The evolution of mitochondrial genome and proteome in animals

Xiujuan Wang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Ecology and Evolutionary Biology Commons

Recommended Citation
The evolution of mitochondrial genome and proteome in animals

by

Xiujuan Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Genetics; Bioinformatics and Computational Biology

Program of Study Committee:
Dennis V. Lavrov, Co-major Professor
Xun Gu, Co-major Professor
Karin S. Dorman
Oliver Eulenstein
Stephen Proulx

Iowa State University
Ames, Iowa
2010
TABLE OF CONTENTS

ABSTRACT ....................................................................................................................... iv

CHAPTER 1. GENERAL INTRODUCTION ....................................................................1
  Introduction .................................................................................................................. 1
  Dissertation Organization .......................................................................................... 8
  References ................................................................................................................. 9

CHAPTER 2. MITOCHONDRIAL GENOME OF THE HOMOSCLEROMORPH
OSCARELLA CARMELA (PORIFERA, DEMOSPONGIAE) REVEALS
UNEXPECTED COMPLEXITY IN THE COMMON ANCESTOR OF SPONGES
  Abstract ................................................................................................................. 14
  Introduction .............................................................................................................. 15
  Materials and Methods ............................................................................................ 18
  Results ...................................................................................................................... 21
  Discussion and Conclusion ...................................................................................... 29
  Acknowledgements .................................................................................................. 31
  Literature Cited ....................................................................................................... 32

CHAPTER 3. SEVENTEEN NEW COMPLETE MTDNA SEQUENCES REVEAL
EXTENSIVE MITOCHONDRIAL GENOME EVOLUTION WITHIN THE
DEMSPOGNGIAE ............................................................................................................ 48
  Abstract ................................................................................................................. 48
  Introduction .............................................................................................................. 49
  Results ...................................................................................................................... 51
  Discussion .............................................................................................................. 60
  Methods .................................................................................................................. 63
  Acknowledgements ................................................................................................ 66
  References ............................................................................................................. 67

CHAPTER 4. GENE RECRUITMENT – A COMMON MECHANISM IN THE
EVOLUTION OF TRANSFER RNA GENE FAMILIES ................................................... 80
  Abstract ................................................................................................................. 80
  Introduction .............................................................................................................. 81
  Methods .................................................................................................................. 83
  Results .................................................................................................................... 85
  Discussion .............................................................................................................. 90
  Conclusion .............................................................................................................. 94
  Acknowledgements ............................................................................................... 95
  Literature Cited .................................................................................................... 96
Mitochondria, found in nearly all eukaryotes, are indispensable double membrane organelles that play pivotal roles in several cellular processes. While diversity of mitochondrial genomes among eukaryotes has been recognized, it was thought that animal mitochondrial genomes are small circular molecules with little variation in size and gene content. However, this picturing of animal mitochondrial genomes was based on a biased sampling drawn primarily from bilaterian animals. In order to explore the diversity and understand the evolution of mitochondrial genomes in animals, we sequenced and analyzed mitochondrial genomes from all 14 orders of demosponges, the biggest class within sponges (phylum Porifera). Comparative genome analysis shows that a large variation in mitochondrial genome architecture is present within this group exceeding that found within Bilateria. Phylogenomic analyses based on mtDNA data support demosponges as a monophyletic group and suggest that the last common ancestor of animals might have had a tissue-level organization. Although transfer RNA (tRNA) genes are generally conserved in these genomes, evidences were found for horizontal evolution of some tRNA genes that cautioned the use of tRNA phylogeny to infer genetic code evolution. While animal mitochondrial genomes only encode a handful of proteins, the complex functions of mitochondria require over a thousand of proteins that more than 98% are nuclear encoded. Comparative gene family analyses for nuclear encoded mitochondrial proteins demonstrate that protein subcellular relocalization enabled the retention and gain of function of genes after duplications and provided a way for recruiting mitochondrial proteins. In addition, mitochondrial proteome also expanded through subfunctionalization mechanism after gene duplications.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Mitochondria, found in nearly all eukaryotes, are cellular powerhouses that produce ATP as energy source of cell through respiration/oxidative phosphorylation. They also play key roles in many other cellular processes including Krebs cycle, heme biosynthesis, cell metabolism, apoptosis, and Fe/S cluster biosynthesis. Proteins involved in these functions are encoded by mitochondrial genome as well as nuclear genome (Bolender et al. 2008). Mutations in these genes might cause mitochondria dysfunction, disease as well as trigger programmed cell death. The communication and regulation between nucleus and mitochondria allow the evolution of specific tissues that require different amount of energy and allow control of mitochondrial function (Scarpulla 2006; Cannino et al. 2007). In addition, mitochondria in different tissues can perform specific functions such as detoxification of ammonia in liver (Wallace 2005). Thus mitochondria are essential for modern-day eukaryotes and it is important to understand their origin and evolution.

It has been widely accepted that mitochondrion originated from a free-living α-proteobacterium through an endosymbiotic process around 2 billion years ago (Lang et al. 1999a). However, most modern-day mitochondrial genomes contain less than 100 genes, a small portion of genes found in free living α-proteobacterium (Adams and Palmer 2003). This difference in gene content is due to gene loss from mitochondrial genomes and gene transfer to nuclear genomes, a so called "endosymbiotic gene transfer" process (Martin and Herrmann 1998; Martin 2003). The evolutionary trajectories of mitochondrial genomes among different groups of eukaryotes vary substantially and resulted in a large diversity of
genome sizes and architectures as well as gene contents. In particular, there is a striking
difference between mitochondrial genome organization in plants and animals. While animal
mitochondrial genomes are streamlined with compact gene content and very limited
intergenic regions, those of plants contain more genes and a large proportion of noncoding
DNA. Furthermore, the evolutionary rates of plant and animal mitochondrial sequences differ
as much as 100-fold (Lynch 2007). In my dissertation, I strived to advance our understanding
of animal mtDNA evolution by focusing on mitochondrial genomes of non-bilateral animals.
My fork was conducted on mitochondrial genomes of demosponges (chapter 2 and 3),
unusual evolutionary trajectories of tRNA genes (chapter 4), and comparative analysis of
nuclear-encoded mitochondrial gene families (chapter 5).

Mitochondrial genome (mtDNA) evolution in animals
Although the astounding diversity of mitochondrial genome architectures has been found
among eukaryotes (Lang et al. 1999b; Lang et al. 2004), the mtDNA within animals,
especially bilateral animals, is extremely conserved in gene content, size and genome
architecture (Boore 1999). With a few exceptions (Armstrong et al. 2000; Helfenbein et al.
2004; Shao et al. 2006), a typical animal mtDNA is a small circular molecule (~16kb) that
contains 13 protein coding genes, 2 ribosomal RNA genes and 22 tRNA genes. In addition,
these mtDNAs usually display high evolutionary rates comparing to nuclear genes and to the
mtDNA of Monosiga brevicollis, the closest unicellular relative of animals (Lynch 2007;
Burger et al. 2003).
However, our notion of a “typical” animal mtDNA is highly biased due to the overrepresentation of mtDNAs from bilaterial animals among sequenced genomes. While several hundreds of bilateral animal mtDNAs were sequenced at the time I started the study, only a few sequences were available from non-bilateral animals (phyla Cnidaria, Ctenophora and Porifera), including eight from Cnidarian (Beaton et al. 1998; Beagley et al. 1998; van Oppen et al. 2002; Fukami and Knowlton 2005) and three from Poriferan (Lavrov and Lang 2005a; Lavrov et al. 2005; Lavrov and Lang 2005b). However, these few sequeces have shown that more variations in gene content, genome architecture and sequence evolutionary rates are present in non-bilateral animal mtDNAs. For example, mtDNAs of Cnidaria are around 18kb and contain only 2 tRNAs while that of Porifera are even bigger in size (about 20 kb) and contain 24-25 tRNA genes and an extra protein coding gene comparing to bilateral animal mtDNAs (Lavrov 2007).

Even with the limited sampling from non-bilateral mtDNAs, it was inferred that the two transitions in animal evolution correlated not only with morphological changes (unicellular to multicellular, and to bilateral), but also with the changes in mtDNAs (Lavrov 2007). For example, along with the development of multicellularity, mtDNAs lost all ribosomal protein genes and most introns and intergenic regions. The appearance of bilateral symmetry, on the other hand, was accompanied by further compaction of mtDNA through the loss of several tRNA genes, as well as the emergence of novel genetic code. However, one has to remember that the above inference was based on extremely biased sampling from animal mtDNAs. The first part of my dissertation was focused on rectifying this situation by determining and
analyzing complete mitochondrial genomes from all major lineages of demosponges, the major group within the phylum Poifera.

**The evolution of mtDNAs in demosponges**

Sponges, a group of animals that are separate from all other animals (Eumetazoa), are exclusively aquatic living with predominantly filter-feeding style. There are approximately 15,000 extant species that belong to three groups: the Hexactinellida (glass sponge), the Calcarea (calcareous sponges), and the Demospongiae (demosponges). Class Demospongiae (demosponges) is the largest (>85% of species) and the most morphologically diversified group in this phylum (Sollas 1885). There are total 14 orders within the extant demosponges, which encompass 88 families, 500 genera and more than 8000 described species. The relationships within the group and to other animals were largely unresolved due to their highly diverse in morphological features and living environments.

At the time of this study started, there were three mtDNAs of demosponges published by our group (Lavrov and Lang 2005a; Lavrov et al. 2005; Lavrov and Lang 2005b). Comparative analysis of these mtDNAs to those of unicellular species (eg. *Monosiga brevicollis*) and bilateral animals (eg. *Homo sapiens*) showed demosponge mtDNAs resembled that of most animals in genome size, compact genome organization and lack of ribosomal proteins, but also showed some distinct features as present in unicellular species such as the presence of an extra protein coding gene (Lavrov et al. 2005; Lavrov 2007). This result indicates that mtDNAs of demosponges might represent the intermediate state of mtDNA evolution from unicellular species to Eumetazoa. To test this inference about animal mtDNA evolution, it
was essential to determine mitochondrial genome sequences from additional demosponge species. A comparative study of mtDNAs from all 14 recognized orders of demosponges is presented in chapter 3.

One interesting subclass within demosponges is Homoscleromorpha, which contains 7 genera and about 77 species in total (LÈvi C., 1957, Syst Zool, 6, 174-183). This group has distinct morphological, cytological and embryological features comparing to other sponges including a tissue-grade level of organization (Gaino et al. 1987; Boute et al. 1996; Ereskovsky and Boury-Esnault 2002; Muricy and Diaz 2002). Further the traditional view of this group as demosponges has been challenged by phylogenetic reconstruction using different molecular datasets (Ereskovsky et al. 2009). The positioning of this group directly affects the monophyly of demosponges and thus affects our understanding of the features of last common ancestor of animals. Here, by sequencing mtDNAs from this group, we compared the mtDNAs of this group to that of remaining demosponges and other animals. This comparative analysis, presented in chapters 2 and 3, provided us insights to the relationship of this group to demosponges and to Eumetazoa.

**The value of using mtDNAs to reconstruct phylogeny of animals, especially deep branch lineages**

Molecular data, as we know today, provide unprecedented power in resolving phylogenies comparing to morphological characters. However, controversies are often raised with the application of different genes or proteins. This problem gets worse when researchers try to use a single gene or protein to establish phylogenetic positions of deep branch lineages.
Recent studies have shown that mtDNAs are valuable for phylogenetic analysis for various reasons (Lavrov et al. 2005). First of all, since mitochondrion is believed to be originated only once from $\alpha$-proteobacteria and there were rare cases of horizontal gene transfer (Rot et al. 2006), the phylogeny of mtDNAs can be used to infer the phylogeny of eukaryotes. Second, large amount of sequences can minimize the sampling errors in sequence based phylogenetic analysis. Third, mtDNAs harbor additional rare genomic characters such as indels in the coding sequences, variations in the genetic code and changes in secondary structures of rRNA and tRNA, which are useful in inferring phylogeny as well. Finally, in addition to protein coding gene datasets, gene orders of mtDNAs can be used to reconstruct phylogeny and hence to validate the phylogeny based on concatenated mitochondrial protein data.

The phylogenetic reconstruction using mtDNAs is especially beneficial in obtaining the relationship of deep branches (i.e., non-bilateral animals) on an animal phylogenetic tree. This is due to two facts that 1) these groups generally lack sequence information from nuclear genomes, and 2) these mtDNAs, at least as demonstrated in the cases of sponges (Lavrov et al. 2005), have relatively slow evolutionary rates. Hence, we sequenced mtDNAs of demosponges to reconstruct the phylogeny within this group, especially to locate the phylogenetic relationship of Homoscleromorpha to demosponges and other animals. This phylogeny provided us an insight of sponge relationship to other animals and hence inferred the features of the last common ancestor of animals. The results of this study are presented in chapter 2 and chapter 3 of the thesis.
The sequenced mtDNAs could also provide us data for reconstructing phylogeny of gene families, in particular transfer RNA (tRNA) gene families. Our group reconstructed the phylogeny of tRNA genes in three demosponge mtDNAs and found evidences of horizontal evolution of tRNA genes (Lavrov and Lang 2005b). This type of evolution was first demonstrated in an *in vitro* study of tRNA genes in *Escherichia coli* (Saks et al. 1998). Given these results, the common application of alloacceptor tRNA phylogenetic relationships to elucidate the evolutionary history of the genetic code and the phylogeny of organisms could be potentially misleading. However, no further study was conducted on this horizontal evolution of tRNA gene families after these two investigations to ask how commonly tRNA gene recruitment events occur in nature. With the assembled mtDNA datasets of all orders of demosponges, we investigated how common the tRNA gene recruitment occurred in mitochondrial genomes. Further this investigation was expanded outside of the organellar genomes by analyzing the tRNA gene families from nuclear genomes. This part of investigation is presented in chapter 4 of the thesis.

**The evolution of mitochondrial proteome in animals**

As mentioned above, animal mtDNAs only contain 13-15 protein-coding genes. However, the complexity of mitochondrial functions requires a proteome composed of as many as over a thousand of proteins, more than 98% of which are encoded by nuclear genes (Ryan and Hoogenraad 2007). There are about 1500 nuclear encoded mitochondrial proteins in human and 700 in yeast (Ryan and Hoogenraad 2007). Studies of mitochondrial proteomes showed that they are variable not only in size but also in content potentially due to the diversely metabolic functions among different phyla (Gabaldon and Huynen 2004). For example,
certain expansion of mitochondrial proteome in animals reflects their adaptation to multicellularity and tissue differentiation (Mootha et al. 2003). A comparative study of human and yeast mitochondrial proteomes has shown that they share only about 20% mitochondrial proteins (Gabaldon and Huynen 2003).

During the evolutionary course, mitochondria have lost their genes or transferred them to nucleus through the "endosymbiotic gene transfer" process (Kurland and Andersson 2000). As we can expect, some of those transferred genes can redirect their coded proteins to mitochondria once the genes obtained proper mitochondrial targeting sequences. However, mitochondria certainly need to recruit new proteins continuously especially considering the emerging new functions of mitochondria such as Fe/S cluster biogenesis (Gabaldon and Huynen 2004). So an interesting question here is to study from where and how the mitochondrial proteins evolve. We investigated this question by analyzing mitochondrial protein families in human, the species with most available mitochondrial protein data, and presented the study of mitochondrial proteome and its evolution in chapter 5.

**Dissertation Organization**

This dissertation consists of the general introduction (this chapter), four journal papers (chapters 2-5), and general conclusions (chapter 6). Chapter 2, published in the Molecular Biology and Evolution (2007, 24:363-373), presented the study of mtDNA of a species in Homoscleromorpha and its phylogenetic relationship to other animals. Chapter 3, published in PLoS ONE (2008, 3(7):e2723), displayed the comparative study of mtDNAs sequenced from all orders of demosponges and the phylogenetic relationship within the group. Chapter
4, submitted to Genome Biology and Evolution, demonstrated evidences of horizontal evolution of tRNA gene families in mitochondrial genomes, as well as in nuclear genomes. Chapter 5, accepted to BMC Evolutionary Biology, showed the study of mitochondrial proteome and its evolution in animals. For the work presented in all four journal papers (chapters 2-5), I conceived research, analyzed data and wrote manuscripts. Chapter 6 contains the summaries of the thesis and recommendations for future research.

References


Helfenbein KG, Fourcade HM, Vanjani RG, and Boore JL. 2004. The mitochondrial genome of Paraspadella gotoi is highly reduced and reveals that chaetognaths are a sister group to protostomes. Proc Natl Acad Sci USA. 101: 10639-43.


The mitochondrial genome of Acropora tenuis (Cnidaria; Scleractinia) contains a large group I intron and a candidate control region. J Mol Evol. 55: 1-13.

CHAPTER 2. MITOCHONDRIAL GENOME OF THE HOMOSCLEROMORPH

OSCARELLA CARMELA (PORIFERA, DEMOSPONGIAE) REVEALS
UNEXPECTED COMPLEXITY IN THE COMMON ANCESTOR OF SPONGES
AND OTHER ANIMALS

A paper published in Molecular Biology and Evolution

Xiujuan Wang and Dennis V. Lavrov

Abstract

Homoscleromorpha is a small group in the phylum Porifera (Sponges) characterized by several morphological features (basement membrane, acrosomes in spermatozoa, and cross-striated rootlets of the flagellar basal apparatus) shared with eumetazoan animals but not found in most other sponges. To clarify the phylogenetic position of this group, we determined and analyzed the complete mitochondrial DNA (mtDNA) sequence of the homoscleromorph sponge Oscarella carmela (Porifera, Demospongiae). O. carmela mtDNA is 20,327 bp and contains the largest complement of genes reported for animal mtDNA including a putative gene for the C subunit of the twin-arginine translocase (tatC) that has never been reported in animal mtDNA. The genes in O. carmela mtDNA are arranged in two clusters with opposite transcriptional orientations, a gene arrangement reminiscent of those in several cnidarian mtDNAs, but unlike those reported in sponges. At the same time, phylogenetic analyses based on concatenated amino acid sequences from twelve mitochondrial protein genes strongly support the phylogenetic affinity between the
Homoscleromorpha and other demosponges. Altogether, our data suggest that homoscleromorphs are demosponges that have retained ancestral features in both mitochondrial genome and morphological organization lost in other taxa, and that the most recent common ancestor of sponges and other animals was morphologically and genetically more complex than previously thought.

Introduction

Sponges (phylum Porifera) are an exclusively aquatic and predominantly filter-feeding group of animals consisting of approximately 15,000 extant species in three distinct groups, the Hexactinellida (glass sponges), the Calcarea (calcareous sponges) and the Demospongiae ( demosponges) (Hooper and Van Soest 2002). Morphologically, sponges are built around an aquiferous system of canals and chambers, connected to the surrounding environment by multiple pores (hence the name Porifera). Histologically, the sponge body consists of two primary layers of cells (pinacoderm and choanoderm) and an inner cellular region (mesohyl) (Harrison and De Vos 1991). Pinacoderm, the outer layer of cells, lines the surface of the sponge and continues into internal canals where it is eventually replaced by the choanoderm, a layer of characteristic flagellated cells (choanocytes) surrounding the chambers. Choanocytes make up the principle “pump” and “filter” of the system, driving water through the sponge, trapping and phagocytizing suspended bacteria and other particulate food (De Vos et al. 1991). It is generally accepted that neither pinacoderm nor choanoderm constitutes the true epithelium (Woollacott and Pinto 1995; Tyler 2003). In fact, it is habitually stated that sponges do not possess any true tissues and thus represent an early stage in the evolution of animal multicellularity (Brusca and Brusca 2002). Consequently,
sponges have been often placed in the subkingdom Parazoa, separate from the true animals - Eumetazoa.

One group of sponges that challenges the view on Porifera as an ancestral animal phylum that never reached the tissue grade of organization is the subclass Homoscleromorpha. This small group (containing only 7 genera and about 60 species) is characterized by several unusual features, including unique cinctoblastula larvae that form by a unique process of multipolar egression; a basement membrane underlying both choanoderm and pinacoderm; flagellated pinacocytes; and distinctive morphology of aquiferous system and spicules (when present) (Gaino et al. 1987; Boute et al. 1996; Ereskovsky et al. 2002; Muricy and Diaz 2002). Recently, it has been shown that the basement membrane previously observed in adult homoscleromorphs is also lining the epithelial cells in homoscleromorph larvae and that these cells meet all criteria of true epithelia in higher animals: cell polarization, apical cell junctions, and a basement membrane (Boury-Esnault et al. 2003). Thus at least one group of sponges has clearly reached the tissue-grade of organization in its evolution. Interestingly, in addition to true epithelia, homoscleromorphs also share with “higher” animals the presence of acrosomes in spermatozoa (Baccetti et al. 1986; Boury-Esnault and Jamieson 1999) and (together with calcareous sponges) the presence of cross-striated rootlets in the flagellar basal apparatus of larval cells (Boury-Esnault et al. 2003; Maldonado 2004).

