proteins that are specific to lineages and could carry out specific functions. Identifying such proteins in pathogenic eukaryotic organisms would be potential drug targets as they would likely be unique.

Once the complete NE proteome of LECA is identified, tracing their prokaryotic origins would help in understanding the evolution of nucleus. In this study, we find that most of the NE proteins have evolved in the LECA. The contribution of bacterial domains to the nuclear envelope proteome is found to be higher than the archaea. Recent studies point to the presence of several eukaryotic specific proteins in Asgard archaea. Though Asgard archaea have been included in this study, due to the relatively poor quality of the sequence data, true homologs may have been missed out. As more archaeal genome sequences of better quality become available, we would be able to better understand the origin of nucleus.

Appendix

A-1: Construction of C-terminal GFP tag for genes of interest

The six ORFs arbitrarily chosen to experimentally test the predicted localisation were cloned into the pUG23 vector containing a C-terminal GFP sequence. In order to clone the genes of our interest, each gene along with the promoter sequence was amplified (primer sequences are given in Table 3). The amplified gene products were then digested with SacI-SalI and inserted at the SacI-SalI digested pUG23 vector (Figure A.1). Clones obtained were confirmed by sequencing.

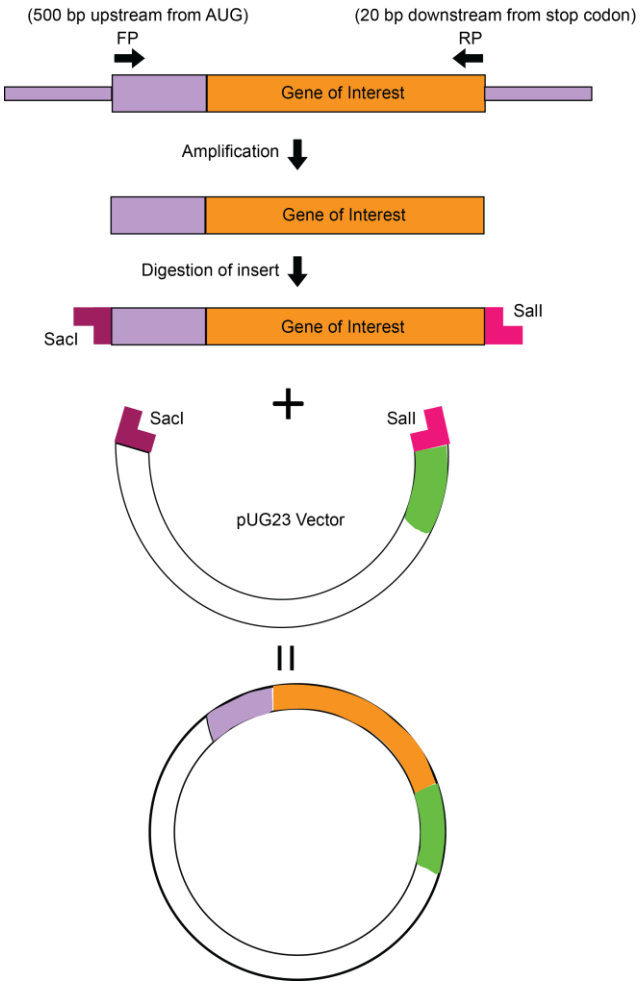

Figure A.1: Construction of C-terminal GFP tagged genes of interest. Schematic representation of cloning of genes into pUG23 vector.

Table A.1: Homologs of Ebp2 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	Organism	Protein Accession	Short name
01	Caenorhabditis elegans	NP_495125.1	Cele
02	Drosophila melanogaster	NP_651850.1	Dmel
03	Anopheles gambiae	XP_317941.4	Agam
04	Ciona intestinalis	XP_009859197.1	Cint
05	Danio rerio	NP_001003840.1	Drer
06	Takifugu rubripes	XP_011613408.1	Trub
07	Anolis carolinensis	XP_003220244.1	Acar
08	Gallus gallus	XP_422396.1	Ggal
09	Ornithorhynchus anatinus	XP_007663065.1	Oana
10	Monodelphis domestica	XP_001376638.1	Mdom
11	Canis lupus familiaris	XP_853475.1	Cfam
12	Sus scrofa	XP_003128129.1	Sscr
13	Mus musculus	NP_081208.1	Mmus
14	Pan troglodytes	XP_009454032.1	Ptro
15	Homo sapiens	NP_006815.2	Hsap
16	Strongylocentrotus purpuratus	XP_794612.1	Spur1
17	Strongylocentrotus purpuratus	XP_001176368.1	Spur2
18	Amphimedon queenslandica	XP_003388417.2	Aque
19	Arthroderma otae	XP_002845253.1	Aota
20	Aspergillus nidulans	XP_657678.1	Anid
21	Neosartorya fischeri	XP_001259469.1	Nfis
22	Leptosphaeria maculans	XP_003839984.1	Lmac
23	Parastagonospora nodorum	XP_001800276.1	Pnod
24	Botrytis cinerea	XP_001545713.1	Bcin
25	Sclerotinia sclerotiorum	XP_001587563.1	Sscl
26	Chaetomium globosum	XP_001229087.1	Cglo
27	Thielavia terrestris	XP_003648933.1	Tter
28	Neurospora crassa	XP_960774.1	Ncra
29	Tuber melanosporum	XP_002836722.1	Tmel
30	Saccharomyces cerevisiae	NP_012749.1	Scer
31	Candida glabrata	XP_448788.1	Cgla
32	Zygosaccharomyces rouxii	XP_002498791.1	Zrou
33	Kluyveromyces lactis	XP_451299.1	Klac
34	Schizosaccharomyces pombe	NP_593575.1	Spom
35	Agaricus bisporus	XP_006462807.1	Abis
36	Schizophyllum commune	XP_003034061.1	Scom
37	Trametes versicolor	XP_008037701.1	Tver
38	Auricularia delicata	XP_007341367.1	Adel
39	Cryptococcus neoformans	XP_012049407.1	Cneo
40	Ustilago maydis	XP_011391671.1	Umay
41	Puccinia graminis	XP_003326528.1	Pgra1
42	Puccinia graminis	XP_003332495.1	Pgra2
43	Batrachochytrium dendrobatidis	XP_006682324.1	Bden
44	Entamoeba histolytica	XP_001913443.1	Ehis1
45	Entamoeba histolytica	XP_656280.1	Ehis2
46	Acytostelium subglobosum	XP_012753003.1	Asub
47	Dictyostelium discoideum	XP_638424.1	Ddis
48	Polysphondylium pallidum	XP_020433302.1	Ppal

49	Trypanosoma brucei	XP_011777122.1	Tbru
50	Leishmania major	XP_003722573.1	Lmaj
51	Naegleria gruberi	XP_002674987.1	Ngru
52	Trichomonas vaginalis	XP_001305814.1	Tvag1
53	Trichomonas vaginalis	XP_001308555.1	Tvag2
54	Giardia lamblia	XP_001708180.1	Glam
55	Plasmodium falciparum	XP_001347561.1	Pfal
56	Theileria parva	XP_766086.1	Tpar
57	Babesia bovis	XP_001608930.1	Bbov
58	Toxoplasma gondii	XP_018634934.1	Tgon
59	Cryptosporidium hominis	XP_665806.1	Chom
60	Paramecium tetraurelia	XP_001423519.1	Ptet1
61	Paramecium tetraurelia	XP_001346830.1	Ptet2
62	Tetrahymena thermophila	XP_012656518.1	Tthe
63	Phytophthora infestans	XP_002899337.1	Pinf
64	Phaeodactylum tricornutum	XP_002176982.1	Ptri
65	Thalassiosira pseudonana	XP_002292294.1	Tpse
66	Bigelowiella natans	aug1.24_g8848	Bnat
67	Cyanophora paradoxa	Contig6949	Cpar
68	Chondrus crispus	XP_005715224.1	Ccri
69	Cyanidioschyzon merolae	XP_005535407.1	Cmer
70	Chlamydomonas reinhardtii	XP_001692490.1	Crei
71	Volvox carteri	XP_002950380.1	Vcar
72	Physcomitrella patens	XP_001762236.1	Ppat
73	Marchantia polymorpha	OAE23983.1	Mpol
74	Amborella trichopoda	XP_020527000.1	Atri
75	Oryza sativa	NP_001059623.1	Osat
76	Zea mays	NP_001183926.1	Zmay1
77	Zea mays	NP_001168353.1	Zmay2
78	Arabidopsis thaliana	NP_188905.1	Atha
79	Glycine max	XP_003556506.1	Gmax1
80	Glycine max	NP_001241039.1	Gmax2

Table A.2: Homologs of Rrs1 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Caenorhabditis elegans	NP_506573.1	Cele
02	Drosophila melanogaster	NP_729108.2	Dmel
03	Anopheles gambiae	XP_308939.3	Agam
04	Ciona intestinalis	XP_002125582.1	Cint
05	Danio rerio	NP_956356.1	Drer
06	Takifugu rubripes	XP_003968019.1	Trub
07	Gallus gallus	XP_015138233.2	Ggal
08	Ornithorhynchus anatinus	XP_007660407.1	Oana
09	Monodelphis domestica	XP_001379079.1	Mdom
10	Canis lupus familiaris	XP_005638081.1	Cfam
11	Sus scrofa	XP_003125652.1	Sscr
12	Mus musculus	NP_067486.2	Mmus
13	Pan troglodytes	XP_016815019.1	Ptro
14	Homo sapiens	NP_055984.1	Hsap

15	Strongylocentrotus purpuratus	XP_781111.1	Spur
16	Amphimedon queenslandica	XP_003382917.1	Aque
17	Arthroderma otae	XP_002843610.1	Aota
18	Aspergillus nidulans	XP_661349.1	Anid
19	Neosartorya fischeri	XP_001261365.1	Nfis
20	Leptosphaeria maculans	XP_003840559.1	Lmac
21	Parastagonospora nodorum	XP_001794249.1	Pnod
22	Botrytis cinerea	XP_001556485.1	Bcin
23	Sclerotinia sclerotiorum	XP_001586192.1	Sscl
24	Chaetomium globosum	XP_001220295.1	Cglo
25	Thielavia terrestris	XP_003652476.1	Tter
26	Neurospora crassa	XP_963880.1	Ncra
27	Tuber melanosporum	XP_002836930.1	Tmel
28	Saccharomyces cerevisiae	NP_014937.1	Scer
29	Candida glabrata	XP_446689.1	Cgla
30	Zygosaccharomyces rouxii	XP_002498658.1	Zrou
31	Kluyveromyces lactis	XP_452112.1	Klac
32	Schizosaccharomyces pombe	NP_595844.1	Spom
33	Agaricus bisporus	XP_006456780.1	Abis
34	Schizophyllum commune	XP_003035670.1	Scom
35	Trametes versicolor	XP_008035531.1	Tver
36	Auricularia delicata	XP_007338271.1	Adel
37	Cryptococcus neoformans	XP_012053372.1	Cneo
38	Ustilago maydis	XP_011391629.1	Umay
39	Puccinia graminis	XP_003324729.2	Pgra
40	Batrachochytrium dendrobatidis	XP_006678309.1	Bden
41	Encephalitozoon intestinalis	XP_003073440.1	Eint
42	Entamoeba histolytica	XP_652536.1	Ehis
43	Acytostelium subglobosum	XP_012748565.1	Asub
44	Dictyostelium discoideum	XP_644082.1	Ddis
45	Polysphondylium pallidum	XP_020432459.1	Ppal
46	Trypanosoma brucei	XP_011773986.1	Tbru
47	Leishmania major	XP_001684652.1	Lmaj
48	Naegleria gruberi	XP_002680317.1	Ngru
49	Trichomonas vaginalis	XP_001324446.1	Tvag
50	Giardia lamblia	XP_001704078.1	Glam
51	Plasmodium falciparum	XP_001347930.1	Pfal
52	Theileria parva	XP_766467.1	Tpar
53	Babesia bovis	XP_001610749.1	Bbov
54	Toxoplasma gondii	XP_002369902.1	Tgon
55	Cryptosporidium hominis	XP_665358.1	Chom
56	Tetrahymena thermophila	XP_001011682.1	Tthe
57	Phytophthora infestans	XP_002906015.1	Pinf
58	Thalassiosira pseudonana	XP_002297333.1	Tpse
59	Bigelowiella natans	aug1.107_g20721	Bnat
60	Cyanophora paradoxa	Contig39528	Cpar
61	Chondrus crispus	XP_005718572.1	Ccri
62	Cyanidioschyzon merolae	XP_005539057.1	Cmer
63	Chlamydomonas reinhardtii	XP_001689510.1	Crei
64	Volvox carteri	XP_002952055.1	Vcar
65	Physcomitrella patens	XP_001775921.1	Ppat1
66	Physcomitrella patens	XP_001775247.1	Ppat2
67	Marchantia polymorpha	OAE25402.1	Mpol

68	Amborella trichopoda	XP_006844882.1	Atri
69	Oryza sativa	NP_001054315.1	Osat
70	Zea mays	NP_001131598.1	Zmay1
71	Zea mays	XP_008655683.1	Zmay2
72	Arabidopsis thaliana	NP_565878.1	Atha
73	Glycine max	XP_003549765.1	Gmax1
74	Glycine max	XP_003542685.1	Gmax2

