lysate, with or without added protease inhibitors. The FLAG–ACT–DHOD–MYC was decreased time-dependently, in the absence of inhibitors, and 35-kDa FLAG–ACT and 35-kDa DHOD–MYC appeared (Fig. 3A and B). Addition of FMN did not affect the increase in FLAG–ACT and DHOD–MYC. The processing from 70-kDa ACT–DHOD to independent ACT and DHOD was completely blocked by the high concentration of protease inhibitor cocktail with Triton-X (Fig. 3A and B). These results indicated that the 35-kDa DHOD arose from the fused ACT–DHOD in *B. saliens*.

Bodo saliens DHOD domain of ACT–DHOD was determined by alignment with trypanosomatid DHODs. The size of the purified recombinant DHOD (V334–E648) is about 35-kDa, and the protein showed yellow color, suggesting the presence of co-factor FMN. Its activity was determined by measuring orotate production in the presence of electron acceptors (with fumarate at 300 nm or with ubiquinone-1 at 287 nm), as described [4,5]. The recombinant protein revealed the high DHOD activity despite of a partial protein of ACT–DHOD (Table 2), depending on fumarate as electron acceptor that resembles the recombinant *T. cruzi* DHOD activity.

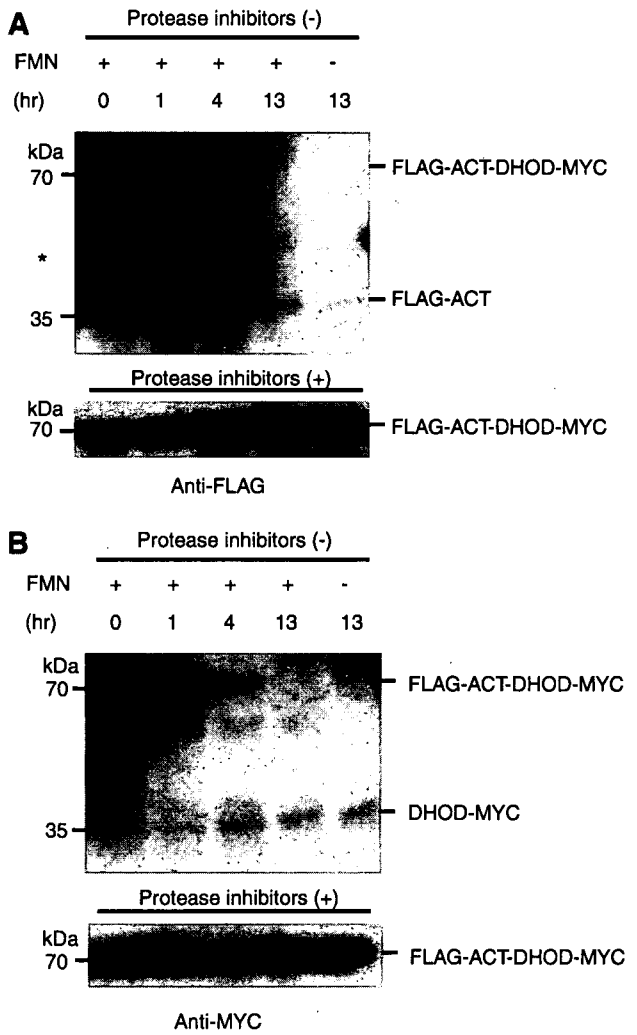


Fig. 3. In vitro processing of the recombinant FLAG–ACT–DHOD–MYC. The recombinant protein was incubated with *Bodo saliens* lysate in the absence or presence of the protease inhibitors. FMN, flavin mononucleotide. Aliquots (25 μg proteins) were withdrawn at 1, 4, and 13 h, and subjected to Western blot analysis using anti-FLAG (A) or anti-MYC (B) antibody. The asterisk is a non-specific band common to kinetoplastids.

Table 2
Comparison of activities of *Bodo saliens* and *Trypanosoma cruzi* recombinant DHODs

Enzyme source	Activities ($nmol^{-1}$ min^{-1} mg^{-1}) with fumarate (500 μM)	Activities ($nmol^{-1}$ min^{-1} mg^{-1}) with ubiquinone-1 (20 μM)
B. saliens (V334-E648)	3318 ± 1.7	987 ± 3.6
T. cruzi	2950 ± 2.1	840 ± 3.3

Assays were performed at 25 °C. Enzyme activities are shown as means ± SD of three independent measurements. *T. cruzi* DHOD activities are cited from [4].

Discussion

We have shown here that, in the kinetoplastid protist *B. saliens*, the *ACT–DHOD* gene is transcribed to a single *ACT–DHOD* mRNA, that its primary translation product is ACT–DHOD, that 35-kDa DHOD arose from ACT–DHOD in an in vitro processing assay, and finally that post-translational processing results in the production of N-terminal blocked mature DHOD. To our knowledge, the *ACT–DHOD* fused gene product and its maturation process is reported here for the first time, but the final product of *B. saliens* 35-kDa DHOD is highly homologous to the *T. cruzi* DHOD [8], the latter showing no such a maturation process.

Western blotting of *B. saliens* extract with anti-ACT-peptide and anti-DHOD-peptide antibodies resulted in a strong 70-kDa band and weak 60- and 35-kDa bands with the former, and a strong 35-kDa band and weak 70- and 85-kDa bands with the latter. These faint bands may carry DHOD or ACT protein. Immunoprecipitation and MS/MS analysis demonstrated the presence of DHOD in the 70- and 85-kDa bands and ACT in the 70-kDa band. The ACT- and DHOD-specific signal intensities may be roughly normalized using the 70-kDa band as standard, suggesting that the quantity of the 35-kDa mature DHOD in *B. saliens* may be much larger than the quantity of the DHOD domains in the primary 70-kDa ACT–DHOD translation product. When we applied this method to *B. saliens* ACT, we found that the quantity of the ACT domain in 70-kDa ACT–DHOD was relatively small, suggesting that the level of 35-kDa ACT protein is very low. Alternatively, a small amount of the native 35-kDa ACT raises the possibility that it is susceptible to endogenous proteolysis [17,18]. Probably because of a small quantity of the 85-kDa protein, our extensive searches only detected a DHOD-specific, but not ACT-specific, polypeptide with weak signal intensities (Fig. 4S, C). There are two

possibilities of the components of 85-kDa protein, one is ACT–DHOD (70-kDa) with some moiety (15-kDa), and the other DHOD (35-kDa) with some moiety (50-kDa). In this context, however, we found consensus sequence of the covalent binding with SUMO (small ubiquitin-related modifier) on *B. saliens* ACT domain (Annoura, unpublished). Interestingly, neither *T. cruzi* nor *Leishmania*, which carries independent *ACT* and *DHOD* genes, possesses such a binding site on ACT, suggesting the need for further study of the processing and maturation of ACT, the biological significance of the primary translation product, ACT–DHOD protein, and of independent DHOD and ACT in *B. saliens*.

We attempted to prepare the full-length recombinant ACT–DHOD protein in *Escherichia coli*, resulting in inclusion bodies that were not suitable for the enzymatic assay. However, high levels of 35-kDa mature DHOD exist in *B. saliens* cytosolic fraction (Figs. 1B and 2B) with N-terminal amino acid blocked. An in vitro processing assay clearly indicated that ACT–DHOD was processed by some component(s) in the *B. saliens* lysate in a short time, yielding 35-kDa DHOD. Moreover, the kinetic properties of *B. saliens* recombinant DHOD (V334-E648) is similar to those of the *T. cruzi* DHOD [4]. These results strongly suggested that *B. saliens* 35-kDa DHOD, which resembles *T. cruzi* enzyme, is a functional mature protein and may play an important role for both pyrimidine biosynthesis and fumarate reduction in controlling the cellular redox state.

Acknowledgments

This work was supported in part by Grants-in-Aid for scientific research (Nos. 18890188, 17390123, and 17590377) from the Ministry of Education, Sports, Culture, Science, and Technology of Japan. T. Annoura and T. Aoki were supported by a Grant-in-Aid for 21st Century COE Research from the Ministry of Education, Sports, Culture, Science, and Technology of Japan.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.04.102.