Three explanations are possible for these intriguing findings:

1) True epithelium, acrosomes, and cross-striated rootlets evolved independently in Homoscleromorpha and Eumetazoa;
2) These shared characters evolved in the common ancestor of sponges and higher animals but were lost in most sponges;

3) Demosponges are not monophyletic; Homoscleromorpha shares a more recent common ancestor with Eumetazoa.

The choice among these alternative explanations has important implications for our understanding of the evolution of sponges and animals in general but requires knowledge of the phylogenetic position of the Homoscleromorpha. The latter, however, remains controversial.

Because of their distinct morphology and relatively simple anatomical organization, Homoscleromorpha has been traditionally regarded as one of the most primitive groups of demosponges (Lévi 1957), although a relationship to calcareous (calcaronean) sponges has also been proposed (Soest 1984; Grothe 1989; Grothe and Reitner 1990). Recent studies utilizing small (SSU) and large (LSU) subunit rRNA sequences found no support for the inclusion of homoscleromorphs in the Demospongiae (Borchiellini et al. 2004; Nichols 2005), while the sister group relationship with Calcarea received some support from the Bayesian analysis of LSU rDNA data (Nichols 2005). The author of the latter study, however, downplays the significance of this association and points to the need of independent loci for the analysis of sponge relationships. Here we describe the complete mitochondrial DNA sequence from the homoscleromorph *Oscarella carmela* and analyze it in an attempt to clarify the phylogenetic position of this group.

Animal mitochondrial DNA (mtDNA) is typically a small (~16 kb), circular-mapping molecule that contains 37 genes coding for 13 proteins, 2 ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs) (Boore 1999). These genes are usually compactly arrayed, have no
introns, and their order is often stable over long evolutionary time. MtDNAs of bilaterian animals are further distinguished by multiple deviations in the genetic code, unusual and/or reduced rRNA and tRNA primary and secondary structures, and the presence of a single large non-coding region [reviewed in (Wolstenholme 1992)]. Demosponge mtDNA resemble that of most other animals in their compact organization, lack of introns, and a well-conserved gene order, but at the same time contain several extra genes, encode bacterial-like ribosomal and transfer RNAs, and use a minimally derived genetic code in protein synthesis (Lavrov et al. 2005). Mitochondrial genomic data provides an excellent dataset to investigate homoscleromorph relationships. In addition to the large amount of sequence data, which minimize the sampling error in sequence-based phylogenetic analysis, mtDNA harbors additional rare genomic characters useful for phylogenetic inference, including indels in the coding sequences, variations in the genetic code, changes in secondary structures of encoded transfer and ribosomal RNAs, and gene rearrangements. The use of mitochondrial data is especially advantageous for the reconstruction of demosponge relationships because mitochondrial sequences evolve relatively slowly in this group, while the rate of gene rearrangements is relatively high (Lavrov and Lang 2005a; Lavrov et al. 2005).

Material and Methods

Specimen collection, DNA extraction, mtDNA amplification, cloning, and sequencing.

A specimen of Oscarella carmela Muricy and Pearse, 2004 (Class Demospongiae: Subclass Homoscleromorpha: Order Homosclerophorida: Family Plakinidae) (Muricy and Pearse 2004) was a gift from Scott A. Nichols (University of California, Berkeley). Total DNA was
extracted from about 0.2 grams of tissue fixed in 95% ethanol with a phenol-chloroform method modified from Saghai-Maroof (1984). Regions of mitochondrial \textit{cox2} and \textit{nad5} were amplified and sequenced using degenerate primers designed in our lab, checked against the Genbank database to exclude the possibility of contamination, and used to design specific primers for these regions:

\textbf{Os-cox2-f1:} 5’-CATATATGGTTCTACTTCAGATC-3’
\textbf{Os-cox2-r1:} 5’-TTAACACCTAAAGATGGTACTGC-3’
\textbf{Os-nad5-f1:} 5’-GCGATAAACGAATATCTCGACC-3’
\textbf{Os-nad5-r1:} 5’-TAGACCTAGTTGAGCTGATTTC-3’

Complete \textit{Oscarella carmela} mtDNA was amplified in two overlapping fragments (~6 and 15 kbp in size) using the TaKaRa LA-PCR kit under recommended conditions. Random clone libraries were constructed from the purified PCR products using the TOPO Shotgun Subcloning Kit from Invitrogen. Plasmid preparation and sequencing was done at the Iowa State University Office of Biotechnology DNA facility. In addition to mtDNA, the nuclear small subunit rRNA gene was amplified by PCR using modified versions of the universal eukaryotic primers A and B (Medlin et al. 1988) and used to confirm the proper taxonomic identification of the sample.

\textbf{Assembly, gene identification and sequence analysis}

Sequencing reads were assembled using the STADEN software package (Staden 1996). To assure the quality of the final sequence, we manually checked the final assembly for sequencing errors and made sure that all genomic regions have either sequencing reads in both directions or at least 3 different reads in the same direction. Problematic and
underrepresented regions in the assembly were sequenced directly from PCR products by primer-walking. Transfer RNA genes were identified by the tRNAscan-SE program (Lowe and Eddy 1997); rRNA and protein genes were identified by similarity searches in local databases using the FASTA program (Pearson 1994), and in GenBank at NCBI using BLAST network service (Benson et al. 2003). The secondary structures of rRNA genes were manually folded by analogy to published rRNA structures, and drawn with the RnaViz 2 program (De Rijk et al. 2003).

**Phylogenetic analysis**

Concatenated alignment of 2812 amino acids deduced from twelve protein genes was created with ClustalW 1.82 (Thompson et al. 1994) and SOAP (Löytynoja and Milinkovitch 2001) programs via a previously described procedure (Lavrov et al. 2005). We performed a maximum likelihood (ML) search for the best tree with the TreeFinder (May 2006) program (Jobb et al. 2004) using the mtREV model of amino-acid substitutions and 4 gamma categories. Bayesian inferences (MB) were conducted with MrBayes 3.1.1 (Ronquist and Huelsenbeck 2003). We used the mtREV model of amino acid substitutions with gamma+invariant distributed rates and ran four Markov Chain Monte Carlo (MCMC) chains for 1,100,000 generations. Trees were sampled every 1000th cycle after the first 100,000 burn-in cycles. Molecular distances were calculated with the TREE-PUZZLE 5.2 program (Strimmer and von Haeseler 1996) and the same substitution model as for the Bayesian analysis. The distance tree topology was inferred with the WEIGHBOR program (Bruno et al. 2000). For the bootstrap analysis of the distance data, a dataset of 1000 replicates was generated by the SEQBOOT program in the PHYLIP package (Felsenstein 2005) and the
distances for each dataset were calculated using the “puzzleboot script” by Mike Holder and Andrew Roger (http://hades.biochem.dal.ca/Rogerlab/Software/software.html) and the programs listed above. The consensus bootstrap tree was calculated by the CONSENSE program of PHYLIP.

Results

Genome Organization: the largest set of genes in animal mtDNA, unusual gene order, and high coding density

The mitochondrial genome of Oscarella carmela is a circular-mapping molecule 20,327 bp in size, and contains 15 protein coding genes, 2 rRNA genes, and 27 tRNA genes; the largest complement of genes found in animal mtDNAs (fig. 1). In addition to the 37 genes typical for bilaterian mtDNAs (Wolstenholme 1992), genes for subunit 9 of ATP synthase (atp9), twin-arginine translocase component C (tatC), 3 extra tRNAs [trnI(cau), trnR(ucu), trnM(cau)e], as well as two duplicated tRNA genes are present in O. carmela mtDNA. While four of these genes [atp9, trnI(cau), trnR(ucu), trnM(cau)e] have been previously described in other demosponge mt. genomes (Lavrov et al. 2005), this is the first report of tatC in animal mtDNA.

The genes in O. carmela mtDNA are arranged into two clusters with opposite transcriptional orientations that subdivide the genome into two nearly equal parts of 9842 and 10485 bp. The change in the transcriptional polarity occurs between cox1 and cox2 and between trnM(cau)e and cob, putative transcription initiation and termination sites, respectively (Fig. 1). The arrangement of genes into two transcriptional units is unique among demosponge mtDNAs, where all genes are typically transcribed from the same
mtDNA strand (Lavrov and Lang 2005a; Lavrov et al. 2005; unpublished data), but has been found in several other animal groups. In particular, mtDNA from two cnidarians, the moon jelly *Aurelia aurita* (Shao et al. 2006) and the octocoral *Sarcophyton glaucum* (Beaton et al. 1998) have similar arrangements of genes with transcription polarity changing at the same gene junction between *cox1* and *cox2*. Aside from \(-cox1+cox2\) gene boundaries, several other mitochondrial gene arrangements are shared between *O. carmela* and other animals (fig. 2), indicating a moderate number of rearrangements in this genome.

The *O. carmela* mtDNA is a compact genome that contains only 1275 non-coding base pairs (6.27% of the genome sequence). These non-coding nucleotides are distributed among 43 intergenic regions 1 - 130 bp in size. Twenty of the intergenic regions contain more than 20 bp, and three, located between *nad2* and *nad5*, *trnF* and *cox3*, *cox1* and *cox2*, are larger than 100 bp. We found no significant similarity between any of these regions and existing sequences in Genbank.

**Nucleotide composition and codon usage: Prevalence of selection over mutational biases**

The A+T content of *O. carmela* mtDNA is 66.4%, similar to those of other demosponge mitochondrial genomes. However, in contrast to the other genomes, the two strands of *O. carmela* mtDNA do not differ significantly in nucleotide composition (the total AT- and GC-skews are 0.02 and 0, respectively). This lack of strand asymmetry in *O. carmela* mtDNA is the result of opposite nucleotide biases in part I (*cox2-cob*) and part II (*cox1-M*) of this genome [coding strands in both parts have positive GC-skews (0.13, 0.12 respectively) and negative AT-skews (-0.02, -0.06 respectively), which cancel each other
when the whole sequence is considered]. Among different types of genes, protein genes and tRNA genes display negative AT skews, while rRNA genes show positive AT skews; all types of genes display positive GC- skews (Table 1). Among individual genes, only *atp8* deviates from the described pattern and has a negative GC skew and a positive AT skew (supplementary table 1). Interestingly, in the case of protein-coding genes the GC-skew is strongly positive (.38) at the first codon positions, negative (-.17) at the second and weakly positive (.09) at the third. Similarly, AT-skew is weakly negative (-0.02) at the first position, strongly negative (-0.37) at the second, but positive (0.06) at the third (table 1). Thus selection for specific amino acids appears to play a dominant role in shaping the nucleotide skews between the two strands of *O. carmela* mtDNA. At the same time, proposed cytosine deamination in the process of asymmetrical replication and transcription (Lobry 1996; Francino et al. 1996; Francino and Ochman 1997; Frank and Lobry 1999) may also play some role in strand asymmetry as evident from the presence of nucleotide skews at the third codon position and in the intergenic regions (Table 1).

Synonymous codon usage largely correlates with the nucleotide biases in the coding strand: codons ending with A or T are clearly preferred (80.9%), while those ending with C are the least frequent (supplementary table 2). Out of 62 codons expected to specify an amino acid, one (CGC) is not found in the mitochondrial protein genes of *O. carmela*, as well as other demosponges (Lavrov et al. 2005). No significant differences were found in the codon usage in protein genes encoded by part I and part II of the mtDNA.

**Protein genes: the first report of tatC in animal mtDNA**
We identified 15 protein-coding genes in the *O. carmela* mitochondrial genome. Fourteen of them (*atp6, atp8-9, cob, cox1-3, nad1-6, nad4L*) have been previously reported in demosponge mtDNA (Lavrov et al. 2005). These genes, coding for protein subunits involved in respiration and oxidative phosphorylation, are similar in sizes to their homologues in other demosponges mtDNAs (± <7 %, except for *atp8* which is 21% smaller than in *Tethya actinia* mtDNA) and share with them on average 69.8% (27.2-89.3%) of inferred amino-acid identity (table 2). As expected, more variation in size and lower sequence identity were found in comparisons of *O. carmela* mitochondrial protein-coding genes with their homologues in *Metridium senile* (table 2) and other animals (not shown).

In addition to the protein genes described above, an ORF has been found in the *O. carmela* mitochondrial sequence and identified as *tatC* based on sequence similarity searches, presence of conserved domains, and predicted secondary structure (fig. 3, see below). *TatC* (also known as *ymf16* and *mttB*) codes for the largest and usually the most conserved subunit of the twin-arginine transport (Tat) pathway (Bogsch et al. 1998), which exists in prokaryotic organisms, chloroplasts and some mitochondria, and functions in the transport of fully folded proteins and enzyme complexes across membranes [for a comprehensive review see Berks et al., (2003)]. Previously *tatC* has been reported in mtDNA of plants and protists (including closely related to animals choanoflagellate *Monosiga brevicolis*), but has never been found in either animal or fungal mtDNA [(Yen et al. 2002); note that *Thraustochytrium aurelum* identified as “marine fungus” in the cited paper is actually a stramenopile alga].

The inferred size of *tatC* in *O. carmela* is 759 bp, typical for homologous genes in other organisms (Yen et al. 2002). This size estimate is based on our assignment of TTG as
the initiation codon for tatC (the closest in-frame ATG codon is 252 nt downstream).
Although this start codon is unusual, it has been reported as an initiation codon in other organisms (Golderer et al. 1995; Ko and Smith 1999; Baar et al. 2003) as well as in mitochondrial protein coding genes (Okimoto et al. 1990) and may be used to regulate the expression of the tatC relative to other mitochondrially-encoded genes (Okimoto et al. 1990; Golderer et al. 1995). The derived amino acid sequence of O. carmela TatC is 27% and 19% identical with those of Reclinomonas americana and Monosiga brevicollis, respectively.

Blast searches against the raw sequences from the nuclear genome of the demosponge Amphimedon queenslandica (Hooper and Van Soest, 2006) and the complete nuclear genomes of other animals identified a tatC-like sequence only in the demosponge genome (NCBI trace archive database; reads 922482408, 922482312; 25% of inferred amino acid identity). This finding suggests that the fate of this gene may have been different in sponges than in other animals (transfer to the nucleus vs. loss). Alternatively, it is possible the gene still exists in the nuclear genomes of other animals but has evolved beyond recognition.

rRNA genes (rns, rnl) encode well-conserved rRNA molecules

Genes for the small and large subunit ribosomal RNAs (rns and rnl, respectively) have been found in O. carmela mtDNA and the secondary structures of encoded rRNAs have been modeled by analogy with homologous molecules (supplementary figs. 1 and 2). Rns and rnl are located more than 5 kbp apart in the genome and have opposite transcriptional polarities (fig. 1). Such arrangement is unusual for demosponge mtDNA where rns and rnl are usually separated by two tRNA genes (rns-trnG-trnV-rnl) (Lavrov et al. 2005; unpublished data), and relatively rare in animal mtDNA in general. Based on secondary
structure modeling, we deduced the length of *rns* as 1281 bp (making it the largest mt-*rns* described for animals) and the length of *rnl* as 2520 bp (similar to homologous genes in other demosponges). The larger size of *O. Carmela rns* is mostly due to the better conservation of stem 39, which has a similar size in *O. Carmela* and *E. coli*, but is reduced in other demosponges (supplementary fig. 1). The loop at the end of stem 33 also has a 15 nt insertion in *O. Carmela* mt SSU-RNA. In contrast, only small indels (less than 10 nt) were found in *O. Carmela rnl*. These include insertions in stem 54 and loops adjacent to stem 25, 52 and 101. The primary sequence of *O. Carmela rns* and *rnl* are well conserved and share on average 65.8% and 68.1% of sequence identity with homologous genes in *Geodia neptuni* and *Tethya actinia*, 43.2% and 50.6% of sequence identity with their homologues in the choanoflagellate *Monosiga brevicolis*, and 44.1% and 47.7% of sequence identity with those in *E. coli*, respectively.

**Duplicated tRNA genes, canonical tRNA structures, and unusual trnP(ugg)**

Twenty-seven tRNA genes have been identified in *O. Carmela* mtDNA and their inferred secondary structures are shown in supplementary figure 3. These genes include the same set of 24 tRNA genes found in three other demosponges (Lavrov and Lang 2005b), an additional gene for an elongator tRNA\textsubscript{Met\_CAU} previously reported only in *Tethya actinia* among demosponges (Lavrov et al. 2005), and duplicated genes for tRNA\textsubscript{Val\_UAC} and tRNA\textsubscript{Thr\_UGU}. The two copies of tRNA\textsubscript{Thr\_UGU} have only one nucleotide difference and are located about 5kb away in the same transcription strand (part II, *cox1-M*), while those of tRNA\textsubscript{Val\_UAC} have a 4 nucleotide difference and are located in different transcriptional strands. The duplicated copies of genes for tRNA\textsubscript{Thr\_UGU} (T') and tRNA\textsubscript{Val\_UAC} (V') are adjacent in *O. Carmela* mtDNA.
(fig 1). The duplication mechanism for these genes is unknown, but is unlikely to be due to the commonly invoked duplication-random loss model (Boore 2000) because it would explain neither the clustering of the duplicated genes in the same region of the genome nor the change in the transcriptional polarity of trnV(uac).

The primary sequences of O. carmela tRNA genes share 55.4% to 91.7% (average = 73.8%) sequence identity with homologous genes in Geodia neptuni and Tethya actinia. The most conserved tRNA is tRNA$_{UCA}^{Tr}$ (the average identity with the two demosponges is 84.5%) and the least conserved is tRNA$_{UGG}^{Glu}$ (the average identity is 59.6%). The consensus primary sequences and secondary structures for type 1 (with short variable arm) and type 2 (with long variable arm) O. carmela mt-tRNAs are shown in figure 4. As can be seen from this figure, most nucleotides involved in tRNA tertiary interactions (including G18-U55 and G19-C56 interactions between D- and T-loops) are well conserved in O. carmela mt-tRNAs.

Interestingly, we found a highly unusual A11-T24 pair in O. carmela tRNA$_{UGG}^{Pro}$, similar to the animal-specific R11-Y24 pair in tRNA$_{UCA}^{Trp}$ (Wolstenholme 1992; Lavrov et al. 2005). The A11-T24 pair is also present in mt-trnP(ugg) genes from other demosponges (Lavrov and Lang 2005b) as well as the placozoan Trichoplax adhaerens (Dellaporta et al. 2006), but is not found in homologous genes of either the outgroups Monosiga brevicolis and Amoebidium parasiticum, or bilaterian animals (fig. 4B). Because the R11-Y24 base pair is an important recognition element for initiator tRNA, it is usually strongly counter-selected in elongator tRNAs (Marck and Grosjean 2002), and its presence in tRNA$_{UGG}^{Pro}$ of demosponges and T. adhaerens may be phylogenetically significant (see below).
Sequence-based phylogenetic analysis supports the demosponge affinity of the Homoscleromorpha

Phylogenetic analysis based on the concatenated amino acid sequences inferred from twelve mitochondrial protein genes recovers an overall conventional tree of eukaryotic relationships, but with “lower” animals (phyla Porifera, Cnidaria, and Placozoa) forming a monophyletic group (fig. 5). This clustering of lower animals has been previously explained by elevated rates of mitochondrial evolution in Bilateria, which would pull the latter group towards the base of metazoan tree (Lavrov et al. 2005). However, the presence of a highly unusual A11-T24 pair in mt-tRNA \textsuperscript{Pro} \text{UGG} of demosponges and T. adhaerens (cnidarians do not encode this tRNA in mtDNA) provide an additional character supporting this clade. Thus further studies are clearly needed to investigate these contentious relationships.

Within Metazoa, Oscarella carmela groups with other demosponges with 100% support in ML, Weighted NJ, and Bayesian analyses (fig. 5). Furthermore, this relationship received 98% bootstrap support in MP analysis using original data and 94% bootstrap support in MP analysis where individual amino-acids were recoded into the six Dayhoff categories as in Embley \textit{et al.} (2003). The results of our analyses are robust with respect to taxa selection and do not change when preliminary sequences from several mitochondrial genes from glass and calcareous sponges are included in the dataset or when bilaterian taxa are removed from the analysis (not shown). The sister-group relationship between \textit{O. carmela} and other demosponges is also recovered by the ML analysis for 8 out of 12 individual mitochondrial protein genes (\textit{atp6, cob, cox2, cox3, nad1, nad2, nad4, nad5}). The analyses of the remaining genes produced four different placements for \textit{O. carmela} (supplementary fig. 4).
Discussion and Conclusion

Our analysis of Oscarella carmela mtDNA revealed several genomic features potentially informative for understanding the phylogenetic position of the Homoscleromorpha. Here we discuss these features with respect to three possible phylogenetic hypotheses: 1) Homoscleromorpha diverged from the animal lineage prior to other demosponges; 2) Homoscleromorpha is more closely related to demosponges than to Eumetazoa; 3) Homoscleromorpha is more closely related to Eumetazoa than to demosponges. It should be noted that because of the scarcity of data from glass and calcareous sponges, we could not test the monophyly of the “Demospongiae” in the present study.