Table A.3: Homologs of Hmg1 and Hmg2 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Caenorhabditis elegans	NP_498626.2	Cele
02	Drosophila melanogaster	NP_732900.1	Dmel
03	Anopheles gambiae	XP_307890.5	Agam
04	Ciona intestinalis	XP_002131486.1	Cint
05	Danio rerio	XP_005165572.1	Drer1
06	Danio rerio	NP_001014314.1	Drer2
07	Takifugu rubripes	XP_003974515.1	Trub1
08	Takifugu rubripes	XP_011603017.1	Trub2
09	Anolis carolinensis	XP_003216303.1	Acar
10	Gallus gallus	NP_989816.2	Ggal
11	Ornithorhynchus anatinus	XP_007661894.1	Oana
12	Monodelphis domestica	XP_001368743.1	Mdom
13	Canis lupus familiaris	XP_536323.3	Cfam
14	Sus scrofa	NP_001116460.1	Sscr
15	Mus musculus	NP_032281.2	Mmus
16	Pan troglodytes	XP_009447288.1	Ptro
17	Homo sapiens	NP_000850.1	Hsap
18	Strongylocentrotus purpuratus	NP_999724.1	Spur
19	Amphimedon queenslandica	XP_003382388.2	Aque
20	Arthroderma otae	XP_002845802.1	Aota1
21	Arthroderma otae	XP_002849525.1	Aota2
22	Aspergillus nidulans	XP_661421.1	Anid1
23	Aspergillus nidulans	XP_659197.1	Anid2
24	Neosartorya fischeri	XP_001265930.1	Nfis1
25	Neosartorya fischeri	XP_001264646.1	Nfis2
26	Neosartorya fischeri	XP_001264202.1	Nfis3
27	Leptosphaeria maculans	XP_003834814.1	Lmac
28	Parastagonospora nodorum	XP_001800116.1	Pnod
29	Botrytis cinerea	XP_001559959.1	Bcin
30	Sclerotinia sclerotiorum	XP_001593096.1	Sscl
31	Chaetomium globosum	XP_001220531.1	Cglo
32	Thielavia terrestris	XP_003656898.1	Tter
33	Neurospora crassa	XP_964546.1	Ncra
34	Tuber melanosporum	XP_002840504.1	Tmel
35	Saccharomyces cerevisiae	NP_013636.1	Scer1
36	Saccharomyces cerevisiae	NP_013555.1	Scer2
37	Candida glabrata	XP_449268.1	Cgla
38	Zygosaccharomyces rouxii	XP_002495578.1	Zrou
39	Kluyveromyces lactis	XP_451740.1	Klac

40	Schizosaccharomyces pombe	NP_588235.1	Spom
41	Agaricus bisporus	XP_006463978.1	Abis
42	Schizophyllum commune	XP_003028889.1	Scom
43	Trametes versicolor	XP_008041304.1	Tver
44	Auricularia delicata	XP_007337478.1	Adel
45	Cryptococcus neoformans	XP_774842.1	Cneo
46	Ustilago maydis	XP_011389590.1	Umay
47	Puccinia graminis	XP_003326671.2	Pgra
48	Batrachochytrium dendrobatidis	XP_006680643.1	Bden
49	Encephalitozoon intestinalis	XP_003073854.2	Eint
50	Acytostelium subglobosum	XP_012758330.1	Asub1
51	Acytostelium subglobosum	XP_012751872.1	Asub2
52	Acytostelium subglobosum	XP_012750020.1	Asub3
53	Dictyostelium discoideum	XP_643058.1	Ddis1
54	Dictyostelium discoideum	XP_646489.1	Ddis2
55	Polysphondylium pallidum	XP_020435570.1	Ppal1
56	Polysphondylium pallidum	XP_020430477.1	Ppal2
57	Trypanosoma brucei	XP_845571.1	Tbru
58	Leishmania major	XP_001684927.1	Lmaj
59	Naegleria gruberi	XP_002670914.1	Ngru1
60	Naegleria gruberi	XP_002668160.1	Ngru2
61	Phaeodactylum tricornutum	XP_002185302.1	Ptri
62	Thalassiosira pseudonana	XP_002289576.1	Tpse
63	Cyanophora paradoxa	Contig26118	Cpar
64	Physcomitrella patens	XP_001751461.1	Ppat1
65	Physcomitrella patens	XP_001771547.1	Ppat2
66	Physcomitrella patens	XP_001763417.1	Ppat3
67	Marchantia polymorpha	OAE31678.1	Mpol
68	Amborella trichopoda	XP_006849945.1	Atri
69	Oryza sativa	NP_001063541.1	Osat1
70	Oryza sativa	NP_001062221.1	Osat2
71	Oryza sativa	NP_001173136.1	Osat3
72	Zea mays	XP_008677153.1	Zmay1
73	Zea mays	NP_001130818.1	Zmay2
74	Zea mays	XP_008646164.1	Zmay3
75	Zea mays	NP_001169411.1	Zmay4
76	Zea mays	NP_001142036.1	Zmay5
77	Zea mays	XP_008666319.1	Zmay6
78	Arabidopsis thaliana	NP_179329.1	Atha1
79	Arabidopsis thaliana	NP_177775.2	Atha2
80	Glycine max	XP_003547886.1	Gmax1
81	Glycine max	XP_003534226.1	Gmax2
82	Glycine max	XP_003517117.1	Gmax3
83	Glycine max	XP_003537699.1	Gmax4
84	Glycine max	XP_003519474.1	Gmax5
85	Glycine max	XP_003545556.1	Gmax6
86	Glycine max	XP_006605576.1	Gmax7

Table A.4: Homologs of Pct1 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	Organism	Protein Accession	Short name
01	Caenorhabditis elegans	NP_001033540.1	Cele1
02	Caenorhabditis elegans	NP_871893.1	Cele2
03	Caenorhabditis elegans	NP_493826.1	Cele3
04	Drosophila melanogaster	NP_647621.1	Dmel1
05	Drosophila melanogaster	NP_647622.1	Dmel2
06	Anopheles gambiae	XP_003436074.1	Agam
07	Ciona intestinalis	XP_002130773.1	Cint
08	Danio rerio	NP_001032451.1	Drer1
09	Danio rerio	NP_001017634.1	Drer2
10	Danio rerio	NP_001018571.1	Drer3
11	Danio rerio	NP_001076415.1	Drer4
12	Takifugu rubripes	XP_011606103.1	Trub1
13	Takifugu rubripes	XP_003971327.1	Trub2
14	Takifugu rubripes	XP_003975731.1	Trub3
15	Anolis carolinensis	XP_008105542.1	Acar1
16	Anolis carolinensis	XP_003229397.1	Acar2
17	Gallus gallus	XP_422725.3	Ggal1
18	Gallus gallus	XP_416793.4	Ggal2
19	Ornithorhynchus anatinus	XP_001517894.2	Oana
20	Monodelphis domestica	XP_007493467.1	Mdom1
21	Monodelphis domestica	XP_007493927.1	Mdom2
22	Canis lupus familiaris	XP_005641268.1	Cfam1
23	Canis lupus familiaris	XP_005639655.1	Cfam2
24	Sus scrofa	XP_005673564.1	Sscr1
25	Sus scrofa	XP_013837801.1	Sscr2
26	Mus musculus	NP_808214.1	Mmus1
27	Mus musculus	NP_001156631.1	Mmus2
28	Pan troglodytes	XP_520980.3	Ptro1
29	Pan troglodytes	XP_526433.3	Ptro2
30	Homo sapiens	NP_004836.2	Hsap1
31	Homo sapiens	NP_005008.2	Hsap2
32	Strongylocentrotus purpuratus	XP_011660498.1	Spur1
33	Strongylocentrotus purpuratus	XP_011660508.1	Spur2
34	Amphimedon queenslandica	XP_003383158.1	Aque
35	Arthroderma otae	XP_002842674.1	Aota
36	Aspergillus nidulans	XP_658961.1	Anid
37	Neosartorya fischeri	XP_001264835.1	Nfis
38	Leptosphaeria maculans	XP_003835273.1	Lmac
39	Parastagonospora nodorum	XP_001795785.1	Pnod
40	Botrytis cinerea	XP_001545301.1	Bcin
41	Sclerotinia sclerotiorum	XP_001590426.1	Sscl
42	Thielavia terrestris	XP_003650871.1	Tter
43	Neurospora crassa	XP_956553.1	Ncra
44	Tuber melanosporum	XP_002835165.1	Tmel
45	Saccharomyces cerevisiae	NP_011718.1	Scer
46	Candida glabrata	XP_445145.1	Cgla
47	Zygosaccharomyces rouxii	XP_002498493.1	Zrou
48	Kluyveromyces lactis	XP_454698.1	Klac

49	Schizosaccharomyces pombe	NP_588548.2	Spom
50	Agaricus bisporus	XP_006455757.1	Abis
51	Schizophyllum commune	XP_003028356.1	Scom1
52	Schizophyllum commune	XP_003032666.1	Scom2
53	Trametes versicolor	XP_008040555.1	Tver1
54	Trametes versicolor	XP_008038886.1	Tver2
55	Auricularia delicata	XP_007350850.1	Adel
56	Cryptococcus neoformans	XP_012046709.1	Cneo
57	Ustilago maydis	XP_011386063.1	Umay
58	Puccinia graminis	XP_003323220.2	Pgra
59	Encephalitozoon intestinalis	XP_003073841.1	Eint
60	Acytostelium subglobosum	XP_012753001.1	Asub
61	Dictyostelium discoideum	XP_635846.1	Ddis
62	Polysphondylium pallidum	XP_020437796.1	Ppal
63	Trypanosoma brucei	XP_011778710.1	Tbru
64	Leishmania major	XP_001682542.1	Lmaj
65	Naegleria gruberi	XP_002676662.1	Ngru
66	Plasmodium falciparum	XP_001349902.1	Pfal
67	Theileria parva	XP_765394.1	Tpar
68	Babesia bovis	XP_001610042.1	Bbov
69	Toxoplasma gondii	XP_002371001.1	Tgon
70	Cryptosporidium hominis	XP_665462.1	Chom
71	Paramecium tetraurelia	XP_001453620.1	Ptet1
72	Paramecium tetraurelia	XP_001440495.1	Ptet2
73	Tetrahymena thermophila	XP_001021465.1	Tthe
74	Phytophthora infestans	XP_002898492.1	Pinf1
75	Phytophthora infestans	XP_002898494.1	Pinf2
76	Phaeodactylum tricornutum	XP_002178614.1	Ptri
77	Physcomitrella patens	XP_001780977.1	Ppat1
78	Physcomitrella patens	XP_001757801.1	Ppat2
79	Physcomitrella patens	XP_001778860.1	Ppat3
80	Marchantia polymorpha	OAE28237.1	Mpol
81	Amborella trichopoda	XP_006854102.1	Atri
82	Oryza sativa	NP_001046040.1	Osat1
83	Oryza sativa	NP_001065503.1	Osat2
84	Oryza sativa	NP_001061052.1	Osat3
85	Zea mays	NP_001130060.1	Zmay1
86	Zea mays	NP_001150315.1	Zmay2
87	Zea mays	NP_001141606.2	Zmay3
88	Zea mays	XP_008662841.1	Zmay4
89	Arabidopsis thaliana	NP_180785.1	Atha1
90	Arabidopsis thaliana	NP_193249.5	Atha2
91	Glycine max	XP_014621648.1	Gmax1
92	Glycine max	NP_001242526.1	Gmax2
93	Glycine max	XP_003533684.1	Gmax3
94	Glycine max	XP_003545206.1	Gmax4
95	Glycine max	NP_001240060.1	Gmax5

Table A.5: Homologs of Cse1 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	Organism	Protein Accession	Short name
01	Caenorhabditis elegans	NP_490716.1	Cele
02	Drosophila melanogaster	NP_523588.2	Dmel
03	Anopheles gambiae	XP_311424.1	Agam
04	Ciona intestinalis	XP_002130634.1	Cint
05	Danio rerio	NP_958858.1	Drer
06	Takifugu rubripes	XP_003973266.1	Trub
07	Anolis carolinensis	XP_003220667.1	Acar
08	Gallus gallus	XP_417389.3	Ggal
09	Ornithorhynchus anatinus	XP_001506734.2	Oana
10	Monodelphis domestica	XP_001369476.1	Mdom
11	Canis lupus familiaris	XP_853206.3	Cfam
12	Sus scrofa	NP_001230144.1	Sscr1
13	Sus scrofa	XP_005654303.1	Sscr2
14	Mus musculus	NP_076054.1	Mmus
15	Pan troglodytes	XP_001166085.1	Ptro
16	Homo sapiens	NP_001307.2	Hsap
17	Strongylocentrotus purpuratus	XP_011667197.1	Spur1
18	Strongylocentrotus purpuratus	XP_786014.4	Spur2
19	Strongylocentrotus purpuratus	XP_011667204.1	Spur3
20	Strongylocentrotus purpuratus	XP_011674880.1	Spur4
21	Amphimedon queenslandica	XP_003386477.1	Aque
22	Arthroderma otae	XP_002847643.1	Aota
23	Aspergillus nidulans	XP_664195.1	Anid
24	Neosartorya fischeri	XP_001257606.1	Nfis
25	Leptosphaeria maculans	XP_003840505.1	Lmac
26	Parastagonospora nodorum	XP_001792310.1	Pnod
27	Botrytis cinerea	XP_001555078.1	Bcin
28	Sclerotinia sclerotiorum	XP_001598480.1	Sscl
29	Chaetomium globosum	XP_001224834.1	Cglo
30	Thielavia terrestris	XP_003649730.1	Tter
31	Neurospora crassa	XP_960866.1	Ncra
32	Tuber melanosporum	XP_002837203.1	Tmel
33	Saccharomyces cerevisiae	NP_011276.1	Scer
34	Candida glabrata	XP_447138.1	Cgla
35	Zygosaccharomyces rouxii	XP_002499336.1	Zrou
36	Kluyveromyces lactis	XP_451037.1	Klac
37	Schizosaccharomyces pombe	NP_595530.1	Spom
38	Agaricus bisporus	XP_006455464.1	Abis
39	Schizophyllum commune	XP_003032086.1	Scom
40	Trametes versicolor	XP_008039825.1	Tver
41	Auricularia delicata	XP_007350393.1	Adel
42	Cryptococcus neoformans	XP_012052925.1	Cneo
43	Ustilago maydis	XP_011389453.1	Umay
44	Puccinia graminis	XP_003320509.2	Pgra
45	Batrachochytrium dendrobatidis	XP_006681999.1	Bden
46	Entamoeba histolytica	XP_650094.1	Ehis
47	Acytostelium subglobosum	XP_012758813.1	Asub
48	Dictyostelium discoideum	XP_629951.1	Ddis