References

[1] A.G. Simpson, J. Lukes, A.J. Roger, The evolutionary history of kinetoplastids and their kinetoplasts, Mol. Biol. Evol. 19 (2002) 2071–2083.

[2] A.G. Simpson, J.R. Stevens, J. Lukes, The evolution and diversity of kinetoplastid flagellates, Trends Parasitol. 22 (2006) 168–174.

[3] T. Nara, T. Hshimoto, T. Aoki, Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes, Gene 257 (2000) 209–222.

[4] E. Takashima, D.K. Inaoka, A. Osanai, T. Nara, M. Odaka, T. Aoki, K. Inaka, S. Harada, K. Kita, Characterization of the dihydroorotate dehydrogenase as a soluble fumarate reductase in *Trypanosoma cruzi*, Mol. Biochem. Parasitol. 122 (2002) 189–200.

[5] I. Sariego, T. Annoura, T. Nara, M. Hashimoto, A. Tsubouchi, K. Iizumi, T. Makiuchi, E. Murata, K. Kita, T. Aoki, Genetic diversity and kinetic properties of *Trypanosoma cruzi* dihydroorotate dehydrogenase isoforms, Parasitol. Int. 55 (2006) 11–16.

[6] T. Nara, Y. Hirayama-Noguchi, G. Gao, E. Murai, T. Annoura, T. Aoki, Diversity of aspartate carbamoyltransferase genes of *Trypanosoma cruzi*, Int. J. Parasitol. 33 (2003) 845–852.

[7] G. Gao, T. Nara, J. Nakajima-Shimada, T. Aoki, Novel organization and sequences of five genes encoding all six enzymes for de novo pyrimidine biosynthesis in *Trypanosoma cruzi*, J. Mol. Biol. 285 (1999) 149–161.

[8] T. Annoura, T. Nara, T. Makiuchi, T. Hashimoto, T. Aoki, The origin of dihydroorotate dehydrogenase genes of kinetoplastids, with special reference to their biological significance and adaptation to anaerobic, parasitic conditions, J. Mol. Evol. 60 (2005) 113–127.

[9] A. Stechmann, T. Cavalier-Smith, Rooting the eukaryote tree by using a derived gene fusion, Science 297 (2002) 89–91.

[10] S. Shallom, K. Zhang, L. Jiang, P.K. Rathod, Essential protein-protein interactions between Plasmodium falciparum thymidylate synthase and dihydrofolate reductase domains, J. Biol. Chem. 274 (1999) 37781–37786.

[11] D.R. Evans, H.I. Guy, Mammalian pyrimidine biosynthesis: fresh insights into an ancient pathway, J. Biol. Chem. 279 (2004) 33035–33038.

[12] J.N. Davidson, K.C. Chen, R.S. Jamison, L.A. Musmanno, C.B. Kern, The evolutionary history of the first three enzymes in pyrimidine biosynthesis, Bioessays 15 (1993) 157–164.

[13] L.R. Livingstone, M.E. Jones, The purification and preliminary characterization of UMP synthase from human placenta, J. Biol. Chem. 262 (1987) 15726–15733.

[14] M. Suchi, H. Mizuno, Y. Kawai, T. Tsuboi, S. Sumi, K. Okajima, M.E. Hodgson, H. Ogawa, Y. Wada, Molecular cloning of the human UMP synthase gene and characterization of point mutations in two hereditary orotic aciduria families, Am. J. Hum. Genet. 60 (1997) 525–539.

[15] M. Hashimoto, J. Nakajima-Shimada, T. Aoki, *Trypanosoma cruzi* posttranscriptionally up-regulates and exploits cellular FLIP for inhibition of death-inducing signal, Mol. Biol. Cell 16 (2005) 3521–3528.

[16] R. Mineki, H. Taka, T. Fujimura, M. Kikkawa, N. Shindo, K. Murayama, In situ alkylation with acrylamide for identification of cysteinyl residues in proteins during one- and two-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis, Proteomics 2 (2002) 1672–1681.

[17] T. Asai, W.J. O'Sullivan, M. Kobayashi, A.M. Gero, M. Yokogawa, M. Tatibana, Enzymes of the de novo pyrimidine biosynthetic pathway in *Toxoplasma gondii*, Mol. Biochem. Parasitol. 7 (1983) 89–100.

[18] I.A. Mejias-Torres, B.H. Zimmermann, Molecular cloning, recombinant expression and partial characterization of the aspartate transcarbamoylase from *Toxoplasma gondii*, Mol. Biochem. Parasitol. 119 (2002) 191–201.

Accepted for publication in *Protist*, 4 Feb 2008

Evolutionary analysis of synteny and gene fusion for pyrimidine biosynthetic enzymes in Euglenozoa: An extraordinary gap between kinetoplastids and diplonemids

Takashi Makiuchi [a], Takeshi Annoura [a], Tetsuo Hashimoto [b], Eri Murata [a], Takashi Aoki [a], and Takeshi Nara [a,1]

[a] Department of Molecular and Cellular Parasitology, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan
[b] Institute of Biological Sciences, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-8572, Japan

Running title: Evolution of synteny and gene fusion in Euglenozoa

[1]Corresponding author;
fax +81 3 5800 0476
e-mail tnara@med.juntendo.ac.jp (T. Nara).

Abstract

Unique genome architecture in the parasitic protists trypanosomatids is large-scale synteny. We addressed the evolutionary trait of synteny in the eukaryotic group, Euglenozoa, which consists of euglenoids (earliest branching), diplonemids, and kinetoplastids (trypanosomatids and bodonids). Synteny of the pyrimidine biosynthetic (*pyr*) gene cluster, which constitutes part of the large syntenic cluster in trypanosomatids and includes four separate genes (*pyr1-pyr4*) and one fused gene (*pyr6/pyr5* fusion), was conserved in the bodonid, *Parabodo caudatus*. In the diplonemid, *Diplonema papillatum*, we identified *pyr4* and *pyr6* genes. Phylogenetic analyses of pyr4 and pyr6 showed the separate origin of each in kinetoplastids and euglenoids/diplonemids and suggested that kinetoplastids have acquired these genes via lateral gene transfer (LGT). Because replacement of genes by non-orthologs within the syntenic cluster is highly unlikely, we concluded that, after separation of the line leading to diplonemids, the syntenic *pyr* gene cluster was established in a common ancestor of kinetoplastids, preceded by their acquisition via LGT. Notably, we found that diplonemid *pyr6* is a stand-alone gene, inconsistent with both euglenoid *pyr5/pyr6* and kinetoplastid *pyr6/pyr5* fusions. Our findings provide insights into the evolutionary gaps within Euglenozoa and into the evolutionary trait of rearrangement of gene fusion in this lineage.

Key Words: Euglenozoa; Gene cluster; Kinetoplastids; Lateral gene transfer; Pyrimidine biosynthetic enzymes; Synteny.

Introduction

Trypanosomatids are flagellated parasitic protists and include medically important pathogens, such as those causing Chagas' disease, African sleeping sickness, and leishmaniasis. The phylogenetic position of trypanosomatids has been extensively studied using molecular phylogeny. Species of trypanosomatids are monophyletic, and this clade is nested in the kinetoplastid clade with bodonids, the sister group of trypanosomatids. Kinetoplastids, together with euglenoids and diplonemids, are assembled into a large monophyletic group, Euglenozoa, which is characterized by distinctive mitochondria with discoid cristae (Cavalier-Smith 1981). Of the three branches of Euglenozoa, euglenoids constitute the earliest branch, followed by separation of the diplonemid and kinetoplastid lineages (Simpson et al. 2002; Simpson and Roger 2004).

Synteny, the preserved order of genes, is often observed in the genomes of phylogenetically related eukaryotic species. Conserved synteny is utilized as an evolutionary marker, which can indicate not only different species to be descent from a common ancestor, but also functional and/or evolutionary relationships of the clustered genes (Bennetzen and Freeling 1997; Nadeau 1989).