The first of these hypotheses is supported by the presence of an extra protein gene (tatC) in Oscarella carmela mtDNA. Because this gene is mitochondrially encoded in multiple outgroups, including the choanoflagellate Monosiga brevicolis, but is absent in other animal mtDNA, the most parsimonious reconstruction is a single loss of tatC after the divergence between the Homoscleromorpha and other animals. Unfortunately, it is well known that the lack of mitochondrial genes is not a reliable phylogenetic character, and that parallel independent losses from organellar DNA are common (Martin et al. 1998). Our finding of a tatC-like sequence in the nuclear genome of Amphimedon queenslandica but not other animals hints to such independent events in tatC evolution and suggests that the fate of this gene was different in demosponges and bilaterian animals (transfer to the nucleus vs. loss). The only other feature supporting the basal position of the Homoscleromorpha is the conservation of some helices in O. carmela mt SSU-RNA secondary structure – not a strong
phylogenetic character either. Overall, we regard the support for the first hypothesis as weak.

The second hypothesis, the inclusion of Homoscleromorpha within the monophyletic Demospongiae and/or Porifera, is supported by the phylogenetic analysis of the mitochondrial sequence data. The grouping of *O. carmela* with other demosponges is robust with respect to different selections of genes, taxa, models, and phylogenetic methods. Furthermore we searched for, but could not identify any potential biases in either nucleotide composition or rates of sequence evolution that would cause this association. Therefore we posit that our results reflect a genuine phylogenetic signal present in the mitochondrial dataset rather than an artifact of phylogenetic reconstruction.

Finally, our data provide no support for the closer phylogenetic relationship of the Homoscleromorpha to the Eumetazoa rather than the Demospongiae. It may appear that similar mitochondrial gene arrangements in *O. carmela* and several cnidarians support this phylogenetic hypothesis, but this is not the case. The reported similarities can be equally parsimoniously explained by these arrangements being plesiomorphic for all animals (or for non-bilaterian animals if the latter group is indeed monophyletic). Unfortunately, we are not able to distinguish between these possibilities due to the lack of informative outgroups outside the Metazoa.

If Homoscleromorpha forms a monophyletic group with demosponges (and potentially other sponges), then the finding of morphological features shared between this group and Eumetazoa (acrosomes in spermatozoa, true epithelia with basal lamina, and cross-striated rootlets) is most easily explained by the presence of these features in the common
ancestor of sponges and other animals and their subsequent loss in most (but not all\(^1\)) sponge lineages. If this is indeed the case, then the common ancestor of sponges and other animals should have been morphologically more complex than modern sponges, which may represent an adaptive simplification to their sessile and filter-feeding life style. An alternative explanation would need to involve an independent origin of similar morphological characters in several animal lineages, an unlikely scenario in our view. Interestingly, a similar deduction has been made recently by Manuel Maldonado (2004), based on an independent re-assessment of embryological and histological data from sponges.

**Supplementary Materials**

Amino-acid alignment used in phylogenetic analyses, Supplementary Tables S1, S2, and Supplementary Fig. S1, S2, S3, S4 are available at *Molecular Biology and Evolution* online ([http://www.mbe.oxfordjournals.org/](http://www.mbe.oxfordjournals.org/)). *Oscarella carmela* mitochondrial genome sequence has been deposited in the GenBank database under the accession number EF081250.

**Acknowledgements**

We thank Scott Nichols for the samples of *Oscarella carmela*, Alexander Ereskovsky, Karri Haen, Sally Leys, and Scott Nichols for valuable comments on an earlier version of this

---

1 Some of the morphological features discussed above have been reported in other demosponges (mostly from the order Poecilosclerida): a network of collagen fibers underlining larval ciliated cells has been described in *Crambe crambe* (Maldonado 2004), the sperms with acrosomes have been found in *Crambe crambe* and *Crellomima imparidens* (Ereskovsky 2005), striated rootlets of the basal body – in *Mycal contarenii* larva (Lévi 1964).
manuscript, and the College of Liberal Arts and Sciences at Iowa State University for funding.

**Literature Cited**


Nichols SA. 2005. An evaluation of support for order-level monophyly and interrelationships within the class Demospongiae using partial data from the large subunit rDNA and cytochrome oxidase subunit I. Mol Phylogenet Evol 34:81-96.


### Table 1 Nucleotide composition for the coding strands of *Oscarella carmela* mtDNA

<table>
<thead>
<tr>
<th>Coding sequences</th>
<th>Codon position</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>part I ((\text{cox}2)-(\text{cob})) total</th>
<th>part II ((\text{cox}1)-(\text{trnM})) total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>Second</td>
<td>Third</td>
<td>Total</td>
<td>rRNA genes</td>
<td>tRNA genes</td>
<td>Inter-genic</td>
<td></td>
</tr>
<tr>
<td>% G</td>
<td>26.8</td>
<td>15.1</td>
<td>10.4</td>
<td>17.5</td>
<td>22.2</td>
<td>23.8</td>
<td>15.8</td>
<td>19.0</td>
</tr>
<tr>
<td>% A</td>
<td>29.9</td>
<td>20.2</td>
<td>42.9</td>
<td>31.0</td>
<td>35.4</td>
<td>27.0</td>
<td>33.9</td>
<td>32.5</td>
</tr>
<tr>
<td>% T</td>
<td>31.3</td>
<td>43.5</td>
<td>37.9</td>
<td>37.6</td>
<td>26.8</td>
<td>31.1</td>
<td>35.1</td>
<td>33.9</td>
</tr>
<tr>
<td>% C</td>
<td>11.9</td>
<td>21.1</td>
<td>8.7</td>
<td>13.9</td>
<td>15.6</td>
<td>18.0</td>
<td>15.1</td>
<td>14.6</td>
</tr>
<tr>
<td>%A+T</td>
<td>61.2</td>
<td>63.7</td>
<td>80.9</td>
<td>68.6</td>
<td>62.2</td>
<td>58.2</td>
<td>69.1</td>
<td>66.4</td>
</tr>
<tr>
<td>AT-skew</td>
<td>-0.02</td>
<td>-0.37</td>
<td>0.06</td>
<td>-0.10</td>
<td>0.14</td>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>GC-skew</td>
<td>0.38</td>
<td>-0.17</td>
<td>0.09</td>
<td>0.11</td>
<td>0.18</td>
<td>0.14</td>
<td>0.03</td>
<td>0.13</td>
</tr>
<tr>
<td>Total (bp)</td>
<td>4420</td>
<td>4420</td>
<td>4420</td>
<td>13260</td>
<td>3801</td>
<td>1991</td>
<td>1275</td>
<td>9842</td>
</tr>
</tbody>
</table>
Table 2: Comparison of mitochondrial protein genes in *Oscarella carmela* (OC) with those of demosponges *Geodia neptuni* (GN) and *Tethya actinia* (TA), cnidarian *Metridium senile* (MS), and choanoflagellate *Monosiga brevicollis* (MB).

<table>
<thead>
<tr>
<th>Gene</th>
<th>OC</th>
<th>GN</th>
<th>TA</th>
<th>MS</th>
<th>MB</th>
<th>OC/GN</th>
<th>OC/TA</th>
<th>OC/MS</th>
<th>OC/MB</th>
<th>Initiation</th>
<th>Stop</th>
<th>Codons</th>
<th>% amino acid identity</th>
<th>Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>atp6</em></td>
<td>245</td>
<td>244</td>
<td>244</td>
<td>229</td>
<td>252</td>
<td>65.7</td>
<td>69.4</td>
<td>64.1</td>
<td>51.6</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>atp8</em></td>
<td>66</td>
<td>63</td>
<td>80</td>
<td>72</td>
<td>99</td>
<td>37.9</td>
<td>27.2</td>
<td>39.4</td>
<td>17.2</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>atp9</em></td>
<td>76</td>
<td>78</td>
<td>78</td>
<td>-</td>
<td>73</td>
<td>83.3</td>
<td>78.2</td>
<td>-</td>
<td>69.2</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cob</em></td>
<td>400</td>
<td>381</td>
<td>381</td>
<td>393</td>
<td>380</td>
<td>67.8</td>
<td>67.0</td>
<td>65.4</td>
<td>62.2</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cox1</em></td>
<td>521</td>
<td>520</td>
<td>522</td>
<td>530</td>
<td>534</td>
<td>84.8</td>
<td>89.3</td>
<td>85.1</td>
<td>70.9</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cox2</em></td>
<td>249</td>
<td>247</td>
<td>243</td>
<td>248</td>
<td>256</td>
<td>78.7</td>
<td>77.9</td>
<td>76.6</td>
<td>54.8</td>
<td>ATG</td>
<td>TAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cox3</em></td>
<td>261</td>
<td>262</td>
<td>262</td>
<td>262</td>
<td>263</td>
<td>78.6</td>
<td>78.6</td>
<td>79.0</td>
<td>60.8</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad1</em></td>
<td>328</td>
<td>327</td>
<td>338</td>
<td>334</td>
<td>343</td>
<td>72.5</td>
<td>76.9</td>
<td>71.8</td>
<td>62.4</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad2</em></td>
<td>465</td>
<td>465</td>
<td>481</td>
<td>385</td>
<td>546</td>
<td>62.4</td>
<td>58.3</td>
<td>53.9</td>
<td>34.3</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad3</em></td>
<td>118</td>
<td>118</td>
<td>118</td>
<td>118</td>
<td>118</td>
<td>67.8</td>
<td>72.9</td>
<td>72.9</td>
<td>62.7</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad4</em></td>
<td>495</td>
<td>481</td>
<td>482</td>
<td>491</td>
<td>498</td>
<td>67.9</td>
<td>70.3</td>
<td>69.1</td>
<td>57.7</td>
<td>ATG</td>
<td>TAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad4L</em></td>
<td>99</td>
<td>106</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>68.9</td>
<td>83.8</td>
<td>74.7</td>
<td>60.6</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad5</em></td>
<td>633</td>
<td>603</td>
<td>622</td>
<td>600</td>
<td>688</td>
<td>71.1</td>
<td>74.4</td>
<td>69</td>
<td>51.9</td>
<td>ATG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nad6</em></td>
<td>197</td>
<td>185</td>
<td>187</td>
<td>202</td>
<td>228</td>
<td>62.3</td>
<td>61.6</td>
<td>51.4</td>
<td>37.3</td>
<td>GTG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>tatC</em></td>
<td>253</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>234</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>18.7</td>
<td>TTG</td>
<td>TAA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Data for *G. neptuni* and *T. actinia* are from (Lavrov et al. 2005); for *M. senile* from (Beagley et al. 1998); for *M. brevicollis* from (Burger et al. 2003).
Table 3 Codon usage among the 14 genes coding for protein subunits involved in respiration and oxidative phosphorylation and, separately, *tatC*

<table>
<thead>
<tr>
<th></th>
<th>A.a</th>
<th>tatC</th>
<th></th>
<th>A.a</th>
<th>tatC</th>
<th></th>
<th>A.a</th>
<th>tatC</th>
<th></th>
<th>A.a</th>
<th>tatC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>269</td>
<td>20</td>
<td>Ser</td>
<td>98</td>
<td>8</td>
<td>Tyr</td>
<td>159</td>
<td>16</td>
<td>Cys</td>
<td>TGT</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0</td>
<td>Ser</td>
<td>11</td>
<td>2</td>
<td>Tyr</td>
<td>25</td>
<td>3</td>
<td>Cys</td>
<td>TGC</td>
<td>6</td>
</tr>
<tr>
<td>Leu</td>
<td>404</td>
<td>26</td>
<td>Ser</td>
<td>88</td>
<td>1</td>
<td>TER</td>
<td>12</td>
<td>1</td>
<td>Trp</td>
<td>TGA</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>6</td>
<td>Ser</td>
<td>10</td>
<td>1</td>
<td>TAG</td>
<td>2</td>
<td>0</td>
<td>Trp</td>
<td>TGG</td>
<td>14</td>
</tr>
<tr>
<td>Leu</td>
<td>CTT</td>
<td>72</td>
<td>Pro</td>
<td>56</td>
<td>4</td>
<td>His</td>
<td>59</td>
<td>3</td>
<td>Arg</td>
<td>CGT</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>Pro</td>
<td>19</td>
<td>3</td>
<td>CAC</td>
<td>21</td>
<td>0</td>
<td>CGT</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>63</td>
<td>2</td>
<td>Pro</td>
<td>71</td>
<td>2</td>
<td>CAA</td>
<td>78</td>
<td>4</td>
<td>CGA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1</td>
<td>Pro</td>
<td>6</td>
<td>0</td>
<td>CAG</td>
<td>8</td>
<td>1</td>
<td>CGG</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ile</td>
<td>156</td>
<td>18</td>
<td>Thr</td>
<td>87</td>
<td>6</td>
<td>Asn</td>
<td>97</td>
<td>8</td>
<td>Ser</td>
<td>AGT</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>2</td>
<td>Thr</td>
<td>17</td>
<td>2</td>
<td>AAT</td>
<td>28</td>
<td>1</td>
<td>Ser</td>
<td>AGC</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>272</td>
<td>27</td>
<td>Thr</td>
<td>118</td>
<td>1</td>
<td>Lys</td>
<td>102</td>
<td>8</td>
<td>Arg</td>
<td>AGA</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>5</td>
<td>Thr</td>
<td>11</td>
<td>1</td>
<td>AAA</td>
<td>13</td>
<td>3</td>
<td>Arg</td>
<td>AGG</td>
<td>13</td>
</tr>
<tr>
<td>Met</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>118</td>
<td>5</td>
<td>Ala</td>
<td>119</td>
<td>3</td>
<td>Asp</td>
<td>74</td>
<td>4</td>
<td>Gly</td>
<td>GGT</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0</td>
<td>Ala</td>
<td>52</td>
<td>1</td>
<td>Asp</td>
<td>29</td>
<td>0</td>
<td>Gly</td>
<td>GGC</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>129</td>
<td>3</td>
<td>Ala</td>
<td>109</td>
<td>0</td>
<td>Gly</td>
<td>94</td>
<td>8</td>
<td>Gly</td>
<td>GGA</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>0</td>
<td>Ala</td>
<td>26</td>
<td>1</td>
<td>Gly</td>
<td>30</td>
<td>2</td>
<td>Gly</td>
<td>GGG</td>
<td>78</td>
</tr>
</tbody>
</table>
Figure 1. Genetic map of *Oscarella carmela* mtDNA. Protein and ribosomal genes (gray) are *atp6, atp8-9*: subunits 6, 8 and 9 of F<sub>0</sub> adenosine triphosphatase (ATP) synthase; *cox1-3*: cytochrome c oxidase subunits 1-3; *cob*: apocytochrome b; *nad1-6* and *nad4L*: NADH dehydrogenase subunits 1-6 and 4L; *tatC*: twin-arginine translocase component C; *rns* and *rnl*: small and large subunit rRNAs. tRNA genes (black) are identified by the one-letter code for their corresponding amino acid; subscripts denote different genes for isoacceptor tRNAs; apostrophes (T' and V') indicate duplicated tRNA genes. Genes are transcribed in two directions: clockwise (*cox2-cob*) and counterclockwise (*cox1-Me*).
Figure 2. Comparison of gene arrangements in the mtDNAs of the homoscleromorph *Oscarella carmela*, demosponge *Geodia neptuni* and cnidarian *Sarcophyton glaucum*. Genes are not drawn to scale; protein and rRNA genes are indicated by larger boxes, tRNA genes by smaller boxes. Open boxes indicate transcriptional direction from left to right, filled boxes – from right to left. Conserved blocks of genes shared between different organisms are underlined and interconnected with arrows. All abbreviations and other symbols are as in fig. 1.
Figure 3. Predicted secondary structures of mitochondrially encoded TatC protein in *Oscarella carmela* (a) and *Monosiga brevicollis* (b). The secondary structure and transmembrane regions were analyzed on the TMHMM server v. 2.0 (Krogh et al. 2001; Sonnhammer, von Heijne, and Krogh 1998). The X-axis designates amino acid positions in each protein. The Y-axis shows posterior probabilities for each prediction.
Figure 4. Consensus secondary structures for *Oscarella carmela* type 1 (with short variable arm) and type 2 (with long variable arm) mt-tRNAs (A) and comparison of 11-24 base pair in tRNA proline (B). Numbering of nucleotides is based on the convention used for yeast tRNA phenylalanine (Robertus et al. 1974). Open circles with numbers, nucleotides are present in all tRNAs group; open circles with letters, nucleotide combinations present in all tRNAs; filled black circle, nucleotides or nucleotide combinations that are described as invariant or semi-invariant in prokaryotic and eukaryotic nuclear-encoded tRNAs with frequencies (percentages) for *O. carmela* type 1 mt-tRNAs shown by accompanying numbers (all these nucleotides are 100% conserved in type 2 tRNAs); filled gray circles, nucleotides present in some but not all tRNAs.

<table>
<thead>
<tr>
<th>Porifera/Placozoa</th>
<th>11 - 24 pair</th>
<th>Bilateria/Outgroups</th>
<th>11 - 24 pair</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oscarella carmela</em></td>
<td>A - T</td>
<td><em>Amoebidium parasiticum</em></td>
<td>C - G</td>
</tr>
<tr>
<td><em>Geodia nepi</em></td>
<td>A - T</td>
<td><em>Monosiga brevicollis</em></td>
<td>T - A</td>
</tr>
<tr>
<td><em>Teilrys aciina</em></td>
<td>A - T</td>
<td><em>Drosophila yakuba</em></td>
<td>T - A</td>
</tr>
<tr>
<td><em>Axinella corrugata</em></td>
<td>A - T</td>
<td><em>Homo sapiens</em></td>
<td>T - A</td>
</tr>
<tr>
<td><em>Trichoplax adhaerens</em></td>
<td>A - T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 3. SEVENTEEN NEW COMPLETE MTDNA SEQUENCES REVEAL
EXTENSIVE MITOCHONDRIAL GENOME EVOLUTION WITHIN THE
DEMOSPONGIAE.

A paper published in *PLoS ONE*

Xiujuan Wang and Dennis V. Lavrov

**Abstract**

Two major transitions in animal evolution – the origins of multicellularity and bilaterality – correlate with major changes in mitochondrial DNA (mtDNA) organization. Demosponges, the largest class in the phylum Porifera, underwent only the first of these transitions and their mitochondrial genomes display a peculiar combination of ancestral and animal-specific features.

To get an insight into the evolution of mitochondrial genomes within the Demospongiae, we determined 17 new mtDNA sequences from this group and analyzing them with five previously published sequences. Our analysis revealed that all demosponge mtDNAs are 16- to 25-kbp circular molecules, containing 13-15 protein genes, 2 rRNA genes, and 2-27 tRNA genes. All but four pairs of sampled genomes had unique gene orders, with the number of shared gene boundaries ranging from 1 to 41. Although most demosponge species displayed low rates of mitochondrial sequence evolution, a significant acceleration in evolutionary rates occurred in the G1 group (orders Dendroceratida, Dictyoceratida, and Verticillitida). Large variation in mtDNA organization was also
observed within the G0 group (order Homosclerophorida) including gene rearrangements, loss of tRNA genes, and the presence of two introns in *Plakortis angulospiculatus*. While introns are rare in modern-day demosponge mtDNA, we inferred that at least one intron was present in *cox1* of the common ancestor of all demosponges.

Our study uncovered an extensive mitochondrial genomic diversity within the Demospongiae. Although all sampled mitochondrial genomes retained some ancestral features, including a minimally modified genetic code, conserved structures of tRNA genes, and presence of multiple non-coding regions, they vary considerably in their size, gene content, gene order, and the rates of sequence evolution. Some of the changes in demosponge mtDNA, such as the loss of tRNA genes and the appearance of hairpin-containing repetitive elements, occurred in parallel in several lineages and suggest general trends in demosponge mtDNA evolution.

**Introduction**

Two major evolutionary events occurred early in animal history and shaped the majority of animals, as we know them today: the origin of multicellularity and the origin of bilateral symmetry. The phylogenetic boundaries of these events are well defined among extant taxa and correspond to the traditional groups Metazoa (multicellular animals) and Bilateria (all animal phyla except Porifera, Placozoa, Cnidaria, and Ctenophora). Multiple genomic changes must have occurred in association with these morphological transitions, and current genome sequencing projects give us the first glimpses into these changes [1,2].
Surprisingly, the transitions to multicellular and bilaterally symmetrical animals also correlate with multiple changes in mitochondrial genome architecture [3], although the main function of mitochondria themselves remained unchanged. In particular, the origin of animal multicellularity is associated with the loss of all ribosomal protein genes from mtDNA, the disappearance of most introns, and a large reduction in the amount of non-coding DNA [3]. The origin of bilaterality correlates with further compaction of mtDNA, multiple changes in the genetic code and the associated losses of some tRNA genes, along with the appearance of several genetic novelties [4]. Obviously, the picture presented above is an extrapolation of our knowledge of extant organisms into the ancient past and as such can be affected by artifacts of ancestral state reconstruction [5]. It is also based on a relatively limited sampling of mitochondrial genomes, especially from non-bilaterian animals, and additional data from Cnidaria, Ctenophora, Porifera, as well as the closely related lineages of eukaryotes (e.g., Choanozoa) are essential to support, expand, or refute it.