49	Polysphondylium pallidum	XP_020429487.1	Ppal
50	Trypanosoma brucei	XP_011774263.1	Tbru
51	Leishmania major	XP_001684948.1	Lmaj
52	Naegleria gruberi	XP_002678129.1	Ngru
53	Trichomonas vaginalis	XP_001300686.1	Tvag
54	Giradia lamblia	XP_001706751.1	Glam
55	Plasmodium falciparum	XP_001352194.1	Pfal
56	Theileria parva	XP_764176.1	Tpar
57	Babesia bovis	XP_001609952.1	Bbov
58	Toxoplasma gondii	XP_018638171.1	Tgon
59	Cryptosporidium hominis	XP_667321.1	Chom
60	Paramecium tetraurelia	XP_001456414.1	Ptet1
61	Paramecium tetraurelia	XP_001451000.1	Ptet2
62	Tetrahymena thermophila	XP_001016036.1	Tthe
63	Phytophthora infestans	XP_002906560.1	Pinf
64	Phaeodactylum tricornutum	XP_002181619.1	Ptri
65	Thalassiosira pseudonana	XP_002297489.1	Tpse
66	Bigelowiella natans	estExt_Genewise1Plus.C_20002	Bnat1
67	Bigelowiella natans	estExt_Genewise1Plus.C_150165	Bnat2
68	Cyanophora paradoxa	Contig53071	Cpar
69	Chondrus crispus	XP_005716824.1	Ccri
70	Cyanidioschyzon merolae	XP_005537205.1	Cmer
71	Chlamydomonas reinhardtii	XP_001692097.1	Crei
72	Volvox carteri	XP_002952992.1	Vcar
73	Physcomitrella patens	XP_001762342.1	Ppat1
74	Physcomitrella patens	XP_001770614.1	Ppat2
75	Marchantia polymorpha	OAE29098.1	Mpol
76	Amborella trichopoda	XP_006850097.1	Atri
77	Oryza sativa	NP_001042522.1	Osat
78	Zea mays	XP_008665295.1	Zmay1
79	Zea mays	XP_020393477.1	Zmay2
80	Arabidopsis thaliana	NP_182175.1	Atha
81	Glycine max	XP_003548351.1	Gmax1
82	Glycine max	XP_003528788.1	Gmax2

Table A.6: Homologs of Ntf2 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Drosophila melanogaster	NP_609878.1	Dmel1
02	Drosophila melanogaster	NP_608422.1	Dmel2
03	Anopheles gambiae	XP_003437010.1	Agam1
04	Anopheles gambiae	XP_308748.2	Agam2
05	Ciona intestinalis	XP_002129876.1	Cint
06	Danio rerio	NP_001006000.2	Drer1
07	Danio rerio	NP_001003598.1	Drer2
08	Takifugu rubripes	XP_003977797.1	Trub
09	Anolis carolinensis	XP_003225422.1	Acar
10	Gallus gallus	NP_001025733.2	Ggal
11	Ornithorhynchus anatinus	XP_001519586.2	Oana
12	Monodelphis domestica	XP_001365121.1	Mdom1

13	Monodelphis domestica	XP_007487894.1	Mdom2
14	Monodelphis domestica	XP_001373183.2	Mdom3
15	Monodelphis domestica	XP_007486325.1	Mdom4
16	Canis lupus familiaris	XP_536812.1	Cfam
17	Sus scrofa	XP_003126970.2	Sscr1
18	Sus scrofa	XP_005658965.2	Sscr2
19	Mus musculus	NP_080808.1	Mmus1
20	Mus musculus	XP_001474007.1	Mmus2
21	Pan troglodytes	XP_001166045.1	Ptro1
22	Pan troglodytes	XP_009427497.1	Ptro2
23	Homo sapiens	NP_005787.1	Hsap
24	Strongylocentrotus purpuratus	XP_797612.1	Spur
25	Amphimedon queenslandica	XP_003389400.1	Aque
26	Arthroderma otae	XP_002850678.1	Aota
27	Aspergillus nidulans	XP_662546.1	Anid
28	Neosartorya fischeri	XP_001263412.1	Nfis
29	Leptosphaeria maculans	XP_003834538.1	Lmac
30	Parastagonospora nodorum	XP_001798316.1	Pnod
31	Botrytis cinerea	XP_001558550.1	Bcin
32	Sclerotinia sclerotiorum	XP_001592408.1	Sscl
33	Thielavia terrestris	XP_003650599.1	Tter
34	Neurospora crassa	XP_960292.2	Ncra
35	Tuber melanosporum	XP_002837974.1	Tmel
36	Saccharomyces cerevisiae	NP_010925.1	Scer
37	Candida glabrata	XP_447218.1	Cgla
38	Zygosaccharomyces rouxii	XP_002498169.1	Zrou
39	Kluyveromyces lactis	XP_453665.1	Klac
40	Schizosaccharomyces pombe	XP_001713065.1	Spom
41	Agaricus bisporus	XP_007326939.1	Abis
42	Schizophyllum commune	XP_003036222.1	Scom
43	Trametes versicolor	XP_008037000.1	Tver
44	Auricularia delicata	XP_007345768.1	Adel
45	Cryptococcus neoformans	XP_572414.1	Cneo
46	Ustilago maydis	XP_011389172.1	Umay
47	Puccinia graminis	XP_003333128.1	Pgra1
48	Puccinia graminis	XP_003336738.1	Pgra2
49	Batrachochytrium dendrobatidis	XP_006681297.1	Bden
50	Entamoeba histolytica	XP_656712.1	Ehis
51	Acytostelium subglobosum	XP_012759834.1	Asub
52	Dictyostelium discoideum	XP_643125.1	Ddis
53	Polysphondylium pallidum	XP_020428684.1	Ppal
54	Trypanosoma brucei	XP_847259.1	Tbru
55	Naegleria gruberi	XP_002677191.1	Ngru
56	Trichomonas vaginalis	XP_001312172.1	Tvag
57	Theileria parva	XP_764619.1	Tpar
58	Babesia bovis	XP_001612213.1	Bbov
59	Toxoplasma gondii	XP_002368194.1	Tgon
60	Cryptosporidium hominis	XP_665716.1	Chom
61	Paramecium tetraurelia	XP_001444291.1	Ptet1
62	Paramecium tetraurelia	XP_001448920.1	Ptet2
63	Paramecium tetraurelia	XP_001453310.1	Ptet3
64	Phytophthora infestans	XP_002895567.1	Pinf
65	Phaeodactylum tricornutum	XP_002183658.1	Ptri

66	Thalassiosira pseudonana	XP_002290343.1	Tpse
67	Bigelowiella natans	e_gw1.5.135.1	Bnat
68	Cyanophora paradoxa	Contig6778	Cpar
69	Chondrus crispus	XP_005711458.1	Ccri
70	Cyanidioschyzon merolae	XP_005539421.1	Cmer
71	Chlamydomonas reinhardtii	XP_001700802.1	Crei
72	Volvox carteri	XP_002954251.1	Vcar
73	Physcomitrella patens	XP_001769890.1	Ppat1
74	Physcomitrella patens	XP_001753949.1	Ppat2
75	Physcomitrella patens	XP_001765346.1	Ppat3
76	Marchantia polymorpha	OAE30006.1	Mpol
77	Amborella trichopoda	XP_006848276.1	Atri1
78	Amborella trichopoda	XP_006826470.3	Atri2
79	Oryza sativa	NP_001062338.1	Osat1
80	Oryza sativa	NP_001044479.1	Osat2
81	Zea mays	NP_001131358.1	Zmay1
82	Zea mays	XP_008678399.1	Zmay2
83	Arabidopsis thaliana	NP_174051.1	Atha1
84	Arabidopsis thaliana	NP_174118.1	Atha2
85	Arabidopsis thaliana	NP_001154326.1	Atha3
86	Glycine max	XP_003538542.1	Gmax1
87	Glycine max	NP_001240272.1	Gmax2
88	Glycine max	XP_003524901.1	Gmax3

Table A.7: Homologs of Trm1 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Caenorhabditis elegans	NP_506513.1	Cele
02	Drosophila melanogaster	NP_609566.1	Dmel
03	Anopheles gambiae	XP_318825.4	Agam
04	Ciona intestinalis	XP_002130419.1	Cint
05	Danio rerio	XP_009304430.1	Drer
06	Takifugu rubripes	XP_003972324.1	Trub
07	Anolis carolinensis	XP_003225255.1	Acar
08	Monodelphis domestica	XP_007489184.1	Mdom
09	Canis lupus familiaris	XP_013977674.1	Cfam
10	Sus scrofa	NP_001230379.1	Sscr1
11	Sus scrofa	XP_013846160.1	Sscr2
12	Sus scrofa	XP_013847081.1	Sscr3
13	Mus musculus	XP_006530874.1	Mmus
14	Pan troglodytes	XP_009433043.1	Ptro
15	Homo sapiens	XP_011526426.1	Hsap
16	Strongylocentrotus purpuratus	XP_003727616.1	Spur1
17	Strongylocentrotus purpuratus	XP_011671215.1	Spur2
18	Amphimedon queenslandica	XP_011404147.1	Aque
19	Arthroderma otae	XP_002849846.1	Aota
20	Aspergillus nidulans	XP_682675.1	Anid
21	Neosartorya fischeri	XP_001259178.1	Nfis
22	Leptosphaeria maculans	XP_003834388.1	Lmac
23	Parastagonospora nodorum	XP_001798253.1	Pnod

24	Botrytis cinerea	XP_001555575.1	Bcin
25	Sclerotinia sclerotiorum	XP_001585873.1	Sscl
26	Chaetomium globosum	XP_001223110.1	Cglo
27	Thielavia terrestris	XP_003657149.1	Tter
28	Neurospora crassa	XP_962370.2	Ncra
29	Tuber melanosporum	XP_002839151.1	Tmel
30	Saccharomyces cerevisiae	NP_010405.3	Scer
31	Candida glabrata	XP_445051.1	Cgla
32	Zygosaccharomyces rouxii	XP_002496732.1	Zrou
33	Kluyveromyces lactis	XP_455951.1	Klac
34	Schizosaccharomyces pombe	NP_596547.1	Spom
35	Agaricus bisporus	XP_006458094.1	Abis
36	Schizophyllum commune	XP_003036161.1	Scom
37	Trametes versicolor	XP_008036501.1	Tver
38	Auricularia delicata	XP_007338595.1	Adel
39	Cryptococcus neoformans	XP_012046176.1	Cneo
40	Ustilago maydis	XP_011390691.1	Umay
41	Puccinia graminis	XP_003326593.2	Pgra
42	Batrachochytrium dendrobatidis	XP_006682751.1	Bden
43	Encephalitozoon intestinalis	XP_003073435.1	Eint
44	Entamoeba histolytica	XP_657464.1	Ehis
45	Acytostelium subglobosum	XP_012754426.1	Asub
46	Dictyostelium discoideum	XP_638511.1	Ddis
47	Polysphondylium pallidum	XP_020438416.1	Ppal
48	Trypanosoma brucei	XP_011779597.1	Tbru
49	Leishmania major	XP_001681838.1	Lmaj
50	Naegleria gruberi	XP_002672847.1	Ngru
51	Trichomonas vaginalis	XP_001309981.1	Tvag1
52	Trichomonas vaginalis	XP_001305624.1	Tvag2
53	Trichomonas vaginalis	XP_001308959.1	Tvag3
54	Trichomonas vaginalis	XP_001308958.1	Tvag4
55	Giardia lamblia	XP_001705693.1	Glam
56	Plasmodium falciparum	XP_001349929.1	Pfal
57	Theileria parva	XP_765091.1	Tpar
58	Babesia bovis	XP_001610090.1	Bbov
59	Toxoplasma gondii	XP_002368816.1	Tgon
60	Cryptosporidium hominis	XP_668628.1	Chom
61	Paramecium tetraurelia	XP_001439107.1	Ptet
62	Tetrahymena thermophila	XP_001032347.2	Tthe1
63	Tetrahymena thermophila	XP_012656207.1	Tthe2
64	Phytophthora infestans	XP_002902693.1	Pinf
65	Phaeodactylum tricornutum	XP_002184592.1	Ptri
66	Bigelowiella natans	estExt_fgenesh1_pg.C_390054	Bnat1
67	Bigelowiella natans	e_gw1.4.49.1	Bnat2
68	Cyanophora paradoxa	Contig9001	Cpar1
69	Cyanophora paradoxa	Contig12367	Cpar2
70	Chondrus crispus	XP_005713566.1	Ccri
71	Cyanidioschyzon merolae	XP_005538801.1	Cmer1
72	Cyanidioschyzon merolae	XP_005538823.1	Cmer2
73	Chlamydomonas reinhardtii	XP_001702763.1	Crei1
74	Chlamydomonas reinhardtii	XP_001702714.1	Crei2
75	Volvox carteri	XP_002950042.1	Vcar1
76	Volvox carteri	XP_002954049.1	Vcar2