Comparative genomics of three trypanosomatids, *Trypanosoma cruzi, T. brucei*, and *Leishmania major*, have highlighted large-scale synteny of polycistronic gene clusters as a feature of their unique genome architecture (El-Sayed et al. 2005). That is, protein-encoding genes, most of which are functionally unrelated, are tandemly arrayed on either strand of DNA as syntenic gene clusters and constitute polycistronic transcription units (Bonen 1993; Liang et al. 2003; Martínez-Calvillo et al. 2004). Due largely to the lack of genomic information on other euglenozoan groups, however, the origin of conserved synteny and gene clustering in trypanosomatids has not yet been determined (Dávila and Lukeš 2003; Jackson et al. 2006).

The de novo pyrimidine biosynthetic pathway is one of the essential catalytic activities in organisms, which produces uridine-5'-monophosphate (UMP) for incorporation into DNA and RNA. This pathway consists of six enzymes: pyr1 (EC 6.3.5.5, carbamoyl-phosphate synthetase II), pyr2 (EC 2.1.3.2, aspartate carbamoyltransferase), pyr3 (EC 3.5.2.3, dihydroorotase), pyr4 (EC 1.3.3.1, dihydroorotate dehydrogenase), pyr5 (EC 2.4.2.10, orotate phosphoribosyltransferase), and pyr6 (EC 4.1.1.23, orotidine-5'-monophosphate decarboxylase), in their order of reaction.

We previously showed that all *pyr* genes are clustered in the genomes of two trypanosomatid species, *T. cruzi* (Gao et al. 1999) and *L. mexicana* (GenBank™ Accession number AB029444), in both of which five genes, *pyr1, pyr3, pyr6/pyr5* fused gene, *pyr2,* and *pyr4* are juxtaposed in this order on chromosomal DNA. Similarly, *T. brucei* possesses the *pyr* gene cluster, although *pyr3* is not annotated in this cluster (Berriman et al. 2005). The *pyr* gene cluster is the only known clustering of genes that encode all enzymes in an essential metabolic pathway in eukaryotes, while there are examples in filamentous fungi of the clustering of genes encoding enzymes catalyzing secondary metabolites (Saikia et al. 2007; Young et al. 2006).

Clustering of genes responsible for a metabolic pathway is structurally similar to bacterial operons, which may be advantageous for concerted expression of functionally related enzymes at the appropriate times in these bacteria. However, in trypanosomatids, no regulatory mechanism of transcription initiation is found and regulation of expression seems to occur at the post-transcriptional level. Thus, the biological significance of the occurrence of the *pyr* gene cluster in trypanosomatids is still unknown.

The *pyr* gene cluster appears to constitute part of a large polycistronic gene cluster in trypanosomatids and, importantly, includes genes acquired via lateral gene transfer (LGT). Indeed, trypanosomatid genomes contain large numbers of genes thought to have acquired via LGT (Opperdoes and Michels 2007). Phylogenetic analyses of pyr4 and pyr6 have shown that both genes have a prokaryotic origin, not only in trypanosomatids but in bodonids (Annoura et al. 2005; Makiuchi et al. 2007; Nara et al. 2000). Thus, LGT events are likely to have preceded the establishment of the *pyr* gene cluster, as well as contributing to it.

In the present study, we regarded the *pyr* gene cluster as model synteny and addressed whether this cluster is present in non-trypanosomatid kinetoplastids, i.e. bodonids. We demonstrate clustering of the *pyr* genes in the bodonids, *Parabodo caudatus* (formerly *Bodo caudatus*) and *Neobodo saliens* (formerly *B. saliens*), suggesting that this gene synteny emerged in a common ancestor of kinetoplastids. In addition, we found that *pyr4* and *pyr6* genes in the diplonemid, *Diplonema papillatum*, had a different origin from the kinetoplastid genes. Phylogenetic and gene organization analyses suggested that the stand-alone *pyr6* in diplonemids might represent a transitional status from the fused *pyr5/pyr6* in euglenoids to the inversely fused *pyr6/pyr5* in kinetoplastids. Our findings provide insights into an evolutionary gap between kinetoplastids and non-kinetoplastid groups in Euglenozoa.

Results

The *pyr* gene cluster in the *P. caudatus* genome

The *pyr1* and *pyr4* genes are the 5'- and 3'-terminal genes, respectively, of the *pyr* gene cluster in trypanosomatids (Gao et al. 1999). Therefore, sequence analysis of the regions downstream to *pyr1* and upstream to *pyr4*, would allow us to detect the putative *pyr* gene cluster in the bodonid genome.

By a series of PCR amplifications using *P. caudatus pyr*-specific primers and the genomic DNA as template, we identified the *pyr* gene cluster in the *P. caudatus* genome (Fig. 1. See also Supplementary Fig. S1). *P. caudatus pyr* gene cluster consists of 11,285 bp and shows the same gene order and organization as the trypanosomatid ones. We found that, compared

with the parasitic trypanosomatids, free-living *P. caudatus* has shorter untranslated regions between the *pyr* genes. Southern blot analysis showed that all *pyr* genes are single-copy genes and are juxtaposed on chromosomal DNA (Supplementary Fig. S1).

We also obtained a partial *pyr* gene cluster in *N. saliens* by screening the genomic DNA library using *pyr1*-specific DNA probe, which carried the partial *pyr1, pyr3,* and *pyr6/pyr5* but lacked *pyr2/pyr4* (Fig. 1). In contrast to *P. caudatus*, we could not detect the full *pyr* gene cluster in *N. saliens.* Amplification of *N. saliens* genomic DNA using primers for *pyr1* and *pyr2/pyr4* or hybridization of the *pyr1* positive clone with the probe corresponding to the 2 kb upstream region of *pyr2/pyr4* did not yield any positive signals (data not shown).

The common evolutionary origin of the *pyr* gene cluster in trypanosomatids and bodonids

We previously showed that *pyr2*, *pyr4* and *pyr6/pyr5* have the same evolutionary origin in trypanosomatids and bodonids (Annoura et al. 2005; Makiuchi et al. 2007). We further assessed whether the other *pyr* genes, *pyr1* and *pyr3*, have a common evolutionary origin in kinetoplastids.

Pyr1 of both trypanosomatids and bodonids were found to have the same protein structure, comprised of an N-terminal GAT domain, a linker, and a C-terminal CPS domain (Supplementary Fig. S1). Phylogenetic reconstruction of the CPS domain of pyr1 showed monophyly of the trypanosomatid and bodonid pyr1, with strong bootstrap support (99% in the maximum likelihood (ML) method, Supplementary Fig. S2).

In the pyr3 tree, a monophyletic grouping of kinetoplastids was also supported (85% in ML, Supplementary Fig. S3). Importantly, amino acid sequence alignment of pyr3 from various species showed kinetoplastid-specific amino acid insertions (Fig. 2). These results indicate that the *pyr1* and *pyr3* genes, as well as *pyr2*, *pyr4* and *pyr6/pyr5*, have the same origin in trypanosomatids and bodonids.

Although the organization of *pyr1*, *pyr2*, and *pyr3* in non-kinetoplastid euglenozoans is unclear, these genes are likely to be independent, as well as kinetoplastids. We found that *Naegleria gruberi,* which belongs to Heterolobosea, the closest group to Euglenozoa, possesses independent *pyr3*. In addition, independent *pyr2* was found in the *N. gruberi* genome sequence database (data not shown). These findings are consistent with the notion that gene fusion of pyr1/pyr3/pyr2 (CAD) are present in unikonts but absent from bikonts (Stechmann and Cavalier-Smith 2003), with exception of a red algal CAD (Nozaki et al. 2005).

The different origin of the *pyr4* gene in kinetoplastids and diplonemids

We addressed whether the *pyr* gene cluster is present in diplonemids, which are most closely related to kinetoplastids. Although we performed cDNA-PCR cloning using various sets of degenerated primers for all *pyr* genes in the diplonemid, *D. papillatum,* we obtained PCR products only for pyr4 and pyr6.

Pyr4 enzymes have been classified into three types, family-1A, -1B, and -2 enzymes. Family-2 pyr4 is present in the euglenoid, *Euglena gracilis,* and in many eukaryotic groups, whereas family-1A pyr4, which originated via LGT, is present in kinetoplastids (Annoura et al. 2005). Using *D. papillatum* RNA as template, we obtained cDNA-PCR products using primers specific for family-2, but not for family-1A, pyr4.

