The Mechanisms of Codon Reassignments in Mitochondrial Genetic Codes

Supratim Sengupta, Xiaoguang Yang, and Paul G. Higgs

Supplementary Information

Methods

The list of species used in our study in each of the principal eukaryotic taxa is given in Table S1. In taxa with few complete mitochondrial genomes, all the available species were included. In taxa with many complete genomes (particularly metazoa), representative species were used. Table S1 also gives information on the set of amino acids for which tRNAs are present in each genome, and the reference number of the genetic code used by each species following the NCBI genetic code numbering system. This system is summarized in Table S2 for reference.

For the plants/algae set we used concatenated protein alignments of cox1,2,3,cob,atp6,9,nad1,3,4,4l,5, having a total length of 3385 amino acids. The genes for atp8, nad2 and nad6 were found to be substantially more divergent than the rest, and were excluded to avoid introducing noise into the data. For the fungi, most of the nad genes are missing. Therefore, we used an alignment of cox1,2,3,cob,atp6,9 with a total length of 1701 amino acids. In a few cases when a specific gene was missing in any species (e.g. *Candida albicans* does not have the cob gene), the entry for that gene was filled by gaps. Multiple sequence alignments were carried out using T-COFFEE (Notredame *et al.* 2000) and the alignment was edited to remove columns having more than 20% gaps. Phylogenetic analysis was performed using the MCMC program available in the PHASE package (Jow *et al.* 2003). The mtREV24 amino acid substitution matrix with a gamma distribution for site variation with four categories was used. In order to compare the Bayesian posterior probabilities from MCMC with bootstrap values, we also generated 100 bootstrap replicate data sets. Maximum likelihood pairwise distance matrices were calculated for each replicate using the same evolutionary model, and trees were constructed using the neighbour joining (NJ) method using the PHYLIP package (Felsenstein 1989). Bootstrap percentages were obtained from this set of trees.

We also carried out independent phylogenetic analysis with a combined alignment of large and small subunit rRNA genes to ascertain if the trees derived from the proteins and rRNAs are consistent with one another. The rRNA sequences were aligned initially using ClustalX (Thompson *et al.* 1997). Variable regions of the molecules were excluded from the analysis by deleting sites with more than 10% gaps in the alignment. The lengths of the alignments were 2063 and 3829 for the fungi and plants/algae. We analyzed the rRNA data sets using a method that accounts for conserved RNA structure. A consensus secondary structure was added to the alignments. The general reversible model was used for the unpaired sites, and a model specifically treating compensatory substitutions was used for the paired sites (model 7A in Savill *et al.* 2001). Parameters for both models are optimized simultaneously by the MCMC program (see Hudelot *et al.* 2003 and the documentation to the PHASE package available at http://www.bioinf.man.ac.uk/resources/phase/). For both models, variation of rates across sites was accounted four using 4 gamma-distributed categories.

Phylogenetic discussion

Although there is general agreement on the definitions of the major eukaryotic taxa, the relationships between groups in the very earliest parts of the eukaryotic tree have proved difficult to resolve (Lang *et al.* 1999; Baldauf *et al.* 2000; Van der Peer *et al.* 2000; Baldauf 2003; Philippe *et al.* 2004). Our aim here was not to resolve these early branches, but to obtain trees that would allow for accurate positioning of the codon reassignments.

The grouping of fungi and metazoa, proposed by Baldauf & Palmer (1993), is supported by all the studies cited in the previous paragraph. *Amoebidium parasiticum* and *Monosiga brevicollis* are thought to be the closest unicellular relatives of the metazoa, and are therefore also included in this group. *Dictyostelium discoideum* and *Acanthamoeba castellanii* are thought to be more distant relatives of this group that diverge before the split of fungi and metazoa (Baldauf *et al.* 2000; Van der Peer *et al.* 2000; Lang *et al.* 2002). These are therefore included in this subset as outgroups.

