Environmental Microbiology (2008) **10**(1), 65–74 doi:10.1111/j.1462-2920.2007.01430.x

Environmental PCR survey to determine the distribution of a non-canonical genetic code in uncultivable oxymonads

Audrey P. de Koning, Geoffrey P. Noble,
Aaron A. Heiss, Jensen Wong and
Patrick J. Keeling*
Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada.

Summary

The universal genetic code is conserved throughout most living systems, but a non-canonical code where TAA and TAG encode glutamine has evolved in several eukaryotes, including oxymonad protists. Most oxymonads are uncultivable, so environmental RT-PCR and PCR was used to examine the distribution of this rare character. A total of 253 unique isolates of four protein-coding genes were sampled from the hindgut community of the cockroach, *Cryptocercus punctulatus*, an environment rich in diversity from two of the five subgroups of oxymonad, saccinobaculids and polymastigids. Four α-tubulins were found with non-canonical glutamine codons. Environmental RACE confirmed that these and related genes used only TGA as stop codons, as expected for the non-canonical code, whereas other genes used TAA or TAG as stop codons, as expected for the universal code. We characterized α-tubulin from manually isolated *Saccinobaculus ambloaxostylus*, confirming it uses the universal code and suggesting, by elimination, that the non-canonical code is used by a polymastigid. HSP90 and EF-1α phylogenies also showed environmental sequences falling into two distinct groups, and are generally consistent with previous hypotheses that polymastigids and *Streblomastix* are closely related. Overall, we propose that the non-canonical genetic code arose once in a common ancestor of *Streblomastix* and a subgroup of polymastigids.

Introduction

The genetic code is the set of rules that guides the fidelity of translation and is one of the most highly conserved characteristics of all life, with the so-called 'universal code' employed by virtually all known living systems. The high degree of conservation is due to strong stabilizing selection acting on the code: unlike other mutations, those altering the code affect not just one protein, but all proteins. Regardless of the apparent complexity involved in changing the genetic code, subtle changes to the code have been implemented in a number of independent lineages, and a handful of theories have been proposed to explain how this might take place without affecting protein sequences (Osawa *et al*., 1992; Andersson and Kurland, 1995; Yarus *et al*., 2005). Mitochondrial genomes have changed their code more often than any other kind of genome (Yokobori *et al*., 2001), but variants have also evolved in the bacterial genus *Mycoplasma* (Yamao *et al*., 1985) and several nuclear lineages (Knight *et al*., 2001). In the nucleus, three different codes have evolved once each, and one non-canonical code has evolved at least five times independently. In this code, TAA and TAG (together designated TAR) no longer direct the termination of protein synthesis (stop codons), but instead encode the amino acid glutamine (Q), so that glutamine is encoded by four codons rather than two, and only one stop codon remains (TGA). This code occurs in several ciliates (where it has evolved multiple times, although the exact number is not known for certain; Lozupone *et al*., 2001), in hexamitid diplomonads (Keeling and Doolittle, 1996; 1997), and in dasycladacean green algae (Schneider and de Groot, 1991), and has been most recently discovered in the oxymonad flagellate *Streblomastix strix* (Keeling and Leander, 2003).

This latest discovery is also the least well defined in terms of distribution, because oxymonads are among the most poorly studied of all eukaryotes. Oxymonads are anaerobes or microaerophiles that are distinguished by the presence of distinctive ultrastructural characteristics, including an axostyle, and a particular arrangement of basal bodies and flagella (Brugerolle and Lee, 2000). They are not widely distributed in nature, and are only found within the guts of animals. A few oxymonad species live in vertebrate guts (Nie, 1950; Kulda and Nohynkova, 1978), but the vast majority of them have been found in highly complex microbial communities in the hindguts of

Received 24 January, 2007; accepted 4 August, 2007. *For correspondence. E-mail pkeeling@interchange.ubc.ca; Tel. (+1) 604 8224906; Fax (+1) 604 8226089.

lower termites and the wood-eating cockroach, *Cryptocercus* (Yamin, 1979). Here, oxymonads are still not particularly common, but are a part of diverse assemblages of anaerobic protists, bacteria and archaea in many species of host. In these environments, they are thought to participate in complex metabolic interactions with other species, and highly adapted symbioses between oxymonads and prokaryotes from these communities are known (Brugerolle and Lee, 2000; Leander and Keeling, 2004). With one exception, oxymonads have not been cultivated, probably because the complexity of their interactions with the diverse microbial communities in which they are found makes it difficult to viably extract one species. The single exception is one of the morphologically simpler species found in vertebrates (chinchilla), *Monocercomonoides* strain PA203 (Hampl *et al*., 2005).

Oxymonads are currently considered to comprise five major lineages: the Polymastigidae (represented by *Monocercomonoides*), the Pyrsonymphidae (represented by *Pyrsonympha* and *Dinenympha*), the Oxymonadidae (represented by *Oxymonas*), the Saccinobaculidae (represented by *Saccinobaculus*) and the Streblomastigidae (represented by *Streblomastix*). Molecular data are sparse for all five groups, but sufficient to draw a few conclusions about the distribution of genetic codes. The genetic code of *Streblomastix* has been studied in the greatest detail, and the presence of a non-canonical code is supported by TAR codons at positions otherwise conserved for glutamine in all protein-coding genes identified so far by PCR, RT-PCR, and from an environmental expressed sequence tag (EST) project (Keeling and Leander, 2003; Slamovits and Keeling, 2006b). Moreover, in a genome using this code, all genes should terminate with TGA, which is the case for all 19 *Streblomastix* genes characterized to date (Keeling and Leander, 2003; Slamovits and Keeling, 2006b). Eight protein-coding genes have been sequenced from *Monocercomonoides* strain PA203 (Hampl *et al*., 2005): none uses non-canonical codons for glutamine, and all three canonical termination codons are found in at least one gene, together making a strong case that this species uses the universal code. Limited data are available for the other groups, but a gene for elongation factor 1α (EF-1α) that is attributed to an unidentified oxymonad is known (Moriya *et al*., 1998), and both EF-1α and α-tubulin genes are known from the closely related genera *Pyrsonympha* and *Dinenymphya* (Moriya *et al*. 2001), and all of these genes lack evidence for a non-canonical code.