We found that an open reading frame of *D. papillatum* pyr4 cDNA encoded for 395 amino acids, with the predicted amino acid sequence sharing a higher identity with that of *E. gracilis* (51%) than those of kinetoplastids (18-26%). The amino acid residues of *D. papillatum* pyr4 used for binding of flavin mononucleotide and orotate were of family-2 type (Fig. 3A). These results clearly indicate that *D. papillatum* possesses the family-2 *pyr4* gene. Phylogenetic analysis of family-2 pyr4 showed monophyly of euglenoids and diplonemids, whereas the bootstrap support was moderate (74% in ML, Fig. 3B). In conjunction with the monophyletic grouping of kinetoplastid pyr4 (Annoura et al. 2005), we conclude that the origin of *pyr4* is different between kinetoplastids and euglenoids/diplonemids.

Different organization of *pyr5* and *pyr6* genes in Euglenozoa

The *pyr5* and *pyr6* genes have a different gene structure in euglenoids and kinetoplastids. *E. gracilis* possesses a fused *pyr5/pyr6* gene, whereas, in kinetoplastids, the genes are inversely fused as *pyr6/pyr5* (Makiuchi et al. 2007).

We found that *D. papillatum* pyr6 cDNA encoded for 271 amino acids and retained spliced leader and poly(A) sequences at its 5'-

and 3'-termini, respectively (data not shown). These results indicate that the cloned *D. papillatum* pyr6 cDNA is not truncated, but is a mature form of the transcript. In addition, we could not detect any clone in which pyr6 was fused with the pyr5 sequence, suggesting that *D. papillatum* has stand-alone genes for pyr6 and probably for pyr5.

Phylogenetic analysis of pyr6 reconstructed a monophyletic grouping of *D. papillatum* pyr6 and the pyr6 domain of *E. gracilis* pyr5/pyr6 within a eukaryotic clade with 82% bootstrap support (ML analysis) (Fig. 4). Interestingly, Apicomplexa and Fungi, the eukaryotic groups phylogenetically distantly related from diplonemids, also possesses a stand-alone *pyr6* (Makiuchi et al. 2007). The apicomplexan pyr6 clustered with the kinetoplastid pyr6/pyr5 in the bacterial clade to the exclusion of *D. papillatum* pyr6 with strong bootstrap support (99% in ML, Fig. 4), indicating that the origin of *pyr6* is different between Apicomplexa and diplonemids.

D. papillatum and fungal pyr6 were nested in the eukaryotic clade. To determine whether diplonemid and fungal *pyr6* have the same origin, we performed an approximately unbiased (AU) test in terms of statistical comparisons of the alternative trees, in which diplonemid or fungal *pyr6* was grafted to either branch in the tree in Fig. 4. The alternative trees grafting the fungal sequence to the diplonemid branch and vise versa were statistically significantly different form the tree in Fig. 4 with p = 0.006 and 0.003, respectively, indicating that the diplonemid and fungal *pyr6* do not have the same origin. Thus, these findings strongly suggest that the euglenoid and diplonemid *pyr6* genes share a common evolutionary origin, despite of different gene organization, and that the secondary split of the *pyr5/pyr6* gene likely occurred on the line leading to *D. papillatum* after its separation from the line leading to euglenoids, although the nature of the diplonemid *pyr5* gene has not yet known.

Discussion

The unity of kinetoplastids, diplonemids, and euglenoids in Euglenozoa has been evidenced by their morphology, protein phylogeny, and mitochondrial properties. For example, all euglenozoans share a unique mitochondrial DNA structure, with the mitochondrial genome comprised of multiple copies of circular DNAs (Marande et al. 2005; Roy et al. 2007). Although these findings empathize uniqueness of Euglenozoa among the eukaryotic groups, paradoxically, the taxonomic boundaries of the euglenozoan groups became rather obscure. Thus, our results shed new light on the critical evolutionary gap between kinetoplastids and diplonemids/euglenoids by analyses of the synteny and gene fusion of *pyr* genes.

Conserved synteny of *pyr* genes in kinetoplastids

We have shown evidence of the *pyr* gene cluster in the bodonid, *P. caudatus*, strongly suggesting that the *pyr* gene cluster was established in a common ancestor of trypanosomatids and bodonids. This finding is also supported by the partial clustering of the *pyr1, pyr3*, and *pyr6/pyr5* genes and the occurrence of *pyr2/pyr4* gene fusion, in *N. saliens* (Fig. 1). The presence of the fused *pyr2/pyr4* gene appears to reflect an evolutionary trait, which may have resulted from the loss of the untranslated region between *pyr2* and *pyr4* genes (Annoura et al. 2005). Failure to detect the full *pyr* gene cluster in *N. saliens* may be due to the presence of the long intervening sequence between *pyr6/pyr5* and *pyr2/pyr4* genes, a sequence too long for PCR amplification. Another possibility is the occurrence of the rearrangement of the *pyr* gene cluster, as in *T. cruzi*, in which the partially amplified *pyr* gene clusters were frequently detected (Nara et al. 2003).

Genome reduction has been frequently observed in endosymbionts and parasitic microbes, accompanied by loss of both genes and non-coding regions (Dobrindt et al. 2004; Gross et al. 2003; Moran 2002; van Ham et al. 2003). We have shown here that, compared with the parasitic trypanosomatids, free-living *P. caudatus* has shorter untranslated regions between the *pyr* genes (Fig. 1). This discrepancy may be explained by the possible compact nature of the bodonid genomes. Comparative genomics between trypanosomatids and free-living bodonids would clarify this point.

The timing of lateral gene transfer of *pyr* genes pinpointed

We were able to identify *pyr4* and *pyr6* genes in *D. papillatum,* thus allowing us to examine their molecular phylogeny in the euglenozoan groups.

The *pyr4* gene: Since family-1A pyr4 of trypanosomatids has different biochemical properties than family-2 pyr4 of canonical eukaryotes, including human, pyr4 may be a promising target for chemotherapy against

trypanosomiasis (Nara et al. 2005). The family-1A *pyr4* gene in kinetoplastids was shown to have originated from anaerobic microbes via LGT, whereas an ancestral eukaryote may have acquired the family-2 *pyr4* from mitochondria (Annoura et al. 2005). We found that euglenoid and diplonemid pyr4 were both nested within the family-2 clade to the exclusion of the kinetoplastid sequences of family-1A type. Together with the finding that most eukaryotic groups, including Heterolobosea, the closest group to Euglenozoa, possess family-2 pyr4 (Arisue et al. 2005; Nara et al. 2000; Rodríguez-Ezpeleta et al. 2007; Simpson et al. 2006a), these results pinpoint the timing of LGT of *pyr4* to a common ancestor of kinetoplastids after separation of the lines leading to euglenoids and diplonemids (Fig. 1).

The *pyr6* gene: Similar to *pyr4*, the kinetoplastid *pyr6* gene was shown to have a prokaryotic origin and to have been acquired via LGT (Makiuchi et al. 2007; Nara et al. 2000). In *D. papillatum,* the presence of a stand-alone *pyr6* gene was unexpected, because euglenoids and the group Heterolobosea, which share a common ancestor, possess fused *pyr5/pyr6* genes. In our phylogenetic analysis, the stand-alone *pyr6* gene of *D. papillatum* was grouped with the pyr6 domain of euglenoid *pyr5/pyr6*, but not with the stand-alone genes of Apicomplexa and Fungi (Fig. 4). It is important to note that apicomplexan *pyr6* has been shown to originate via LGT (Makiuchi et al. 2007). In addition, our AU test rejected the monophyly of diplonemid and fungal pyr6 (this study). These findings indicate that the diplonemid *pyr6* was not derived from apicomplexan or fungal *pyr6* via LGT. Therefore, it is likely that the diplonemid *pyr6* has the same origin as euglenoid *pyr5/pyr6*. Thus, we conclude that LGT of *pyr6* occurred in a common ancestor of kinetoplastids after separation of the lines leading to euglenoids and diplonemids (Fig. 1).