In Figure 1 we focus on the phylogeny within the fungi and include only a few representative metazoa. The tree from the mitochondrial proteins is very well resolved. Almost all nodes have 100% posterior probability (PP), and most also have high bootstrap percentage (BP). The branching order of the species within Ascomycota agrees with Bullerwell *et al.* (2003a) and Kurtzman and Robnett (2003), and that within Chytridiomycota agrees with Bullerwell *et al.* (2003b). An exception is *A. macrogynus*. In the NJ analysis, the most common position of this species is with the Chytridiomycota, but it appears as basal to the Zygomycota/Basidiomycota/Ascomycota according the MCMC analysis in Figure 1. When the Zygomycota are excluded, *A. macrogynus* moves to the base of the Chytridiomycota in the MCMC tree. The uncertainty in the position of this species does not affect any of our conclusions regarding changes in the genetic code.

The mitochondrial rRNA tree topology for the fungi is identical to the protein tree in Figure 1, with the exception of the position of *Candida stellata*. In the rRNA tree *C. stellata* pairs with *Yarrowia* (PP = 98%, BP = 72%). In the protein tree, *C. stellata* branches after *Yarrowia* (PP = 100%), although the BP is 0% for this arrangement because in the protein NJ consensus tree, it branches before *Yarrowia*. The positioning of *C. stellata* is therefore not clear. Nevertheless, whatever its position, *C. stellata* is separate from *C. albicans* + *C. parapsilosis*, and these are all separate from *C. glabrata*. Thus, it is clear that *Candida* is paraphyletic. Kurzman & Robnett (2003) and Diezmann *et al.* (2004) have considered the phylogeny of these groups using nuclear genes, and find that *Candida*, *Kluyveromyces, Saccharomyces* and *Pichia* are all paraphyletic.

The second data set contains Rhodophyta (red algae), Chlorophyta (green algae) and Streptophyta (plants). *Malwimonas jakobiformis* and *Reclinomonas americana* are added as outgroups. These are found to be related to the plants and algae by Lang *et al.* (2002) and Forget *et al.* (2002). The protein sequence phylogeny for this group (Figure 2) is again quite well resolved. The Rhodophyta, Chlorophyta and Streptophyta are all monophyletic, with Rhodophyta basal to the other two, which is consistent with results of Lang *et al.* (2002). However, the grouping of *Reclinomonas* with the rhodophytes, instead of being an outgroup to rhodophytes + chlorophytes + streptophytes is most likely an artefact. The positioning of *Nephroselmis* and *Prototheca* is also somewhat questionable because these species appear in different positions in the NJ tree and hence they have low BP in Figure 2. The rRNA phylogeny is the same as Figure 2 with the following exceptions. In this case *Reclinomonas* and *Malawimonas* are indeed outgroups as expected, and there is strong support for rhodophytes + chlorophytes + streptophytes (100% PP). However, in the rRNA tree, *Nephroselmis* shifts to being basal to Streptophyta instead of being basal to the other Chlorophyta. Turmel *et al.* (1999)

find that the position of *Nephroselmis* shifts according to the method and sequences used, but conclude that it is the basal species within Chlorophyta. Figure 2 is also equivalent to the tree obtained by Pombert *et al*. (2004). We therefore conclude that, for the study in the remainder of this paper, the protein tree in Figure 2 is the best estimate of the phylogeny of this group, with the interpretation that *Reclinomonas* should be an outgroup. The position of *Reclinomonas*, and the uncertainty regarding *Nephroselmis* do not affect the conclusions regarding codon reassignments that we make later in the paper.