Based on these data, the genetic code must have changed at some point within the oxymonad lineage, but our knowledge of its distribution is hampered by a poor understanding of oxymonad phylogeny and a lack of molecular data in genera. In particular, there is a total absence of verifiable protein-coding gene sequences from any member of the Saccinobaculidae or Oxymonadidae (the unattributed EF-1α is from *Reticulitermes speratus*, from which members of neither subgroup have been observed; Yamin, 1979). The Saccinobaculidae represents an especially difficult taxon to sample because they are restricted to single source in nature: the gut of the wood-eating cockroach, *Cryptocercus*. Despite this, the cell biology of *Saccinobaculus* has been studied in some detail because of its unique and visually arresting characteristics (Cleveland *et al*., 1934; Hollande and Carruette-Valentin, 1970; McIntosh, 1973): it is distinguished by a highly motile axostyle (its name is a translation of ‘snake in a bag’, reflecting the writhing of this structure within the cytoplasm; Cleveland *et al*., 1934). Within *Cryptocercus*, *Saccinobaculus* is relatively abundant and makes up one of the most conspicuous components of the gut flora. It is also highly diverse: early microscopic studies suggested as many as nine different species (Cleveland *et al*., 1934), exhibiting a great range of morphological characteristics (*e.g.* the size range is 14–170 μm). Besides this diverse population of *Saccinobaculus* (and many parabasalid protists and prokaryotes), *Cryptocercus* also contains polymastigid oxymonads. One species of *Monoceromonoides* (*M. globus*) has been described (Cleveland *et al*., 1934), and later argued to be two distinct species, *M. globus* and *Paranotila lata* (Cleveland, 1966). However, the distinction between the two is very weak [*P. lata* represents what were originally considered to be life-cycle stages of *M. globus* (Cleveland, 1966)] and *P. lata*-like cells have only been reported to be rare, in a small proportion of host individuals.

To overcome the technical limitations imposed by the inability to cultivate most of oxymonad diversity, we have used environmental PCR, RT-PCR and 3′ RACE to examine the distribution of the non-canonical genetic code in oxymonads from *Cryptocercus* (*i.e.* all known Saccinobaculidae and one or two species of Polymastigidae). We characterized a total of 253 distinct copies of the four protein-coding genes used to examine the non-canonical code in *Streblomastix*: α-tubulin, β-tubulin, EF-1α, and heat-shock protein 90 (HSP90). The vast majority of sequences apparently use the universal genetic code; however, a small number of α-tubulin genes were found to use the same code as *Streblomastix*. No member of the Streblomastigidae inhabits this environment, so the non-canonical code is therefore more broadly distributed than previously believed. We used single-cell isolation to characterize α-tubulin from the type species of *Saccinobaculus*, *S. ambloaxostylus*, and show that *Saccinobaculus* uses the universal code. By elimination, we hypothesize that the non-canonical code is used by a polymastigid, and propose that the code evolved once in the common ancestor of *Streblomastix* and a subgroup of polymastigids.

Results and discussion

Environmental sampling of protein-coding genes from oxymonads

The *Cryptocercus* gut environment is home to a unique diversity of oxymonads and is the only place to find one of the major groups, represented by *Saccinobaculus*. However, none of these organisms are cultivable, so molecular data are not easily obtained. To characterize the oxymonad molecular diversity of the *Cryptocercus* gut environment, and to determine whether the non-canonical genetic code is employed by any organisms in this environment, we used environmental RT-PCR and PCR to rapidly generate a large body of protein-coding data from this oxymonad population. To be sure that protein-coding genes can be accurately attributed to oxymonads and not parabasalia, we used genes that had been previously sampled in other oxymonads. Four protein-coding genes have been sampled from more than one oxymonad: α-tubulin, β-tubulin, EF-1α and HSP90. These four genes were also those used to characterize the genetic code in *Streblomastix* (Keeling and Leander, 2003). Amplifications from whole-gut RNA resulted in products of the expected size representing most of the conserved coding regions of all four genes (87–96% of α-tubulin, 83% of β-tubulin, 94% of EF-1α, and 80% of HSP90). These products were cloned, and 161, 25, 69 and 5 individual clones were sequenced on both strands, for α-tubulin, β-tubulin, EF-1α and HSP90 respectively. Amplifications from total DNA also yielded a product of the expected size for HSP90, from which 23 individual clones were sequenced. One α-tubulin and two β-tubulin clones were similar to insect genes, and specifically related to *Cryptocercus* (see below), whereas the remaining clones were clearly of protistan origin. *Streblomastix* has been found to have a relatively high frequency of introns in some genes (Slamovits and Keeling, 2006a), but not in the four genes sampled here (Keeling and Leander, 2003). In our sample, only one HSP90 clone amplified from DNA was found to contain a 49 bp canonical spliceosomal intron, and an apparently unspliced intron of 90 bp with canonical splice sites was detected in an EF-1α cDNA clone.

Most of the diversity we observed was at synonymous sites, although 126 unique protein sequences were found in total (the few sequences found to be identical to another sequence at the DNA level were removed from subsequent analyses). All sequences were examined for TAR codons, but only three clones were found to contain any premature stop codons. Interestingly, these three clones were closely related α-tubulin genes. One encoded a TAA at position 9 (numbered according to *Streblomastix*), one encoded a TAA at position 154, and one encoded a TAA at both positions (Fig. 1). Both positions are conserved for glutamine in other oxymonads and indeed in most other eukaryotic α-tubulins, strongly suggesting that these genes come from an organism using the same genetic code as *Streblomastix*.

Phylogeny of oxymonad genes

Phylogenetic trees, including a broad sampling of eukaryotic diversity, confirmed that the vast majority of the sequences branched within the oxymonad lineage, as expected. Only three sequences did not, and these all branched with *Cryptocercus*, confirming their origin from the host (not shown).

The diversity and relationships between the oxymonad sequences were assessed by inferring phylogenies of all four genes. The position of the root of the oxymonad tree has been discussed in detail elsewhere (Dacks *et al*., 2001; Moriya *et al*., 2003; Heiss and Keeling, 2006), and when we included the outgroup *Trimastix* (Dacks *et al*., 2001), we found the root to fall in any one of the major oxymonad subgroups, depending on the gene or analysis used (not shown). This suggests that the position of the root is likely not reliably determined from these data, so we restricted our subsequent analyses to oxymonads.

The relationships between many subgroups in the HSP90 and EF-1α phylogenies were resolved with high

	9	154	211
Oxymonad clone 6	CLEHGI*PDGQMP	SFTVYPSPQISNAVVEP	TNLNRLIAQVISSLTAS
Oxymonad clone 1	CLEHGIQPDGQMP	SFTVYPSP*ISNAVVEP	TNLNRLIAQVISSLTAS
Oxymonad clone 4	CLEHGI*PDGQMP	SFTVYPSP*ISNAVVEP	TNLNRLIAQVISSLTAS
Oxymonad clone R12		SFTVYPSPQISNAVVEP	TNLNRLIA*VISSLTAS
Saccinobaculus	CLEHGIQPDGQMP	SFTVYPSPQISNAVVEP	TNLNRLISQVISSLTAS
Streblomastix	CLEHGIQPDGQMP	SFTVYPSPQIATAVVEP	TNLNRLIGQVISSLTAS
Dinenympha	CQEHGIQPDGQMP	SFTVYPSPQISNAVVEP	TNLNRLISQVISFLTAS
Trimastix	CLEHGIQPDGQMP	SFTVYPSPQISTAVVEP	TNINRLIAQVISSLTAS
Trichomonas	CLEHGIQPDGQLP	EFTVYPSPQVSTAIVEP	TNLNRLIGQVVSSLTAS
Giardia	CLEHGIQHDGQMP	EFVVYPSPQIATAVVEP	TNLNRLIAQCISSITAS
Euglena	CLEHGIQPDGSMP	GFTIYPSPQISTAVVEP	TNLNRLIAQVISSLTAS
Trypanosoma	CLEHGIQPDGAMP	GYTVYPSPQVSTAVVEP	TNLNRLIGQVVSSLTAS
Paramecium	CLEHGIQPDGQMP	GFTIYPSPQVSTAVVEP	TNLNRLIAQVISSLTAS
Phytophthora	CLEHGIQPDGQMP	GFTIYPSPQVSTAVVEP	TNLNRLIAQVISSLTAS
Heterocapsa	CLEHGIQPDGQMP	SFTVWACPQVATAVVEP	TNLNRLLAQIISSLTAS
Bigelowiella	CLEHGIQPDGQMP	GFTVYPSPQVSTAVVEP	TNLNRLIAQVISSLTAS
Arabidopsis	CLEHGIQPDGQMP	GFTVYPSPQVSTSVVEP	TNLNRLVSQVISSLTAS
Monosiga	CLEHGIQPDGQMP	EFAVYPAPQVSTAVVEP	TNLNRLIAQVVSSVTAS
Homo	CLEHGIQPDGQMP	EFAIYPAPQVSTAVVEP	TNLNRLIGQIVSSITAS
Spizellomyces	CLEHGIQPDGQMP	EFSVYPAPQVSTAVVEP	TNLNRLIAQVVSSITAS