An evolutionary gap between kinetoplastids and diplonemids

Due primarily to the lack of sequence information on the *pyr1*, *pyr2*, *pyr3*, and *pyr5* genes, it is unclear if the *pyr* gene cluster is present in diplonemids. Regarding to the organization of *pyr1*, *pyr2*, and *pyr3* in diplonemids and euglenoids, these genes are likely to be independent, as well as in kinetoplastids, because independent *pyr2* and *pyr3* was found in Heterolobosea, which shares a common ancestor with Euglenozoa.

Clustering of *pyr1*, *pyr2*, and *pyr3* may be present in the genomes of diplonemids and euglenoids, which should be clarified by further sequence analysis of *pyr* genes in these groups. On the other hand, clustering of all of *pyr* genes is likely to have occurred in a common ancestor of kinetoplastids after separation of the line leading to diplonemids. Because the *pyr4* and *pyr6* genes of diplonemids and kinetoplastids do not share high sequence similarity at both nucleotide and amino acid levels, it is difficult to suppose that homologous recombination has occurred between the original and acquired genes within the already established *pyr* gene cluster.

Another possibility for the origin of clustering of *pyr* genes in kinetoplastids is that the *pyr* gene cluster may have originated in the other organism(s) and transferred laterally to kinetoplastids. This explanation seems more parsimonious and cannot be entirely excluded. However, it is rather difficult to postulate such hypothetical eukaryotes, which may possess the same mosaic nature of *pyr* genes in their gene organization and evolutionary origin as kinetoplastids do. Likewise, there is no prokaryote having *pyr* genes of a eukaryotic origin in its *pyr* operon. Thus, the *pyr* gene cluster is likely to have established in a common ancestor of kinetoplastids after separation of the line leading to diplonemids (Fig. 1).

Evolutionary implications of synteny in Euglenozoa

Although the *pyr* gene cluster constitutes a part of a large syntenic block in trypanosomatids, it is not yet known whether large-scale synteny is conserved in non-trypanosomatid euglenozoans. More importantly, the occurrence of large-scale synteny, composed of dozens of functionally unrelated genes, raises fundamental questions of how large-scale synteny is maintained and what is its biological significance. Presently, synteny is thought to be significant only as an important evolutionary signal (Bennetzen and Freeling 1997; Nadeau 1989). In general, we recognize phylogenetic relationships between taxa by the presence of conserved synteny. However, lack of conserved synteny does not necessarily indicate their phylogenetically distant relation.

Because genome architecture varies from species to species, synteny becomes scarce or is lost by accumulation of frequent, species-specific rearrangement of chromosomal DNA, including meiotic DNA recombination.

Accordingly, the genomes of trypanosomatids, and probably bodonids, may be more resistant than the genomes of other eukaryotes to chromosomal rearrangements caused by transpositions and reversals of DNA regions, resulting in conservation of large-scale synteny. Actually, conservation of clustering of whole *pyr* genes in kinetoplastids appears to indicate that there is strong selection pressure to maintain both the number and order of these genes in a syntenic cluster in the genome within this group of organisms. Although the mechanisms used to maintain their genome architecture are not yet known, this assumption is also supported by the occurrence of large polycistronic transcription units in trypanosomatids, which may be interfered by the reversals of regions of chromosomal DNA.

Another factor affecting synteny includes the timing of lineage divergence. In trypanosomatids, the estimate time of the origin of *T. cruzi* is ≥95 million years ago when calibrated by the phylogeny of insects, which are the sole hosts of trypanosomatids (Gaunt and Miles 2002). However, an estimate of the age of divergence of the euglenozoan branches had not been demonstrated, due exclusively to the lack of appropriate calibration methods for the timing of their branching. The evolutionary gap between kinetoplastids and non-kinetoplastid euglenozoans could be explained by their extremely ancient origins, which should be proven using reliable means of estimate for the divergence time of the euglenozoan branches.

Complex evolutionary events of gene rearrangement in Euglenozoa

Evolutionarily, the presence of stand-alone *pyr6* in diplonemids, which appears to have the same origin as its euglenoid counterpart, may provide important insights into the evolutionary process of rearrangement of gene fusion. Our previous analyses suggested that the kinetoplastid *pyr6/pyr5* emerged through three evolutionary events: 1) splitting of the original *pyr5/pyr6*, 2) LGT-based acquisition of the *pyr6* gene, and 3) re-fusion between the acquired *pyr6* and the resident *pyr5* genes in the reverse order, although the temporal order of gene splitting and LGT is not clear.

Importantly, diplonemids constitute an intermediate branch between euglenoids and kinetoplastids in Euglenozoa. Therefore, it is likely that the splitting of *pyr5/pyr6* occurred first, in a common ancestor of diplonemids and kinetoplastids, and that, after separation of the diplonemid lineage, a common ancestor of kinetoplastids acquired *pyr6* via LGT, followed by its subsequent re-fusion into *pyr6/pyr5*. Nevertheless, we cannot exclude the possibility that rearrangement of *pyr5/pyr6* occurred independently in the diplonemid and kinetoplastid lineages (Fig. 1).

Structural rearrangements of *pyr5/pyr6* have been shown to accompany LGT events of either or both genes, suggesting that the acquired genes have contributed to the establishment of the rearranged gene structure (Makiuchi et al. 2007). It is important to determine whether diplonemid and euglenoid *pyr5* have common or different evolutionary origins. Further cloning and phylogenetic analysis of the *pyr5* gene in diplonemids would clarify the evolutionary process involved in the rearrangement of gene fusion.

Concluding Remarks

In the present study, we pinpointed the timing of synteny of the *pyr* gene cluster and LGT of *pyr4* and *pyr6* to a common ancestor of kinetoplastids. We propose the putative evolutionary steps in the transition from *pyr5/pyr6* fusion to inverted *pyr6/pyr5* fusion. Our findings also emphasize the phylogenetic and molecular biological gaps between kinetoplastids and diplonemids/euglenoids. Comparative genomics of these groups are necessary to understand the nature and evolutionary traits of diversification in Euglenozoa.

Materials and Methods

Organisms: Monoxenic cultures of *P. caudatus* (ATCC 50361; American Type Culture Collection, Manassas, VA, USA) or *N. saliens* (ATCC 50358) with *Klebsiella pneumoniae* subsp. (ATCC 27889) were routinely maintained as described (Annoura et al. 2005). Axenic cultures of *D. papillatum* (ATCC 50162) were maintained at 25°C in ATCC medium 1728; the cells were collected by centrifugation at 3,000 rpm for 10 min at 4°C, and washed 3 times by suspension in artificial seawater and centrifugation at 3,000 rpm for 10 min at 4°C.

Nucleic acid extraction: Total RNA was extracted from freshly prepared cells of *P. caudatus, N. saliens,* or *D. papillatum* (4.9×10^8 cells) using TRIZOL® reagent (Invitrogen, Carlsbad, San Diego, CA, USA) as described (Annoura et al. 2005). Protist poly(A) RNA was isolated from total RNA using GenElute™ mRNA

Miniprep Kits (Sigma-Aldrich Japan, K.K., Japan). For isolation of *P. caudatus* genomic DNA, 1.0 x 10^9 freshly harvested cells were suspended in TNE buffer (10 mM Tris-HCl, 100 mM NaCl, 2 mM EDTA, pH 7.5) and lysed by adding sodium dodecyl sulfate (SDS) and Proteinase K (Roche Diagnostics K.K., Minato-ku, Tokyo, Japan) to final concentrations of 0.6% and 200 μg/ml, respectively, and incubating the mixtures at 42°C for 90 hours. Cell lysate was extracted with an equivalent volume of phenol saturated with TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5) for 15 min at room temperature with gentle shaking, followed by centrifugation at 3,000 rpm for 10 min at room temperature. The lysate was extracted with chloroform using the same procedure, and total DNA was precipitated with ethanol, dissolved in TE containing 10 μg/ml RNase A, and stored at 4°C until use.