Class Demospongiae [6] is the largest (>85% of species) and most morphologically diverse group in the phylum Porifera. It contains sponges of various shapes and sizes that occupy both freshwater and marine environments from shallow to abysmal depths and includes such oddities as carnivorous sponges [7]. Within the extant Demospongiae 14 orders are recognized that encompass 88 families, 500 genera and more than 8000 described species [8,9]. Although traditionally three subclasses have been distinguished, two of them do not appear to be monophyletic. Instead, recent molecular studies [10,11] provide strong support for five major clades within the Demospongiae: Homoscleromorpha (G0) (Homosclerophorida), Keratosa (G1) (Dictyoceratida + Dendroceratida), Myxospongiae (G2)
(Chondrosida, Halisarcida, and Verongida), Marine Haplosclerida (G3), and all the remaining groups (G4) (Figure 1). Our knowledge of mtDNA diversity within the demosponges has been rudimentary, with only five sequences representing 3 of the 5 major groups available [12-15]. Previous studies revealed that demosponge mtDNA resembles that of most other animals in its compact organization, lack of introns, and well-conserved gene order, but at the same time contains extra genes, including \textit{atp9}, \textit{trnI(cau)}, \textit{trnR(ucu)}, encodes bacterial-like ribosomal and transfer RNAs, and uses a minimally derived genetic code in protein synthesis [12]. Furthermore, additional unusual features found in the mitochondrial genomes of \textit{Oscarella carmela} [14] and \textit{Amphimedon queenslandica} [15] suggested that more mitochondrial genomic diversity might exist among the demosponges. Here we describe complete mitochondrial sequences from 17 species of demosponges and analyze them with five previously published mitochondrial genomes from this group that were available at the time this study was conducted. Taken together, our sampling covers all recognized order-level diversity within the Demospongiae and provides the first analysis of general evolutionary trends in mitochondrial genome organization for this group. Such a comprehensive approach to the analysis of demosponge mtDNA is needed because, at least in the fossil record, the evolution of demosponges closely mirrors the evolution of all bilaterian animals with the first demosponge fossils appearing in Precambrian deposits and a major radiation occurring in the Lower Cambrian [16,17].

Results

Genome organization and nucleotide composition
All twenty-two analyzed mtDNAs of demosponges were circular-mapping molecules, each containing a conserved set of thirteen protein-coding and two rRNA genes identical to that found in the mtDNA of most bilaterian animals [18]. In addition, atp9, a gene for subunit 9 of ATP synthase was identified in mtDNA of all demosponges except Amphimedon queenslandica [15], and tatC, a gene for twin arginine translocase subunit C, was found in Oscarella carmela [14]. The number of tRNA genes showed more variation. Although 24 or 25 tRNA genes were present in most analyzed demosponge mitochondrial genomes, as few as 2 and as many as 27 tRNA genes were found in mitochondrial genomes of some demosponge species (Figure 2, see below). In addition, a sequence with a potential to form a tRNA-like structure, named trnX, was located downstream of cox1 in Xestospongia muta and Ephydatia muelleri mtDNA. Inferred tRNA(X) had a well-conserved primary (65.3% nucleotide identity) and secondary structure, except for the putative anticodon arm, which differed both in length and in sequence between the two species.

The sampled demosponge mitochondrial genomes displayed moderate size variation (16-26kb; mean = 19.7kb), most of which could be attributed to the expansions of non-coding regions usually caused by the presence of repetitive elements (Figure 2). We detected no obvious phylogenetic pattern associated with this variation, and no similarity in the sequence of repetitive elements among different species. Most demosponge mitochondrial genomes were larger than their counterparts in bilaterian animals. However, even the largest demosponge mitochondrial genomes were dwarfed in comparison to those in the choanoflagellate Monosiga brevicollis and the placozoan Trichoplax adhaerens, which have
a much higher percentage of non-coding DNA and, in the case of *M. brevicollis*, an expanded gene set (Figure 2).

All analyzed mitochondrial genomes were relatively uniform in the overall nucleotide composition (A+T content between 56-72%) and, on average, displayed negative AT- and positive GC-skews of the coding strand (Figure 3). The sense strand of protein and tRNA genes had a negative AT-skew in all species, that of rRNA genes had a positive AT-skew, while non-coding regions and 3rd codon positions showed a large variation in AT-skew both among and within major demosponge groups (Figure 3B). All types of sequences in demosponge mtDNAs showed positive GC-skews except for the tRNA genes in *Igernella notabilis* and the non-coding regions in *Ephydatia muelleri* and *Aplysina fistularis*. The genomic values for AT- and GC-skews correlated more strongly with those for protein genes ($R^2$=0.89 and 0.95, respectively) and rRNA genes ($R^2$=0.61 and 0.93) than those for tRNA genes ($R^2$=0.06 and 0.57) and non-coding regions ($R^2$=0.13 and 0.34), while genomic A+T content correlated most strongly with that of rRNA genes ($R^2$=0.89) comparing to non-coding regions ($R^2$=0.78), tRNA genes ($R^2$=0.65), and protein genes ($R^2$=0.44). Interestingly, non-coding regions and 3rd codons (that are usually assumed to experience similar mutational pressure) showed little correlation in all three types of measurements ($R^2$ values are 0.05, 0.3 and 0.58 for A+T content, AT- and GC-skews, respectively).

All but four pairs of sampled mitochondrial genomes had unique gene orders, with the number of shared gene boundaries between individual genomes ranging from 1 to 41. The extent of gene order variation and the type of gene rearrangements differed among major groups of demosponges (Figure 4). Gene arrangements of protein and rRNA genes were
generally well conserved within G2, G3 and G4 and the predominant type of change within these groups was tRNA transposition. By contrast, more rearrangements were found within G0 (13% of shared boundaries between two sampled genomes) as well as within G1 (59% of shared boundaries among three genomes). Still, most of the rearrangements were transpositions and only two inversions were found in the whole dataset (in Oscarella carmela and Aplysina fistularis).

**Protein coding genes**

The protein coding genes identified in all 22 demosponge mtDNAs showed 0.33-11.81% variation in size and 31.9-87.3% average pairwise identity calculated based on inferred amino acid sequences (Table S1). *Atp8* was the least conserved gene both in terms of size (11.81% variation), pairwise sequence identity among demosponges (31.9% on average, range 8.5-85.7%), and genetic distance to cnidarian homologues (Figure 5), followed by *nad6*. By contrast, *atp9*, a gene encoding another subunit of the ATP-synthase complex, was the most conserved, with an average pairwise identity of 87.3% (range 76.9%-100%). Other genes were relatively uniform both in their average pairwise identities across the demosponges and the calculated rates of sequence evolution (Figure 5).

Codon usage in all analyzed demosponge mitochondrial genomes was consistent with the minimally modified genetic code inferred in our previous study [12]. All 22 mtDNAs share similar codon usage bias with an effective number of codons equivalent to 41.8± 3.5. Synonymous codons ending with A or T were clearly preferred (56-85% for individual
species; 73.6% on average), while the codon CGC was not used at all in mitochondrial coding sequences of 12 species. *Tethya actinia* displayed the most biased mitochondrial codon usage with no AAC, CGC, CTC, CTG, and TGC codons present.

ATG was the most common initiation codon, followed by GTG, which occurred frequently in *nad6* (15 out of 22 species) and occasionally in other genes (Table S2). The unusual start codon ATT was inferred for *cox2* in *Hippospongia lachne*, *nad3* in *Cinachyrella kuekenthali* and *nad6* in *Vaceletia* sp. and a TGG start codon was inferred for *nad2* in *Ephydatia muelleri*, *nad6* in *Tethya actinia*, *Axinella corrugata, Amphimedon queenslandica* and *tatC* in *Oscarella carmela* (Table S2). Such initiation codons are common in mitochondrial coding sequences of bilaterian animals [4], but are rare, although not unprecedented, in non-bilaterian animals and non-animal outgroups [19,20]. The stop codons TAA and TAG were inferred for all coding sequences except *nad5* in *Amphimedon compressa, Ectyoplasia ferox, Ephydatia muelleri, and Callyspongia plicifera* as well as *nad4L* in *Cinachyrella kuekenthali*. No standard or abbreviated stop codons were found for the latter genes and the mechanism of their translational termination remains unclear.

Among the five major clades within the Demospongiae (G0-G4), a significant acceleration in the rates of evolution was found in G1, especially in the lineage leading to *Vaceletia* sp. and *Hippospongia lachne* (Figure 5; RRTree P=1.00E-07). We tested whether the G1 accelerated rates could have been the result of positive selection as suggested by Bazin et al. [21] but did not find significant support for this hypothesis by either the M1-M2 test in PAML or by the synonymous vs. non-synonymous substitution rate test with the DNASP program [22].
Introns in cox1

Although introns are common in mtDNA of two groups of non-bilaterian animals, Cnidaria and Placozoa, only one mitochondrial intron (in cox1 of Tetilla sp.) has been reported so far in demosponges [23]. Among the 22 demosponge mitochondrial genomes analyzed for this study, we found two additional group I introns, both of them in cox1 of Plakortis angulospiculatus. These introns were 388 bp and 1118 bp in size (henceforth intron 1 and 2, respectively), and separated by only 9 nucleotides (3 codons) in the gene. Intron 2 in P. angulospiculatus was found after position 726 in cox1, at the same location as the intron reported for Tetilla sp. [23]. Intron 2 in P. angulospiculatus and its counterpart in Tetilla sp. share 81.2% nucleotide sequence identity, have a similar secondary structure, and both contain an ORF homologous to LAGLIDADG-type homing endonuclease with identical LAGLIDADG motifs (LAGLIEGDG and LAGFLDADG). By contrast, introns 1 and 2 in P. angulospiculatus share only 43.5% sequence identity in the aligned overlap regions and intron 1 does not contain any ORF.

Recently, group I introns highly similar to, and in the same position as intron 2 in P. angulospiculatus and its homolog in Tetilla sp. were reported in cox1 of 20 scleractinian corals [24]. Phylogenetic analysis of amino-acid sequences derived from intronic LAGLIDADG ORFs in P. angulospiculatus, Tetilla sp., scleractinian corals, and several outgroup taxa grouped introns found in Tetilla sp. and P. angulospiculatus with 72% bootstrap support and placed them as a sister group to Scleractinian corals with 100% bootstrap support (Figure S2). The results of this analysis are consistent with the vertical
evolution of this intron in cnidarians and sponges and suggest that its sporadic presence among sampled taxa is due to independent losses rather than the horizontal intron transfer proposed earlier [23]. This inference is reinforced by the observations that the genetic distance between LAGLIDADG ORFs in *P. angulospiculatus* and *Tetilla* sp. is similar to that between their host genes and that both ORFs contain a TGA codon at the same position (data not shown). The latter finding makes it highly unlikely that the two introns have been transferred in parallel from the nucleus, because TGA signifies a stop codon in cytoplasmic translation.

**rRNA genes**

Genes for the small and large subunit ribosomal RNAs (*rns* and *rnl*) were located in close proximity of each other (separated by 1-3 tRNA genes) in most analyzed genomes, with the most common gene order being +*rns*+*trnG*+*trnV*+*rnl* (Figure 4). The two exceptions to this pattern were found in *Igerella notabilis*, where the two genes were separated by *atp9*, and *Oscarella carmela*, where *rnl* and *rns* were separated by multiple genes and had opposite transcriptional orientations. The size of *rns* ranged between 828 (Hippospongia lachne) and 1516 bp (Ephydatia muelleri), with the average size being 1224 bp. The size of *rnl* varied between 2166 (Hippospongia lachne) and 3487 bp (Axinella corrugata), with the average size being 2589 bp. The size differences in rRNA genes were due to two factors. First, some helices outside the core region of each rRNA were shortened or lost in several lineages, especially G1 (Figure 6). Second, unusual repetitive elements (see
below) were inserted in rRNA genes in several distantly related species, in particular *Axinella corrugata*, *Ephydatia muelleri*, *Igernella notabilis*, and *Vaceletia* sp. (Figure S3).

tRNA genes

Sampled demosponge mitochondrial genomes contained as few as 2 and as many as 27 tRNA genes. The variation in the number of tRNA genes was due to the loss of all but two mitochondrial tRNA genes (*trnM(cau)* and *trnW(uca)*) in G1, partial losses of tRNA genes in *Agelas schmidti* (at least one gene), *Amphimedon queenslandica* (at least 7 genes), and *Plakortis angulospiculatus* (at least 18 genes), the sporadic presence of *trnM(cau)e* among sampled species, and duplication of *trnT(ugu)* and *trnV(uac)* in *Oscarella carmela* mtDNA. Given that at least 24 species of tRNAs are needed for mitochondrial translation in demosponges [12], we expect that the loss of tRNA genes from mtDNA is compensated by the import of required tRNAs from the cytoplasm.

In accord with our previous study [12], tRNA genes in all studied demosponge mtDNA were well conserved in terms of size, primary sequence and inferred secondary structure. All inferred mt-tRNA structures had well conserved D- and T-loops (7-11 and 7 nucleotides in length, respectively) with a potential to form the standard tertiary interactions G18-U55 and G19-C56. Variable or semi-invariable nucleotide positions, and secondary and tertiary interactions known for prokaryotic and nuclear tRNAs were also well conserved (Figure 7). At the same time, an unusual A11-T24 pair in tRNA<sub>Pro<sup>UGG</sup></sub> and an unusual G11-C24 pair in tRNA<sub>Trp<sup>UCA</sup></sub> were present among all sampled demosponges. The first of them is
characteristic for demosponges, glass sponges, and placozoans [12,14], while the second – for all bilaterian animals [25]. The R11-Y24 pair is otherwise a distinctive feature of bacterial, archaeal, and organellar initiator tRNA$^{\text{Met}}_{\text{CAU}}$ that is strongly counter-selected in elongator tRNAs [26].

Among individual tRNA genes, trnW(uca) had the most conserved primary structure (84.9% pairwise sequence identity on average) while trnS(uga) was the least conserved (66.7% identity on average). The inferred gene for elongator tRNA(M) (trnM(cau)e) that is present in 11 out of 22 analyzed genomes also displayed high sequence conservation (average pairwise identity 79.6%), an observation that suggests its intermittent occurrence among sampled genomes is due to multiple losses rather than de novo evolution through gene duplication and/or recruitment [e.g., 27]. Interestingly, the other gene for methionine tRNA (trnM(cau)f) is more conserved among the species where trnM(cau)e is present, than among species were it is absent (78.1% vs. 67.8% pairwise identity on average).

Our previous analysis discovered several cases of tRNA gene recruitment in Axinella corrugata [27]. The more expanded dataset of demosponge mitochondrial tRNA genes assembled for this study revealed several additional instances of tRNA gene recruitment in demosponge mtDNA (to be described elsewhere).

**Intergenic regions and repeats**

The combined size of non-coding regions in the 22 demosponge mtDNAs analyzed in this study varied from 371 bp in Geodia neptuni to 6077 bp in Axinella corrugata or from 2
to 24% of the total genome size. In contrast to bilaterian animals, the distribution of non-coding nucleotides was more even in demosponge mtDNA, with the largest intergenic region usually containing <15%, and at most 39% (in *Iotrochota birotulata*), of all non-coding nucleotides. We found little conservation in the position of the largest intergenic regions among the sampled genomes, even for the species that share identical gene arrangements, such as *Chondrilla nucula* and *Haliscarca dujardini*, *Geodia neptuni* and *Cinachyrella kuekenthali*, and *Hippospongia lachne* and *Vaceletia* sp. Furthermore, we detected little sequence conservation either among individual regions within each mtDNA or between identically located non-coding regions in different species, except for the presence of repetitive elements in some genomes, as described below.

Multiple repetitive elements were found in several analyzed genomes. Repeats larger than 100 bp were found only in *Vaceletia* sp., with the two biggest repetitive elements (229 bp) located in the intergenic regions that flank *nad2*, while 20-100 bp repeats were discovered in multiple species. The most abundant repeats were found in *Vaceletia* sp., *Igernella notabilis*, *Ephydatia muelleri*, and *Axinella corrugata*, where they have been located in most intergenic regions, as well as in ribosomal RNA genes and even some protein coding genes. The presence of repeated elements was very sporadic in respect to phylogeny, with repeats often present/absent in closely related species. Overall, repeats were very rare in sampled species from G0, G2 and G3, but more common in G1 and G4.

**Discussion**
Our analysis of 22 complete mtDNA sequences representing all 14 orders of demosponges revealed both remarkable conservation and also an extensive diversity in mitochondrial genome organization within this group. Among the features shared among all sampled demosponge mitochondrial genomes are compact organization of the genetic material, similar gene content, well conserved structures of encoded tRNAs, a minimally modified genetic code for mitochondrial translation, and the absence of a single large “control” region characteristic of mtDNA in bilaterian animals. Genomic features that showed substantial variation include the number of tRNA genes, rRNA structures, the presence/absence of introns, and gene arrangements. In particular, two groups clearly stand out in our analysis with respect to their genome organization: G0 (order Homosclerophorida) and G1 (orders Dictyoceratida, Dendroceratida, and Verticillitida).

As reported previously, the mitochondrial genome of the homosclerophorid *Oscarella carmela* contains 44 genes – the largest complement of genes in animal mtDNA – including *tatC*, a gene for subunit C of the twin arginine translocase that has not been found in any other animal mtDNA, and genes for 27 tRNAs [14]. By contrast, the mtDNA sequence of the homosclerophorid *Plakortis angulospiculatus* determined for this study contains only 20 genes and lacks *tatC* as well as 19 of the 25 tRNA genes typical for demosponges. Other differences between these two genomes include distinct gene arrangements (only 4 shared gene boundaries) and the presence of two group I introns in *P. angulospiculatus cox1*. Furthermore, the estimated genetic distances between these two species are greater than those between many orders of demosponges, indicating an ancient radiation and the presence of extensive genetic diversity within this group.
Mitochondrial genomes of the three species within the G1 group are also unusual. These genomes lack all but two tRNA genes (for methionine and tryptophan tRNAs) – a feature previously associated with cnidarian mtDNA [28]. Furthermore, this is the only group of demosponges where a significant acceleration in the rates of mitochondrial sequence evolution has been detected. There appears to be no causal connection between these two observations, as the loss of all but two tRNA genes is shared by all three species in the group, while the accelerated sequence evolution is much more pronounced in Dictyoceratida and Verticillitida. The retention of \( \text{trnW(uca)} \) and \( \text{trnM(cau)} \) as the only tRNA genes in the genome supports our previous inference [29] that these genes are difficult to replace because of the unique role of their products in mitochondrial translation: \( \text{tRNA}_{\text{CAU}}^{\text{Met}} \) is used for the initiation of mitochondrial translation with formylmethionine [30] while \( \text{tRNA}_{\text{UCA}}^{\text{Trp}} \) must translate the TGA in addition to the TGG codons as tryptophan. The presence of such constraints can cause a parallel genomic evolution in independent lineages.

An unusual mitochondrial genome has been previously reported for the haplosclerid demosponge \textit{Amphimedon queenslandica} [15]. This genome lacks \textit{atp9} and at least seven tRNA genes, contains deletions in several protein coding genes, and displays accelerated rates of sequence evolution in both protein and RNA genes. Our analysis of three additional species from the same order, \textit{Amphimedon compressa}, \textit{Callyspongia plicifera}, and \textit{Xestospongia muta}, found no similar features in the latter taxa. These results most likely indicate that \textit{A. queenslandica} mitochondrial genome has undergone an unusual evolution and is a poor representative of the G3 group, although incorporation of nuclear sequences, such as nuclear Numts [31], in the mtDNA assembly cannot be ruled out. Given that \textit{A.
Queenslandica has become a model system for the study of demosponge biology, the evolution of its unusual mtDNA should be investigated in more details.

Another interesting result that came out of this study is the discovery of two group I introns in *cox1* of *P. angulospiculatus*. Several lines of evidence, including phylogenetic analysis, the identical location in *cox1*, a similar extent of genetic divergence to their host genes, and the presence of TGA codons at the same position, support the vertical evolution of one of these introns from the common ancestor shared not only with Tetilla sp. (order Spirophorida), but also with scleractinian corals. This in turn suggests that the absence of this intron in most demosponge lineages is due to massive parallel loss. While examples of such losses are well known in nuclear genomes [32-37], an interesting question posed by this result is why mitochondrial introns are retained so scarcely in demosponges but so commonly in cnidarians?

Finally, this study is interesting in what we did not find – any structures and/or sequences potentially involved in the maintenance and expression of mtDNA. Obviously, replication and transcription initiation/termination signals do exist in these genomes, but they were not detected by our comparative genomic analysis. Further data collection and experimental work will be essential to elucidate the mechanisms of these processes in demosponge mitochondria.