Fig. 1. TAA and TAG codons at positions conserved for glutamine (Q) in α-tubulin genes from oxymonads in *Cryptocercus*. Blocks from a protein alignment are shown with TAA and TAG codons in oxymonads represented by asterisks (*). Positions are numbered according to the Streblomastix sequence.

support, but the relationships between various groups in the tubulin trees were generally not resolved with any significance. In the case of α-tubulin (Fig. 2), the overall poor support may be because the gene is too highly conserved to contain sufficient information to resolve such close relationships. However, the α-tubulin tree is still interesting because of the sequences with the non-canonical genetic code. The great majority of the environ-

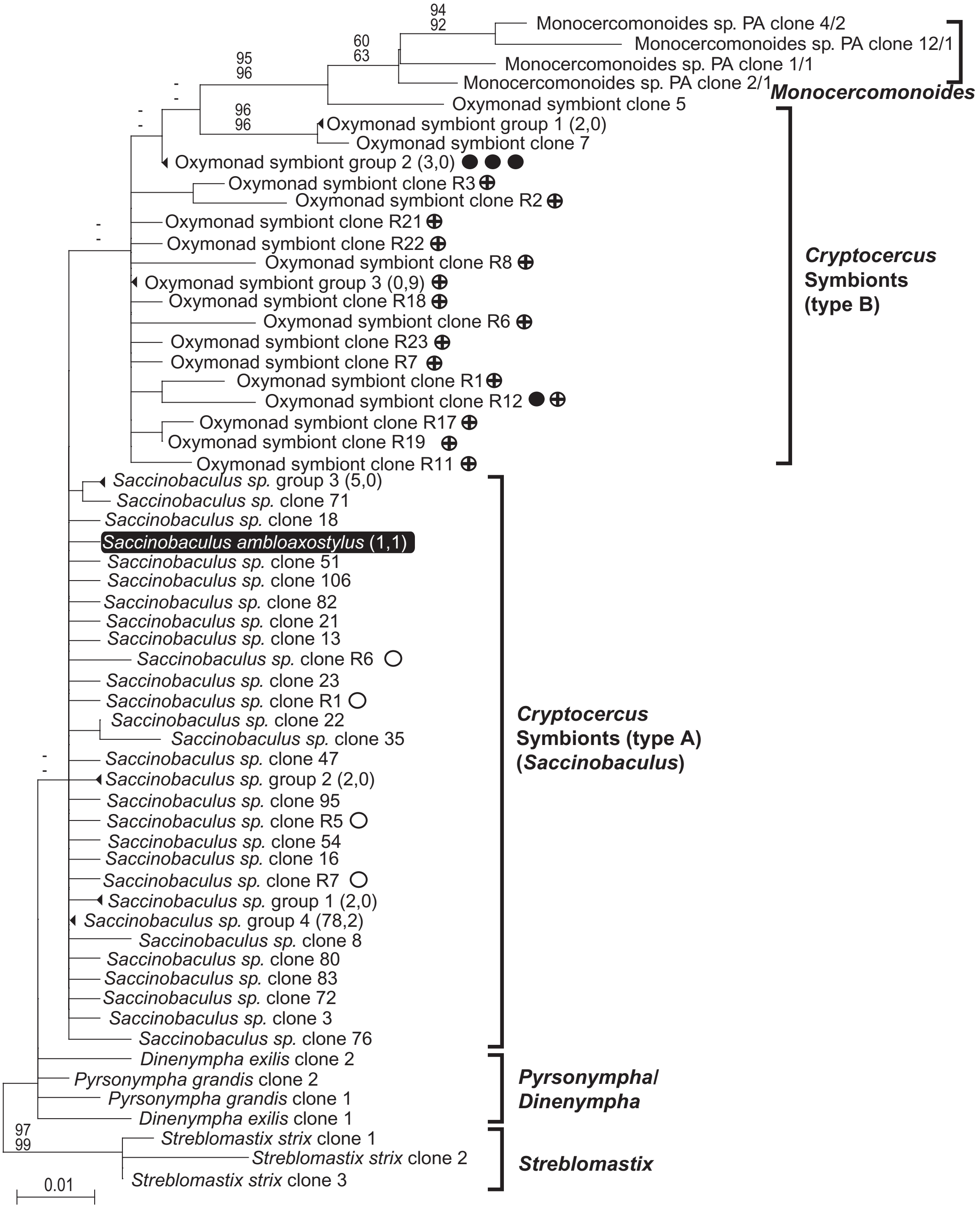

94
92
60
63
95
96
-
-
96
96
-
-
-
-
-
-
97
99
Monocercomonoides sp. PA clone 4/2
Monocercomonoides sp. PA clone 12/1
Monocercomonoides sp. PA clone 1/1
Monocercomonoides sp. PA clone 2/1
Monocercomonoides
Oxymonad symbiont clone 5
Oxymonad symbiont group 1 (2,0)
Oxymonad symbiont clone 7
Oxymonad symbiont group 2 (3,0) ● ● ●
Oxymonad symbiont clone R3 ⊕
Oxymonad symbiont clone R2 ⊕
Oxymonad symbiont clone R21 ⊕
Oxymonad symbiont clone R22 ⊕
Oxymonad symbiont clone R8 ⊕
Oxymonad symbiont group 3 (0,9) ⊕
Oxymonad symbiont clone R18 ⊕
Oxymonad symbiont clone R6 ⊕
Oxymonad symbiont clone R23 ⊕
Oxymonad symbiont clone R7 ⊕
Oxymonad symbiont clone R1 ⊕
Oxymonad symbiont clone R12 ● ⊕
Oxymonad symbiont clone R17 ⊕
Oxymonad symbiont clone R19 ⊕
Oxymonad symbiont clone R11 ⊕
Cryptocercus Symbionts (type B)
Saccinobaculus sp. group 3 (5,0)
Saccinobaculus sp. clone 71
Saccinobaculus sp. clone 18
Saccinobaculus ambloaxostylus (1,1)
Saccinobaculus sp. clone 51
Saccinobaculus sp. clone 106
Saccinobaculus sp. clone 82
Saccinobaculus sp. clone 21
Saccinobaculus sp. clone 13
Saccinobaculus sp. clone R6 ○
Saccinobaculus sp. clone 23
Saccinobaculus sp. clone R1 ○
Saccinobaculus sp. clone 22
Saccinobaculus sp. clone 35
Saccinobaculus sp. clone 47
Saccinobaculus sp. group 2 (2,0)
Saccinobaculus sp. clone 95
Saccinobaculus sp. clone R5 ○
Saccinobaculus sp. clone 54
Saccinobaculus sp. clone 16
Saccinobaculus sp. clone R7 ○
Saccinobaculus sp. group 1 (2,0)
Saccinobaculus sp. group 4 (78,2)
Saccinobaculus sp. clone 8
Saccinobaculus sp. clone 80
Saccinobaculus sp. clone 83
Saccinobaculus sp. clone 72
Saccinobaculus sp. clone 3
Saccinobaculus sp. clone 76
Cryptocercus Symbionts (type A) (Saccinobaculus)
Dinenympha exilis clone 2
Pyrsonympha grandis clone 2
Pyrsonympha grandis clone 1
Dinenympha exilis clone 1
Pyrsonympha/ Dinenympha
Streblomastix strix clone 1
Streblomastix strix clone 2
Streblomastix strix clone 3
Streblomastix
0.01

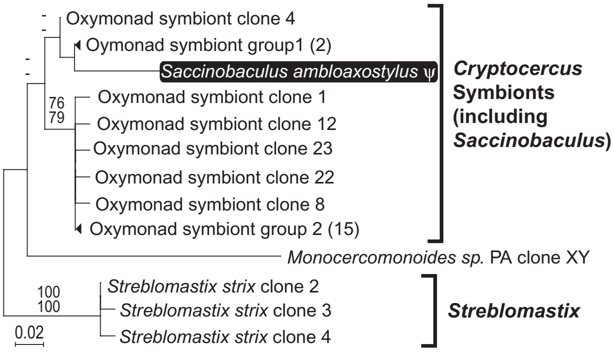
Fig. 2. Protein maximum likelihood phylogeny of α-tubulin from oxymonads. Environmental sequences that contain TAR codons for glutamine are indicated by filled circles, whereas open circles indicate the presence of TAA or TAG stop codons and circles with a cross indicate only TGA stop codons. Sequence from manually isolated cells is highlighted in black. Triangles indicate groups of clones whose DNA sequences conceptually translate into identical amino acid sequences. The numbers in brackets following the name of these groups indicate the number of cDNA and 3′ RACE clones respectively. Numbers at major nodes indicate per cent bootstrap support from PhyML (top) and ProML (bottom), with values less than 50 represented by a dash (-).

mental clones are virtually identical to one another (we will refer to these as type A sequences: Fig. 2), which form a large comb on the tree. The remaining sequences (which we will refer to as type B sequences: Fig. 2) form a unique but unsupported clade in all analyses with *Monocercomonoides* (although most of these relationships are not supported, clone 5 specifically branches with *Monocercomonoides* strain PA203 with 95–96% support).