The third data set contains species that are clearly not associated with either the fungi/metazoa or plants/algae. Basal branches in this tree have been left as multifurcating. This set may not be monophyletic with respect to the groups in Figures 1 and 2, *i.e.* it is possible that the fungi/metazoa or plants/algae may lie within the tree of Figure 3. If this is the case, it does not affect the conclusions regarding the genetic code. In order to position some of the genetic code changes in this set it is necessary to include species for which complete mitochondrial genomes are not available. We therefore cannot not show a tree derived only from mitochondrial genes. Stramenopiles and alveolates are two groups that are thought to be monophyletic according to Wolters (1991) and Saunders *et al*. (1995). The combination of these two groups is sometimes called Chromalveolata, and is supported by Van der Peer *et al.* (2000), Baldauf *et al.* (2000), and Philippe *et al.* (2004). For the arrangement of the species within the Stramenopiles, we used the results of a much larger scale study with the nuclear small subunit rRNA gene (Van der Peer *et al.* 1996). For the species in which mitochondrial genomes were available, we carried out protein and rRNA phylogenies as with the previous two sets. This was not very informative due to the small number of species available. However, one useful point that emerged was that *N. gruberi* appeared as sister group to the Alveolates with 100% PP from both the protein and rRNA analysis. We have therefore placed it in that position on Figure 3. We are not confident of this result due to the poor species sampling and the possibility of long-branch artefacts involving the Alveolates. Repositioning this species might lead to an increase in the inferred number of times that tRNA import has evolved, but it would not affect the positioning of the codon reassignments. The Haptophyceae are a separate group from Alveolates and Stramenopiles. For the species in this group we followed Hayashi-Ishimaru et.al. (1997). The complete genome of *E. huxleyi* has become available since that study, and has been added in Figure 3. We confirmed its relationship to *I. galbana* using the cox1 protein sequence only.

Figure 4 shows the metazoa. This is a well studied group. We have followed Mallatt & Winchell (2002) and Halanych (2004) for the relationship between the animal phyla.

References

Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. USA* 90:11558–11562.

Baldauf SL, Roger AJ, Wenk-Siefert I, Dolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science. 290: 972-977.

Baldauf SL (2003) The deep roots of eukayotes. *Science* 300: 1703-1706.

Bullerwell CE, Leigh J, Forget L, Lang BF (2003a). A comparison of three fission yeast mitochondrial genomes. *Nucl. Acids Res.* 31: 759-768.

Bullerwell CE, Forget L, Lang BF (2003b) Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. *Nucl. Acids Res.* 31: 1614-1623.

Diezmann S, Cox CJ, Schonian G, Vilgalys RJ, Mitchell TG (2004) Phylogeny and evolution of medical species of Candida and related taxa: a multigenic analysis. *J Clin Microbiol.* 42:5624-5635.

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2) Cladistics. 5:164-166. See also Felsenstein, J (2005) PHYLIP (Version 3.6) Distributed by author. Department of Genome Sciences, University of Washington, Seattle.

Forget L, Ustinova J, Wang Z, Huss VAR, Lang BF (2002) *Hyaloraphidium curvatum*: a linear mitochondrial genome, tRNA editing, and an evolutionary link to the lower fungi. *Mol. Biol. Evol.* 19: 310-319.

Halanych KM (2004) The new view of animal phylogeny. *Annu. Rev. Ecol. Evol. Syst.* 35: 229-256.

Hayashi-Ishimaru Y, Ohama T, Kawatsu Y, Nakamura K, Osawa S (1996) UAG is a sense codon in several chlorophycean mitochondria. Curr. Genet. 30:29-33.

Hudelot C, Gowri-Shankar V, Jow H, Rataray M, Higgs PG (2003) Bayesian phylogenetics using an RNA-based phylogenetic methods: Applications to mammalian mitochondrial RNA sequences. *Mol. Phyl. Evol.* 28: 241- 252.

Jow H, Hudelot C, Rattray M, Higgs PG (2002) Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. 19:1591-1601.

Kurtzman CP, Robnett CJ (2003) Phylogenetic relationships among yeasts of the Saccharomyces complex determined from multigene sequence analyses. FEMS Yeast Res. 3: 417-432.

Lang BF, Seif E, Gray MW, O'Kelly CJ, Burger G (1999) A comparative genomics approach to the evolution of eukaryotes and their mitochondria. *J Eukaryot Microbiol.* 46:320-326.