Methods

**Genome sequencing and phylogenetic analysis**

**Annotation and analysis of coding sequences**

We used flip (http://megasun.bch.umontreal.ca/ogmp/ogmpid.html) to predict ORFs in assembled sequences; similarity searches in local databases and in GenBank using FASTA [39] and NCBI BLAST network service [40], respectively, to identify them. Protein-coding genes were aligned with their homologues from other species and their 5’ and 3’ ends inspected for alternative start and stop codons. Inferred amino acid sequences of encoded proteins were aligned with ProbCons [41] using default parameters. Genetic distances between demosponges and four species of cnidarians (*Briareum asbestinum*, *Metridium senile*, *Montastraea annularis* and *Ricordia florida*) were calculated with the TREE-PUZZLE program [42], using the mtREV matrix, estimated frequencies of amino acids and 8
gamma rate categories. Effective numbers of codons [43] were calculated with the chips program within the EMBOSS package [44].

**Annotation and analysis of RNA genes**

Genes for small and large subunit ribosomal RNAs (*rns* and *rnl*, respectively) were identified based on their similarity to homologous genes in other species, and their 5’ and 3’ ends were predicted based on sequence and secondary structure conservation. The secondary structures of selected rRNA genes were manually folded by analogy to published rRNA structures, and drawn with the RnaViz 2 program [45].

Transfer RNA genes were identified by the tRNAscan-SE program [46] and aligned manually in MacGDE 2.3 [47] using their secondary structure as a guide. This alignment was used to calculate sequence conservation at each position and average pairwise identity values for individual tRNAs. For the latter calculation we excluded all tRNAs from *Plakortis angulospiculatus, Amphimedon queenslandica* and all species in G1, which encode incomplete sets of tRNAs in their mtDNA.

**Intronic sequences**

We used intron prediction programs RNAweasel [48] and Rfam [49] to search for introns in coding sequences. The exact positions of two introns found in *cox1* of *Plakortis angulospiculatus* were adjusted based on *cox1* alignments with homologous sequences from
other demosponges. The inferred amino acid sequence of the large ORF found in one of the
*P. angulospiculatus* introns was aligned with the sequences of LAGLIDADG ORFs analyzed
by Rot *et al.* [23] and Fukami *et al.* [24] and used for a phylogenetic analysis. We selected
the best model for these ORFs with the ProtTest program [50] and performed a maximum
likelihood search and bootstrap analysis in TREEFINDER [51], using the WAG model of
sequence evolution, estimated amino acid frequencies and 4 gamma categories.

**Intergenic regions and repeated sequences.**

Intergenic regions were extracted from each genome with the PEPPER program
([http://megasun.bch.umontreal.ca/ogmp/ogmpid.html](http://megasun.bch.umontreal.ca/ogmp/ogmpid.html)) and searched for similarity using
FASTA. In addition, we searched for interspersed identical repeats in individual genomes
using FINDREP ([http://megasun.bch.umontreal.ca/ogmp/ogmpid.html](http://megasun.bch.umontreal.ca/ogmp/ogmpid.html)) with minimum repeat
subsequence lengths of 20 bp and 100 bp respectively.

**Acknowledgements**

We thank Alexander Ereskovsky, Michelle Kelly, Franz Lang, Sally Leys, Shirley
Pomponi, Antonio Solé-Cava, Robert Thacker, and Gert Wörheide for samples contributed to
this project; Natalia Frishman, Megan Rasmuson, Brandi Sigmon and Zhiyong Shao, for
assistance in laboratory work; Karri Haen and Ehsan Kayal for their help with earlier
versions of this manuscript.
References


Figure 1. Phylogenetic analysis of demosponge relationships using mitochondrial genomic data. Posterior majority-rule consensus tree obtained from the analysis of 2,558 aligned amino acid positions under the CAT+F+Γ model is shown. Other methods of phylogenetic reconstruction produced similar topologies [11]. The numbers at each node are Bayesian posterior probabilities. Nodes with ≥ 95% support are marked with an asterisk. For simplicity, non-demosponge clades were collapsed to triangles. The full tree is presented in Figure S1.
Figure 2. The size (A) and gene content (B) of demosponge mtDNA. Demosponge species are subdivided into five major groups (G0-G5). Selected species from other animal groups and the outgroup Monosiga brevicollis are included for comparison. Species are abbreviated as following: mb, M. brevicollis; ta, Trichoplax adhaerens; ms, Metridium senile; hs, Homo sapiens; oc, Oscarella carmela; pa, Plakortis angulospiculatus; hl, Hippospongia lachne; in, Igernella notabilis; vs, Vaceletia sp.; af, Aplysina fistularis; cn, Chondrilla nucula; hd, Halisarca dujardini; ac, Amphimedon compressa; aq, Amphimedon queenslandica; cp, Callyspongia plicifera; xm, Xestospongia muta; ax, Agelas schmidti; ck, Cinachyrella kuekenthali; ef, Ectyoplasia ferox; em, Ephydatia muelleri; gn, Geodia neptuni; to, Topsentia ophiraphidites; ib, Iotrochota birotulata; pw, Ptilocaulis walpersi; ax, Axinella corrugata; te, Tethya actinia.
Figure 3. Nucleotide composition of mtDNA in five major groups of demosponges. (A) A+T content; (B) AT-skew; (C) GC-skew. The values are shown for the sense (non-template) strand of the whole genome (genome), its concatenated genetic components (protein genes, rRNA genes, and tRNA genes), 3rd codon positions in protein genes, and for the corresponding strand in intergenic regions. Colored bars indicate the mean value for each group of demosponges; error bars show standard deviation.
Figure 4. Mitochondrial gene arrangements in demosponges. Protein and rRNA genes (larger boxes) are: atp6, 8-9 – subunits 6, 8 and 9 of the F0 ATPase, coxl-3 – cytochrome c oxidase subunits 1-3, cob – apocytochrome b (cob), nadl-6 and nad4L – NADH dehydrogenase subunits 1-6 and 4L, mrs and rnl – small and large subunit rRNAs, tatC – twin-arginine translocase component C. tRNA genes (smaller boxes) are abbreviated using the one-letter amino acid code. The two arginine, isoleucine, leucine, and serine tRNA genes are differentiated by subscripts with trnR(ucg) marked as R1, trnR(ucu) – as R2, trnL(ugu) – as L1, trnL(uaa) – as L2, trnS(ucu) as S1, and trnS(uga) as S2. All genes are transcribed from left to right except those underlined to indicate an opposite transcriptional orientation. Genes are not drawn to scale and intergenic regions are not shown.
Figure 5. Relative rates of evolution of individual species (A) and individual genes (B). Rates are estimated by average genetic distances to orthologous genes from four cnidarians. Each boxplot represents data for 13 individual genes in (A) and 22 demosponge species in (B). Lower horizontal bar, non-outlier smallest observation; lower edge of rectangle, 25 percentile; central bar within rectangle, median; upper edge of rectangle, 75 percentile; upper horizontal bar, non-outlier largest observation; open circle, outlier.
Figure 6. Inferred secondary structure of *Hippospongia lachne* mitochondrial small subunit RNA in comparison to that of *Oscarella carmela*. The helices are numbered in boldface as in Brimacombe *et al.* [52]. Structural regions present in *O. carmela* srRNA but absent in *H. lachne* srRNA are shown in red.
Figure 7. Secondary structures and consensus sequences of demosponge mitochondrial tRNAs. The secondary structure of each type of tRNAs was folded based on sequence and structure alignment. Nucleotides in uppercase letters indicate >90% sequence conservation, lowercase letters indicate >75% sequence conservation, and the dots represent < 75% conservation.
CHAPTER 4. GENE RECRUITMENT – A COMMON MECHANISM IN THE
EVOLUTION OF TRANSFER RNA GENE FAMILIES

A paper submitted to Molecular Biology and Evolution

Xiujuan Wang and Dennis V. Lavrov

Abstract

The evolution of alloacceptor transfer RNAs (tRNAs) has been traditionally thought to occur vertically and reflect the evolution of the genetic code. Yet there have been several indications that a tRNA gene could evolve horizontally, from a copy of an alloacceptor tRNA gene in the same genome. Earlier, we provided the first unambiguous evidence for the occurrence of such “tRNA gene recruitment” in Nature – in the mitochondrial (mt) genome of the demosponge Axinella corrugata. Yet the extent and the pattern of this process in the evolution of tRNA gene families remained unclear. Here we analyzed tRNA genes from 21 mt genomes of demosponges as well as nuclear genomes of rhesus macaque, chimpanzee and human. We found four new cases of alloacceptor tRNA gene recruitment in mt genomes and eleven cases in the nuclear genomes. In most of these cases we observed a single nucleotide substitution at the middle position of the anticodon, which resulted in the change of not only the tRNA’s amino-acid identity but also the class of the amino-acyl tRNA synthethases (aaRS) involved in amino-acylation. We hypothesize that the switch to a different class of aaRS may have prevented the conflict between anticodon and amino-acid identities of recruited tRNAs. Overall our results suggest that gene recruitment is a common phenomenon in tRNA multigene family evolution and should be taken into consideration when tRNA evolutionary history is reconstructed.
Introduction

Transfer RNA (tRNA) is a small RNA molecule that plays a central role in protein biosynthesis. Each tRNA carries a specific amino acid to the ribosome and recognizes one or several specific codons in mRNA, functioning as a liaison between DNA encoded genetic information and its expression in proteins (Crick 1958). Based on their aminoacylation identity, tRNAs are subdivided into 20 amino acid accepting groups (alloacceptors), each of them comprising one to several tRNAs (isoacceptors) that translate synonymous codons (Saks et al. 1998). The number of tRNA genes varies from organism to organism. There are 86 tRNA genes in *Escherichia coli* DH10B, 286 in *Saccharomyces cerevisiae*, 298 in *Drosophila melanogaster*, 513 in human and 630 in *Arabidopsis thaliana* (Schattner et al. 2005). By contrast mitochondrial genomes do not typically encode multiple isoacceptor tRNAs and are able to translate all codons with as few as 22 tRNAs (Marck and Grosjean 2002). Although it is clear that tRNA gene families have diverged extensively among different organisms, the details of this evolution remain elusive.

The traditional view of tRNA evolution presumes that alloacceptor tRNAs coevolve with the genetic code while isoacceptor tRNA genes evolve by gene duplication from a common ancestor having the same amino-acid identity (Fitch and Upper 1987; Wong 1975; Xue et al. 2003). If this is indeed the case, then the modern phylogenetic relationships of alloacceptor tRNAs can be used to elucidate the evolution of the genetic code (Fitch and (1989, Eigen et al.;1975Wong ;1987Upper and/or to infer phylogenetic relationships among early diverging groups of organisms (2003, Xue et al.;1993Kumazawa and Nishida. However, several previous studies reported an unexpectedly high similarity among alloacceptor tRNA genes in some modern organisms and suggested that at least some tRNAs
could have evolved independently of the genetic code from duplicated genes for alloacceptor tRNAs (Cedergren et al. 1980; Burger et al. 1995). The latter hypothesis, named “gene recruitment”, received support from experimental studies in *Escherichia coli*, in which a tRNA$_{UCU}^{Arg}$ with a point mutation that changed its anticodon from UCU to UGU replaced the function of the tRNA$_{UGU}^{Thr}$ (Saks et al. 1998) and our recent analysis of alloacceptor tRNA genes in mitochondrial genomes of three demosponges (Lavrov and Lang 2005b). In the latter study, we provided several lines of evidence for the evolution of three tRNA genes in mtDNA of the demosponge *Axinella corrugata* from alloacceptor tRNA genes in the same genome. We also found a strong indication for the presence of gene recruitment in mitochondrial genomes of the green alga *Scenedesmus obliquus* and ichthyosporean *Amoebidium parasiticum*. Although these results have clearly shown that some tRNA genes in modern organisms evolved by gene recruitment, the importance of this process in tRNA evolution remains uncertain and is the focus of the present study.

In order to evaluate the importance of gene recruitment in tRNA evolution, we analyzed two different datasets. First, we expanded our mitochondrial genome dataset to 21 species of demosponges representing all currently recognized orders of demosponges except two (Dendroceratida and Dictyoceratida) that experienced significant tRNA gene loss (Wang and Lavrov 2008). Second, we analyzed all identified tRNA genes in the nuclear genomes of human, chimpanzee and rhesus macaque (Chan and Lowe 2009). Each of these datasets has its unique advantages. Demosponge mitochondrial genomes are a convenient system to study the tRNA evolution due to their relatively small size (~19 kb), conserved set of tRNA genes, conventional secondary structures of encoded tRNAs, and slow rate of sequence evolution (Wang and Lavrov 2008). In addition, the set of tRNA genes in demosponges does
not usually contain any duplicates; hence any observed gene recruitment events would likely be functional. By contrast, nuclear genomes contain multiple isoacceptor tRNA genes and, in some cases, multiple copies of each isoacceptor. Thus, at least in theory, they should be more prone to gene recruitment, although the functional significance of observed changes may remain uncertain.

**Methods**

**Construction of mitochondrial and nuclear tRNA gene datasets**

For the first part of this study, we selected 21 complete mitochondrial genome sequences of demosponges that did not undergo significant loss of tRNA genes (Wang and Lavrov 2008): *Agelas schmidti* (NC_010213), *Amphimedon compressa* (NC_010201), *Amphimedon queenslandica* (NC_008944), *Aplysina fulva* (NC_010203), *Axinella corrugata* (NC_006894), *Callyspongia plicifera* (NC_010206), *Ectyoplasia ferox* (NC_010210), *Ephydatia muelleri* (NC_010202), *Halisarca dujardini* (NC_010212), *Iotrochota birotulata* (NC_010207), *Negombata magnifica* (NC_010171), *Ptilocaulis walpersi* (NC_010209), *Topsentia ophiraphidites* (NC_010204), *Xestospongia muta* (NC_010211), *Oscarella carmela* (NC_009090), *Plakortis angulospiculatus* (NC_010217), *Chondrilla aff. nucula CHOND* (NC_010208), *Cinachyrella kuekenthali* (NC_010198), *Geodia neptuni* (NC_006990), *Suberites domuncula* (NC_010496), *Tethya actinia* (NC_006991). There were a total of 500 tRNA genes in these 21 demosponge mitochondrial genomes. For the second part, we chose to analyze nuclear genomes of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and rhesus monkey (*Macaca mulatta*) because they are closely related species with well annotated genomic information. The tRNA gene sequences were downloaded from
the Genomic tRNA Database (Chan and Lowe 2009). Pseudogenes and 100% identical copies were removed from the dataset so that the final nuclear dataset consisted of 901 nuclear tRNA genes.

**Sequence alignment and pairwise identity analysis**

The mitochondrial and nuclear tRNA gene sequences were manually aligned based on the inferred tRNA secondary structures. For each dataset, pairwise sequence similarities between alloacceptor tRNAs within the same species and between isoacceptor tRNAs among different species were calculated as the percentages of matched nucleotides in the alignments. The distributions of such identities were plotted for each group of tRNAs using the R package (http://www.r-project.org/). The probabilities of observing high pairwise sequence identities for candidate tRNA gene recruitment cases were calculated in R package based on normal distributions. Normal distribution assumptions were verified using Q-Q plots.

**Phylogenetic analysis of mitochondrial and nuclear tRNA genes**

Neighbor-joining analysis with 100 bootstrap replicates was conducted for each dataset based on uncorrected (‘p’) pairwise distances using the PAUP* 4.0b10 program (Swofford 2002). Additional phylogenetic analyses using other methods and/or models (eg. corrected distances) were also conducted but not shown due to problematic overparameterization for short sequences (Sullivan and Joyce 2005). Sequences of anticodons, variable-length portions of the extra loop, and introns (for nuclear tRNAs) were excluded from all phylogenetic analyses. Both mitochondrial and nuclear tRNA gene phylogenies were arbitrarily rooted using trnP as an outgroup.
Results

Four new cases of alloacceptor tRNA gene recruitment in demosponge mitochondrial genomes

Mitochondrial genomes of most demosponges contain a conserved set of 24-25 tRNA genes comprised of two isoacceptor genes for arginine, isoleucine, leucine, and serine, one or two genes for methionine, and a single gene for every other amino acid (Wang and Lavrov 2008). We analyzed tRNA genes from 21 mitochondrial genomes of demosponges representing all but two recognized orders in this group that lost all but two tRNA genes. The neighbor-joining tree generated using uncorrected p-distances showed that most of the equivalent tRNA genes (with the same amino-acid and anticodon identities) from different species form well-defined clades, an indication of their orthologous relationship. However, several genes appeared at unexpected positions on the tree – a potential indication of gene recruitment (Figure 1). In addition to the three recruitment cases in Axinella corrugata that have been reported earlier (Lavrov and Lang 2005b), four new putative cases have been found, including two cases of trnT(ugu) and trnR(ucu) (in Ptilocaulis walpersi and Ectyoplasia ferox), one case of trnR(ucg) and trnY(auga) in Negombata magnifica, and one case of trnI(gau) and trnN(guu) in Agelas schimdtii. All newly identified cases had moderately high bootstrap support for the unexpected groupings of tRNAs (51-78). Similar results were obtained in phylogenetic analyses using other methods and/or models of sequence evolution, although overparameterization (Sullivan and Joyce 2005) appeared to be a problem when more complex models of evolution were applied to such short sequences.
Although the results of a phylogenetic analysis provide an important indication of potential gene recruitments, an unexpected position of a gene on a phylogenetic tree may result from various additional factors (Brinkmann et al. 2005). Hence, we investigated these potential recruitment cases using two additional criteria identified in our previous study (Lavrov and Lang 2005b): 1) high sequence similarity with an alloacceptor gene in the same genome and 2) change in the position of the recruited gene in the genome.

A recent gene recruitment event should manifest itself through unexpectedly high sequence similarity between genes for alloacceptor tRNAs in the same genome and/or through unexpectedly low sequence similarity between genes for equivalent tRNAs among different genomes. In demosponges, the sequence similarities among equivalent tRNAs in different genomes are usually much higher than among alloacceptor tRNAs in the same genome and these measures form two largely non-overlapping normal distributions with means of 74.5% (standard deviation=10.1%) and 46.5% (standard deviation =7.5%), respectively (Figure 2A). However, our analysis revealed the presence of outliers in both of these distributions. These observations support the occurrence of alloacceptor tRNA gene recruitments in the demosponge mitochondrial genomes. We found that all tRNA genes identified as candidate genes for gene recruitment also had unexpectedly high similarities with genes for alloacceptor tRNAs in the same genomes (p<<0.0001, Table 1). Among the four new cases, three had a single point mutation in the middle position of the anticodon while one had all three anticodon nucleotides replaced (Table 1). However, some tRNA genes found at an unexpected position in our phylogenetic analysis (but with little bootstrap support) did not share high sequence similarity with any alloacceptor genes in the same
genome (not shown). These tRNAs may have evolved by a more ancient gene recruitment event or had some other unusual evolutionary patterns and are not discussed here.

Based on the tRNA duplication/remolding mechanism proposed in our previous study (Lavrov and Lang 2005b) we expected to find recruited tRNA genes in close proximity to their donor genes. However, none of the newly identified cases of alloacceptor tRNA gene recruitment showed such gene arrangement, although all recruited genes underwent gene rearrangements. This observation suggests that different mechanisms of gene duplication have been probably involved (see below).

**Multiple cases of alloacceptor tRNA gene recruitment in mammal nuclear genomes**

To evaluate whether the process of gene recruitment occurs outside the organellar genomes, we downloaded nuclear tRNA genes of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and rhesus macaque (*Macaca mulatta*) from the GtRNA database (Chan and Lowe 2009). After removing pseudogenes reported in the GtRNA database and identical tRNA genes, we aligned the remaining 901 tRNA genes manually based on secondary structures and constructed a neighbor-joining tree using uncorrected p-distances. In the resulting phylogeny, tRNA genes grouped mostly by the anticodon and amino acid identities of encoded tRNAs, except a few cases where they were found in unexpected positions supported by high bootstrap values (Figure 3). As nuclear genomes contain multiple tRNA isoacceptors, it is difficult to verify tRNA gene recruitment based on the patterns of gene rearrangement. However, one can still expect that recruited tRNA genes would have unexpectedly high sequence similarities with their donor genes.
Indeed, the sequence identity scores among equivalent tRNA genes in the three nuclear genomes and alloacceptor tRNA genes within each of these genomes formed two largely non-overlapping distributions (Figure 2B). The average sequence identity values for the two groups of genes were 76.5% and 46.6% with standard deviations of 16.6% and 8%, respectively. Deviation from normality was observed only for sequence identity scores among equivalent genes in different genomes, with the distribution of these scores showing a long tail of low sequence identities (Figure 2B). We further observed a number of outliers in alloacceptor identity distribution that had extremely high sequence identity values. The observed pattern supports the presence of alloacceptor gene recruitment in these genomes. Among the tRNA genes located at unexpected positions in our phylogenetic analyses, eleven had unexpectedly high sequence identities (78% to 95%) with alloacceptor tRNA genes in the same genome (p<<0.0001, Table 1) and we propose these tRNAs as putative cases of alloacceptor tRNA gene recruitment.