cDNA cloning of *pyr* genes from Euglenozoa: Using total RNA from *P. caudatus* or *N. saliens*, cDNA was synthesized using SuperScript™ III RNase H⁻ Reverse Transcriptase (Invitrogen), and oligo(dT) primer. The CPS domain of eukaryotic pyr1 was amplified using the outer primer set, pyr1-F1 (5'-CARGCIGGIGARTTYGAYTA-3', according to the IUB format) and pyr1-R2 (5'-GTIWSIACIGTYTCIGGRTT-3'), followed by a second PCR step (nested PCR) using the inner primer set, pyr1-F2 (5'-GGIGGICARACIGCIYTIAA-3') and pyr1-R1 (5'-GGRSAICCIGTIGCYTT-3'). The 674 bp PCR product was subcloned in a TA cloning vector and sequenced using an automated DNA sequencer (CEQ™ 8000 Genetic Analysis System, BECKMAN COULTER, Fullerton, CA, USA). *D. papillatum* cDNA, synthesized as above, was used for nested PCR. *D. papillatum* pyr4 cDNA was PCR amplified using the primers pyr4-F1 (5'-GAATTCGCIGCIGGITTYGAYAA-3') and pyr4-R2 (5'-GTCGACACIARISWIGCICCIGC-3'), followed by a second amplification using the primers pyr4-F2 (5'-GAATTCATYAAYMGITAYGGITT-3') and pyr4-R1 (5'-GTCGACARICCIGGIGTRTTIGG-3'). *D. papillatum* pyr6 cDNA was cloned by PCR amplification using the primers pyr6-F1 (5'-TTYGARGAYMGIAARYTIGC-3') and pyr6-R2 (5'-DWWIAYICCICKICCIAC-3'), followed by a second amplification using the primers pyr6-F2 (5'-GCIGAYATHGGIMAIRYIGT-3') and pyr6-R1 (5'-GGISWRTYRTAYTGYTGICC-3'). The PCR products were subcloned in pCRII-TOPO vector (Invitrogen) and sequenced.

Rapid amplification of cDNA ends (RACE) of *pyr* genes in Euglenozoa: *P. caudatus* pyr1 cDNA was synthesized using the specific antisense primer (5'-GTGTGAATTCCCATCGGATC-3'). 5'-RACE was carried out using the sense primer (5'-ACTTACGCTATAAAAGATACAGT-3'), which is specific to the spliced leader (SL) sequence of *P. caudatus* (Sturm et al. 2001) and the antisense primer (5'-CTTGACGGACATCTCGATGA-3'). Full-length cDNA of *D. papillatum* pyr4 and pyr6 were cloned by 5'- and 3'-RACE. For 5'-RACE of *D. papillatum* pyr4, cDNA was synthesized using the pyr4-specific antisense primer (5'-ACACGTTGATGACGATGTAG-3'), followed by PCR amplification using the *D. papillatum* SL-specific primer (5'-GCTACAGTTTCTGTACTTTATTG-3') (Sturm et al. 2001) and the antisense primer (5'-ACCGGGGTCAGGTTGTATTT-3'). For 5'-RACE of *D. papillatum* pyr6, cDNA was synthesized using the pyr6-specific antisense primer (5'-AACCCTGCGTGGTCGGTGAT-3'), followed by PCR amplification using the above SL-specific primer and the antisense primer (5'-ACAATGTGCGCGTTGGTCAC-3'). For 3'-RACE, *D. papillatum* cDNA was synthesized using the oligo(dT)-anchor primer (5'-AATAAAGCGGCCGCGGATCCAATTTTTTT TTTTTTTTTVN-3'), followed by PCR amplification using a single sense primer (5'-GCTTCGTTGAAATCGGCACCGTGAC-3' for pyr4 or 5'-TGATGCTGCAGTACGAGCACGGCCA-3' for pyr6) to quantify the cDNA template. The resulting cDNA was PCR amplified using the anchor primer (5'-AATAAAGCGGCCGCGGATCCAA-3') and the primer specific to pyr4 (5'-CGCATGTGGCGATTGACCGA-3') or pyr6 (5'-AAATGTCGTCGAAGGGGACGCTCGC-3'). Finally, the full-length pyr4 cDNA was PCR amplified using the specific primers, (5'-CACCATGTTCACTCGGCTGGCTGTGGCT GGGGGA-3') and (5'-CACCATGCCCGAGCTCTCTTTCCAAGAAC -3'), and full length pyr6 cDNA was PCR amplified using the specific primers, (5'-CTAGGCCCCCCGCAGGTCACTTGGCATC AG-3') and (5'-CTACTGAAACATGGCCTTCGGGTAC-3'). Each full-length cDNA was subcloned and sequenced completely. Note that, although sequencing of full-length pyr1 cDNA of *N. saliens* was incomplete because of failure of 5'-RACE, partial *N. saliens* pyr1 had the same domain structure as the other kinetoplastid ones.

Identification of the *pyr* gene cluster in *P.*

***caudatus*:** The *pyr1* and *pyr4* genes are the 5'- and 3'-terminal genes, respectively, of the *pyr* gene cluster in trypanosomatids. If the syntenic *pyr* gene cluster is conserved in bodonids, the other *pyr* genes are likely to be located between *pyr1* and *pyr4*. To identify the *pyr* gene cluster in *P. caudatus*, we separately amplified its 5'-half, including the putative *pyr1-pyr3-pyr6/pyr5*, and its 3'-half including *pyr6/pyr5-pyr2-pyr4*. PCR amplification was performed using the genomic DNA as template and the pyr1-specific sense primer (5'-CAGATCGGCGAGCATGTTGC-3') and the pyr6-specific antisense primer (5'-TGAGCCTCTTGAGCCGGATC-3') for *pyr1-pyr3-pyr6/pyr5* or the pyr5-specific sense primer (5'-CCGTGTTCGAGCAGTCGCTC-3') and the pyr4-specific antisense primer (5'-GACACGAAAGGTTCAGCTCC-3') for *pyr6/pyr5-pyr2-pyr4*. The 7 kb and 3 kb DNA fragments were obtained and found to contain *pyr3* and *pyr2*, respectively. The overlapping 170 bp nucleotides between these 7 kb and 3 kb PCR fragments showed 100% sequence identity. Finally, the *pyr* gene cluster was completely sequenced.

Phylogenetic analyses: All sequence data, with the exception of those cloned by us and first reported here, were collected from public sequence databases by taxonomic and BLAST searches. The sequences reported in this paper appear in the DDBJ/EMBL/GenBank databases with the accession numbers AB307736 for the *Parabodo caudatus* pyrimidine biosynthetic gene cluster encoding pyr1, pyr3, pyr6/pyr5 fusion protein, pyr2, and pyr4; AB307737 for the *Neobodo saliens* pyrimidine biosynthetic gene cluster encoding pyr1, pyr3, and pyr6/pyr5; AB307738 for *Diplonema papillatum* pyr4 mRNA; and AB307739 for *D. papillatum* pyr6 mRNA. Unpublished sequences were obtained from various genome project databases: DOE Joint Genome Institute (http://genome.jgi-psf.org/) for the green alga, *Chlamydomonas reinhardtii*, the diatom, *Thalassiosira pseudonana,* the heterolobosea, *Naegleria gruberi*, and the oomycete, *Phytophthora sojae*; the *Cyanidioschyzon merolae* Genome Project (http://merolae.biol.s.u-tokyo.ac.jp/) for the red alga, *Cyanidioschyzon merolae*; and ToxoDB (http://toxodb.org/) for the apicomplexan *Toxoplasma gondii*. Multiple alignments for pyr1, pyr3, pyr4, and pyr6 sequences were obtained using CLUSTAL W (Thompson et al. 1994), with the alignments corrected by manual inspection. The amino acid sequence alignments used in the present study are available upon request. Unambiguously aligned positions were selected and used for phylogenetic analyses. The maximum likelihood (ML), distance matrix (DM), and maximum parsimony (MP) methods for protein phylogeny were applied to the data sets using the CODEML program in PAML3.1 (Yang 1997) and the PROML, PROTDIST, NEIGHBOR, PROTPARS, SEQBOOT, and CONSENSE programs in PHYLIP3.6a, distributed by Dr. Joseph Felsenstein, University of Washington. In ML analysis, an initial tree search was performed by applying PROML with the JTT model for amino acid substitution, assuming homogeneous rates across sites. Based on the optimal tree obtained, the Γ-shape parameter (α) of the discrete Γ-distribution with four categories that approximated site rates was estimated using CODEML. Using this α-value, a further tree search was performed with the JTT model with four site-rate categories using PROML with the global rearrangement option, producing the final optimal tree. In DM analysis, ML estimates for pairwise distances among the sequences analyzed were calculated using PROTDIST, based on the JTT model with rate variation allowed among sites. Then the neighbor-joining (NJ) tree was reconstructed from the distances using NEIGHBOR. In MP analysis, the MP tree was searched using PROTPARS. Bootstrap analysis for each of the three methods was performed in the same way by applying PROML, PROTDIST + NEIGHBOR, or PROTPARS to the resampled data sets produced by SEQBOOT. One hundred and 1,000 resamplings were performed for ML and for DM and MP analyses, respectively. A consensus tree was generated using the CONSENSE program based on the bootstrap analysis of the ML method. The approximately unbiased (AU) test (Shimodaira 2002) in the CONSEL program (Shimodaira and Hasegawa 2001) was used for statistical comparisons among the alternative pyr6 trees.