The presence of two distinct types of α-tubulin sequences, one including all TAR codon-containing genes, immediately begs the question of which type belongs to which oxymonad. *Cryptocercus* is home to two subgroups of oxymonad, the Saccinobaculidae (represented by the abundant and species-rich genus *Saccinobaculus*) and the Polymastigidae (represented by the more rare single species *Monocercomonoides globus* and the taxonomically contentious, closely related and very rare *P. lata*). Polymastigids are not amenable to manual isolation because they are relatively rare in *Cryptocercus*, and so small that they are not easily distinguished from the similarly small species of *Saccinobaculus* under the conditions used to isolate cells. In contrast, the type species of Saccinobaculidae, *S. ambloaxostylus*, is large, abundant and readily distinguishable from all other species (Cleveland *et al*., 1934; Heiss and Keeling, 2006). Accordingly, we sought to identify which of the two types of α-tubulin corresponds to *Saccinobaculus*. We manually isolated 40 *S. amblaxostylus* cells and amplified the α-tubulin gene directly from these cells, resulting in a single protein sequence that branched with the type A sequences. In fact, the sequence was identical at the amino-acid level to one of the sequences isolated from environmental PCR (black box in Fig. 2) and differed at no more than three amino acids from any of the type A sequences. This confirms the identity of type A sequences as belonging to the Saccinobaculidae. By elimination, we hypothesize that the type B, TAR codon-containing genes come from polymastigids, which is also consistent with their affinity in all analyses with the polymastigid *Monocercomonoides*.

If type B sequences use the non-canonical code, they might also be expected to closely related to *S. strix*, as it too uses this code. However, in the α-tubulin tree the type B sequences are not closely related to those of *S. strix*, although there is no support for any branch separating them. If the origin of a new code is a very complex and rare event, we must assume it evolved only once within a small group like oxymonads unless there is compelling evidence to the contrary. We therefore directly compared alternative topologies of the α-tubulin tree to see whether such evidence exists in these data. Representatives of each major group (*S. strix* clone 3, *P. grandis* clone 2, *S. ambloaxostylus*, type B group 2, and *Monocercomonoides* clone 2/1) were chosen, and all possible topologies were compared using approximately unbiased (AU) tests. Overall, there was no strong pattern among those topologies that were rejected: from 15 possible trees, 6 were not rejected and in 3 of these, type B and *S. strix* formed a clade as a sister group to *Monocercomonoides*. The phylogeny of α-tubulin therefore does not exclude the possibility that oxymonads with the non-canonical code are sisters, and therefore α-tubulin phylogeny is not evidence for multiple origins of this code.

The remaining three phylogenies are not as informative as α-tubulin, because no *Cryptocercus* sequences have been linked to cells by isolation, the trees are poorly resolved, or poorly sampled, or all three. In β-tubulin (Fig. 3), no TAR codon-containing sequences were found, the gene is poorly sampled for other oxymonad groups, and the phylogeny not well resolved. A β-tubulin gene was also retrieved from the 40 isolated cells of *S. ambloaxostylus*, but all clones characterized were unfortunately determined to be from a single pseudogene that contains a large deletion. Aside from the deletion, the sequence differed at only a handful of positions from one clone from the environmental sample (Fig. 3). This suggests that at least the smaller subgroup of β-tubulin genes is from saccinobaculids. EF-1α (Fig. 4) is better sampled, and all new sequences form a single, strongly supported clade (type A, with 98–100% support) with one exception (type B, clone 23), which branches with *Pyrsonympha*, *Dinenympha* and an unidentified termite symbiont with low support (62–64%). Lastly, in HSP90 phylogeny (Fig. 5), new sequences also forms two distinct and well-supported groups: one large clade (type A, with 86–100%) that is sister to *Monocercomonoides* strain PA203, and a smaller clade (type B, with 98% support) that is sister to *Streblomastix*. Because of their position in the tree, the sequences designated type B in HSP90 and EF-1α are