77	Physcomitrella patens	XP_001754787.1	Ppat1
78	Physcomitrella patens	XP_001754216.1	Ppat2
79	Marchantia polymorpha	OAE35175.1	Mpol1
80	Marchantia polymorpha	OAE24340.1	Mpol2
81	Marchantia polymorpha	OAE24341.1	Mpol3
82	Amborella trichopoda	XP_006856396.2	Atri1
83	Amborella trichopoda	XP_020521827.1	Atri2
84	Oryza sativa	NP_001051491.1	Osat1
85	Oryza sativa	NP_001064424.1	Osat2
86	Oryza sativa	NP_001055198.1	Osat3
87	Zea mays	NP_001340373.1	Zmay1
88	Zea mays	XP_008665644.1	Zmay2
89	Zea mays	XP_008681549.1	Zmay3
90	Zea mays	NP_001144117.1	Zmay4
91	Arabidopsis thaliana	NP_186881.2	Atha1
92	Arabidopsis thaliana	NP_197085.2	Atha2
93	Arabidopsis thaliana	NP_191192.1	Atha3
94	Glycine max	XP_006603927.1	Gmax1
95	Glycine max	XP_006593702.1	Gmax2
96	Glycine max	XP_003528018.1	Gmax3

Table A.8: Homologs of Slp1 protein identified across the eukaryotes considered in this study and their short names used in the tree.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Caenorhabditis elegans	NP_491321.1	Cele
02	Drosophila melanogaster	NP_724277.2	Dmel
03	Anopheles gambiae	XP_001237874.2	Agam
04	Ciona intestinalis	XP_002127085.1	Cint
05	Danio rerio	XP_686501.6	Drer
06	Takifugu rubripes	XP_011612880.1	Trub1
07	Takifugu rubripes	XP_011614974.1	Trub2
08	Anolis carolinensis	XP_008123692.1	Acar1
09	Anolis carolinensis	XP_008116027.1	Acar2
10	Gallus gallus	XP_004943242.1	Ggal
11	Ornithorhynchus anatinus	XP_007667083.1	Oana
12	Monodelphis domestica	XP_007480853.1	Mdom
13	Canis lupus familiaris	XP_013970657.1	Cfam
14	Sus scrofa	XP_013835333.1	Sscr1
15	Sus scrofa	XP_013842175.1	Sscr2
16	Mus musculus	XP_006496826.1	Mmus
17	Pan troglodytes	XP_009436576.1	Ptro
18	Homo sapiens	NP_057311.3	Hsap
19	Strongylocentrotus purpuratus	XP_001186094.2	Spur
20	Amphimedon queenslandica	XP_011403188.1	Aque
21	Arthroderma otae	XP_002844949.1	Aota
22	Aspergillus nidulans	XP_659868.1	Anid
23	Neosartorya fischeri	XP_001265989.1	Nfis
24	Leptosphaeria maculans	XP_003844542.1	Lmac
25	Parastagonospora nodorum	XP_001799493.1	Pnod
26	Botrytis cinerea	XP_001560808.1	Bcin

27	Sclerotinia sclerotiorum	XP_001598225.1	Sscl
28	Chaetomium globosum	XP_001229882.1	Cglo
29	Thielavia terrestris	XP_003655336.1	Tter
30	Neurospora crassa	XP_957243.2	Ncra
31	Tuber melanosporum	XP_002839343.1	Tmel
32	Saccharomyces cerevisiae	NP_014797.1	Scer
33	Zygosaccharomyces rouxii	XP_002497932.1	Zrou
34	Kluyveromyces lactis	XP_452295.1	Klac
35	Schizosaccharomyces pombe	NP_596096.1	Spom
36	Agaricus bisporus	XP_006462805.1	Abis
37	Trametes versicolor	XP_008044945.1	Tver
38	Auricularia delicata	XP_007337143.1	Adel
39	Cryptococcus neoformans	XP_012048773.1	Cneo
40	Ustilago maydis	XP_011391104.1	Umay
41	Puccinia graminis	XP_003329428.2	Pgra
42	Batrachochytrium dendrobatidis	XP_006681670.1	Bden
43	Entamoeba histolytica	XP_656924.2	Ehis
44	Acytostelium subglobosum	XP_012753510.1	Asub
45	Dictyostelium discoideum	XP_638007.1	Ddis
46	Polysphondylium pallidum	XP_020431163.1	Ppal
47	Trypanosoma brucei	XP_011773950.1	Tbru
48	Leishmania major	XP_001684621.1	Lmaj
49	Naegleria gruberi	XP_002671598.1	Ngru
50	Plasmodium falciparum	XP_001348546.2	Pfal
51	Theileria parva	XP_766102.1	Tpar
52	Babesia bovis	XP_001608804.1	Bbov
53	Toxoplasma gondii	XP_002371997.1	Tgon
54	Paramecium tetraurelia	XP_001456944.1	Ptet1
55	Paramecium tetraurelia	XP_001426451.1	Ptet2
56	Tetrahymena thermophila	XP_001023338.2	Tthe
57	Phytophthora infestans	XP_002896539.1	Pinf
58	Phaeodactylum tricornutum	XP_002179468.1	Ptri
59	Thalassiosira pseudonana	XP_002293328.1	Tpse
60	Bigelowiella natans	estExt_fgenesh1_pg.C_260176	Bnat
61	Cyanophora paradoxa	Contig37983	Cpar
62	Chondrus crispus	XP_005717049.1	Ccri
63	Chlamydomonas reinhardtii	XP_001696958.1	Crei
64	Physcomitrella patens	XP_001775438.1	Ppat1
65	Physcomitrella patens	XP_001758570.1	Ppat2
66	Marchantia polymorpha	OAE31126.1	Mpol1
67	Marchantia polymorpha	OAE24071.1	Mpol2
68	Amborella trichopoda	XP_006836858.1	Atri1
69	Amborella trichopoda	XP_020527955.1	Atri2
70	Oryza sativa	NP_001044969.1	Osat1
71	Oryza sativa	NP_001043487.1	Osat2
72	Zea mays	NP_001183941.1	Zmay1
73	Zea mays	XP_008674756.1	Zmay2
74	Zea mays	NP_001147071.1	Zmay3
75	Arabidopsis thaliana	NP_177292.4	Atha1
76	Arabidopsis thaliana	NP_683323.2	Atha2
77	Arabidopsis thaliana	NP_194126.5	Atha3
78	Glycine max	XP_003534427.1	Gmax1
79	Glycine max	XP_006592993.1	Gmax2

80	Glycine max	XP_006591521.2	Gmax3
81	Glycine max	XP_006574312.2	Gmax4
82	Glycine max	XP_014632378.1	Gmax5
83	Glycine max	XP_003522822.1	Gmax6
84	Glycine max	XP_003526394.1	Gmax7

Table A.9: Homologs of HMG-CoA reductases protein identified across the eukaryotes and selected prokaryotes and their short names used in the tree. The class II HMG-CoA reductases are highlighted in bold.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Caenorhabditis elegans	NP_498626.2	Cele
02	Drosophila melanogaster	NP_732900.1	Dmel
03	Anopheles gambiae	XP_307890.5	Agam
04	Ciona intestinalis	XP_002131486.1	Cint
05	Danio rerio	XP_005165572.1	Drer1
06	Danio rerio	NP_001014314.1	Drer2
07	Takifugu rubripes	XP_003974515.1	Trub1
08	Takifugu rubripes	XP_011603017.1	Trub2
09	Anolis carolinensis	XP_003216303.1	Acar
10	Gallus gallus	NP_989816.2	Ggal
11	Ornithorhynchus anatinus	XP_007661894.1	Oana
12	Monodelphis domestica	XP_001368743.1	Mdom
13	Canis lupus familiaris	XP_536323.3	Cfam
14	Sus scrofa	NP_001116460.1	Sscr
15	Mus musculus	NP_032281.2	Mmus
16	Pan troglodytes	XP_009447288.1	Ptro
17	Homo sapiens	NP_000850.1	Hsap
18	Strongylocentrotus purpuratus	NP_999724.1	Spur
19	Amphimedon queenslandica	XP_003382388.2	Aque1
20	Amphimedon queenslandica	XP_011403190.1	Aque2
21	Arthroderma otae	XP_002845802.1	Aota1
22	Arthroderma otae	XP_002849525.1	Aota2
23	Aspergillus nidulans	XP_661421.1	Anid1
24	Aspergillus nidulans	XP_659197.1	Anid2
25	Neosartorya fischeri	XP_001265930.1	Nfis1
26	Neosartorya fischeri	XP_001264646.1	Nfis2
27	Neosartorya fischeri	XP_001264202.1	Nfis3
28	Leptosphaeria maculans	XP_003834814.1	Lmac
29	Parastagonospora nodorum	XP_001800116.1	Pnod
30	Botrytis cinerea	XP_001559959.1	Bcin
31	Sclerotinia sclerotiorum	XP_001593096.1	Sscl
32	Chaetomium globosum	XP_001220531.1	Cglo
33	Thielavia terrestris	XP_003656898.1	Tter
34	Neurospora crassa	XP_964546.1	Ncra
35	Tuber melanosporum	XP_002840504.1	Tmel
36	Saccharomyces cerevisiae	NP_013636.1	Scer1
37	Saccharomyces cerevisiae	NP_013555.1	Scer2
38	Candida glabrata	XP_449268.1	Cgla
39	Zygosaccharomyces rouxii	XP_002495578.1	Zrou
40	Kluyveromyces lactis	XP_451740.1	Klac

41	Schizosaccharomyces pombe	NP_588235.1	Spom
42	Agaricus bisporus	XP_006463978.1	Abis
43	Schizophyllum commune	XP_003028889.1	Scom
44	Trametes versicolor	XP_008041304.1	Tver
45	Auricularia delicata	XP_007337478.1	Adel
46	Cryptococcus neoformans	XP_774842.1	Cneo
47	Ustilago maydis	XP_011389590.1	Umay
48	Puccinia graminis	XP_003326671.2	Pgra
49	Batrachochytrium dendrobatidis	XP_006680643.1	Bden
50	Encephalitozoon intestinalis	XP_003073854.2	Eint
51	Acytostelium subglobosum	XP_012758330.1	Asub1
52	Acytostelium subglobosum	XP_012751872.1	Asub2
53	Acytostelium subglobosum	XP_012750020.1	Asub3
54	Dictyostelium discoideum	XP_643058.1	Ddis1
55	Dictyostelium discoideum	XP_646489.1	Ddis2
56	Polysphondylium pallidum	XP_020435570.1	Ppal1
57	Polysphondylium pallidum	XP_020430477.1	Ppal2
58	Trypanosoma brucei	XP_845571.1	Tbru
59	Leishmania major	XP_001684927.1	Lmaj
60	Naegleria gruberi	XP_002670914.1	Ngru1
61	Naegleria gruberi	XP_002668160.1	Ngru2
62	Phaeodactylum tricornutum	XP_002185302.1	Ptri
63	Thalassiosira pseudonana	XP_002289576.1	Tpse
64	Cyanophora paradoxa	Contig26118	Cpar
65	Physcomitrella patens	XP_001751461.1	Ppat1
66	Physcomitrella patens	XP_001771547.1	Ppat2
67	Physcomitrella patens	XP_001763417.1	Ppat3
68	Marchantia polymorpha	OAE31678.1	Mpol
69	Amborella trichopoda	XP_006849945.1	Atri
70	Oryza sativa	NP_001063541.1	Osat1
71	Oryza sativa	NP_001062221.1	Osat2
72	Oryza sativa	NP_001173136.1	Osat3
73	Zea mays	XP_008677153.1	Zmay1
74	Zea mays	NP_001130818.1	Zmay2
75	Zea mays	XP_008646164.1	Zmay3
76	Zea mays	NP_001169411.1	Zmay4
77	Zea mays	NP_001142036.1	Zmay5
78	Zea mays	XP_008666319.1	Zmay6
79	Arabidopsis thaliana	NP_179329.1	Atha1
80	Arabidopsis thaliana	NP_177775.2	Atha2
81	Glycine max	XP_003547886.1	Gmax1
82	Glycine max	XP_003534226.1	Gmax2
83	Glycine max	XP_003517117.1	Gmax3
84	Glycine max	XP_003537699.1	Gmax4
85	Glycine max	XP_003519474.1	Gmax5
86	Glycine max	XP_003545556.1	Gmax6
87	Glycine max	XP_006605576.1	Gmax7
88	**Rhizophagus irregularis**	**XP_025182392.1**	**Rirr**
89	**Batrachochytrium dendrobatidis**	**XP_006682657.1**	**Bden2**
90	**Giardia lamblia**	**XP_001708651.1**	**Glam**
91	**Phytophthora infestans**	**XP_002897124.1**	**Pinf**
92	**Paramecium tetraurelia**	**XP_001438919.1**	**Ptet**
93	**Tetrahymena thermophila**	**XP_001032083.2**	**Tthe**