Alloacceptor gene recruitment appeared to occur in all three analyzed nuclear genomes as well as in the common ancestor of humans and chimps. Most (9 out of 11) of the nuclear tRNA gene recruitment cases had a single point mutation in the anticodon while the remaining two tRNA genes had two point mutations in the anticodon region. Furthermore, 7 out of the 9 single anticodon mutations occurred at the second position of the anticodon (Table 1). Several tRNA genes showed additional evidence of gene recruitment. Among the candidates for gene recruitment, trnV(uac) in chimps contains an intron as found in its inferred donor gene trnI(uau) but not in any other trnV(uac). In contrast, the putatively recruited trnY(gua) in human and chimpanzee did not have an intron as seen in the equivalent gene in rhesus macaque. The potentially most informative cases of gene recruitment were
observed for trnA(ggc) in rhesus macaque and trnR(gcg) in chimps as they were the only identified genes coding for tRNAs with these anticodons in each genome. This observation suggests that at least these two recruited genes are likely functional. While this analysis was conducted, the GtRNA database was updated and trnL(aag) and trnY(gua) of chimpanzee were removed from it. However, we kept these genes in our analysis because of their high similarities to the equivalent tRNAs in human. No other findings of candidate tRNA recruitment cases have been affected by the update.

The cases of isoacceptor tRNA recruitment in analyzed genomes

In addition to the potential cases of alloacceptor gene recruitment described above, we identified one candidate of isoacceptor tRNA gene recruitment in mitochondrial genomes of demosponges: trnL(uag) was recruited from trnL(uaa) in Agelas schimdti. The two isoacceptors in Agelas schimdti were grouped together in the mitochondrial tRNA phylogeny with 98% bootstrap support. Furthermore, they showed an extremely high sequence identity (90.3%) comparing to the average sequence identity of trnL(uaa) and trnL(uag) (55.6%±5.08%) in the remaining mitochondrial genomes (p=5e-12). Several additional examples of isoacceptor gene recruitment have been reported previously in other species (Rawlings et al. 2003; Higgs et al. 2003).

More putative cases of isoacceptor gene recruitment were detected in the nuclear genomes, and for some tRNA genes, these isoacceptor tRNA gene recruitments appear to occur repeatedly. For example, while the majority of trnV(uac)s formed a single clade in our phylogenetic analysis, the remaining few trnV(uac)s were located within the clusters of
trnV(aac) and trnV(cac) genes, suggesting that these genes were recruited from copies of their isoacceptors. Similarly, trnR(ccg) family in these nuclear genomes was subdivided into two sets: one grouped with trnR(ucg) and the other with trnR(ccu) with high bootstrap support (71 and 78 respectively). The sequence and phylogeny comparison indicates that both recruitment events from trnR(ucg) and trnR(ccu) occurred before the divergence of human, chimp and macaque.

Discussion

A special evolutionary pattern of threonine and arginine tRNA genes in demosponge mitochondrial genomes

Among the four newly identified candidate cases of tRNA gene recruitment in demosponge mitochondrial genomes, two involved trnT(ugu) and trnR(ucu). This is an unusual pair of tRNA genes in demosponge mtDNAs that showed an unexpectedly high sequence similarity in our previous studies despite their conserved position in most genomes (Lavrov and Lang 2005b; Wang and Lavrov 2008). These tRNAs are also known to be functionally convertible with a single nucleotide substitution in E. coli (Saks et al. 1998). Furthermore, in the present study we found that trnR(ucu) and trnT(ugu) were the only pair of alloacceptor tRNA genes that clustered together with greater than 50% bootstrap support (not shown). Even excluding the two pairs of tRNA genes implicated in gene recruitment, the average pairwise similarity between these two tRNAs across the remaining demosponges is abnormally high (79.5%±6.1%), very similar to the average pairwise similarities for each of these genes (81.7%±6.1% and 78%±6.4% for trnT(ugu) and trnR(ucu), respectively). In addition, the pattern of sequence conservation between trnT(ugu) and trnR(ucu) within each
species was highly unusual, with most of the differences concentrated in the acceptor stem. These differences are mostly attributed to changes in acceptor stems of trnR, while acceptor stem sequences of trnT are well conserved across different species (not shown). Although one may think that this pattern can be explained by gene recruitment of trnR(ucu) from a copy of trnT(ugu) before the divergence of poriferans, the presence of species-specific nucleotides in both of these tRNAs (Lavrov and Lang 2005b) indicates that the situation is more complex than this. Nevertheless, the finding of two recent independent gene recruitments in the T->R direction suggests a bias in favor of this particular event in demosponge mitochondria.

**Mechanisms of tRNA gene recruitment in mitochondrial and nuclear genomes**

In our previous study (Lavrov and Lang 2005b), we proposed a mechanism for tRNA gene recruitment in mitochondrial genomes that involves a tandem duplication of a tRNA gene followed by gene remodeling. This mechanism leaves a genomic signature in the mitochondrial genomes in a form of a specific gene rearrangement, where the remolded tRNA gene is located adjacent to its donor gene. Indeed, such close proximity between the two genes is found in all three cases of gene recruitment in Axinella corrugata mtDNA (Lavrov and Lang 2005a; Lavrov and Lang 2005b) and in the case of the trnL(uag) -> trnL(uaa) recruitment in Agelas schmidtii. However none of the four new cases of alloacceptor tRNA recruitment identified in this study in demosponge mitochondrial genomes is associated with similar gene rearrangement. Thus other types of gene duplication were probably involved in these tRNA gene recruitments. For example, a frequently invoked duplication-loss mechanism for gene rearrangements (Boore 2000) may provide duplicated
gene sequences required for gene recruitment, but would not necessarily result in the close proximity between the two copies of a tRNA gene. Yet another (currently unidentified) mechanism is likely responsible for single tRNA transpositions commonly found in animal mtDNA and may create duplicated tRNA genes spaced far apart in the genome (as observed, for example, in arthropods (Boore et al. 1998) or Oscarella carmela mtDNA (Wang and Lavrov 2007)). We also observed an interesting rearrangement pattern in the case of trnT(ugu) -> trnR(ucu) gene recruitment in Ptilocaulis wilhelmi and Ectyoplasia ferox, in which two tRNA genes “switched” positions (Figure 4). We propose the following mechanism for this observation: First a copy of trnT(ugu) is inserted between the genes cox1 and trnS(uga) in the common ancestor of the two species. Then the original trnT(ugu) was recruited as trnR(ucu) in Ptilocaulis wilhelmi, while the duplicated trnT(ugu) was recruited as trnR(ucu) in Ectyoplasia ferox. During the course of evolution, the original trnR(ucu) was lost in both genomes. Such a scenario provides an easy explanation for an unusual “switching” of the two genes in the genomes with otherwise identical gene arrangements.

Our examination of 11 putative cases of tRNA gene recruitments in the nuclear genomes of human, chimpanzee, and rhesus macaque revealed that a duplication-mutation-recruitment mechanism might also be involved. Based on the sequence comparison and chromosomal locations of tRNA genes implicated in gene recruitment, we found at least 4 alloacceptor recruitment events consistent with this mechanism. The donor and recruited tRNA genes were not only located close to each other on the same chromosome, but also had similar sequences in upstream and/or downstream regions (Table 2). These recruitments might be induced through local duplications followed by mutations. We also found a case of a duplication followed by inversion that caused the donor and recruited tRNA genes to have
opposite transcriptional orientation (Table 2). The multiple possibilities of duplications in nuclear genomes include segmental duplication, local duplication, retroposition and unequal crossover. In other candidate cases of gene recruitment, the two genes were present on different chromosomes. For example, we found that recruited gene $\text{trnV(uac)}$ and its inferred donor gene $\text{trnI(uau)}$ are located at chromosome 16 and 6 in the chimpanzee genome, respectively. The two genes share exactly the same sequence in the downstream region (tgtttcttc) and the recruited $\text{trnV(uac)}$ contains an intron like those found in $\text{trnI(uau)}$ but not in any other $\text{trnVs}$.

While gene duplication is an indispensable step for recruitment events in the mitochondrial genomes, it may not be an immediate prerequisite in the nuclear genomes due to the common presence of multiple isoacceptors and multiple copies of each isoacceptor in nuclear genomes. So one of the isoacceptors may be remolded into a tRNA with a different amino-acid identity. The recruitment of $\text{trnQ(uug)}$ from $\text{trnR(ucg)}$ before the divergence of humans and chimps is probably an example of such a process, as no donor $\text{trnR(ucg)}$ could be identified in either of these genomes. Yet $\text{trnQ(uug)}$ is extremely well conserved in the two species both in its sequences as well as the chromosomal location (chromosome 4) and is surrounded by identical upstream and downstream regions. This direct recruitment mechanism can also be visualized from a phylogenetic tree where the recruited tRNA genes are not grouped with the potential donor tRNA genes in the same genome but with the tRNA genes from other genomes (e.g., $\text{trnA(ggc)}$ in rhesus macaque and $\text{trnL(caa)}$ in human; Figure 3).

The emerging patterns in the identity of tRNA genes involved in gene recruitment
Most of tRNA gene recruitment cases identified in this study (3 for mitochondrial genes and 7 for nuclear genes) involved only a single nucleotide mutation in the middle position of the anticodon (position 35, Robertus et al. 1974). In vitro and in vivo studies have demonstrated that the anticodon and acceptor arm, which are in peripheral regions of folded tRNA, are often the most important elements in the tRNA recognition by aminoacyl-tRNA synthetases (aaRSs) (2008 Giege and that the nucleotide at the middle position of the anticodon is always important for recognition while the nucleotides at the first and third position are less often used as recognition elements (Vasil'eva and Moor 2007). In fact, the middle position of the anticodon is the main recognition element for the class Iα aaRSs (Fukai et al. 2003). Interestingly, for 9 out of 10 cases with a single nucleotide substitution at the middle of the anticodon, this change would result in a switch between the two classes of aaRSs responsible for tRNA amino-acylation (the only exception being trnR(acg) -> trnL(aag) in human and chimp). We speculate that such a switch may protect organisms from a mismatch between the anticodon and aminoacyl identities of recruited tRNAs. One major difference between the two classes lies in the mode of binding of the acceptor arm of tRNA. Class I aaRSs bind to the minor groove side of the acceptor stem, while class II aaRSs bind to the major groove side (Ibba et al. 2000). This distinction, in principle, can also explain the accumulation of differences between trnT(ugu) and trnR(ucu) in demosponge mitochondrial genomes primarily in the acceptor stems.

**Conclusion**

To evaluate the significance of gene recruitment in tRNA evolution, we analyzed the phylogenetic relationships among tRNA genes from 21 mitochondrial genomes of
demosponges and from the nuclear genomes of rhesus macaque, chimpanzee and human. We found that alloacceptor gene recruitment is relatively common in demosponge mitochondrial genomes, with strong evidence for this process found in 5 out of 21 analyzed species. Similarly, all three nuclear genomes showed evidence of alloacceptor tRNA gene recruitment with 11 cases showing a strong signal for this pattern of evolution. Rather surprisingly, isoacceptor gene recruitments do not appear to happen more frequently than alloacceptor gene recruitments, at least in mtDNA: only one such case has been found in Agelas schmidtii (trnL(uua), trnL(uag)). This may be due to the presence of only one isoacceptor tRNA gene for most codon families in mitochondrial genomes.

In most cases of gene recruitment, the new tRNAs differed from its donor by only one nucleotide at the middle position of anticodon sequence. Such a change almost always leads to a switch between the two classes of aaRSs responsible for tRNA amino-acylation that may protect organisms from a mismatch between the anticodon and aminoacyl identities of recruited tRNAs. The finding that tRNA gene recruitment is common both in organellar and nuclear genomes suggests that horizontal evolution is an important player in evolution of tRNA gene families and should be taken into account when tRNA sequences are used in evolutionary studies.

Acknowledgements

We thank Karri M. Haen, Eshan Kayal and Walker Pett for valuable comments on an earlier version of this manuscript, and Pioneer Graduate Fellowship and the Assembling Tree of Life project for funding.
Literature Cited


Table 1. Alloacceptor tRNA gene recruitment cases in the mitochondrial and nuclear genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Donor tRNA</th>
<th>Recruited tRNA</th>
<th>Species</th>
<th>Sequence similarity (%)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitochondrion</td>
<td>trnT(uug)</td>
<td>trnR(ucu)</td>
<td><em>Ptilocaulis walpersi</em></td>
<td>94.7</td>
<td>8.00E-11</td>
</tr>
<tr>
<td></td>
<td>trnT(uug)</td>
<td>trnR(ucu)</td>
<td><em>Ectyoplasia ferox</em></td>
<td>83.8</td>
<td>3.76E-07</td>
</tr>
<tr>
<td></td>
<td>trnR(uug)</td>
<td>trnY(aua)</td>
<td><em>Negombata magnifica</em></td>
<td>75.7</td>
<td>5.40E-05</td>
</tr>
<tr>
<td></td>
<td>trnL(gau)</td>
<td>trnN(guu)</td>
<td><em>Agelas schimdti</em></td>
<td>73.6</td>
<td>0.00016</td>
</tr>
<tr>
<td></td>
<td>trnK(uuu)</td>
<td>trnN(guu)</td>
<td><em>Axinella corrugata</em></td>
<td>88.9</td>
<td>9.27E-09</td>
</tr>
<tr>
<td></td>
<td>trnT(uug)</td>
<td>trnA(ugc)</td>
<td><em>Axinella corrugata</em></td>
<td>91.8</td>
<td>9.25E-10</td>
</tr>
<tr>
<td></td>
<td>trnK(uuu)</td>
<td>trnC(gca)</td>
<td><em>Axinella corrugata</em></td>
<td>73.6</td>
<td>0.00016</td>
</tr>
<tr>
<td></td>
<td>trnT(agu)</td>
<td>trnA(ggc)</td>
<td><em>Macaca mulatta</em></td>
<td>78.9</td>
<td>2.52E-05</td>
</tr>
<tr>
<td></td>
<td>trnR(acg)</td>
<td>trnL(aag)</td>
<td><em>Homo sapiens, Pan</em></td>
<td>82.4</td>
<td>4.99E-06</td>
</tr>
<tr>
<td></td>
<td>trnR(ucg)</td>
<td>trnQ(uug)</td>
<td><em>Homo sapiens, Pan</em></td>
<td>78.9</td>
<td>4.28E-05</td>
</tr>
<tr>
<td></td>
<td>trnM(cau)</td>
<td>trnT(cgu)</td>
<td><em>Homo sapiens</em></td>
<td>84.9</td>
<td>7.69E-07</td>
</tr>
<tr>
<td></td>
<td>trnL(uau)</td>
<td>trnV(uac)</td>
<td><em>Pan troglodytes</em></td>
<td>91.5</td>
<td>8.79E-09</td>
</tr>
<tr>
<td></td>
<td>trnL(aau)</td>
<td>trnS(acu)</td>
<td><em>Pan troglodytes</em></td>
<td>92</td>
<td>6.10E-09</td>
</tr>
<tr>
<td></td>
<td>trnM(cau)</td>
<td>trnL(caa)</td>
<td><em>Homo sapiens</em></td>
<td>81.8</td>
<td>4.99E-06</td>
</tr>
<tr>
<td></td>
<td>trnH(gug)</td>
<td>trnR(gcg)</td>
<td><em>Pan troglodytes</em></td>
<td>94.5</td>
<td>9.24E-10</td>
</tr>
<tr>
<td></td>
<td>trnG(ccc)</td>
<td>trnV(cac)</td>
<td><em>Homo sapiens</em></td>
<td>93</td>
<td>2.90E-09</td>
</tr>
<tr>
<td></td>
<td>trnC(gca)</td>
<td>trnY(gua)</td>
<td><em>Homo sapiens, Pan</em></td>
<td>85.9</td>
<td>4.07E-07</td>
</tr>
<tr>
<td>Nuclear</td>
<td>trnA(ugc)</td>
<td>trnE(cuc)</td>
<td><em>Macaca mulatta</em></td>
<td>94.5</td>
<td>9.24E-10</td>
</tr>
</tbody>
</table>

*a*: The three recruitment cases in *Axinella corrugata* were reported earlier.

*b*: The sequence similarities were the averages of similarities in the two genomes.
Table 2. Physical and adjacent sequence information of donar and recruited tRNA genes

<table>
<thead>
<tr>
<th>Species</th>
<th>Donor or recruited tRNA</th>
<th>Chromosome</th>
<th>Sequence begin</th>
<th>Sequence end</th>
<th>Upstream region</th>
<th>Downstream region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pan troglodytes</td>
<td>trnS(acu)</td>
<td>chr6</td>
<td>27795391</td>
<td>27795464</td>
<td>atttgtagtc</td>
<td>ttgacttttc</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>trnI(aau)</td>
<td>chr6</td>
<td>27220462</td>
<td>27220389</td>
<td>tccccagcgg</td>
<td>cgggcttttc</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>trnR(gcg)</td>
<td>chr15</td>
<td>42375257</td>
<td>42375328</td>
<td>ccggtagctc</td>
<td>tttggaac</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>trnH(gug)</td>
<td>chr15</td>
<td>42372801</td>
<td>42372730</td>
<td>ccggtagctc</td>
<td>tttggaac</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>trnV(cac)</td>
<td>chr1</td>
<td>147561360</td>
<td>147561290</td>
<td>ctggggagag</td>
<td>agaggttctt</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>trnG(ccc)</td>
<td>chr1</td>
<td>16745091</td>
<td>16745021</td>
<td>ttacgcagag</td>
<td>agaggttctt</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>trnG(ccc)</td>
<td>chr1</td>
<td>17061003</td>
<td>17061073</td>
<td>ttacgcagag</td>
<td>aaaggttctt</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>trnE(cuc)</td>
<td>chr4</td>
<td>28530898</td>
<td>28530827</td>
<td>nnnnnnnnnn</td>
<td>agtgctggtt</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>trnA(ugc)</td>
<td>chr4</td>
<td>28545136</td>
<td>28545065</td>
<td>ctttaagcaa</td>
<td>agtgatggtt</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>trnQ(uug)</td>
<td>chr4</td>
<td>41284523</td>
<td>41284451</td>
<td>ggagagagga</td>
<td>aatgtttccc</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>trnQ(uug)</td>
<td>chr4</td>
<td>40603572</td>
<td>40603500</td>
<td>ggagagagga</td>
<td>aatgtttccc</td>
</tr>
</tbody>
</table>
Figure 1. Neighbor-joining analysis based on uncorrected ("p") distances among mitochondrial tRNA genes from 21 species of demosponges. The parts of the tree indicating potential tRNA gene recruitment events are zoomed in with recruited genes highlighted in bold. The numbers above branches indicate percentage of bootstrap support based on 1000 replicates (if >50%). The tips are labeled as “species tRNA anticodon” with the species names abbreviated as following: Ag: Agelas schmidtii, Ac: Amphimedon compressa, Aq: Amphimedon queenslandica, Ap: Aplysina fulva, Ax: Axinella corrugata, Ca: Callyspongia plicifera, Ec: Ectyoplasia ferox, Ef: Ephydatia muelleri, Hd: Halisarca dujardini, Io: Iotrochota birotulata, Ne: Negombata magnifica, Pa: Plakortis angulospiculatus, Pt: Ptilocaulis walpersi, To: Topsentia ophiraphidites, Xm: Xestospongia muta, Os: Oscarella carmela, Ch: Chondrilla aff. nucula, Ci: Cinachyrella kuekenthali, Ge: Geodia neptuni, Su: Suberites domuncula, Te: Tethya actinia.
Figure 2. Distributions of tRNA sequence identity scores in (A) mitochondrial and (B) nuclear genomes. Lower horizontal bar: non-outlier lowest similarity; lower edge of rectangle: 25 percentile; central bar within rectangle: median; upper edge of rectangle: 75 percentile; upper horizontal bar: non-outlier highest similarity; open circle, outlier.
Figure 3. Neighbor-joining analysis based on uncorrected (“p”) distances among nuclear tRNA genes from human, chimpanzee and rhesus macaque. The parts of the tree indicating potential tRNA gene recruitment events are zoomed in with recruited genes highlighted in bold. The numbers above branches indicate percentage of bootstrap support based on 1000 replicates (if >50%). The tips are labeled as “species tRNA anticodon numbers” with the species name abbreviated as following: Ch: Pan troglodytes, Mc: Macaca mulatta, Hm: Homo sapiens. The numbers were assigned to distinguish copies of isoacceptor tRNAs. The two tRNAs marked with asterisks were removed from the Genomic tRNA Database during the course of this study (Chan and Lowe 2009).
Figure 4. Gene recruitment mechanism for trnR(ucu) and trnT(ugu) pairs in Ptilocaulis wilhelmi and Ectyoplasia ferox. TrnT(ugu) first duplicated and inserted between the genes cox1 and trnS(uga). The original trnT(ugu) was recruited to trnR(ucu) in Ptilocaulis wilhelmi while the duplicated trnT(ugu) was recruited to trnR(ucu) in Ectyoplasia ferox. The original trnR(ucu)s were lost in both genomes.
CHAPTER 5. COMPARATIVE STUDY OF HUMAN MITOCHONDRIAL PROTEOME REVEALS EXTENSIVE PROTEIN SUBCELLULAR RELOCALIZATION AFTER GENE DUPLICATIONS

A paper published in BMC Evolutionary Biology

Xiujuan Wang, Yong Huang, Dennis V. Lavrov, Xun Gu

Abstract

Background

Gene and genome duplication is the principle creative force in evolution. Recently, protein subcellular relocalization, or neolocalization was proposed as one of the mechanisms responsible for the retention of duplicated genes. This hypothesis received support from the analysis of yeast genomes, but has not been tested thoroughly on animal genomes. In order to evaluate the importance of subcellular relocalizations for retention of duplicated genes in animal genomes, we systematically analyzed nuclear encoded mitochondrial proteins in the human genome by reconstructing phylogenies of mitochondrial multigene families.