-
-
Oxymonad symbiont clone 4
Oymonad symbiont group1 (2)
-
-
Saccinobaculus ambloaxostylus ψ
76
79
Oxymonad symbiont clone 1
Oxymonad symbiont clone 12
Oxymonad symbiont clone 23
Oxymonad symbiont clone 22
Oxymonad symbiont clone 8
Oxymonad symbiont group 2 (15)
Cryptocercus Symbionts (including Saccinobaculus)
Monocercomonoides sp. PA clone XY
100
100
Streblomastix strix clone 2
Streblomastix strix clone 3
Streblomastix strix clone 4
Streblomastix
0.02

Fig. 3. Protein maximum likelihood phylogeny of β-tubulin from oxymonads. Sequence from manually isolated cells is highlighted in black. Triangles indicate groups of clones whose DNA sequences conceptually translate into identical amino acid sequences. The numbers of clones that are contained in the groups are given in brackets following the group name. Numbers at major nodes indicate per cent bootstrap support from PhyML (top) and ProML (bottom), with values less than 50 represented by a dash (-).

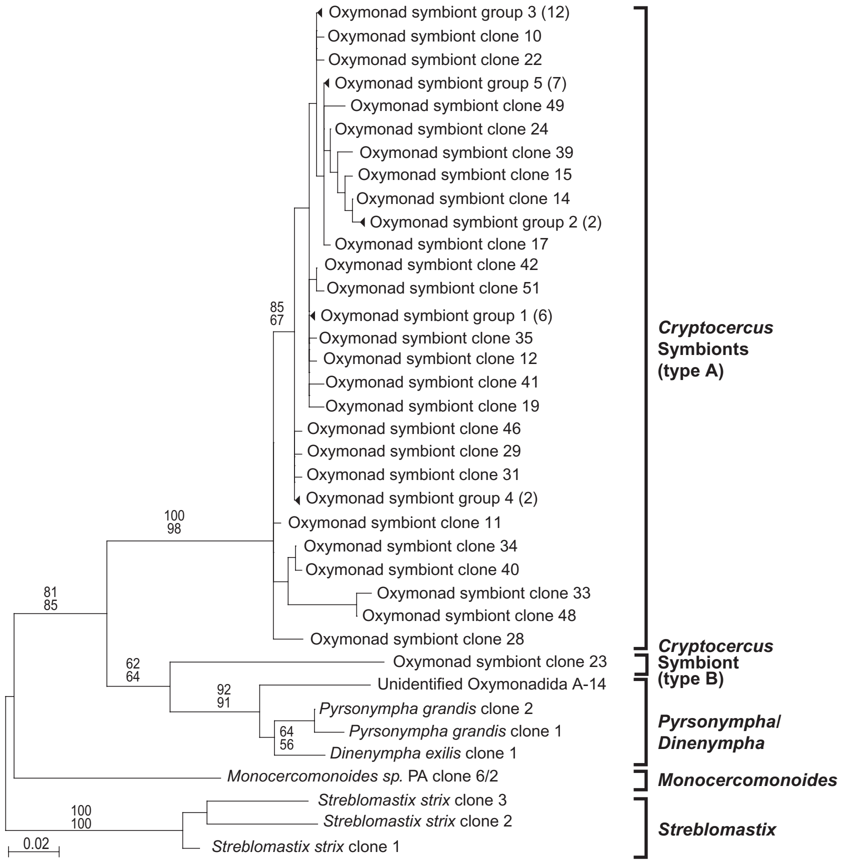

Oxymonad symbiont group 3 (12)
Oxymonad symbiont clone 10
Oxymonad symbiont clone 22
Oxymonad symbiont group 5 (7)
Oxymonad symbiont clone 49
Oxymonad symbiont clone 24
Oxymonad symbiont clone 39
Oxymonad symbiont clone 15
Oxymonad symbiont clone 14
Oxymonad symbiont group 2 (2)
Oxymonad symbiont clone 17
Oxymonad symbiont clone 42
Oxymonad symbiont clone 51
85
67
Oxymonad symbiont group 1 (6)
Oxymonad symbiont clone 35
Oxymonad symbiont clone 12
Oxymonad symbiont clone 41
Oxymonad symbiont clone 19
Oxymonad symbiont clone 46
Oxymonad symbiont clone 29
Oxymonad symbiont clone 31
Oxymonad symbiont group 4 (2)
100
98
Oxymonad symbiont clone 11
Oxymonad symbiont clone 34
Oxymonad symbiont clone 40
81
85
Oxymonad symbiont clone 33
Oxymonad symbiont clone 48
Oxymonad symbiont clone 28
62
64
Oxymonad symbiont clone 23
92
91
Unidentified Oxymonadida A-14
Pyrsonympha grandis clone 2
64
56
Pyrsonympha grandis clone 1
Dinenympha exilis clone 1
Monocercomonoides sp. PA clone 6/2
Streblomastix strix clone 3
100
100
Streblomastix strix clone 2
0.02
Streblomastix strix clone 1
***Cryptocercus* Symbionts (type A)**
***Cryptocercus* Symbiont (type B)**
Pyrsonympha*/ *Dinenympha
Monocercomonoides
Streblomastix

Fig. 4. Protein maximum likelihood phylogeny of EF-1α from oxymonads. Triangles indicate groups of clones whose DNA sequences conceptually translate into identical amino acid sequences. The numbers of clones that are contained in the groups are given in brackets following the group name. Numbers at major nodes indicate per cent bootstrap support from PhyML (top) and ProML (bottom).