94	**Trichomonas vaginalis**	**XP_001329920.1**	**Tvag1**
95	**Trichomonas vaginalis**	**XP_001327732.1**	**Tvag2**
96	**Trichomonas vaginalis**	**XP_001321843.1**	**Tvag3**
97	Chloracidobacterium thermophilum	WP_058866193.1	CtheBac
98	Actinobaculum suis	WP_049619884.1	AsuiBac
99	Corynebacterium amycolatum	WP_005509971.1	CamyBac
100	Blastococcus saxobsidens	WP_014374180.1	BsaxBac
101	Curtobacterium sp. MCBA15_005	WP_071247058.1	CurtBac
102	Actinoplanes subtropicus	WP_034215766.1	AsubBac
103	Streptomyces sp. MMG1121	WP_053666143.1	StreBac
104	Fibrisoma limi	WP_009280789.1	FlimBac
105	Flavobacterium sp. KMS	WP_052259357.1	FlavBac
106	Caldilinea aerophila	WP_014431310.1	CaerBac
107	Brevibacillus laterosporus	WP_022584705.1	BlatBac
108	Sebaldella termitidis	WP_012861397.1	SterBac
109	Actibacterium atlanticum	WP_035251219.1	AatlBac
110	Limnobacter sp. MED105	WP_008252167.1	LimnBac
111	Desulfuromonas soudanensis	WP_053551980.1	DsouBac
112	Vibrio rotiferianus	WP_010451706.1	VrotBac
113	**Parascardovia denticolens**	**WP_006291948.1**	**PdenBac**
114	**Nocardia carnea**	**WP_033241108.1**	**NcarBac**
115	**Streptomyces olivaceus**	**WP_031037692.1**	**SoliBac**
116	**Kordia algicida**	**WP_007095272.1**	**KalgBac**
117	**Lactobacillus helveticus**	**WP_012211578.1**	**LhelBac**
118	**Methylobacterium aquaticum**	**WP_060848312.1**	**MaquBac**
119	**Advenella mimigardefordensis**	**WP_025371539.1**	**AmimBac**
120	**Legionella rubrilucens**	**WP_058530402.1**	**LrubBac**
121	**Cryptosporangium arvum**	**WP_035856380.1**	**CarvBac**
122	**Bdellovibrio bacteriovorus**	**WP_038447072.1**	**BbacBac**
123	**Prolixibacter bellariivorans**	**WP_025865296.1**	**PbelBac**
124	**Staphylococcus warneri**	**WP_002450135.1**	**SwarBac**
125	**Bacillus coagulans**	**WP_061575034.1**	**BcoaBac**
126	**Clostridiales bacterium VE202-03**	**WP_024723639.1**	**CbacBac**
127	**Legionella pneumophila subsp. pneumophila str. Philadelphia 1**	**YP_096068.1**	**LpneBac**
128	**Marinobacter lutaoensis**	**WP_076723081.1**	**MlutBac**
129	**Spirochaeta cellobiosiphila**	**WP_053228073.1**	**ScelBac**
130	**Acholeplasma equifetale**	**WP_026399677.1**	**AequBac**
131	Hyperthermus butylicus	WP_011822671.1	HbutArc
132	Pyrolobus fumarii	WP_014027115.1	PfumArc
133	Halococcus morrhuae	WP_004055536.1	HmorArc
134	Natronobacterium gregoryi	WP_005580701.1	NgreArc
135	Methanothermus fervidus	WP_013413095.1	MferArc
136	Methanobrevibacter arboriphilus	WP_054835208.1	MarbArc
137	Methanotorris formicicus	WP_007044754.1	MforArc
138	Methanospirillum hungatei	WP_011449942.1	MhunArc
139	Methanolobus tindarius	WP_023844161.1	MtinArc
140	Thermococcus litoralis	WP_004069164.1	TlitArc
141	Thermoplasmatales archaeon BRNA1	WP_015492422.1	TarcArc
142	Candidatus Methanomethylophilus alvus	WP_015504677.1	CalvArc

143	**Archaeoglobus profundus**	**WP_012940664.1**	**AproArc**
144	**Aciduliprofundum boonei**	**WP_008083177.1**	**AbooArc**
145	**Methanocella conradii**	**WP_014404884.1**	**MconArc**
146	**Methanosaeta harundinacea**	**WP_014587577.1**	**MharArc**
147	**Picrophilus torridus**	**WP_011177944.1**	**PtorArc**
148	**Candidatus Odinarchaeota archaeon LCB_4**	**OLS17153.1**	**CodiArc**
149	**Candidatus Thorarchaeota archaeon**	**RDE12852.1**	**CthoArc**
150	**Candidatus Lokiarchaeota archaeon**	**RLI66664.1**	**ClokArc**
151	**Candidatus Heimdallarchaeota archaeon**	**MBS85960.1**	**ChemArc**

Table A.10: Homologs of tRNA-methyl transferase identified in eukaryotes and selected prokaryotes and their short names used in the tree.

S. No	Organism	Protein Accession	Short name
01	Caenorhabditis elegans	NP_506513.1	Cele
02	Drosophila melanogaster	NP_609566.1	Dmel
03	Anopheles gambiae	XP_318825.4	Agam
04	Ciona intestinalis	XP_002130419.1	Cint
05	Danio rerio	XP_009304430.1	Drer
06	Takifugu rubripes	XP_003972324.1	Trub
07	Anolis carolinensis	XP_003225255.1	Acar
08	Monodelphis domestica	XP_007489184.1	Mdom
09	Canis lupus familiaris	XP_013977674.1	Cfam
10	Sus scrofa	NP_001230379.1	Sscr1
11	Sus scrofa	XP_013846160.1	Sscr2
12	Sus scrofa	XP_013847081.1	Sscr3
13	Mus musculus	XP_006530874.1	Mmus
14	Pan troglodytes	XP_009433043.1	Ptro
15	Homo sapiens	XP_011526426.1	Hsap
16	Strongylocentrotus purpuratus	XP_003727616.1	Spur1
17	Strongylocentrotus purpuratus	XP_011671215.1	Spur2
18	Amphimedon queenslandica	XP_011404147.1	Aque
19	Arthroderma otae	XP_002849846.1	Aota
20	Aspergillus nidulans	XP_682675.1	Anid
21	Neosartorya fischeri	XP_001259178.1	Nfis
22	Leptosphaeria maculans	XP_003834388.1	Lmac
23	Parastagonospora nodorum	XP_001798253.1	Pnod
24	Botrytis cinerea	XP_001555575.1	Bcin
25	Sclerotinia sclerotiorum	XP_001585873.1	Sscl
26	Chaetomium globosum	XP_001223110.1	Cglo
27	Thielavia terrestris	XP_003657149.1	Tter
28	Neurospora crassa	XP_962370.2	Ncra
29	Tuber melanosporum	XP_002839151.1	Tmel
30	Saccharomyces cerevisiae	NP_010405.3	Scer
31	Candida glabrata	XP_445051.1	Cgla
32	Zygosaccharomyces rouxii	XP_002496732.1	Zrou
33	Kluyveromyces lactis	XP_455951.1	Klac
34	Schizosaccharomyces pombe	NP_596547.1	Spom
35	Agaricus bisporus	XP_006458094.1	Abis

36	Schizophyllum commune	XP_003036161.1	Scom
37	Trametes versicolor	XP_008036501.1	Tver
38	Auricularia delicata	XP_007338595.1	Adel
39	Cryptococcus neoformans	XP_012046176.1	Cneo
40	Ustilago maydis	XP_011390691.1	Umay
41	Puccinia graminis	XP_003326593.2	Pgra
42	Batrachochytrium dendrobatidis	XP_006682751.1	Bden
43	Encephalitozoon intestinalis	XP_003073435.1	Eint
44	Entamoeba histolytica	XP_657464.1	Ehis
45	Acytostelium subglobosum	XP_012754426.1	Asub
46	Dictyostelium discoideum	XP_638511.1	Ddis
47	Polysphondylium pallidum	XP_020438416.1	Ppal
48	Trypanosoma brucei	XP_011779597.1	Tbru
49	Leishmania major	XP_001681838.1	Lmaj
50	Naegleria gruberi	XP_002672847.1	Ngru
51	Trichomonas vaginalis	XP_001309981.1	Tvag1
52	Trichomonas vaginalis	XP_001305624.1	Tvag2
53	Trichomonas vaginalis	XP_001308959.1	Tvag3
54	Trichomonas vaginalis	XP_001308958.1	Tvag4
55	Giardia lamblia	XP_001705693.1	Glam
56	Plasmodium falciparum	XP_001349929.1	Pfal
57	Theileria parva	XP_765091.1	Tpar
58	Babesia bovis	XP_001610090.1	Bbov
59	Toxoplasma gondii	XP_002368816.1	Tgon
60	Cryptosporidium hominis	XP_668628.1	Chom
61	Paramecium tetraurelia	XP_001439107.1	Ptet
62	Tetrahymena thermophila	XP_001032347.2	Tthe1
63	Tetrahymena thermophila	XP_012656207.1	Tthe2
64	Phytophthora infestans	XP_002902693.1	Pinf
65	Phaeodactylum tricornutum	XP_002184592.1	Ptri
66	Bigelowiella natans	estExt_fgenesh1_pg.C_390054	Bnat1
67	Bigelowiella natans	e_gw1.4.49.1	Bnat2
68	Cyanophora paradoxa	Contig9001	Cpar1
69	Cyanophora paradoxa	Contig12367	Cpar2
70	Chondrus crispus	XP_005713566.1	Ccri
71	Cyanidioschyzon merolae	XP_005538801.1	Cmer1
72	Cyanidioschyzon merolae	XP_005538823.1	Cmer2
73	Chlamydomonas reinhardtii	XP_001702763.1	Crei1
74	Chlamydomonas reinhardtii	XP_001702714.1	Crei2
75	Volvox carteri	XP_002950042.1	Vcar1
76	Volvox carteri	XP_002954049.1	Vcar2
77	Physcomitrella patens	XP_001754787.1	Ppat1
78	Physcomitrella patens	XP_001754216.1	Ppat2
79	Marchantia polymorpha	OAE35175.1	Mpol1
80	Marchantia polymorpha	OAE24340.1	Mpol2
81	Marchantia polymorpha	OAE24341.1	Mpol3
82	Amborella trichopoda	XP_006856396.2	Atri1
83	Amborella trichopoda	XP_020521827.1	Atri2
84	Oryza sativa	NP_001051491.1	Osat1
85	Oryza sativa	NP_001064424.1	Osat2
86	Oryza sativa	NP_001055198.1	Osat3
87	Zea mays	NP_001340373.1	Zmay1
88	Zea mays	XP_008665644.1	Zmay2

89	Zea mays	XP_008681549.1	Zmay3
90	Zea mays	NP_001144117.1	Zmay4
91	Arabidopsis thaliana	NP_186881.2	Atha1
92	Arabidopsis thaliana	NP_197085.2	Atha2
93	Arabidopsis thaliana	NP_191192.1	Atha3
94	Glycine max	XP_006603927.1	Gmax1
95	Glycine max	XP_006593702.1	Gmax2
96	Glycine max	XP_003528018.1	Gmax3
97	Aquifex aeolicus VF5	NP_213571.1	AaeoBac
98	Armatimonadetes bacterium GBS	WP_072260006.1	AbacBac
99	Crocosphaera watsonii	WP_007307237.1	CwatBac
100	Halothece sp. PCC 7418	WP_015227420.1	HalspBac
101	Myxosarcina sp. GI1	WP_036483192.1	MyxspBac
102	Spirulina major	WP_072620001.1	SmajBac
103	Acaryochloris sp. CCMEE 5410	WP_010472203.1	AcaspBac
104	Synechococcus sp. BL107	WP_037988485.1	SynspBac
105	Limnothrix rosea	WP_075889949.1	LrosBac
106	Candidatus Korarchaeum cryptofilum	WP_012310153.1	CkorArc
107	Caldisphaera lagunensis	WP_015232806.1	ClagArc
108	Pyrodictium occultum	WP_058370022.1	PoccArc
109	Saccharolobus solfataricus	WP_009992377.1	SsolArc
110	Thermofilum pendens	WP_011752053.1	TpenArc
111	Archaeoglobus sulfaticallidus	WP_015590878.1	AsulArc
112	Aciduliprofundum boonei	WP_008084418.1	AbooArc
113	Haladaptatus sp. R4	WP_066140587.1	HalspArc
114	Haloferax volcanii	WP_004045350.1	HvolArc
115	Natrialba magadii	WP_004217263.1	NmagArc
116	Methanobacterium formicicum	WP_004031240.1	MforArc
117	Methanocaldococcus villosus	WP_004591259.1	MvilArc
118	Methanococcoides methylutens	WP_048206359.1	MmetArc
119	Thermococcus gammatolerans	WP_015859006.1	TgamArc
120	Methanomassiliicoccus luminyensis	WP_019177115.1	MlumArc
121	Thermoplasmatales archaeon BRNA1	WP_015492201.1	TherArc
122	Nitrosopumilus sp. Nsub	WP_067958810.1	NitspArc
123	Candidatus Lokiarchaeota archaeon	RLI66673.1	ClokArc
124	Candidatus Thorarchaeota archaeon	RLI61887.1	CthoArc
125	Candidatus Heimdallarchaeota archaeon	RLI70572.1	CheiArc
126	Candidatus Odinarchaeota archaeon LCB_4	OLS17473.1	CodiArc

Table A.11: Homologs of Ntf2 protein identified in eukaryotes and bacteria and their short names used in the tree.