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G (2002) The closest unicellular relatives of animals. Curr. Biol. 12:1773-1778.

Mallatt J, Winchell CJ (2002) Testing the new animalphylogeny: First use of combined large subunit and small subunit rRNA gene sequences to classify protostomes. *Mol. Biol. Evol.* 19: 289-301.

Notredame C, Higgins D, Heringa J (2000) T-Coffee: A novel method for multiple sequence alignments. J. Mol. Biol. 302:205-217.

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D (2004) Phylogenomics of Eukaryotes: Impact of missing data on large alignments. *Mol. Biol. Evol.* 21: 1740-1752.

Pombert JF, Otis C, Lemieux C, Turmel M The Complete Mitochondrial DNA Sequence of the Green Alga *Pseudendoclonium akinetum* (Ulvophyceae) Highlights Distinctive Evolutionary Trends in the Chlorophyta and Suggests a Sister-Group Relationship Between the Ulvophyceae and Chlorophyceae. *Mol. Biol. Evol.* 21:922–935 (2004).

Saunders GW, Potter D, Paskind MP, Andersen RA (1995) Cladistic analyses of combined traditional and molecular data sets reveal an algal lineage. *Proc. Natl. Acad. Sci. USA* 92:244–248.

Savill NJ, Hoyle DC, Higgs PG (2001) RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum likelihood methods. *Genetics* 157, 399-411.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* 25: 4876-4882.

Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW (1999) The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. *The Plant Cell* 11: 1717-1729.

Van der Peer Y, Van der Auwera G, De Wachter R (1996) The evolution of stramenopiles and alveolates as derived by "substitution rate calibration" of small subunit RNA. *J. Mol. Evol.* 42**:** 201-210.

Van de Peer Y, Baldauf SL, Dolittle WF, Meyer A (2000) An update and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. J. Mol. Evol. 51:565-576.

Wolters J (1991) The troublesome parasites - Molecular and morphological evidence that Apicomplexa belong to the dinoflagellate ciliate clade. *BioSystems* 25:75-83 (1991).

Table S1 - List of species with complete mitochondrial genomes used in this study