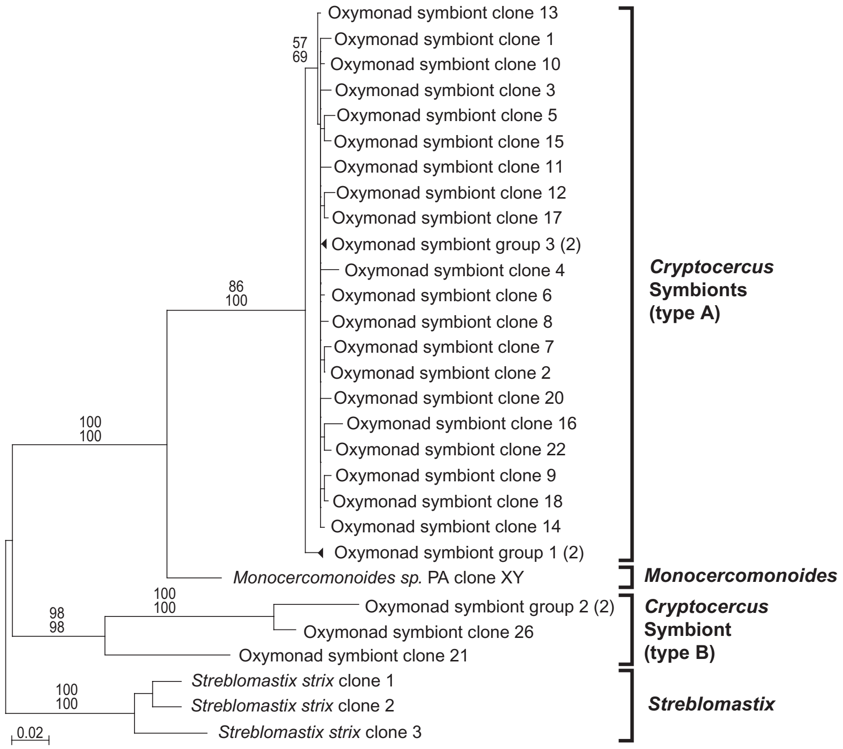

Oxymonad symbiont clone 13
57
69
Oxymonad symbiont clone 1
Oxymonad symbiont clone 10
Oxymonad symbiont clone 3
Oxymonad symbiont clone 5
Oxymonad symbiont clone 15
Oxymonad symbiont clone 11
Oxymonad symbiont clone 12
Oxymonad symbiont clone 17
Oxymonad symbiont group 3 (2)
Oxymonad symbiont clone 4
86
100
Oxymonad symbiont clone 6
Oxymonad symbiont clone 8
Oxymonad symbiont clone 7
Oxymonad symbiont clone 2
Oxymonad symbiont clone 20
100
100
Oxymonad symbiont clone 16
Oxymonad symbiont clone 22
Oxymonad symbiont clone 9
Oxymonad symbiont clone 18
Oxymonad symbiont clone 14
Oxymonad symbiont group 1 (2)
Monocercomonoides sp. PA clone XY
100
100
98
98
Oxymonad symbiont group 2 (2)
Oxymonad symbiont clone 26
Oxymonad symbiont clone 21
Streblomastix strix clone 1
100
100
Streblomastix strix clone 2
0.02
Streblomastix strix clone 3
***Cryptocercus* Symbionts (type A)**
Monocercomonoides
***Cryptocercus* Symbiont (type B)**
Streblomastix

Fig. 5. Protein maximum likelihood phylogeny of HSP90 from oxymonads. Triangles indicate groups of clones whose DNA sequences conceptually translate into identical amino acid sequences. The numbers of clones that are contained in the groups are given in brackets following the group name. Numbers at major nodes indicate per cent bootstrap support from PhyML (top) and ProML (bottom).

interesting candidates for using the non-canonical code, although no TAR codons are found in these sequences. In the four clones representing the three HSP90 sequences, only 48 glutamine codons were sampled, while in the single EF-1α outlier, only 8 glutamines are found. Comparing this with α-tubulin, there are five non-canonical codons out of a sample of 254 glutamine positions in all type B genes, for a frequency of 0.019 (assuming all genes designated as type B come from genomes using the non-canonical code). Not all genes in a genome have the same frequency of canonical and non-canonical codons (just as different genes differ in frequencies of other codon); however, if the frequency was similar, we would expect to find 0.91 and 0.15 non-canonical codons in existing samples of HSP90 and EF-1α respectively. It is therefore possible that these genes may also come from the same source as type B α-tubulin genes, but simply do not happen to utilize any TAA or TAG codons.

Termination codon use by oxymonad α-tubulins

If the TAR codon-containing (type B) α-tubulins use the non-canonical code, then they should all terminate with TGA. Conversely, all other α-tubulins inferred to use the universal code can use all three stop codons. We used environmental 3′ RACE to test this. TAR codon-containing α-tubulins were underrepresented in RNA sampling, so the termination codons of these genes were deliberately targeted using nested, gene-specific 5′ primers. Twenty-two products in the expected size range were cloned and sequenced, all of which were unique. These products all branched with type B sequences in α-tubulin trees (indicated by a circled cross in Fig. 2), and interestingly, one of the products encoded a TAA at the conserved glutamine at position 211 (clone R12, Figs 1 and 2). Most importantly, all 22 products terminated at the anticipated position using a TGA codon, followed by a short UTR and poly-A tail, exactly as predicted if they use the non-canonical code. The number of unique type B mRNAs characterized was large, but not so large as to rule out the possibility that they come from a single species given the source is a complex natural community, and it is also not outside the range of variability previously observed from similar samples taken from the single species, *S. strix* (Keeling and Leander, 2003).

Stop codon use in general was examined by doing 3′ RACE with the degenerate 5′ α-tubulin primer originally used to amplify cDNAs. In this case, all products branched with type A sequences, outside the clade where the non-canonical code is found (Fig. 2). Furthermore, all clones terminated at the expected position with either TAA or TAG, confirming these sequences cannot use the *Streblomastix* code. It is interesting that no TGA codon was found in any of the canonical code-using clones. It is most likely that it was simply not sampled, but it is an intriguing possibility that this codon has become unassigned, which would have interesting implications for codon evolution in the oxymonads as a whole. This seems unlikely, however, because *Monocercomonoides* strain PA203 uses TGA as a terminator (Hampl *et al.*, 2005).

Evolution of the genetic code in oxymonads

We have used environmental PCR to sample sequence diversity of uncultivable oxymonads from the *Cryptocercus* gut environment, and address the distribution of genetic codes in this poorly understood group of protists. The surveys revealed a great diversity of oxymonad sequences, which is in agreement with the obvious and visually arresting diversity of oxymonads evident from light microscopy in this environment, as well as sampling of SSU rRNA gene sequences (Heiss and Keeling, 2006). More surprisingly, however, the sequence survey also showed that the non-canonical genetic code first described among the oxymonads in *Streblomastix* is not restricted to this genus, but is also found in some other oxymonad in the complex *Cryptocercus* community. Two questions are immediately raised by this observation: which oxymonad uses the code, and how is the code distributed in the oxymonad tree?

The sequence of the α-tubulin from the manually isolated cells of *S. ambloaxostylus* confirms that the majority of sequences (type A) most likely come from the genus *Saccinobaculus* and, together with RACE data, this shows that *Saccinobaculus* uses the universal code. The only other oxymonad group reported in this cockroach is the less abundant polymastigids (Cleveland *et al.*, 1934). Recent phylogenies based on SSU rRNA have shown that the polymastigid *Monocercomonoides* strain PA203 is sister to *Streblomastix* (Hampl *et al.*, 2005; Heiss and Keeling, 2006), and this is also seen in the β-tubulin, EF-1α and HSP90 phylogenies reported here (albeit with poor sampling and/or support). Such a relationship has also been proposed previously, based on morphology of *Streblomastix* when it is treated with drugs that kill its surface symbionts (Leander and Keeling, 2004). At the same time, the α-tubulin phylogeny suggests a relationship between the type B, TAR codon-containing sequences and *Monocercomonoides* strain PA203 (Fig. 2). Taking all this together, we speculate that the family Polymastigidae is paraphyletic, with *Streblomastix* evolving from within this lineage. According to this hypothesis (Fig. 6), the non-canonical genetic code evolved within the polymastigids, so that some species (e.g. *Monocercomonoides* strain PA203) still use the universal code, whereas others (e.g. at least one polymastigid in *Cryptocercus*, perhaps the more common *M. globus*) and *Streblomastix* use the non-canonical code. This

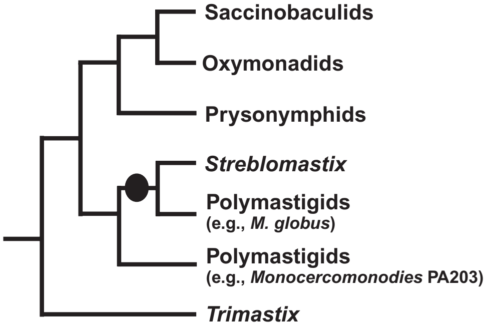


Fig. 6. Hypothetical relationships between oxymonad classes and the evolution of the non-canonical genetic code. The circle indicates the proposed timing of the development of the non-canonical genetic code within the polymastigids.

hypothesis is testable by the further characterization of *M. globus*, other species of polymastigid potentially present in *Cryptocercus*, and other species of *Monocercomonoides*. We predict that the non-canonical code will be found in other polymastigids closely related to *Streblomastix*.