S. No	**Organism**	**Protein Accession**	**Short name**
01	Drosophila melanogaster	NP_609878.1	Dmel1
02	Drosophila melanogaster	NP_608422.1	Dmel2
03	Anopheles gambiae	XP_003437010.1	Agam1
04	Anopheles gambiae	XP_308748.2	Agam2
05	Ciona intestinalis	XP_002129876.1	Cint

06	Danio rerio	NP_001006000.2	Drer1
07	Danio rerio	NP_001003598.1	Drer2
08	Takifugu rubripes	XP_003977797.1	Trub
09	Anolis carolinensis	XP_003225422.1	Acar
10	Gallus gallus	NP_001025733.2	Ggal
11	Ornithorhynchus anatinus	XP_001519586.2	Oana
12	Monodelphis domestica	XP_001365121.1	Mdom1
13	Monodelphis domestica	XP_007487894.1	Mdom2
14	Monodelphis domestica	XP_001373183.2	Mdom3
15	Monodelphis domestica	XP_007486325.1	Mdom4
16	Canis lupus familiaris	XP_536812.1	Cfam
17	Sus scrofa	XP_003126970.2	Sscr1
18	Sus scrofa	XP_005658965.2	Sscr2
19	Mus musculus	NP_080808.1	Mmus1
20	Mus musculus	XP_001474007.1	Mmus2
21	Pan troglodytes	XP_001166045.1	Ptro1
22	Pan troglodytes	XP_009427497.1	Ptro2
23	Homo sapiens	NP_005787.1	Hsap
24	Strongylocentrotus purpuratus	XP_797612.1	Spur
25	Amphimedon queenslandica	XP_003389400.1	Aque
26	Arthroderma otae	XP_002850678.1	Aota
27	Aspergillus nidulans	XP_662546.1	Anid
28	Neosartorya fischeri	XP_001263412.1	Nfis
29	Leptosphaeria maculans	XP_003834538.1	Lmac
30	Parastagonospora nodorum	XP_001798316.1	Pnod
31	Botrytis cinerea	XP_001558550.1	Bcin
32	Sclerotinia sclerotiorum	XP_001592408.1	Sscl
33	Thielavia terrestris	XP_003650599.1	Tter
34	Neurospora crassa	XP_960292.2	Ncra
35	Tuber melanosporum	XP_002837974.1	Tmel
36	Saccharomyces cerevisiae	NP_010925.1	Scer
37	Candida glabrata	XP_447218.1	Cgla
38	Zygosaccharomyces rouxii	XP_002498169.1	Zrou
39	Kluyveromyces lactis	XP_453665.1	Klac
40	Schizosaccharomyces pombe	XP_001713065.1	Spom
41	Agaricus bisporus	XP_007326939.1	Abis
42	Schizophyllum commune	XP_003036222.1	Scom
43	Trametes versicolor	XP_008037000.1	Tver
44	Auricularia delicata	XP_007345768.1	Adel
45	Cryptococcus neoformans	XP_572414.1	Cneo
46	Ustilago maydis	XP_011389172.1	Umay
47	Puccinia graminis	XP_003333128.1	Pgra1
48	Puccinia graminis	XP_003336738.1	Pgra2
49	Batrachochytrium dendrobatidis	XP_006681297.1	Bden
50	Entamoeba histolytica	XP_656712.1	Ehis
51	Dictyostelium discoideum	XP_643125.1	Ddis
52	Polysphondylium pallidum	XP_020428684.1	Ppal
53	Trypanosoma brucei	XP_847259.1	Tbru
54	Naegleria gruberi	XP_002677191.1	Ngru
55	Trichomonas vaginalis	XP_001312172.1	Tvag
56	Theileria parva	XP_764619.1	Tpar
57	Babesia bovis	XP_001612213.1	Bbov
58	Toxoplasma gondii	XP_002368194.1	Tgon

59	Cryptosporidium hominis	XP_665716.1	Chom
60	Paramecium tetraurelia	XP_001444291.1	Ptet1
61	Paramecium tetraurelia	XP_001448920.1	Ptet2
62	Paramecium tetraurelia	XP_001453310.1	Ptet3
63	Phytophthora infestans	XP_002895567.1	Pinf
64	Phaeodactylum tricornutum	XP_002183658.1	Ptri
65	Thalassiosira pseudonana	XP_002290343.1	Tpse
66	Bigelowiella natans	e_gw1.5.135.1	Bnat
67	Cyanophora paradoxa	Contig6778	Cpar
68	Chondrus crispus	XP_005711458.1	Ccri
69	Cyanidioschyzon merolae	XP_005539421.1	Cmer
70	Chlamydomonas reinhardtii	XP_001700802.1	Crei
71	Volvox carteri	XP_002954251.1	Vcar
72	Physcomitrella patens	XP_001769890.1	Ppat1
73	Physcomitrella patens	XP_001753949.1	Ppat2
74	Physcomitrella patens	XP_001765346.1	Ppat3
75	Marchantia polymorpha	OAE30006.1	Mpol
76	Amborella trichopoda	XP_006848276.1	Atri1
77	Amborella trichopoda	XP_006826470.3	Atri2
78	Oryza sativa	NP_001062338.1	Osat1
79	Oryza sativa	NP_001044479.1	Osat2
80	Zea mays	NP_001131358.1	Zmay1
81	Zea mays	XP_008678399.1	Zmay2
82	Arabidopsis thaliana	NP_174051.1	Atha1
83	Arabidopsis thaliana	NP_174118.1	Atha2
84	Arabidopsis thaliana	NP_001154326.1	Atha3
85	Glycine max	XP_003538542.1	Gmax1
86	Glycine max	NP_001240272.1	Gmax2
87	Glycine max	XP_003524901.1	Gmax3
88	Streptomyces sp. NRRL S-350	WP_030242646.1	Stre1Bac
89	Kitasatospora phosalacinea	WP_033217154.1	KphoBac
90	Streptomyces novaecaesareae	WP_033332243.1	SnovBac
91	Kitasatospora sp. MY 5-36	WP_049649899.1	Kita1Bac
92	Kitasatospora sp. MY 5-36	WP_049650885.1	Kita2Bac
93	Kitasatospora sp. MY 5-36	WP_049650888.1	Kita3Bac
94	Streptomyces sp. MJM8645	WP_063349806.1	Stre2Bac
95	Streptomyces uncialis	WP_073786791.1	SuncBac
96	Streptomyces sp. CB02056	WP_074001113.1	Stre3Bac

Legends to Additional data (provided in CD)

Additional data 1: Homologs of yeast NE proteins across eukaryotes.

The homologs of each of the 45 NE proteins across the 74 organisms, along with the GeneID, protein accession, and the various domains (with coordinates) found are shown. Details of each protein are shown in separate sheets of the excel file. The organisms that are not included in RefSeq are represented with their corresponding fasta headers. The homologs in which the conserved regions could not be detected using Pfam, but found using CD-search are mentioned as CDsearch in the brackets following the coordinates. The additional domains found in some of the homologs are bunched under "Others" and their coordinates are mentioned in the respective column.

Additional data 2: HMG-CoA reductases identified across the three domains of life.

All the class I and II HMG-CoA reductases identified across Eukaryotes, Bacteria and Archaea are shown in separate sheets. The class II enzymes are highlighted by filling the corresponding cells in yellow. The first column "short name" refes to the name used in the phylogenetic analysis.

Additional data 3: The tRNA methyltransferases identified across the three domains of life.

All the tRNA methyltransferases identified across Eukaryotes, Bacteria and Archaea are shown in separate sheets. The first column "short name" refers to the name used in the phylogenetic analysis.

Additional data 4: The Perl script and the required input files

The Perl script that was used to predict the localization of proteins and the necessary input files are provided. The sample output files for one of the protein are also provided.

Additional data 5: Localization status of conserved NE homologs.

The subcellular localization data of the homologs of conserved NE proteins is shown in human (*H. sapiens*), mouse (*M. musculus*) and plant (*A. thaliana*) in separate sheets of the excel file. First column gives the conserved NE protein (yeast gene name); the second column lists the gene ID of the homolog of the respective protein found in the respective organism. The following columns give the taxonomy ID of the organism, the gene ID of the homolog (repeated), the gene ontology ID, kind of evidence, gene ontology term and the PubMed ID of the evidence. The rows with text highlighted in red and filled in yellow are the ones with experimental evidence, while those not filled with yellow are annotated as NE/ER although no direct experimental evidence is available.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104-2105.

Ahmed S, Brickner DG, Light WH, Cajigas I, McDonough M, Froyshteter AB, Volpe T, Brickner JH. 2010. DNA zip codes control an ancient mechanism for gene targeting to the nuclear periphery. *Nature cell biology* **12**: 111-118.

Akhtar A, Gasser SM. 2007. The nuclear envelope and transcriptional control. *Nature reviews Genetics* **8**: 507-517.

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.

Anderson DJ, Vargas JD, Hsiao JP, Hetzer MW. 2009. Recruitment of functionally distinct membrane proteins to chromatin mediates nuclear envelope formation in vivo. *J Cell Biol* **186**: 183-191.

Aravind L, Koonin EV. 2001. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* **11**: 1365-1374.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **2**: 28-36.

Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* **90**: 11558-11562.

Basson ME, Thorsness M, Rine J. 1986. Saccharomyces cerevisiae contains two functional genes encoding 3-hydroxy-3-methylglutaryl-coenzyme A reductase. *Proc Natl Acad Sci U S A* **83**: 5563-5567.

Baum DA. 2015. A comparison of autogenous theories for the origin of eukaryotic cells. *Am J Bot* **102**: 1954-1965.

Baum DA, Baum B. 2014. An inside-out origin for the eukaryotic cell. *BMC Biol* **12**: 76.

Bione S, Maestrini E, Rivella S, Mancini M, Regis S, Romeo G, Toniolo D. 1994. Identification of a novel X-linked gene responsible for Emery-Dreifuss muscular dystrophy. *Nat Genet* **8**: 323-327.

Blum T, Briesemeister S, Kohlbacher O. 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC bioinformatics* **10**: 274.

Bonen L, Doolittle WF. 1976. Partial sequences of 16S rRNA and the phylogeny of blue-green algae and chloroplasts. *Nature* **261**: 669-673.

Boucher Y, Doolittle WF. 2000. The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol Microbiol* **37**: 703-716.

Brachner A, Foisner R. 2011. Evolvement of LEM proteins as chromatin tethers at the nuclear periphery. *Biochemical Society transactions* **39**: 1735-1741.

Brown MS, Goldstein JL. 1980. Multivalent feedback regulation of HMG CoA reductase, a control mechanism coordinating isoprenoid synthesis and cell growth. *J Lipid Res* **21**: 505-517.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. *Curr Biol* **22**: 1123-1127.

Bupp JM, Martin AE, Stensrud ES, Jaspersen SL. 2007. Telomere anchoring at the nuclear periphery requires the budding yeast Sad1-UNC-84 domain protein Mps3. *J Cell Biol* **179**: 845-854.

Burke B, Stewart CL. 2002. Life at the edge: the nuclear envelope and human disease. *Nat Rev Mol Cell Biol* **3**: 575-585.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PloS one* **2**: e790.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540-552.

Cavalier-Smith T. 1987a. Eukaryotes with no mitochondria. *Nature* **326**: 332-333.

Cavalier-Smith T. 1987b. The origin of eukaryotic and archaebacterial cells. *Ann N Y Acad Sci* **503**: 17-54.

Cavalier-Smith T. 1988. Origin of the cell nucleus. *Bioessays* **9**: 72-78.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol* **46**: 347-366.

Cavalier-Smith T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol* **52**: 297-354.

Chen P, Johnson P, Sommer T, Jentsch S, Hochstrasser M. 1993. Multiple ubiquitin-conjugating enzymes participate in the in vivo degradation of the yeast MAT alpha 2 repressor. *Cell* **74**: 357-369.

Chong YT, Koh JL, Friesen H, Duffy SK, Cox MJ, Moses A, Moffat J, Boone C, Andrews BJ. 2015. Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **161**: 1413-1424.

Chou K-C, Shen H-B. 2010a. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science* **Vol.02No.10**: 14.

Chou KC, Shen HB. 2010b. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PloS one* **5**: e9931.

Ciska M, Moreno Diaz de la Espina S. 2013. NMCP/LINC proteins: putative lamin analogs in plants? *Plant Signal Behav* **8**: e26669.

Corbett AH, Silver PA. 1996. The NTF2 gene encodes an essential, highly conserved protein that functions in nuclear transport in vivo. *J Biol Chem* **271**: 18477-18484.