Species	Taxon	Code No.	Amino acids for which tRNAs are present
Figure 1			
Acanthamoeba castellanii	Acanthamoebidae	4	A, D, E, F, H, I, K, L(CUN), L(UUR), M, P, Q, W, Y
Dictyostelium discoideum	Mycetozoa	1	All but D, G, R(CGN), S(UCN), S(AGY), T, V
Amoebidium parasiticum	Ichthyosporia	4	Full set
Monosiga brevicollis	Choanoflagellida	4	Full set
Rhizophydium sp. 136	Chytridiomycota	16	K, L(UAG), M, P, Q, W, Y
Spizellomyces punctatus	Chytridiomycota	16	D, K, L(UAG), M, P, Q, W, Y
Hyaloraphidium curvatum	Chytridiomycota	1*	D, E, M, P, Q, W, Y
Monoblepharella sp. JEL15	Chytridiomycota	1	D, E, G, K, M, P, Q, W, Y
Harpochytrium sp. JEL105	Chytridiomycota	1*	D, E, K, M, P, Q, W, Y
Harpochytrium sp. JEL94	Chytridiomycota	1*	D, E, K, M, P, Q, W, Y
Allomyces macrogynus	Chytridiomycota	1*	Full set
Mortierella verticillata	Zygomycota	1*	Full set
Rhizopus oryzae	Zygomycota	1	Full set
Crinipellis perniciosa	Basidiomycota	4	Full set
Schizophyllum commune	Basidiomycota	4	Full set
Penicillium marneffei	Ascomycota	4	Full set
Podospora anserina	Ascomycota	4	Full set
Hypocrea jecorina	Ascomycota	4	Full set
Lecanicillium muscarium	Ascomycota	4	All but A, C
Schizosaccharomyces japonicus	Ascomycota	4	Full set
Schizosaccharomyces octosporus	Ascomycota	4	Full set
Schizosaccharomyces pombe	Ascomycota	4	Full set
Yarrowia lipolytica	Ascomycota	4	All but R(CGN), V
Candida stellata	Ascomycota	4	Full set
Candida albicans	Ascomycota	4	Full set
Candida parapsilosis	Ascomycota	4	Full set
Pichia canadensis	Ascomycota	4	Full set
Ashbya gossypii	Ascomycota	3	All but Q, R(CGN)
Kluyveromyces lactis	Ascomycota	3	All but L(CUN), R(CGN)
Kluyveromyces thermotolerans	Ascomycota	3	All but L(CUN)
Candida glabrata	Ascomycota	3	All but L(CUN), R(CGN)
Saccharomyces cerevisiae	Ascomycota	3	All but L(CUN)
Saccharomyces castellii	Ascomycota	3	All but L(CUN), R(CGN)
Saccharomyces servazzii	Ascomycota	3	All but L(CUN), R(CGN)
Figure 2			
Malwimonas jakobiformis	Malawimonadidae	1	Full set
Reclinomonas americana	Jakobidae	1	All but T
Cyanidioschyzon merolae	Rhodophyta	1	All but T
Chondrus crispus	Rhodophyta	4	All but S(AGY), T
Porphyra purpurea	Rhodophyta	4	All but T
Mesostigma viride	Streptophyta	1	All but T
Chaetosphaeridium globosum	Streptophyta	1	Full set
Chara vulgaris	Streptophyta	1	Full set
Marchantia polymorpha	Streptophyta	1	All but I
Arabidopsis thaliana	Streptophyta	1	C, D, F, G, H, K, M, N, P, Q, S(UCN), S(AGY), W, Y
Nephroselmis olivacea	Chlorophyta	1	Full set
Prototheca wickerhamii	Chlorophyta	1	Full set

Pseudendoclonium akinetum	Chlorophyta	1	Full set
Pedinomonas minor	Chlorophyta	4	C, E, F, H, L(UUR), Q, W, Y
Scenedesmus obliquus	Chlorophyta	22	All but T
Chlamydomonas eugametos	Chlorophyta	1	M, Q, W
Chlamydomonas reinhardtii	Chlorophyta	1	M, Q, W
Figure 3			
Rhodomonas salina	Cryptophyta	1	Full set
Emiliania huxleyi	Haptophyceae	4	Full set
Naegleria gruberi	Heterolobosea	1	All but A, C, G, R(CGN), T, V
Plasmodium falciparum	Alveolata	1*	None
Plasmodium reichenowi	Alveolata	1*	None
Paramecium aurelia	Alveolata	4	F, M, W, Y
Tetrahymena pyriformis	Alveolata	4	F, H, L(UUR), M, Q, W, Y
Tetrahymena thermophila	Alveolata	4	F, H, L(UUR), M, Q, W, Y
Cafeteria roenbergensis	Stramenopiles	4	All but L(CUN), R(CGN), T
Thraustochytrium aureum	Stramenopiles	23	All but R(CGN), R(AGR), T
Phytophthora infestans	Stramenopiles	1	All but T
Saprolegnia ferax	Stramenopiles	1	All but T
Chrysodidymus synuroideus	Stramenopiles	1	All but R(CGN), T
Ochromonas danica	Stramenopiles	1	All but R(CGN), T
Laminaria digitata	Stramenopiles	1	All but R(CGN), T
Pylaiella littoralis	Stramenopiles	1	All but R(CGN), T
Figure 4			
Axinella corrugata	Porifera	4	Full set
Geodia neptuni	Porifera	4	Full set
Metridium senile	Cnidaria	4	M, W
Acropora tenuis	Cnidaria	4	M, W
Limulus polyphemus	Arthropoda	5	Full set
Daphnia pulex	Arthropoda	5	Full set
Drosophila melanogaster	Arthropoda	5	Full set
Caenorhabditis elegans	Nematoda	5	Full set
Trichinella spiralis	Nematoda	5	Full set
Katharina tunicata	Mollusca	5	Full set
Lumbricus terrestris	Annelida	5	Full set
Terebratulina retusa	Brachiopoda	5	Full set
Fasciola hepatica	Platyhelminthes	9	Full set
Schistosoma mansoni	Platyhelminthes	9	Full set
Taenia crassiseps	Platyhelminthes	9	Full set
Paracentrotus lividus	Echinodermata	9	Full set
Asterina pectinifera	Echinodermata	9	Full set
Balanoglossus carnosus	Hemichordata	9	Full set
Saccoglossus kowalewskii	Hemichordata	9a	Full set
Halocynthia roretzi	Urochordata	13	Full set
Cionia intestinalis	Urochordata	13	Full set
Branchiostoma lanceolatum	Cephalochordata	5	Full set
Branchiostoma floridae	Cephalochordata	5	Full set
Epigonichthys maldivensis	Cephalochordata	5	Full set
Myxine glutinosa	Craniata	2	Full set
Homo sapiens	Vertebrata	2	Full set
Didelphis virginiana	Vertebrata	2	All but K