Experimental procedures

Isolation of oxymonads

Wood-eating cockroaches (*Cryptocercus punctulatus*) were collected by C.A. Nelapa in North Carolina (Clifton in Ashe County and Bear Trap Gap in Haywood County), Virginia (Mountain Lake in Giles County) and Georgia (Black Rock Mountain in Rabun County). Animals were killed and the gut dissected to release the protist flora, which were suspended in Trager's Medium U (Trager, 1934). Protist cells were harvested by centrifugation, and total DNA was isolated using QIAamp DNA Mini Kit (Qiagen), following the manufacturer's protocol. Total RNA was isolated by resuspending harvested material in 1 ml of Trizol (Invitrogen) and transferring it to a Kontes Duall 20 tissue homogenizer. Material was ground for 5 min and incubated for 5 min at room temperature without grinding. Lysate was extracted with 200 μl of chloroform : isoamyl alcohol (24:1), and the aqueous phase precipitated with 500 μl isopropanol. Material used to generate all cDNA sequences (all four genes) was taken from Clifton populations. Material used to generate all genomic DNA sequences (HSP90 only) was taken from Mountain Lake populations. Material used to generate all 3′ RACE sequences (α-tubulin only) was taken from Bear Trap Gap and Mountain Lake populations, and were pooled prior to amplification.

Environmental PCR sampling

Four protein-coding genes were amplified to sample oxymonad diversity at several loci using primers known to amplify oxymonad genes but not genes from the other major gut constituent, parabasalia (Keeling and Leander, 2003). Alpha-tubulin was amplified using primers GGGCCCCAG GTCGGCAAYGCNTGYTGG and GGGCCCCGAGAACTC SCCYTCYTCCAT, β-tubulin using primers GCCTGCA GGNCARTGYGGNAAYCA and TCCTCGAGTRAAYTCC ATYTCRTCCAT, EF-1α using primers AACATCGTCGT GATHGGNCAYGTNGA and CTTGATCACNCCNACNGC NACNGT, and HSP90 using primers GGAGCCTGATH ATHAAYACNTTYTA and CGCCTTCATDATNCKYTCCATR TTNGC. For HSP90, partial gene sequences were amplified by PCR using the following conditions: initial denaturation at 95°C, followed by 35 cycles of 95°C for 1 min, 52°C for 1 min and 72°C for 2 min, and then a final elongation step of 5 min. Partial cDNA sequences of all four genes were amplified by one-step RT-PCR using the following conditions: cDNA synthesis at 50°C for 30 min, an initial denaturation at 94°C for 2 min, and then 35 cycles of 94°C for 15 s, annealing for 30 s (at 55°C for α-tubulin and β-tubulin, 50°C for EF-1α, or 42°C for HSP90), 72°C for 2 min, and a final extension at 72°C for 5 min. Products of the expected size (or larger in the case of DNA-based amplifications) were gel isolated and cloned using Topo TA cloning (Invitrogen).

The termination codons of α-tubulin genes were determined by 3′ RACE on total RNA. Gene-specific nested primers ACTGGTCTCCAAGGCTTCTTAGT (outer) and ACAGGTGCTGGTCTTGGATCTCT (inner) were used to amplify fragments that included 923 bp of the coding region, and a degenerate primer, CGCGGCCTCARGTNGGNAAY GCNTGYTGGGA, was used to amplify fragments that included 1295 bp of the coding region. All sequences amplified in this study have been submitted to the GenBank databases under accession numbers DQ924974–DQ925226.

Single-cell isolation

Forty individual cells with distinctive characteristics of *S. ambloaxostylus* were isolated from a single cockroach from the Black Rock Mountain population and DNA isolated as described (Heiss and Keeling, 2006). This was used in a multiplex amplification with primers for all four genes (using primers described above), the product of which was split and each gene was re-amplified using a single primer pair. Products of the expected size were only observed for α-tubulin and β-tubulin, which were cloned and sequenced as above.

Phylogenetic analysis

All new sequences were added to existing protein alignments (Keeling and Leander, 2003). Protein phylogenies were inferred by maximum likelihood using PhyML 2.4.4 (Guindon and Gascuel, 2003) and ProML 3.6 (Felsenstein, 1993). PhyML trees were inferred with the WAG substitution matrix and site-to-site rate variation modelled on a gamma distribution based on four variable rate categories and invariable sites, with α- and i-parameters estimated from the data. For α-tubulin, β-tubulin, EF-1α and HSP90, the α-parameters were 2.04, 0.40, 1.30 and 4.24 respectively. The i-parameters were 0.068, 0.184, 0.218 and 0.554 respectively. One thousand bootstrap replicates were inferred in the same way with parameters estimated from the original data. ProML analyses were carried out using the JTT substitution matrix, global rearrangements and randomized input order and 1000 bootstrap replicates performed in the same way. Data sets consisted of 61 taxa and 243 sites (α-tubulin), 13 taxa and 378 sites (β-tubulin), 38 taxa and 333 sites (EF-1α), and 30 taxa and 518 sites (HSP90).

The relationship between type A (*Saccinobaculus*) and type B (TAR codon-containing) α-tubulins and other oxymonad groups was tested with the AU test using CONSEL 1.19 (Shimodaira and Hasegawa, 2001; Shimodaira, 2002). The lack of support for the overall tree made it impossible to exhaustively test meaningful alternatives using the whole data set, so we selected one sequence from each major group (the manually isolated sequence in the case of *Saccinobaculus* and the slowest evolving one in all other cases), and tested all possible relationships of those sequences. These representatives were: *S. strix* clone 3, *P. grandis* clone 2, *S. ambloaxostylus* (type A), Oxymonad symbiont 2 (type B – TAR codon-containing), and *Monocercomonoides* PA clone 2/1. Tests were also run using the outgroup *Trimastix pyriformis*. Because this limited data set did not include any 3′ RACE products, the entire gene could be used, so both the original data (243 sites) and the whole gene (367 sites) were tested. In total, 15 or 105 trees were generated (depending on whether or not *Trimastix* was included) using PAUP* 4.0b10 (Swofford, 2002) and AU tests performed using α- and i-parameters calculated using PhyML. Trees were sorted according to whether they were rejected at 5% confidence or not, and the two pools of trees were examined manually and by creating a consensus of trees that were either rejected or failed to be rejected.