Acknowledgments

This work was supported in part by Grants-in-Aid for Scientific Research (Nos. 17370086, 17390123, 18890188, and 19590436) and for the 21st Century Center of Excellence Research (to Makiuchi, Murata, and Aoki) from the Ministry of Education, Science, Sports, Culture, and Technology of Japan.

Appendix A. Supplementary materials

Supplementary data are available at Protist online (http://www.elsevier.de/protist).

References

Annoura T, Nara T, Makiuchi T, Hashimoto T, Aoki T (2005) The origin of dihydroorotate dehydrogenase genes of kinetoplastids, with special reference to their biological significance and adaptation to anaerobic, parasitic conditions. J Mol Evol **60**: 113-127

Arisue N, Hasegawa M, Hashimoto T (2005) Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol Biol Evol **22**: 409-420

Bennetzen JL, Freeling M (1997) The unified grass genome: synergy in synteny. Genome Res **7**: 301-306

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science **309**: 416-422

Bonen L (1993) Trans-splicing of pre-mRNA in plants, animals, and protists. FASEB J **7**: 40-46

Cavalier-Smith T (1981) Eukaryote kingdoms: seven or nine? Biosystems **14**: 461-481

Dávila AM, Lukeš J (2003) Towards a framework for the evolutionary genomics of Kinetoplastids: what kind of data and how much? Kinetoplastid Biol Dis **2**: 16

Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol **2**: 414-424

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science **309**: 404-409

Gao G, Nara T, Nakajima-Shimada J, Aoki T (1999) Novel organization and sequences of five genes encoding all six enzymes for de novo pyrimidine biosynthesis in *Trypanosoma cruzi*. J Mol Biol **285**: 149-161

Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol **19**: 748-761

Gross R, Hacker J, Goebel W (2003) The Leopoldina international symposium on parasitism, commensalism and symbiosis--common themes, different outcome. Mol Microbiol **47**: 1749-1758

Jackson AP, Vaughan S, Gull K (2006) Evolution of tubulin gene arrays in Trypanosomatid parasites: genomic restructuring in *Leishmania*. BMC Genomics **7**: 261

Liang XH, Haritan A, Uliel S, Michaeli S (2003) trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot Cell **2**: 830-840

Makiuchi T, Nara T, Annoura T, Hashimoto T, Aoki T (2007) Occurrence of multiple,

independent gene fusion events for the fifth and sixth enzymes of pyrimidine biosynthesis in different eukaryotic groups. Gene **394**: 78-86

Marande W, Lukes J, Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell **4**: 1137-1146

Martínez-Calvillo S, Nguyen D, Stuart K, Myler PJ (2004) Transcription initiation and termination on *Leishmania major* chromosome 3. Eukaryot Cell **3**: 506-517

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell **108**: 583-586

Nadeau JH (1989) Maps of linkage and synteny homologies between mouse and man. Trends Genet **5**: 82-86

Nara T, Hashimoto T, Aoki T (2000) Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes. Gene **257**: 209-222

Nara T, Hirayama-Noguchi Y, Gao G, Mural E, Annoura T, Aoki T (2003) Diversity of aspartate carbamoyltransferase genes of *Trypanosoma cruzi*. Int J Parasitol **33**: 845-852

Nara T, Kamei Y, Tsubouchi A, Annoura T, Hirota K, Iizumi K, Dohmoto Y, Ono T, Aoki T (2005) Inhibitory action of marine algae extracts on the *Trypanosoma cruzi* dihydroorotate dehydrogenase activity and on the protozoan growth in mammalian cells. Parasitol Int **54**: 59-64

Nozaki H, Matsuzaki M, Misumi O, Kuroiwa H, Higashiyama T, Kuroiwa T (2005) Phylogenetic implications of the CAD complex from the primitive red alga *Cyanidioschyzon merolae* (Cyanidiales, Rhodophyta). J Phycol **41**: 652-657

Opperdoes FR, Michels PA (2007) Horizontal gene transfer in trypanosomatids. Trends Parasitol **23**: 470-476

Rodríguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF (2007) Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol **17**: 1420-1425

Roy J, Faktorová D, Lukeš J, Burger G (2007) Unusual mitochondrial genome structures throughout the Euglenozoa. Protist **158**: 385-396

Saikia S, Parker EJ, Koulman A, Scott B (2007) Defining paxilline biosynthesis in *Penicillium paxilli*: functional characterization of two cytochrome P450 monooxygenases. J Biol Chem **282**: 16829-16837

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

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

Simpson AG, Inagaki Y, Roger AJ (2006a) Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol Biol Evol **23**: 615-625

Simpson AG, Lukeš J, Roger AJ (2002) The evolutionary history of kinetoplastids and their kinetoplasts. Mol Biol Evol **19**: 2071-2083

Simpson AG, Roger AJ (2004) Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol Phylogenet Evol **30**: 201-212

Simpson AG, Stevens JR, Lukes J (2006b) The evolution and diversity of kinetoplastid flagellates. Trends Parasitol **22**: 168-174

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

Sturm NR, Maslov DA, Grisard EC, Campbell DA (2001) *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. J Eukaryot Microbiol **48**: 325-331

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**: 4673-4680

van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M,

Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A (2003) Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci U S A **100**: 581-586

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**: 555-556

Young CA, Felitti S, Shields K, Spangenberg G, Johnson RD, Bryan GT, Saikia S, Scott B (2006) A complex gene cluster for indole-diterpene biosynthesis in the grass endophyte *Neotyphodium lolii*. Fungal Genet Biol **43**: 679-693