Corliss JO. 1984. The kingdom Protista and its 45 phyla. *Biosystems* **17**: 87-126.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**: 59-65.

de Souza RF, Iyer LM, Aravind L. 2010. Diversity and evolution of chromatin proteins encoded by DNA viruses. *Biochimica et biophysica acta* **1799**: 302-318.

DeGrasse JA, DuBois KN, Devos D, Siegel TN, Sali A, Field MC, Rout MP, Chait BT. 2009. Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. *Mol Cell Proteomics* **8**: 2119-2130.

Deng M, Hochstrasser M. 2006. Spatially regulated ubiquitin ligation by an ER/nuclear membrane ligase. *Nature* **443**: 827-831.

Derelle R, Lang BF. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* **29**: 1277-1289.

Derelle R, Torruella G, Klimes V, Brinkmann H, Kim E, Vlcek C, Lang BF, Elias M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci U S A* **112**: E693-699.

Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, Sali A, Rout MP. 2004. Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol* **2**: e380.

Doblas VG, Amorim-Silva V, Pose D, Rosado A, Esteban A, Arro M, Azevedo H, Bombarely A, Borsani O, Valpuesta V et al. 2013. The SUD1 gene encodes a putative E3 ubiquitin ligase and is a positive regulator of 3-hydroxy-3-methylglutaryl coenzyme a reductase activity in Arabidopsis. *Plant Cell* **25**: 728-743.

DuBois KN, Alsford S, Holden JM, Buisson J, Swiderski M, Bart JM, Ratushny AV, Wan Y, Bastin P, Barry JD et al. 2012. NUP-1 Is a large coiled-coil nucleoskeletal protein in trypanosomes with lamin-like functions. *PLoS Biol* **10**: e1001287.

Ellis SR, Morales MJ, Li JM, Hopper AK, Martin NC. 1986. Isolation and characterization of the TRM1 locus, a gene essential for the N2,N2-dimethylguanosine modification of both mitochondrial and cytoplasmic tRNA in Saccharomyces cerevisiae. *J Biol Chem* **261**: 9703-9709.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* **440**: 623-630.

Embley TM, van der Giezen M, Horner DS, Dyal PL, Foster P. 2003. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos Trans R Soc Lond B Biol Sci* **358**: 191-201; discussion 201-192.

Eme L, Moreira D, Talla E, Brochier-Armanet C. 2009. A complex cell division machinery was present in the last common ancestor of eukaryotes. *PloS one* **4**: e5021.

Eustice M, Pillus L. 2014. Unexpected function of the glucanosyltransferase Gas1 in the DNA damage response linked to histone H3 acetyltransferases in Saccharomyces cerevisiae. *Genetics* **196**: 1029-1039.

Fang Y, Spector DL. 2005. Centromere positioning and dynamics in living Arabidopsis plants. *Molecular biology of the cell* **16**: 5710-5718.

Feng JM, Tian HF, Wen JF. 2013. Origin and evolution of the eukaryotic SSU processome revealed by a comprehensive genomic analysis and implications for the origin of the nucleolus. *Genome Biol Evol* **5**: 2255-2267.

Ferrero S, Grados-Torrez RE, Leivar P, Antolin-Llovera M, Lopez-Iglesias C, Cortadellas N, Ferrer JC, Campos N. 2015. Proliferation and Morphogenesis of the Endoplasmic Reticulum Driven by the Membrane Domain of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase in Plant Cells. *Plant Physiol* **168**: 899-914.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* doi:10.1093/nar/gkr367.

Fischer T, Strasser K, Racz A, Rodriguez-Navarro S, Oppizzi M, Ihrig P, Lechner J, Hurt E. 2002. The mRNA export machinery requires the novel Sac3p-Thp1p complex to dock at the nucleoplasmic entrance of the nuclear pores. *The EMBO journal* **21**: 5843-5852.

Friesen JA, Rodwell VW. 2004. The 3-hydroxy-3-methylglutaryl coenzyme-A (HMG-CoA) reductases. *Genome Biol* **5**: 248.

Funabiki H, Hagan I, Uzawa S, Yanagida M. 1993. Cell cycle-dependent specific positioning and clustering of centromeres and telomeres in fission yeast. *J Cell Biol* **121**: 961-976.

Gambe AE, Matsunaga S, Takata H, Ono-Maniwa R, Baba A, Uchiyama S, Fukui K. 2009. A nucleolar protein RRS1 contributes to chromosome congression. *FEBS Lett* **583**: 1951-1956.

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B et al. 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* **38**: 285-293.

Gerace L, Burke B. 1988. Functional organization of the nuclear envelope. *Annual review of cell biology* **4**: 335-374.

Grant JR, Katz LA. 2014. Phylogenomic study indicates widespread lateral gene transfer in Entamoeba and suggests a past intimate relationship with parabasalids. *Genome Biol Evol* **6**: 2350-2360.

Graumann K, Vanrobays E, Tutois S, Probst AV, Evans DE, Tatout C. 2014. Characterization of two distinct subfamilies of SUN-domain proteins in Arabidopsis and their interactions with the novel KASH-domain protein AtTIK. *Journal of experimental botany* **65**: 6499-6512.

Gruenbaum Y, Foisner R. 2015. Lamins: Nuclear Intermediate Filament Proteins with Fundamental Functions in Nuclear Mechanics and Genome Regulation. *Annual review of biochemistry* doi:10.1146/annurev-biochem-060614-034115.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**: 948-951.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307-321.

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. *Mol Biol Evol* **24**: 1702-1713.

Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, de Oliveira MC. 2004. Comparative analysis of the complete plastid genome sequence of the red alga Gracilaria tenuistipitata var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J Mol Evol* **59**: 464-477.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci U S A* **106**: 3859-3864.

Han S, Bahmanyar S, Zhang P, Grishin N, Oegema K, Crooke R, Graham M, Reue K, Dixon JE, Goodman JM. 2012. Nuclear envelope phosphatase 1-regulatory subunit 1 (formerly TMEM188) is the metazoan Spo7p ortholog and functions in the lipin activation pathway. *J Biol Chem* **287**: 3123-3137.

Haraguchi T, Kojidani T, Koujin T, Shimi T, Osakada H, Mori C, Yamamoto A, Hiraoka Y. 2008. Live cell imaging and electron microscopy reveal dynamic processes of BAF-directed nuclear envelope assembly. *J Cell Sci* **121**: 2540-2554.

He D, Fiz-Palacios O, Fu CJ, Fehling J, Tsai CC, Baldauf SL. 2014. An alternative root for the eukaryote tree of life. *Curr Biol* **24**: 465-470.

Hermkes R, Fu YF, Nurrenberg K, Budhiraja R, Schmelzer E, Elrouby N, Dohmen RJ, Bachmair A, Coupland G. 2011. Distinct roles for Arabidopsis SUMO protease ESD4 and its closest homolog ELS1. *Planta* **233**: 63-73.

Hetzer MW, Walther TC, Mattaj IW. 2005. Pushing the envelope: structure, function, and dynamics of the nuclear periphery. *Annu Rev Cell Dev Biol* **21**: 347-380.

Hochstrasser M, Mathog D, Gruenbaum Y, Saumweber H, Sedat JW. 1986. Spatial organization of chromosomes in the salivary gland nuclei of Drosophila melanogaster. *J Cell Biol* **102**: 112-123.

Hodge CA, Choudhary V, Wolyniak MJ, Scarcelli JJ, Schneiter R, Cole CN. 2010. Integral membrane proteins Brr6 and Apq12 link assembly of the nuclear pore complex to lipid homeostasis in the endoplasmic reticulum. *J Cell Sci* **123**: 141-151.

Hontz RD, Niederer RO, Johnson JM, Smith JS. 2009. Genetic identification of factors that modulate ribosomal DNA transcription in Saccharomyces cerevisiae. *Genetics* **182**: 105-119.

Hood JK, Silver PA. 1998. Cse1p is required for export of Srp1p/importin-alpha from the nucleus in Saccharomyces cerevisiae. *J Biol Chem* **273**: 35142-35146.

Horigome C, Okada T, Shimazu K, Gasser SM, Mizuta K. 2011. Ribosome biogenesis factors bind a nuclear envelope SUN domain protein to cluster yeast telomeres. *The EMBO journal* **30**: 3799-3811.

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res* **35**: W585-587.

Hu LL, Feng KY, Cai YD, Chou KC. 2012. Using protein-protein interaction network information to predict the subcellular locations of proteins in budding yeast. *Protein Pept Lett* **19**: 644-651.

Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY. 2008. ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC bioinformatics* **9**: 80.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. *Nature* **425**: 686-691.

Istvan ES, Deisenhofer J. 2000. The structure of the catalytic portion of human HMG-CoA reductase. *Biochimica et biophysica acta* **1529**: 9-18.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A* **86**: 9355-9359.

Jaspersen SL, Martin AE, Glazko G, Giddings TH, Jr., Morgan G, Mushegian A, Winey M. 2006. The Sad1-UNC-84 homology domain in Mps3 interacts with Mps2 to connect the spindle pole body with the nuclear envelope. *J Cell Biol* **174**: 665-675.

Jiang JQ, Wu M. 2012. Predicting multiplex subcellular localization of proteins using protein-protein interaction network: a comparative study. *BMC bioinformatics* **13 Suppl 10**: S20.

Jordan-Starck TC, Rodwell VW. 1989. Pseudomonas mevalonii 3-hydroxy-3-methylglutaryl-CoA reductase. Characterization and chemical modification. *J Biol Chem* **264**: 17913-17918.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

Katz LA, Grant JR, Parfrey LW, Burleigh JG. 2012. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst Biol* **61**: 653-660.

Kim Y, Gentry MS, Harris TE, Wiley SE, Lawrence JC, Jr., Dixon JE. 2007. A conserved phosphatase cascade that regulates nuclear membrane biogenesis. *Proc Natl Acad Sci U S A* **104**: 6596-6601.

Klasson L, Kambris Z, Cook PE, Walker T, Sinkins SP. 2009. Horizontal gene transfer between Wolbachia and the mosquito Aedes aegypti. *BMC Genomics* **10**: 33.

Koga Y, Morii H. 2007. Biosynthesis of ether-type polar lipids in archaea and evolutionary considerations. *Microbiol Mol Biol Rev* **71**: 97-120.

Koh JL, Chong YT, Friesen H, Moses A, Boone C, Andrews BJ, Moffat J. 2015. CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae. *G3 (Bethesda)* **5**: 1223-1232.

Koonin EV. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol* **11**: 209.

Koreny L, Field MC. 2016. Ancient Eukaryotic Origin and Evolutionary Plasticity of Nuclear Lamina. *Genome Biol Evol* **8**: 2663-2671.

Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. 2013. Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit Rev Biochem Mol Biol* **48**: 373-396.

Kruger A, Batsios P, Baumann O, Luckert E, Schwarz H, Stick R, Meyer I, Graf R. 2012. Characterization of NE81, the first lamin-like nucleoskeleton protein in a unicellular organism. *Molecular biology of the cell* **23**: 360-370.

Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-Covo E, McInerney JO, Landan G, Martin WF. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**: 427-432.

Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331**: 184-186.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci U S A* **81**: 3786-3790.

Lake JA, Rivera MC. 1994. Was the nucleus the first endosymbiont? *Proc Natl Acad Sci U S A* **91**: 2880-2881.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Lee K, Chuang HY, Beyer A, Sung MK, Huh WK, Lee B, Ideker T. 2008. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res* **36**: e136.

Liu J, Lee KK, Segura-Totten M, Neufeld E, Wilson KL, Gruenbaum Y. 2003. MAN1 and emerin have overlapping function(s) essential for chromosome segregation and cell division in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* **100**: 4598-4603.

Lopez-Garcia P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. *Trends in biochemical sciences* **24**: 88-93.

Lurie-Weinberger MN, Gomez-Valero L, Merault N, Glockner G, Buchrieser C, Gophna U. 2010. The origins of eukaryotic-like proteins in Legionella pneumophila. *Int J Med Microbiol* **300**: 470-481.

Lusk CP, Blobel G, King MC. 2007. Highway to the inner nuclear membrane: rules for the road. *Nat Rev Mol Cell Biol* **8**: 414-420.

Lykidis A. 2007. Comparative genomics and evolution of eukaryotic phospholipid biosynthesis. *Prog Lipid Res* **46**: 171-199.

Mans BJ, Anantharaman V, Aravind L, Koonin EV. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell cycle* **3**: 1612-1637.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI et al. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**: D222-226.

Martin W. 1999. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. *Proceedings of the Royal Society of London Series B: Biological Sciences* **266**: 1387-1395.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**: 41-45.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* **392**: 37-41.

Mekhail K, Moazed D. 2010. The nuclear envelope in genome organization, expression and stability. *Nat Rev Mol Cell Biol* **11**: 317-328.

Mekhail K, Seebacher J, Gygi SP, Moazed D. 2008. Role for perinuclear chromosome tethering in maintenance of genome stability. *Nature* **456**: 667-670.

Moreira D, Lopez-Garcia P. 1998. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol* **47**: 517-530.

Murtas G, Reeves PH, Fu YF, Bancroft I, Dean C, Coupland G. 2003. A nuclear protease required for flowering-time regulation in Arabidopsis reduces the abundance of SMALL UBIQUITIN-RELATED MODIFIER conjugates. *Plant Cell* **15**: 2308-2319.