Table S2 - Summary of Genetic Code Numbering system (following NCBI)

Code No.	Differences from Canonical Code
1	Canonical Code. (Species marked 1* use the canonical code but are erroneously labelled in NCBI.)
4	UGA = Trp.
5	UGA = Trp; AUA = Met; AGR = Ser.
2	UGA = Trp; AUA = Met; AGR = Stop.
11	UGA = Trp; AUA = Met; AGR = Gly.
21	UGA = Trp; AUA = Met; AGR = Ser; AAA = Asn.
3	UGA = Trp; AUA = Met or unassigned; CGN = rare or unassigned; CUN = Thr.
16	UAG = Leu.
22	UAG = Leu; UCA = Stop.
23	UUA = Stop.
9	UGA = Trp; AGR = Ser; AAA = Asn or unassigned
9a	UGA = Trp; AGR = Ser. (This is similar to 9 but does not have a number in the NCBI system)

Table S3: Annotation of tRNA's with CAU anticodon in the Fungi group together with their position on the mitochondrial genome. *Podospora anserina* has two identical tRNA-Met(CAU) elongator genes labelled as 1a and 1b.

Species	Met-Elongator (1)	Met-Initiator (2)	Ile(CAU) (3)
Crinipellis perniciosa	44802-44874	14445-14516	62938-63011
Schizophyllum commune	16251-16324	18809-18881	23325-23398
Schizosaccharomyces japonicus	42182-42255	63581-63652	48100-48171
Schizosaccharomyces octosporus	27317-27391	40910-40983	
Schizosaccharomyces pombe	4591-4665	14081-14153	17542-17613
Candida stellata	18882-18954	22272-22343	22191-22262
Pichia canadensis	21635-21709	19011-19083	19088-19158
Kluyveromyces lactis	6768-6840	14545-14616	
Kluyveromyces thermotolerans	21284-21356	7352-7424	
Ashbya gossypii	5758-5828	18723-18794	
Candida glabrata	13-86	6889-6959	
Saccharomyces cerevisiae	72630-72705	85035-85112	
Saccharomyces castellii	3292-3364	8530-8603	
Saccharomyces servazzii	10527-10600	15211-15284	
Hypocrea jecorina	16538-16609	14475-14545	14691-14763
Lecanicillium muscarium	5774-5845	4898-4968	5116-5188
Penicillium marneffei	23057-23127	22140-22210	22214-22286
Podospora anserina	(1a) 11266-11336 (1b) 26071-26141	9600-9670	9437-9509
Allomyces macrogynus	42129-42203	47535-47605	44168-44241
Candida albicans		33486-33557	26935-27005
Candida parapsilosis		21711-21780	21792-21862
Yarrowia lipolytica		3272-3343	3374-3445