Results

The 456 human mitochondrial proteins selected for this study were clustered into 305 gene families including 92 multigene families. Among the multigene families, 59 (64%) consisted of both mitochondrial and cytosolic (non-mitochondrial) proteins (mt-cy families) while the remaining 33 (36%) were composed of mitochondrial proteins (mt-mt families).
Phylogenetic analyses of mt-cy families revealed three different scenarios of their neolocalization following gene duplication: 1) relocalization from mitochondria to cytosol, 2) from cytosol to mitochondria and 3) multiple subcellular relocalizations. The neolocalizations were most commonly enabled by the gain or loss of N-terminal mitochondrial targeting signals. The majority of detected subcellular relocalization events occurred early in animal evolution, preceding the evolution of tetrapods. Mt-mt protein families showed a somewhat different pattern, where gene duplication occurred more evenly in time. However, for both types of protein families, most duplication events appear to roughly coincide with two rounds of genome duplications early in vertebrate evolution. Finally, we evaluated the effects of inaccurate and incomplete annotation of mitochondrial proteins and found that our conclusion of the importance of subcellular relocalization after gene duplication on the genomic scale was robust to potential gene misannotation.

**Conclusion**

Our results suggest that protein subcellular relocalization is an important mechanism for the retention and gain of function of duplicated genes in animal genome evolution.

**Background**

Gene duplication is an important evolutionary process that plays a key role in generating new genomic information in all the three domains of life: Eubacteria, Archaea and Eukarya [1-5]. Various processes can cause gene duplication on the molecular level, including unequal crossovers, retroposition, or whole chromosome/genome duplication [6]. New functional
genes resulting from gene duplication are retained in the genome through the processes of subfunctionalization and neofunctionalization [9, 10]. Subfunctionalization refers to a situation when each of the daughter genes adopts only a partial function of the parental gene, while neofunctionalization refers to the gain of new functions by the duplicate, usually related to their ancestor’s function [4].

From the evolutionary perspective, neofunctionalization presents more interest than subfunctionalization because it results in the increase of the total genetic information [11]. Several mechanisms have been invoked to explain the gain of novel gene function for duplicated genes such as dosage compensation, epigenetic complementation, moonlighting, and catalytic promiscuity [12-17. Interestingly, functional divergence can in some cases precede (and facilitate) gene duplication through allelic divergence [18]. Recently, protein subcellular relocalization or neolocalization has been proposed as a key event for generating new functional genes after duplication [19, 20]. Such neolocalization can be achieved by the gain or loss of N-terminal targeting peptide sequences that can direct the products of duplicated genes from the cytosol to mitochondria, endoplasmic reticulum and chloroplast or vice versa.

The idea of subcellular relocalization underlying the gain of function for duplicated genes has been tested in two yeast genomes by Marques et al. (2008). The authors demonstrated that about one-third of the duplicated genes retained in the yeast genomes had undergone protein subcellular relocalization following whole genome duplication [20]. A few anecdotal observations suggest that neolocalization after duplication also occurs in animal genomes [21,
22], however the magnitude of this process has not been explored. In this study, we performed a systematic survey of subcellular relocalization following gene duplication in the human genome by analyzing nuclear-encoded mitochondrial protein families.

Mitochondria, cell organelles present in nearly all eukaryotes, are instrumental for the production of ATP through oxidative phosphorylation process, and are also involved in heme biosynthesis, cell metabolism, apoptosis, and Fe/S cluster biosynthesis. The complex functions of mitochondria demand a proteome composed of over a thousand of proteins, more than 98% of which are nuclear encoded, which suggests these organelles should play a major role in the process of neolocalization [23]. Hence, the exploration of nuclear encoded mitochondrial gene families is an ideal system to test subcellular relocalization of duplicated genes in the evolution of animal genomes.

Results and Discussion

Subcellular relocalization as a mechanism underlying protein functional divergence

For this study we retrieved 456 human mitochondrial proteins from MitoP2 database (http://www.mitop.de:8080/mitop2/) that were also annotated as mitochondrial proteins in Swissprot database. Reciprocal blasting and single linkage clustering were carried out to group the proteins into 305 families, among which 195 were single gene families (not considered here) and 110 were multigene families. After removing 18 families with members that appeared to be alternative splicing products or annotation artifacts, we obtained the final dataset of 92 multigene families for further analysis (Table 1). These 92 families can be classified into two categories based on their designated subcellular localizations: 1)
mitochondrial-cytosolic (mt-cy) families that consist of at least one protein member localized in mitochondria and at least one in another (non-mitochondrial) cellular compartment, and 2) mitochondrial only (mt-mt) families that are composed of protein members localized exclusively in mitochondria. The mt-cy category contained 59 families with 144 mitochondrial proteins and 196 non-mitochondrial proteins in humans, while the mt-mt category had 33 families with 79 human mitochondrial proteins. This result suggests that around two thirds of the mitochondrial multigene families have undergone subcellular relocation after duplication.

For each human protein in the 92 gene families, we retrieved orthologs in mouse (*Mus musculus*), chicken (*Gallus gallus*), fish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), mosquito (*Anopheles gambiae*) and nematode (*Caenorhabditis elegans*) from the Homologene database at NCBI. Phylogenetic analyses were conducted for each family and the time of occurrence of gene duplications in relationship to major divergences in animal evolution was evaluated. Because this study is based on the human mitochondrial proteome, only branches of the phylogenetic tree leading to humans were investigated. For mitochondrial-cytosolic (mt-cy) families that underwent several rounds of duplications, we only considered the duplications that were followed by subcellular relocations.

Among the 59 mt-cy families, twenty-seven (45.8%) were inferred to undergo gene duplication prior to the protostome/deuterostome divergence, twenty-nine (49.2%) after the protostome/deuterostome divergence but before that of fish/tetrapods and only three families (5%) within the tetrapod lineage (Table 1). Among the 33 mt-mt families, eight (24%)
underwent gene duplication prior to the protostome/deuterostome split, nineteen (57.6%) prior to the fish/tetrapod split and six (18.4%) within the tetrapod lineage. The observation that the majority of investigated families experienced gene duplication between the protostome/deuterostome and fish/tetrapod divergences correlates well with the two rounds of genome duplication at the early stage of vertebrate evolution \[24,25\]. However, the scarcity of more recent subcellular relocalization events is surprising, especially considering very high rates of gene birth and death in animal genomes [26].

**Bidirectional relocalization of proteins encoded by duplicated genes in mitochondrial-cytosolic (mt-cy) two-gene families**

In order to get insight into the direction of protein subcellular relocalization, we explored mt-cy gene families in which human genes are represented by two copies, one functioning in the cytosol and the other in mitochondria. Among the 24 such families, one third appeared to have its original function in mitochondria with the products of duplicated genes being relocalized to other cellular compartments, another third showed the opposite direction of protein relocalization and for the rest the direction of relocalization could not be determined due to the lack of outgroup information.

The arginase family, encoding enzymes that catalyze the hydrolysis of arginine to ornithine and urea, is an example of neolocalization to the cytosol. Phylogenetic analysis of this family shows that the product of the ancestral gene had an original localization in mitochondria. Following a gene duplication in the lineage leading to vertebrates, type I arginase (ARG1) has relocalized to the cytoplasm while type II arginase (ARG2) retained its ancestral
mitochondrial location (Figure 1). The evolutionary rates remain similar after the divergence of ARG1 and ARG2 genes. Sequence comparisons indicated that N-terminal mitochondrial targeting signals were not found in ARG1 in either human or mouse, but are present in all ARG2 sequences. The loss of N-terminal mitochondrial signals suggests that ARG1 could not be transported into mitochondria and is retained in the cytoplasm.

The type IB subfamily of DNA topoisomerases that includes DNA topoisomerase 1 (TOP1) and mitochondrial DNA topoisomerase 1 (TOP1MT) presents an example of neolocalization to mitochondria. DNA topoisomerases control DNA topological states by catalyzing the transient breaking and rejoining of single strand DNA, allowing DNA strands or double helices to pass through each other [27]. These enzymes are essential in maintaining DNA topology during replication, transcription, recombination and DNA repair. Our phylogenetic analysis of the type IB subfamily revealed a gene duplication that occurred early in vertebrate evolution. Following this duplication, the product of one copy of the gene, TOP1MT, relocalized to mitochondria while the product of another – TOP1, retained its ancestral cytoplasmic and nuclear locations (Figure 2).

Both TOP1 and TOP1MT consist of four domains: a N-terminal localization domain, a core domain, a linker domain and a C-terminal domain [28]. Sequence comparisons show that the N-terminal domain of TOP1MT consists of a mitochondrial targeting signal, while the N-terminus of TOP1 contains a nuclear localization signal. This implies that the change in the N-terminal targeting sequence of TOP1MT helped the protein direct itself to mitochondria and eventually to acquire a new mitochondrial function. It should be noted that the lack of
TOP1MT in invertebrates does not mean that type I topoisomerases are not needed for mitochondrial replication and transcription in this group. Recent studies have shown that DNA topoisomerase IIIα from the type IA subfamily has mitochondrial localization in *Drosophila melanogaster* [29, 30].

Does subcellular relocalization following duplication influence protein evolutionary rates? To answer this question we analyzed the relative evolutionary rates of duplicated genes in the mt-cy two gene families with well-resolved phylogenies and outgroup data. Three hypotheses were investigated: 1) Mitochondrial proteins generally have higher evolutionary rates comparing to their cytosolic counterparts; 2) The proteins involved in neolocalization have higher evolutionary rates; 3) Proteins undergo faster evolution following duplication/neolocalization due to functional relaxation or positive selection, with the evolutionary rates decreasing over time. To test the first two hypotheses, we compared the average branch lengths leading to mitochondrial and nuclear paralogs following a gene duplication [(a+a’)/(b+b’)] in Figure 3A]. For the third hypothesis, we compared branch length ratio of mitochondrial vs. cytosolic proteins before and after the divergence between tetrapods and fish (a/b and a’/b’ in Figure 3A). None of the proposed hypotheses was supported by our data (Figure 3B). Although some families had clearly uneven rates of evolution in mitochondrial vs. cytosolic proteins (e.g., Figure 2), most of the families displayed overall branch length ratios close to 1 regardless the direction of relocalization (Figure 3B). Similarly in some families mitochondrial proteins had higher evolutionary rates earlier in evolution (a/b>1) but lower rates at later stages (a’/b’<1) while in others an opposite pattern was observed (Figure 3B).
Multiple subcellular relocalizations after gene duplications

In addition to the two-gene families discussed above, there are 35 mt-cy families with three or more (8 on average) members. Based on the current cellular component annotation in Swissprot database, we inferred that at least one third of these families had undergone multiple subcellular relocalizations. Class I sirtuin family presents a relatively simple example. In humans this family consists of SIRT1, SIRT2 and SIRT3 that regulate transcriptional repression [31]. SIRT1, located in the nucleus, is a deacetylase that regulates the tumor suppressor p53, NF-κB signaling, and FOXO transcription factors. SIRT2 is a cytoplasmic protein that deacetylates Lys40 of α-tubulin. Finally, SIRT3 is localized to the mitochondrial matrix [32, 33]. The phylogeny of class I sirtuins suggests that the first round of duplication generated two copies with one copy (SIRT1) localized to the nucleus, while the other copy duplicated again resulting in one cytoplasmic copy (SIRT2) and one mitochondrial copy (SIRT3) (Figure 4). The N-terminal sequence analysis indicated the mitochondrial targeting signal was present in SIRT3 in all vertebrates on the tree except Mus musculus and Rattus norvegicus. The loss of the N-terminal mitochondrial targeting signal of SIRT3 in rodents suggests the loss of relocalization to mitochondria, an inference supported by an experimental demonstration that mouse SIRT3 actually locates to cytoplasm [34].

Expansion of mitochondrial proteome by gene duplications

The presence of 33 mt-mt families among the 92 multigene families supports the notion that gene duplication also contributes to mitochondrial proteome expansion [35], although the average family size of mt-mt families is smaller than that of mt-cy families (2.4 vs 5.7). Our
phylogenetic analyses of these mt-mt families showed that such duplications occurred at different stages in animal evolution, predating the divergence of the protostome/deuterostome lineages, within the vertebrate lineage, and within the mammalian lineage. In general, these families consist of proteins with similar functions that have been retained by subfunctionalization as different subunits or isoforms. Furthermore, the expression of these genes often shows tissue specificity such that one copy in the gene family is expressed ubiquitously, while the other(s) is/are expressed in specific tissues. For example, human SCO2 homolog (SCO cytochrome oxidase deficient homolog 2 (yeast)) is expressed ubiquitously while the SCO1 homolog is predominantly expressed in muscle, heart, and brain, the tissues featured by high rates of oxidative phosphorylation [36].

**Evolutionary modifications of relocalized proteins at sequence level**

Proteins synthesized in the cytosol can be directed to organelles such as mitochondria via mitochondrial targeting sequences [37]. While targeting signals in protein sequences can be located at the C-terminus and in internal regions, they are most commonly found at the N-terminus [38]. Hence, we expected that a large fraction of mitochondrial proteins in the analyzed mt-cy multigene families would have a mitochondrial N-terminal targeting sequence comparing to their cytosolic counterparts. We used targetP prediction [38] to analyze the presence of N-terminal mitochondrial targeting sequences in human proteins within the mt-cy families (Figure 5A). While over 80% of non-mitochondrial proteins lack a recognizable mitochondrial N-targeting signal, over 50% of mitochondrial proteins have the signal (chi-square p-value is 3.06e-15). This result meets our expectation that the gain or loss of mitochondrial N-terminal sequences plays an important role in directing protein
subcellular relocalization after duplication. At the same time, the existence of mitochondrial targeting signals in approximately 20% of non-mitochondrial proteins and our inability to find such a signal in 50% mitochondrial proteins indicates that the N-terminal sequence is not the only modification. Similar sequence modifications might have taken place at C-terminal or internal protein regions that are difficult to identify.

We further investigated the differences in protein functional domains between mitochondrial and non-mitochondrial proteins based on a suggestion that protein function and/or protein functional efficiency can be modified upon the change in its subcellular location [39]. Here we compared the distribution of mitochondrial Pfam domains that were previously found only in eukaryotic (excluding human) mitochondrial proteins [40] among members of mt-cy protein families. We found a significant difference in this distribution (figure 5B): 53% of mitochondrial proteins have mitochondrial domains, but only 16% of non-mitochondrial proteins have such domains (chi-square p-value is 3.5e-9). This result indicates that subcellular relocalizations were characterized by the formation of mitochondrial protein domains or their loss in nuclear copies during evolution.

Effects of inaccurate or incomplete cellular component annotation
Knowing accurate protein subcellular localization is important for this study. Although we combined information from several databases to infer protein functional locations, uncertainty still exists in our assignments. To check how these uncertainties would affect our results and conclusions, we reanalyzed all human proteins in this study by applying maestro scores, a scoring system for predicting nuclear encoded mitochondrial proteins in human and
mouse that incorporates eight genomic-scale data sets including targeting sequence prediction, protein domain enrichment, presence of cis-regulatory motifs, yeast homology, ancestry, tandem-mass spectrometry, coexpression and transcriptional induction during mitochondrial biogenesis [40]. The maestro score distributions of mitochondrial and non-mitochondrial proteins in our dataset are separated with some overlaps (Figure 6). By applying the suggested score cutoff (5.65) to assign subcellular locations to the analyzed proteins [40], we observed 11 out of 59 mt-cy families were grouped into mt-mt category; yet 5 out of 33 mt-mt families were classified to mt-cy families. Even if we do not count the 11 potentially ambiguous families into mt-cy category, still more than 50% of all mitochondrial multigene families have undergone subcellular relocalization after gene duplication.

Limited knowledge of protein subcellular location also prevents us from discovering cases of a different form of subcellular relocalization, called sublocalization, in which duplicated genes become targeted to a subset of their ancestral cellular compartments [21]. For example, the glutamate dehydrogenase family was grouped into mt-mt families since both GLUD1 and GLUD2 are located to mitochondria based on the Genbank and Swissprot annotations. However Rosso et al. recently reported that GLUD1 located to both cytoplasm and mitochondria while GLUD2 became specifically localized to mitochondria owing to a single positively selected amino acid substitution at the N-terminal targeting sequence [21]. In addition, if the ancestral protein had dual localizations but only one of them was annotated, then neolocalization would be inferred instead of sublocalization. The latter problem should be especially pronounced if non-model species are used as outgroups since the localizations of proteins in these species are not thoroughly studied. To investigate the potential effect of
this bias, we searched protein localizations in Drosophila melanogaster for the families duplicated in the vertebrate lineage. None of these Drosophila proteins are annotated to have dual localizations. Thus available data suggests that neolocalization rather than sublocalization is the prevalent mode of evolution of duplicated genes studied here.

**Conclusions**

Protein subcellular relocalization was proposed as an evolutionary mechanism for generating new functional genes after gene duplication [19]. This mechanism was studied in yeast genomes but only received support from individual cases/families in animal genomes [20-22]. Here we systematically investigated human mitochondrial protein families and found that around two thirds of multigene families have protein members that underwent subcellular relocalization after gene duplication. These subcellular relocalizations can occur between mitochondria and another subcellular compartment as well as among several compartments. Comparative sequence analyses showed that the subcellular relocalization processes were primarily enabled via the gain or loss of N-terminal mitochondrial targeting sequences.

After evaluating possible effects of incomplete or incorrect annotations, we conclude that our observation of the subcellular relocalization after gene duplication on the genomic scale was robust to misannotations. Surprisingly our results indicate a scarcity of recent subcellular relocalization events and suggest that protein subcellular relocalization was more important in obtaining new functional genes at the early stages of animal genome evolution. The observation that subcellular relocalization rarely follows recent gene duplications needs to be further investigated because our dataset might be biased in terms of annotation availability.
This investigation, together with the previously published finding in yeast [20], indicates that subcellular relocalization is an important mechanism in the retention and gain of function of duplicated genes over the course of eukaryotic genome evolution.

**Methods**

**Retrieving human mitochondrial proteins and identifying the paralogs**

We used the MitoP2 database (http://www.mitop.de:8080/mitop2/) to retrieve human mitochondrial proteins [41]. MitoP2 is a manually annotated database of mitochondrial proteins that integrates computational predictions, proteome mapping, mutant screening, expression profiling, protein-protein interactions and cellular localization. Among 920 inferred human mitochondrial proteins in MitoP2 at the time of our study, we selected 456 that were also annotated as mitochondrial proteins in the Swissprot database. These 456 protein sequences were used to query the Swissprot database using blastp program. The hit sequences with e-value smaller than 0.001 and bit scores equal to or greater than 100 were kept for further analyses. Single linkage clustering was then carried out to cluster the sequences into 305 families [42].

**Ortholog collection and phylogeny construction**

Gene families were removed from analysis if all human proteins in the family were products of alternative splicing of the same gene or the records of the proteins no longer existed in Genbank. For each human protein in the remaining 92 multigene families, we retrieved orthologs in mouse (*Mus musculus*), chicken (*Gallus gallus*), fish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), mosquito (*Anopheles gambiae*) and nematode (*Caenorhabditis*).
*Caenorhabditis elegans* from the Homologene database at NCBI. The subcellular localizations of proteins in each family were inferred from Swissprot and Genbank databases. The protein sequences in each family were aligned with T-coffee using the default settings [43]. The aligned sequences were manually inspected and then imported into the Mega package to construct neighbor joining and parsimony trees with 1000 bootstrap replicates [44]. For each family, we further performed a maximum likelihood search with 100 bootstrap replicates as implemented in the PHYML (v.2.4.4) program using the WAG model with estimated amino acid frequencies and 4 gamma categories. The inferred topologies were congruent among different tree-making methods except for some minor differences (not shown). We selected maximum likelihood trees for illustrations.

**Comparative analyses of protein sequences**

For human proteins in each family, we retrieved information of N-terminal mitochondrial targeting signal (TARGETP), Pfam domains (MITODOMAIN) and maestro scores developed by Calvo and coauthors for predicting mitochondrial proteins [40]. The distributions of these data were plotted in R package (http://www.r-project.org/) for mitochondrial and non-mitochondrial protein members in the analyzed multigene families. The Pearson chi-square tests were applied to test whether the distributional proportions of these genomic criteria were the same for mitochondrial and non-mitochondrial proteins.