Acknowledgements

This work was supported by a grant (227301) from the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank C. Nelapa for providing specimens of *Cryptocercus punctulatus*, and A. Roger for unpublished sequences from *Trimastix pyriformis*. A.P.D.K. is supported by the Canadian Institutes for Health Research (CIHR) and Michael Smith Foundation for Health Research (MSFHR). G.P.N. and J.W. were supported by Summer Scholarships from NSERC. P.J.K. is a Fellow of the Canadian Institute for Advanced Research and a MSFHR Senior Investigator.

References

Andersson, S.G., and Kurland, C.G. (1995) Genomic evolution drives the evolution of the translation system. *Biochem Cell Biol* **73:** 775–787.

Brugerolle, G., and Lee, J.J. (2000) Order Oxymonadida. In *The Illustrated Guide to the Protozoa.* Lee, J., Leedale, G.F. & Bradbury, P. (eds). Lawrence, KS, USA: J. Allen Press, pp. 1186–1195.

Cleveland, L.R. (1966) Nuclear division without cytokinesis followed by fusion of pronuclei in *Paranotila lata* General et sp. nov. *J Protozool* **13:** 132–136.

Cleveland, L.R., Hall, S.R., Sanders, E.P., and Collier, J. (1934) The wood-feeding roach *Cryptocercus*, its protozoa and the symbiosis between protozoa and roach. *Mem Am Acad Arts Sci* **17:** 185–342.

Dacks, J.B., Silberman, J.D., Simpson, A.G., Moriya, S., Kudo, T., Ohkuma, M., and Redfield, R.J. (2001) Oxymonads are closely related to the excavate taxon *Trimastix*. *Mol Biol Evol* **18:** 1034–1044.

Felsenstein, J. (1993) *PHYLIP (Phylogeny Inference Package).* Seattle, WA, USA: University of Washington.

Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.

Hampl, V., Horner, D.S., Dyal, P., Kulda, J., Flegr, J., Foster, P.G., and Embley, T.M. (2005) Inference of the phylogenetic position of oxymonads based on nine genes: support for metamonada and excavata. *Mol Biol Evol* **22:** 2508–2518.

Heiss, A.A., and Keeling, P.J. (2006) The phylogenetic position of the oxymonad *Saccinobaculus* based on SSU rRNA. *Protist* **157:** 335–344.

Hollande, A., and Carruette-Valentin, J. (1970) La lignée des Pyrsonymphines et les caractères infrastructuraux communs aux genres *Opisthomitus*, *Oxymonas*, *Saccinobacculus*, *Pyrsonympha* et *Streblomastix*. *C R Acad Sci Paris* **270:** 1587–1590.

Keeling, P.J., and Doolittle, W.F. (1996) A non-canonical genetic code in an early diverging eukaryotic lineage. *EMBO J* **15:** 2285–2290.

Keeling, P.J., and Doolittle, W.F. (1997) Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol Biol Evol* **14:** 895–901.

Keeling, P.J., and Leander, B.S. (2003) Characterisation of a non-canonical genetic code in the oxymonad *Streblomastix strix*. *J Mol Biol* **326:** 1337–1349.

Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001) Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet* **2:** 49–58.

Kulda, J., and Nohynkova, E. (1978) Flagellates of human intestine and of intestines of other species. In *Parasitic Protozoa*. Kreier, J.P. (ed.). New York, NY, USA: Academic Press, pp. 2–138.

Leander, B.S., and Keeling, P.J. (2004) Symbiotic innovation in the oxymonad *Streblomastix strix*. *J Eukaryot Microbiol* **51:** 291–300.

Lozupone, C.A., Knight, R.D., and Landweber, L.F. (2001) The molecular basis of nuclear genetic code change in ciliates. *Curr Biol* **11:** 65–74.

McIntosh, J.R. (1973) The Axostyle of *Saccinobacculus* II. Motion of the microtubule bundle and a structural comparison of straight and bent axostyles. *J Cell Biol* **56:** 324–339.

Moriya, S., Ohkuma, M., and Kudo, T. (1998) Phylogenetic position of symbiotic protist *Dinenympha* [correction of *Dinemympha*] exilis in the hindgut of the termite *Reticulitermes speratus* inferred from the protein phylogeny of elongation factor 1 alpha. *Gene* **210:** 221–227.

Moriya, S., Tanaka, K., Ohkuma, M., Sugano, S., and Kudo, T. (2001) Diversification of the microtubule system in the early stage of eukaryote evolution: elongation factor 1 alpha and alpha-tubulin protein phylogeny of termite symbiotic oxymonad and hypermastigote protists. *J Mol Evol* **52:** 6–16.

Moriya, S., Dacks, J.B., Takagi, A., Noda, S., Ohkuma, M., Doolittle, W.F., and Kudo, T. (2003) Molecular phylogeny of three oxymonad genera: *Pyrsonympha*, *Dinenympha* and *Oxymonas*. *J Eukaryot Microbiol* **50:** 190–197.

Nie, D. (1950) Morphology and taxonomy of the intestinal protozoa of the guinea-pig *Cavia porcella*. *J Morphol* **86:** 391–493.

Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiol Rev* **56:** 229–264.

Schneider, S.U., and de Groot, E.J. (1991) Sequences of two rbcS cDNA clones of *Batophora oerstedii*: structural and evolutionary considerations. *Curr Genet* **20:** 173–175.

Shimodaira, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51:** 492–508.

Shimodaira, H., and Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17:** 1246–1247.

Slamovits, C.H., and Keeling, P.J. (2006a) Pyruvate-phosphate dikinase of oxymonads and parabasalia and the evolution of pyrophosphate-dependent glycolysis in anaerobic eukaryotes. *Eukaryot Cell* **5:** 148–154.

Slamovits, C.H., and Keeling, P.J. (2006b) A high density of ancient spliceosomal introns in oxymonad excavates. *BMC Evol Biol* **6:** 34.

Swofford, D.L. (2002) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sunderland, MA, USA: Sinauer Associates.

Trager, W. (1934) The cultivation of a cellulose- digesting flagellate. *Trichomonas termopsidis*, and of certain other termite protozoa. *Biol Bull* **66:** 182–190.

Yamao, F., Muto, A., Kawauchi, Y., Iwami, M., Iwagami, S., Azumi, Y., and Osawa, S. (1985) UGA is read as tryptophan in *Mycoplasma capricolum*. *Proc Natl Acad Sci USA* **82:** 2306–2309.

Yamin, M.A. (1979) Flagellates of the orders Trichomonadida Kirby, Oxymonadida Grassé, and Hypermastigida Grassi & Foà reported from lower termites (Isoptera Falilies Mastotermitidae, Kalotermitidae, Hodotermitidae, Termopsidae, Rhinotermitidae, and Serritermididae) and from the wood-feeding roach *Cryptocercus* (Dictyoptera: Ceyptocercidae). *Sociobiology* **4:** 1–120.

Yarus, M., Caporaso, J.G., and Knight, R. (2005) Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem* **74:** 179–198.

Yokobori, S., Suzuki, T., and Watanabe, K. (2001) Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. *J Mol Evol* **53:** 314–326.