Nagai S, Dubrana K, Tsai-Pflugfelder M, Davidson MB, Roberts TM, Brown GW, Varela E, Hediger F, Gasser SM, Krogan NJ. 2008. Functional targeting of DNA damage to a nuclear pore-associated SUMO-dependent ubiquitin ligase. *Science* **322**: 597-602.

Neumann N, Lundin D, Poole AM. 2010. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. *PloS one* **5**: e13241.

Newport JW, Forbes DJ. 1987. The nucleus: structure, function, and dynamics. *Annual review of biochemistry* **56**: 535-565.

O'Reilly AJ, Dacks JB, Field MC. 2011. Evolution of the karyopherin-beta family of nucleocytoplasmic transport factors; ancient origins and continued specialization. *PloS one* **6**: e19308.

Oda Y, Fukuda H. 2011. Dynamics of Arabidopsis SUN proteins during mitosis and their involvement in nuclear shaping. *Plant J* **66**: 629-641.

Ognibene A, Sabatelli P, Petrini S, Squarzoni S, Riccio M, Santi S, Villanova M, Palmeri S, Merlini L, Maraldi NM. 1999. Nuclear changes in a case of X-linked Emery-Dreifuss muscular dystrophy. *Muscle Nerve* **22**: 864-869.

Palancade B, Liu X, Garcia-Rubio M, Aguilera A, Zhao X, Doye V. 2007. Nucleoporins prevent DNA damage accumulation by modulating Ulp1-dependent sumoylation processes. *Molecular biology of the cell* **18**: 2912-2923.

Palancade B, Zuccolo M, Loeillet S, Nicolas A, Doye V. 2005. Pml39, a novel protein of the nuclear periphery required for nuclear retention of improper messenger ribonucleoparticles. *Molecular biology of the cell* **16**: 5258-5268.

Pan X, Roberts P, Chen Y, Kvam E, Shulga N, Huang K, Lemmon S, Goldfarb DS. 2000. Nucleus-vacuole junctions in Saccharomyces cerevisiae are formed through the direct interaction of Vac8p with Nvj1p. *Molecular biology of the cell* **11**: 2445-2457.

Pawlowski J, Burki F. 2009. Untangling the phylogeny of amoeboid protists. *J Eukaryot Microbiol* **56**: 16-25.

Philippe H, Lopez P, Brinkmann H, Budin K, Germot A, Laurent J, Moreira D, Muller M, Le Guyader H. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proceedings Biological sciences* **267**: 1213-1221.

Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B. 2006. Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet* **38**: 1005-1014.

Pillai AN, Shukla S, Rahaman A. 2017. An evolutionarily conserved phosphatidate phosphatase maintains lipid droplet number and endoplasmic reticulum morphology but not nuclear morphology. *Biology open* **6**: 1629-1643.

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber AP, Schwacke R, Gross J, Blouin NA, Lane C et al. 2012. Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science* **335**: 843-847.

Puhler G, Leffers H, Gropp F, Palm P, Klenk HP, Lottspeich F, Garrett RA, Zillig W. 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc Natl Acad Sci U S A* **86**: 4569-4573.

Ren R, Sun Y, Zhao Y, Geiser D, Ma H, Zhou X. 2016. Phylogenetic Resolution of Deep Eukaryotic and Fungal Relationships Using Highly Conserved Low-Copy Nuclear Genes. *Genome Biol Evol* **8**: 2683-2701.

Reyes-Prieto A, Weber AP, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. *Annu Rev Genet* **41**: 147-168.

Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95**: 6239-6244.

Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**: 74-76.

Roberts P, Moshitch-Moshkovitz S, Kvam E, O'Toole E, Winey M, Goldfarb DS. 2003. Piecemeal microautophagy of nucleus in Saccharomyces cerevisiae. *Molecular biology of the cell* **14**: 129-141.

Rochette NC, Brochier-Armanet C, Gouy M. 2014. Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. *Mol Biol Evol* **31**: 832-845.

Rogers JV, McMahon C, Baryshnikova A, Hughson FM, Rose MD. 2014. ER-associated retrograde SNAREs and the Dsl1 complex mediate an alternative, Sey1p-independent homotypic ER fusion pathway. *Molecular biology of the cell* **25**: 3401-3412.

Rothballer A, Kutay U. 2013. The diverse functional LINCs of the nuclear envelope to the cytoskeleton and chromatin. *Chromosoma* **122**: 415-429.

Rusche LN, Kirchmaier AL, Rine J. 2003. The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. *Annual review of biochemistry* **72**: 481-516.

Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol* **14**: 255-274.

Santarella-Mellwig R, Franke J, Jaedicke A, Gorjanacz M, Bauer U, Budd A, Mattaj IW, Devos DP. 2010. The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. *PLoS Biol* **8**: e1000281.

Schirmer EC, Florens L, Guan T, Yates JR, 3rd, Gerace L. 2003. Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science* **301**: 1380-1382.

Schreiner SM, Koo PK, Zhao Y, Mochrie SG, King MC. 2015. The tethering of chromatin to the nuclear envelope supports nuclear mechanics. *Nat Commun* **6**: 7159.

Schwikowski B, Uetz P, Fields S. 2000. A network of protein-protein interactions in yeast. *Nature biotechnology* **18**: 1257-1261.

Scott MS, Calafell SJ, Thomas DY, Hallett MT. 2005. Refining protein subcellular localization. *PLoS computational biology* **1**: e66.

Serpeloni M, Vidal NM, Goldenberg S, Avila AR, Hoffmann FG. 2011. Comparative genomics of proteins involved in RNA nucleocytoplasmic export. *BMC Evol Biol* **11**: 7.

Shen S, Tobery CE, Rose MD. 2009. Prm3p is a pheromone-induced peripheral nuclear envelope protein required for yeast nuclear fusion. *Molecular biology of the cell* **20**: 2438-2450.

Sherban DG, Kennelly PJ, Brandt KG, Rodwell VW. 1985. Rat liver 3-hydroxy-3-methylglutaryl-CoA reductase. Catalysis of the reverse reaction and two half-reactions. *J Biol Chem* **260**: 12579-12585.

Siniossoglou S. 2009. Lipins, lipids and nuclear envelope structure. *Traffic* **10**: 1181-1187.

Siniossoglou S, Santos-Rosa H, Rappsilber J, Mann M, Hurt E. 1998. A novel complex of membrane proteins required for formation of a spherical nucleus. *The EMBO journal* **17**: 6449-6464.

Smirnov A, Nassonova E, Berney C, Fahrni J, Bolivar I, Pawlowski J. 2005. Molecular phylogeny and classification of the lobose amoebae. *Protist* **156**: 129-142.

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**: 173-179.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**: D535-539.

Starr DA, Fridolfsson HN. 2010. Interactions between nuclei and the cytoskeleton are mediated by SUN-KASH nuclear-envelope bridges. *Annu Rev Cell Dev Biol* **26**: 421-444.

Starr DA, Han M. 2003. ANChors away: an actin based mechanism of nuclear positioning. *J Cell Sci* **116**: 211-216.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science* **297**: 89-91.

Stechmann A, Cavalier-Smith T. 2003. The root of the eukaryote tree pinpointed. *Curr Biol* **13**: R665-666.

Stingele J, Habermann B, Jentsch S. 2015. DNA-protein crosslink repair: proteases as DNA repair enzymes. *Trends in biochemical sciences* **40**: 67-71.

Swanson R, Locher M, Hochstrasser M. 2001. A conserved ubiquitin ligase of the nuclear envelope/endoplasmic reticulum that functions in both ER-associated and Matalpha2 repressor degradation. *Genes Dev* **15**: 2660-2674.

Taddei A, Hediger F, Neumann FR, Bauer C, Gasser SM. 2004. Separation of silencing from perinuclear anchoring functions in yeast Ku80, Sir4 and Esc1 proteins. *The EMBO journal* **23**: 1301-1312.

Takemoto A, Kawashima SA, Li JJ, Jeffery L, Yamatsugu K, Elemento O, Nurse P. 2016. Nuclear envelope expansion is crucial for proper chromosomal segregation during a closed mitosis. *J Cell Sci* **129**: 1250-1259.

Talamas JA, Capelson M. 2015. Nuclear envelope and genome interactions in cell fate. *Front Genet* **6**: 95.

Texari L, Dieppois G, Vinciguerra P, Contreras MP, Groner A, Letourneau A, Stutz F. 2013. The nuclear pore regulates GAL1 gene transcription by controlling the localization of the SUMO protease Ulp1. *Mol Cell* **51**: 807-818.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol* **4**: 466-485.

Tkach JM, Yimit A, Lee AY, Riffle M, Costanzo M, Jaschob D, Hendry JA, Ou J, Moffat J, Boone C et al. 2012. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nature cell biology* **14**: 966-976.

Ulbert S, Antonin W, Platani M, Mattaj IW. 2006. The inner nuclear membrane protein Lem2 is critical for normal nuclear envelope morphology. *FEBS Lett* **580**: 6435-6441.

van Hooff JJ, Tromer E, van Wijk LM, Snel B, Kops GJ. 2017. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO reports* **18**: 1559-1571.

Varas J, Graumann K, Osman K, Pradillo M, Evans DE, Santos JL, Armstrong SJ. 2015. Absence of SUN1 and SUN2 proteins in Arabidopsis thaliana leads to a delay in meiotic progression and defects in synapsis and recombination. *Plant J* **81**: 329-346.

Wainright PO, Hinkle G, Sogin ML, Stickel SK. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science* **260**: 340-342.

Wang H, Xu Z, Gao L, Hao B. 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* **9**: 195.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189-1191.

Wilkie GS, Korfali N, Swanson SK, Malik P, Srsen V, Batrakou DG, de las Heras J, Zuleger N, Kerr AR, Florens L et al. 2011. Several novel nuclear envelope transmembrane proteins identified in skeletal muscle have cytoskeletal associations. *Mol Cell Proteomics* **10**: M110 003129.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**: 231-236.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**: 5088-5090.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* **87**: 4576-4579.

Woolfit M, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL. 2009. An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium Wolbachia pipientis. *Mol Biol Evol* **26**: 367-374.

Worman HJ, Ostlund C, Wang Y. 2010. Diseases of the nuclear envelope. *Cold Spring Harb Perspect Biol* **2**: a000760.

Wright R, Basson M, D'Ari L, Rine J. 1988. Increased amounts of HMG-CoA reductase induce "karmellae": a proliferation of stacked membrane pairs surrounding the yeast nucleus. *J Cell Biol* **107**: 101-114.

Xiong H, Rivero F, Euteneuer U, Mondal S, Mana-Capelli S, Larochelle D, Vogel A, Gassen B, Noegel AA. 2008. Dictyostelium Sun-1 connects the centrosome to chromatin and ensures genome stability. *Traffic* **9**: 708-724.

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. 1985. Mitochondrial origins. *Proc Natl Acad Sci U S A* **82**: 4443-4447.

Yofe I, Weill U, Meurer M, Chuartzman S, Zalckvar E, Goldman O, Ben-Dor S, Schutze C, Wiedemann N, Knop M et al. 2016. One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nature methods* **13**: 371-378.

Yoon HS, Grant J, Tekle YI, Wu M, Chaon BC, Cole JC, Logsdon JM, Jr., Patterson DJ, Bhattacharya D, Katz LA. 2008. Broadly sampled multigene trees of eukaryotes. *BMC Evol Biol* **8**: 14.

Yu L, Pena Castillo L, Mnaimneh S, Hughes TR, Brown GW. 2006. A survey of essential gene function in the yeast cell division cycle. *Molecular biology of the cell* **17**: 4736-4747.

Zablen LB, Kissil MS, Woese CR, Buetow DE. 1975. Phylogenetic origin of the chloroplast and prokaryotic nature of its ribosomal RNA. *Proc Natl Acad Sci U S A* **72**: 2418-2422.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**: 353-358.

Zargari A, Boban M, Heessen S, Andreasson C, Thyberg J, Ljungdahl PO. 2007. Inner nuclear membrane proteins Asi1, Asi2, and Asi3 function in concert to maintain the latent properties of transcription factors Stp1 and Stp2. *J Biol Chem* **282**: 594-605.

Zhang W, Neuner A, Ruthnick D, Sachsenheimer T, Luchtenborg C, Brugger B, Schiebel E. 2018. Brr6 and Brl1 locate to nuclear pore complex assembly sites to promote their biogenesis. *J Cell Biol* **217**: 877-894.

Origin and evolution of nuclear envelope proteome - A comparative genomics approach

by Hita Sony Garapati

Submission date: 16-Apr-2019 12:46PM (UTC+0530)
Submission ID: 1113470468
File name: Hita_Thesis_Final_For_Plagiarism.pdf (5.15M)
Word count: 27378
Character count: 153968

Origin and evolution of nuclear envelope proteome - A comparative genomics approach

ORIGINALITY REPORT

20%	17%	19%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	15%
2	Hita Sony Garapati, Krishnaveni Mishra. "Comparative genomics of nuclear envelope proteins", BMC Genomics, 2018 Publication	1%
3	images.nature.com Internet Source	<1%
4	Advances in Experimental Medicine and Biology, 2014. Publication	<1%
5	Caroline T. Meyer, Irma K. Bauer, Martin Antonio, Mitchell Adeyemi et al. "Prevalence of classic, MLB-clade and VA-clade Astroviruses in Kenya and The Gambia", Virology Journal, 2015 Publication	<1%
6	www.biomedcentral.com Internet Source	<1%