**Acknowledgements**

We thank Karri Haen, Ehsan Kayal, and two anonymous reviewers for valuable comments on earlier versions of the manuscript, and Pioneer Computational Biology Fellowship for funding.
References


37. Omura T: **Mitochondria-targeting sequence, a multi-role sorting sequence recognized at all steps of protein import into mitochondria.** *J Biochem* 1998, **123**:1010-1016.


<table>
<thead>
<tr>
<th>Category</th>
<th>Multigene families(a)</th>
<th>Total number of proteins</th>
<th>Phylogenetic interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mitochondrial</td>
<td>non-mitochondrial</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Before vertebrates</td>
<td>At the root of</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>emergence of</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>vertebrates of</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>fish</td>
</tr>
<tr>
<td>Mitochondrial-cytosolic families (mt-cy families)</td>
<td>ACAA2, ALDH1B1, CPT2, GPAM, HMGCS2, NDUF4, SH3BP5, HIBADH, ABCB10, CABC1, ARG2, ATP5A1, BNIP3, CYP11B2, CASA, CPS1, CDS2, CYBSB, DECR1, DGUOK, DLD, DNAJA3, GFM1, TUFIN, FTMT, GK, SHMT2, GPX4, HSPA9, ALAS2, HTRA2, HK1, IDH2, MTIF2, PPA2, AK3, CKMT2, ME2, MFN2, MGST1, MIPEP, NME4, NFS1, OAS2, OPA1, SLC25A15, PCK2, PHB, PPIF, PRDX3, SIRT3, IARS2, SARS2, DNAJC19, TMLHE, TOP1MT, TRAP1, TST, OXR1</td>
<td>59</td>
<td>144</td>
</tr>
<tr>
<td>Mitochondrial-mitochondrial families (mt-mt families)</td>
<td>AIFM1, MAOA, ATP5G2, BCL2L1, D2HGDH, GLUD2, GRPEL1, LETM1, MCART1, PMPCB, MTCCH2, MTERF, ENDOG, OAT, BCKDHA, PDHB, PDK4, PDP2, MTRF1, RHO1, SCO2, TIMM17A, TOMM40, VDAC3, COX4I2, COX7A1, COX6A1, COX6B1, MCCC2, OXCT2, MRPS18A, MRPS10, MRPS30</td>
<td>33</td>
<td>79</td>
</tr>
</tbody>
</table>

\(a\)Families are listed using gene names of mitochondrial members.

\(b\)For families that underwent multiple rounds of duplications, only the duplications followed by subcellular relocalizations were considered.
Figure 1. Maximum likelihood phylogeny of the arginase family. Numbers indicate bootstrap support based on 100 replicates. ARG1: Type I arginases; ARG2: Type II arginases. Colored boxes indicate annotated and/or predicted subcellular locations of the proteins: cytoplasm (yellow) and mitochondria (green). There is no subcellular information for the proteins in Anopheles gambiae and Nematostella vectensis.
Figure 2. Maximum likelihood phylogeny of the DNA topoisomerase typeIB family. Numbers indicate bootstrap support based on 100 replicates. TOP1: DNA topoisomerase 1; TOP1MT: mitochondrial DNA topoisomerase 1. Colored boxes indicate annotated and/or predicted subcellular locations of the proteins: nucleus/cytoplasm (blue) and mitochondria (green). There is no subcellular information for the proteins in Anopheles gambiae and Nematostella vectensis.
Figure 3. Evolutionary rates in mitochondrial vs. non-mitochondrial proteins. (A) A schematic phylogeny of a mt-cy two gene family with gene duplication occurred in the vertebrate lineage. Branch lengths before the divergence between fish and tetrapods are marked as a and b for mitochondrial and cytosolic proteins, respectively. The corresponding average branch lengths after this divergence are marked as a’ and b’. (B) The ratios of branch lengths for mitochondrial vs. nuclear paralogs (a/b, a’/b’, and (a+a’)/(b+b’)) were calculated on the maximum likelihood topologies as illustrated in (A) with the exception of the TST family, for which the divergence between birds (chicken) and mammals was used. TOP1MT, TST, SHMT2 and CDS2 families had undergone relocalization from cytosol to mitochondria, while the remaining 6 families had the opposite direction of relocalizations.
Figure 4. Maximum likelihood phylogeny of the class I sirtuin family. Numbers indicate bootstrap support based on 100 replicates. SIRT1: Sirtuin 1; SIRT2: Sirtuin 2; SIRT3: Sirtuin 3. Colored boxes indicate annotated and/or predicted subcellular locations of the proteins: nucleus (purple), cytoplasm (yellow) and mitochondria (green). SIRT3 in *Rattus norvegicus* and *Mus musculus* lost the mitochondrial N-terminal targeting signal and thus were retained in the cytoplasm.
Figure 5. The presence of mitochondrial N-terminal targeting signals (A) and mitochondrial Pfam domains (B) for human mitochondrial (mt) and non-mitochondrial (nonmt) proteins. N-terminal mitochondrial targeting signals were inferred for proteins in mt-cy families based on targetP predictions [38]. Mitochondrial Pfam domains refer to those domains that were found only in eukaryotic (excluding human) mitochondrial proteins [40].
Maestro score distribution

Figure 6. Maestro score distributions for human mitochondrial and non-mitochondrial proteins. Maestro scoring system incorporates eight genomic-scale data sets (targeting sequence prediction, protein domain enrichment, presence of \textit{cis}-regulatory motifs, yeast homology, ancestry, tandem-mass spectrometry, coexpression and transcriptional induction during mitochondrial biogenesis) for predicting nuclear encoded mitochondrial proteins [40]. The cutoff score of 5.65 is indicated as the vertical bar.
CHAPTER 6. GENERAL CONCLUSIONS

Conclusions

The mitochondrial genomes (mtDNAs) of Metazoa have been considered as remarkably uniform in terms of genome size, gene contents and genome architecture (Boore 1999). However this traditional view was challenged when first mtDNA sequences of non-bilateral animal were determined and showed variations in genome size as well as gene content (Beaton et al. 1998; Beagley et al. 1998; van Oppen et al. 2002; Fukami and Knowlton 2005; Lavrov and Lang 2005a; Lavrov et al. 2005; Lavrov and Lang 2005b). Previous studies showed that the emergence of multicellularity and bilateral symmetry in animals were associated with the changes in animal mtDNAs (Lavrov 2007). However, this inference was based on the limited sampling from non-bilateral animal mtDNAs. In order to understand mtDNA evolution in animals and infer the features of last common ancestor of animals, it is essential to acquire mtDNA samplings from non-bilateral animals. In this dissertation I focus on obtaining mtDNAs from non-bilateral animals and reconstructing animal phylogeny using mtDNAs.

As a part of my dissertation I explored the mitochondrial genome diversity in the Demospongiae (demosponges), the largest group in the phylum Porifera (Sponges). We sequenced mtDNAs from all 14 orders within the Demospongiae and performed comparative analysis of these genomes. Although all sampled mitochondrial genomes retained some ancestral features, including a minimally modified genetic code, conserved structures of tRNA genes, and presence of multiple non-coding regions, they vary considerably in their size (16- to 25-kb), gene content (13-15 protein coding genes, 2 ribosomal RNA genes, and
2-27 tRNA genes), gene order (1-41 shared boundaries), and the rates of sequence evolution. For example, as most of mtDNAs have length around 20kb, those of *Axinenna corrugata* and *Ephydata muelleri* deviated to about 25kb due to the expansion of non-coding regions. Furthermore, the number of tRNAs was reduced to only 2, equivalent to what has been found in Cnidarian, for three species within this group.

This diversity of mtDNA organizations is well illustrated by two species, *Oscarella carmela* and *Plakortis angulospiculatus*, within Homoscleromorpha, a small group traditionally considered to be a part of the Demospongiae. The mtDNA of *O. carmela* is 20,327 bp and contains the largest number of genes (44) reported for animal mtDNA including a putative gene for the C subunit of the twin-arginine translocase (*tatC*) that has never been reported in animal mtDNA. The genes in *O. carmela* mtDNA are arranged in two clusters with opposite transcriptional orientations, a gene arrangement reminiscent of those in several cnidarian mtDNAs but not in sponges. By contrast, the mtDNA of *P. angulospiculatus* contains only 20 genes and lacks *tatC* as well as 19 of the 25 tRNA genes typical for demosponges. Surprisingly, we also discovered two group I introns in *cox1* of *P. angulospiculatus*, which are rare in demosponges and animal mtDNAs in general. The two genomes further displayed distinct gene arrangements that they share only 4 gene boundaries.

The large amount of mitochondrial genomic data collected for demosponges allowed us to explore the phylogeny of this group. Our phylogenetic analysis demonstrated the presence of five major clades within demosponges: Homoscleromorpha (G0 = Homosclerophorida), Keratosa (G1 = Dictyoceratida+Dendroceratida), Myxospongiae (G2=Chondrosida+...
Halisarcida+ Verongida), Marine Haplosclerida (G3), and all the remaining groups (G4).

Although most demosponge mtDNA sequences have low rates of sequence evolution, a significant acceleration in evolutionary rates was observed in the G1 group. Furthermore, the estimated genetic distances between *Oscarella carmela* and *Plakortis angulospiculatus* are greater than those between many orders of demosponges, indicating an ancient radiation and the presence of extensive genetic diversity within this group. These findings of phylogenetic affinity of Homoscleramorpha to other demosponges and the monophyly of demposponges suggest that the most recent common ancestor of animals share combined features of sponges and eumetazoa.

As mentioned above, some of the sequenced demosponge mtDNAs experienced significant loss of tRNA genes while others had duplications. Further phylogenetic analysis of these tRNAs demonstrated that horizontal evolution, including both alloacceptor and isoacceptor recruitments, is relatively common in tRNAs of these organellar genomes. We also showed that these tRNA gene recruitment events were not just limited to mtDNAs. The analysis of three nuclear genomes provided support that tRNA recruitments are also common in nuclear genomes. Interestingly, most recruited tRNA genes, in both mitochondrial and nuclear genome, differed from their donor tRNAs by only one nucleotide at the middle anticodon position. Such change almost always leads to a switch between class I and class II aminoacyl-tRNA synthetases (aaRSs) responsible for tRNA amino-acylation that may protect organisms from a mismatch between the anticodon and aminoacyl identities of recruited tRNAs. Although the recruitment mechanisms in mitochondrial and nuclear genomes might not be same, the finding that tRNA gene recruitment is common both in organellar and
nuclear genomes suggests that horizontal evolution is an important player in evolution of tRNA gene families and should be taken into account when tRNA sequences are used in evolutionary studies.

Since mitochondrion cannot function with only the proteins coded by its own genome, it is important to study the mitochondrial proteins encoded by nuclear genome. Here we analyzed nuclear encoded mitochondrial proteins in human to investigate the origin of mitochondrial proteins and the evolution of mitochondrial proteome. The study revealed that among the mitochondrial multigene families, about two thirds had undergone subcellular relocalizations after gene duplications and the remaining one third of the families generated multiple mitochondrial copies after duplication. Comparative sequence analyses showed that the subcellular relocalization processes were primarily enabled via the gain or loss of N-terminal mitochondrial targeting sequences. The results indicate that mitochondrial proteome was expanded through protein subcellular relocalization after gene duplication as well as subfunctionalization after duplication. Further our investigation indicates subcellular relocalization is an important mechanism in the gain of function of duplicated genes over the course of animal genome evolution.

**Future directions**

Our study of demosponge mtDNAs uncovered an extensive mitochondrial genomic diversity within the group and resolved demosponge phylogenetic relationship with high support. The findings of monophyly of demosponges and the affinity of Homoscleromorpha to demosponges provide insights to the most common ancestor of animals. To further support
our discovery, more data can be collected, especially from Homoscleromorpha. Since sparse sampling in non-bilateral animal groups can cause problems in phylogeny reconstruction using mtDNAs, further mtDNA sampling from other non-bilateral animal groups is needed (Boore 1999; Lang et al. 2002; Lavrov and Lang 2005a). This accumulation of mtDNAs can be used to perform congeneric comparisons, which are useful to identify regulatory regions in mtDNA and investigate mechanisms of mtDNA rearrangement (Gissi et al. 2008). Also new sequence evolutionary models can be developed to overcome the potential tree reconstruction problems using heterogeneous mtDNAs. In addition, a more comprehensive analysis of mitochondrial proteome can be performed to study its evolution once more mitochondrial proteins are available.

References


Supplementary figure 1. Inferred secondary structure for Oscarella carmela mt-srRNA
Supplementary figure 2. Inferred secondary structure for Oscarella carmela mt-rRNA
Supplementary figure 3. Inferred secondary structures for Oscarella carmela mt-tRNAs
Supplementary figure 4. Phylogenetic inference using individual genes (part 1)
Supplementary figure 4 (cont.). Phylogenetic inference using individual genes.
Supplementary figure 4 (cont.). Phylogenetic inference using individual genes
### Supplementary table 1. Nucleotide composition of *Oscarella carmela* protein genes

<table>
<thead>
<tr>
<th></th>
<th>atp6</th>
<th>atp8</th>
<th>atp9</th>
<th>cob</th>
<th>cox1</th>
<th>cox2</th>
<th>cox3</th>
<th>nad1</th>
<th>nad2</th>
<th>nad3</th>
<th>nad4</th>
<th>nad4L</th>
<th>nad5</th>
<th>nad6</th>
<th>tatC</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>% G</td>
<td>17.1</td>
<td>8.5</td>
<td>23.2</td>
<td>17.4</td>
<td>18.5</td>
<td>19.6</td>
<td>18.7</td>
<td>17.2</td>
<td>17.0</td>
<td>17.9</td>
<td>16.9</td>
<td>18.3</td>
<td>17.6</td>
<td>18.9</td>
<td>13.4</td>
<td>17.5</td>
</tr>
<tr>
<td>% A</td>
<td>29.5</td>
<td>41.8</td>
<td>27.6</td>
<td>30.6</td>
<td>29.7</td>
<td>32.3</td>
<td>29.4</td>
<td>31.5</td>
<td>31.6</td>
<td>26.3</td>
<td>31.4</td>
<td>31.0</td>
<td>30.4</td>
<td>32.8</td>
<td>33.9</td>
<td>31.0</td>
</tr>
<tr>
<td>% T</td>
<td>40.4</td>
<td>34.8</td>
<td>34.6</td>
<td>37.2</td>
<td>37.0</td>
<td>33.7</td>
<td>36.8</td>
<td>37.2</td>
<td>36.8</td>
<td>40.6</td>
<td>37.7</td>
<td>38.0</td>
<td>38.3</td>
<td>36.0</td>
<td>42.5</td>
<td>37.6</td>
</tr>
<tr>
<td>% A+T</td>
<td>69.9</td>
<td>76.6</td>
<td>62.3</td>
<td>67.8</td>
<td>66.7</td>
<td>66</td>
<td>66.2</td>
<td>68.7</td>
<td>68.4</td>
<td>67.1</td>
<td>69.2</td>
<td>69</td>
<td>68.7</td>
<td>68.9</td>
<td>76.6</td>
<td>68.6</td>
</tr>
<tr>
<td>AT skew</td>
<td>-0.16</td>
<td>0.09</td>
<td>-0.11</td>
<td>-0.1</td>
<td>-0.11</td>
<td>-0.02</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.21</td>
<td>-0.09</td>
<td>-0.1</td>
<td>-0.11</td>
<td>-0.05</td>
<td>-0.11</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>GC skew</td>
<td>0.14</td>
<td>-0.28</td>
<td>0.23</td>
<td>0.08</td>
<td>0.11</td>
<td>0.15</td>
<td>0.11</td>
<td>0.1</td>
<td>0.07</td>
<td>0.09</td>
<td>0.1</td>
<td>0.18</td>
<td>0.13</td>
<td>0.21</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>Total (bp)</td>
<td>738</td>
<td>201</td>
<td>228</td>
<td>1203</td>
<td>1566</td>
<td>750</td>
<td>786</td>
<td>987</td>
<td>1398</td>
<td>357</td>
<td>1488</td>
<td>300</td>
<td>1902</td>
<td>594</td>
<td>762</td>
<td>13260</td>
</tr>
</tbody>
</table>
Figure S1. Phylogenetic analysis of demosponge relationships using mitochondrial genomic data.
Figure S3. Inferred secondary structure of *Axinella corrugate* mitochondrial large subunit RNA. Inserted repetitive elements are shown in red.
Table S1. Comparison of mitochondrial protein genes in demosponges: number of encoded amino acids.

| Protein-coding genes | Oscarella carmela | Geodia neptuni | Tethya actinia | Cinachyrella kuekenthalii | Chondrilla nucula | Agelas schmidti | Topsentia ophirotretis | Hymeniacidon petiolaris | Amphimedon queenslandica | Halisarca dujardini | Vaceletia sp. | Hippocastania lachne | Igerella notabilis | Plakortis angulospinales |
|----------------------|------------------|----------------|---------------|--------------------------|------------------|-----------------|------------------------|--------------------------|-------------------------|----------------------|--------------------|-------------------|---------------------|
| atp6                  | 245              | 244            | 243           | 241                      | 244              | 244             | 244                    | 248                      | 244                     | 244                  | 244                | 244                | 244                  |
| atp8                  | 66               | 63             | 63            | 63                       | 68               | 68              | 68                     | 68                       | 68                      | 68                   | 68                 | 68                 | 68                  |
| atp9                  | 58               | 56             | 56            | 56                       | 58               | 58              | 58                     | 58                       | 58                      | 58                   | 58                 | 58                 | 58                  |
| cob                   | 199              | 199            | 199           | 199                      | 199              | 199             | 199                    | 199                      | 199                     | 199                  | 199                | 199                | 199                 |
| cox1                  | 509              | 508            | 508           | 508                      | 507              | 507             | 507                    | 507                      | 507                     | 507                  | 507                | 507                | 507                 |
| cox2                  | 249              | 247            | 247           | 247                      | 250              | 250             | 250                    | 250                      | 250                     | 250                  | 250                | 250                | 250                 |
| cox3                  | 261              | 262            | 262           | 262                      | 262              | 262             | 262                    | 262                      | 262                     | 262                  | 262                | 262                | 262                 |
| nad1                  | 386              | 385            | 385           | 385                      | 384              | 384             | 384                    | 384                      | 384                     | 384                  | 384                | 384                | 384                 |
| nad3                  | 328              | 325            | 325           | 325                      | 325              | 325             | 325                    | 325                      | 325                     | 325                  | 325                | 325                | 325                 |
| nad4                  | 495              | 481            | 481           | 481                      | 481              | 481             | 481                    | 481                      | 481                     | 481                  | 481                | 481                | 481                 |
| nad4L                 | 99               | 103            | 103           | 103                      | 103              | 103             | 103                    | 103                      | 103                     | 103                  | 103                | 103                | 103                 |
| nad5                  | 625              | 624            | 624           | 624                      | 624              | 624             | 624                    | 624                      | 624                     | 624                  | 624                | 624                | 624                 |
| nad6                  | 253              | 253            | 253           | 253                      | 253              | 253             | 253                    | 253                      | 253                     | 253                  | 253                | 253                | 253                 |
### Table S2. Comparison of mitochondrial protein genes in demosponges: inferred start codons.

| Protein-coding genes | Oscarella carmela | Geodia neptuni | Tethya actinia | Cinachyrella kuekenthali | Chondrilla nucula | Agelas schmidti | Topsentia ophirap | Hydridites ptiloculata | Axinella corrugata | Callyspongia plicifera | Iotrochota birotula | Amphimedon compressa | Aplysina fistularis | Ectyoplasia ferox | Ephydatia muelleri | Xestospongia muta | Amphimedon queenslandica | Halisarca dujardini | Vaceletia sp. | Hippophaea lachnea | Igerneilla notabilis | Plakortis angulospiculatus |
|----------------------|-------------------|----------------|---------------|--------------------------|-------------------|----------------|-------------------|-------------------------|-------------------|------------------------|-------------------|-------------------------|-------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| atp6                | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| atp8                | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| atp9                | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| cob                 | ATG               | GTG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| com1               | ATG               | GTG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| com2               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| com3               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad1               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad2               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad3               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad4               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | ATG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad4L              | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad5               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad6               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| nad7               | ATG               | ATG            | ATG           | ATG                      | ATG               | ATG            | ATG               | ATG                      | GTG                | ATG                    | ATG               | ATG                    | ATG               | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  | ATG                  |
| tatC               | TTG               | -              | -             | -                        | -                 | -              | -                 | -                        | -                  | -                      | -                 | -                      | -                 | -                    | -                    | -                    | -                    | -                    | -                    | -                    | -                    | -                    | -                    | -                    |
ACKNOWLEDGEMENTS

I would like to thank my advisor, lab mates, friends and family for their support during my graduate studies. First of all, I am grateful to Dr. Dennis Lavrov for providing direction in my professional development. I also thank Dr. Xun Gu for the suggestions he has made to my projects and other members of my POS committee: Drs. Karin Dorman, Oliver Eulenstein, and Stephen Proulx for guidance. Finally, I thank my husband Yi Jia, my parents Zeyun Wang and Guiying Shi, and my sister Xiufeng Wang for their support and encouragement.