Figure Legends

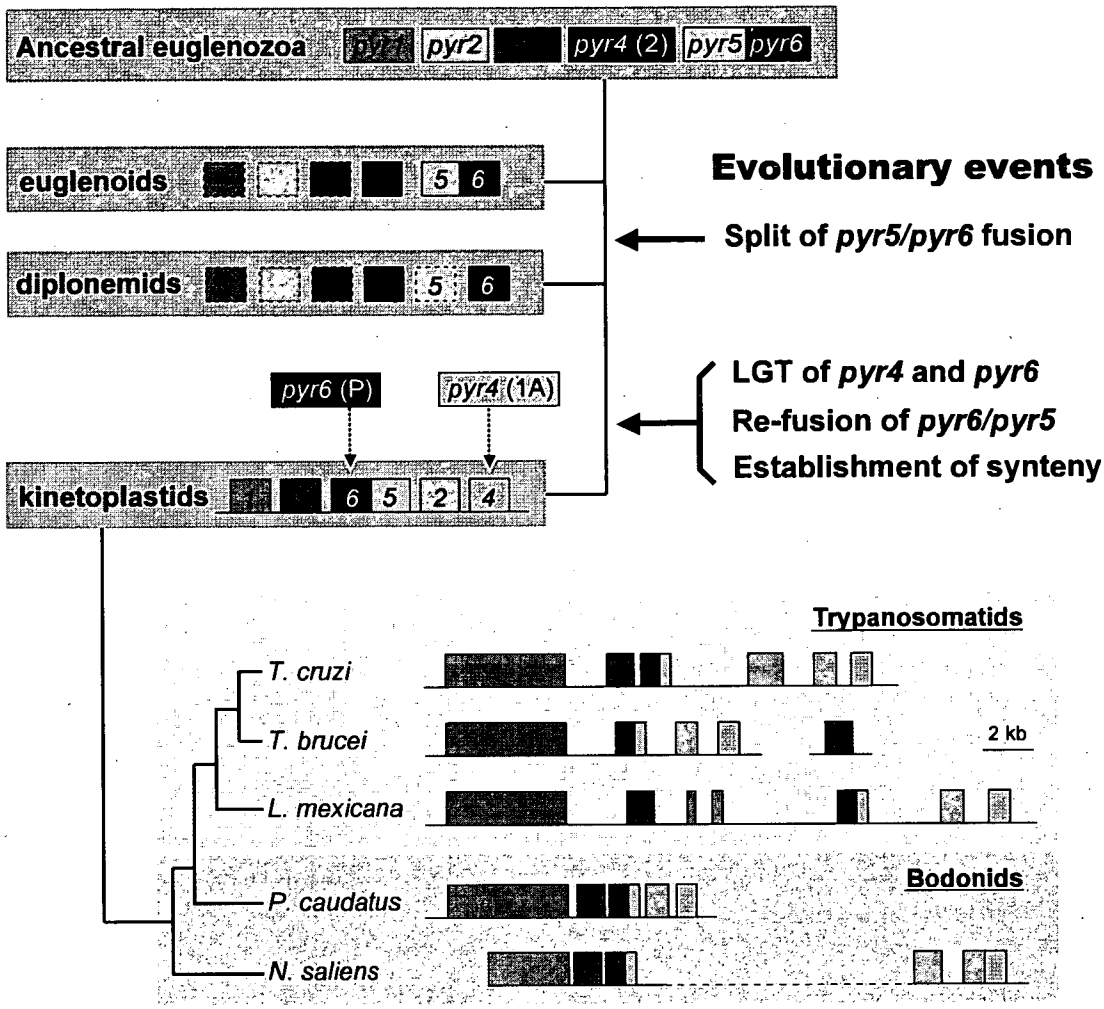
Figure 1. The evolutionary scenario of synteny and gene fusion of pyrimidine biosynthetic (*pyr*) genes in Euglenozoa. The euglenozoan tree is based on a recently proposed model, in which euglenoids constitutes the earliest branch, followed by separation of the diplonemid and kinetoplastid lineages (Simpson et al. 2006b). The de novo pyrimidine biosynthetic pathway is composed of six enzymes and encoded by *pyr1-6* genes. Genes for the first three enzymes, pyr1, pyr2, and pyr3, are fused in animals, Fungi, and Amoebozoa (called unikonts), while they are assumed to be independent in the other eukaryotic groups (bikonts), including Euglenozoa (Stechmann and Cavalier-Smith 2003). In this scheme, an ancestral euglenozoa is assumed to have independent *pyr1, pyr2, pyr3* genes, family-2 type *pyr4* gene, and *pyr5/pyr6* fused gene. Boxes with dotted line indicate as yet unidentified gene in euglenoids and diplonemids. The stand-alone *pyr6* gene in diplonemids was derived by secondary split of *pyr5/pyr6*. An ancestral kinetoplastid acquired both family-1A *pyr4* and *pyr6* of a prokaryotic origin, as indicated by dashed arrows (Annoura et al. 2005; Makiuchi et al. 2007). The gene re-fusion event occurred subsequently between the stand-alone *pyr5* and the acquired prokaryotic *pyr6*. Following these events, the syntenic *pyr* gene cluster was established. The *pyr* gene cluster in the kinetoplastid species (lower panel) is illustrated along with their organismal tree. It is unclear whether all *pyr* genes cluster in *N. saliens*. Gray boxes indicate genes unrelated to pyrimidine synthesis. Note: *pyr4* (2), family-2 *pyr4*; *pyr* (1A), family-1A *pyr4*; *pyr6* (P), *pyr6* of a possible prokaryotic origin.

Figure 2. Alignment of pyr3 sequences among various species. The kinetoplastid-specific insertions are shaded. Within this region, amino acid sequence motif, GXWPHPTXXXP, is fully conserved in kinetoplastids and shown in black boxes.

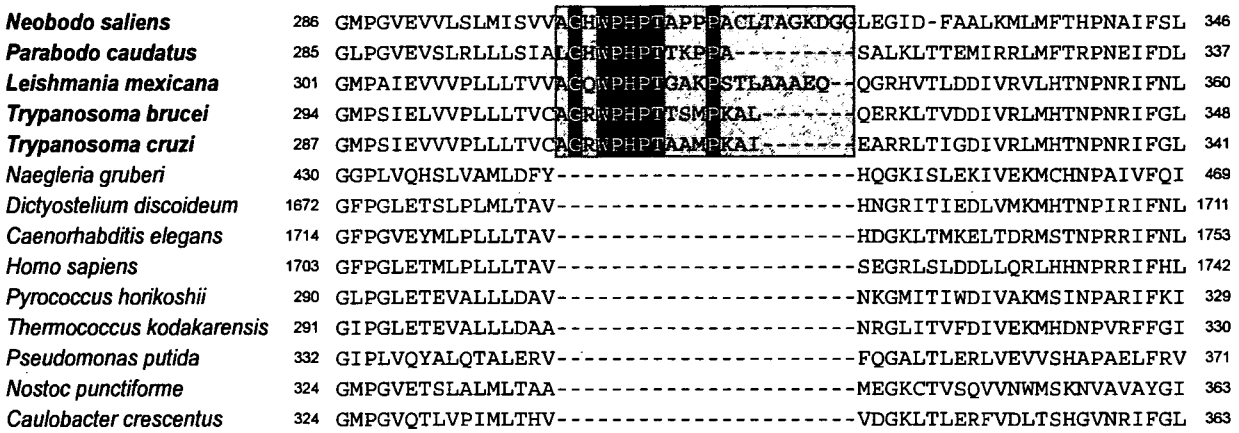
Figure 3. Amino acid sequence comparison and phylogenetic analysis of pyr4. (A) Amino acids indispensable for binding of flavin mononucleotide (FMN) (black boxes) and orotate (gray boxes) in family-2 and -1A type pyr4. (B) Phylogenetic analysis of family-2 pyr4. The consensus tree of the maximum likelihood (ML) method with 100 replicates is shown. The tree was inferred by the JTT model taking across-site rate heterogeneity into consideration. The α-value of the Γ shape parameter used in the analysis was 0.63597. Bootstrap proportion (BP) values are attached to the internal branches. Branches with less than 50% BP support are unmarked. BP values for a diplonemids/euglenoids clade are marked by an asterisk and given by ML, distance matrix (DM) and maximum parsimony (MP) methods. The length of each branch is proportional to the estimated number of substitutions. With 17 taxa, 230 unambiguously aligned amino acid sites were used for analysis, corresponding to residues 81-105, 107-157, 166-174, 179-225, 240-250, 255-260, 265-28w1, 290-308 and 317-361 of the *D. papillatum* sequence.

Figure 4. Bootstrap consensus tree of the ML method with 100 replicates for pyr6. The tree was inferred by the JTT model taking across-site rate heterogeneity into consideration. The α-value of the Γ shape parameter used in the analysis was 1.11259. The eukaryotic species having stand-alone pyr6 are shaded. Methods and labeling are as in Fig. 3. Analysis involved 118 unambiguously aligned amino acid sites with 24 taxa, corresponding to residues 32-39, 43-71, 87-100, 114-125, 142-155, 173-181, 192-200, 202-205 and 219-237 of the *D. papillatum* sequence.

Makiuchi *et al.* Fig. 1



Makiuchi *et al.* Fig. 2

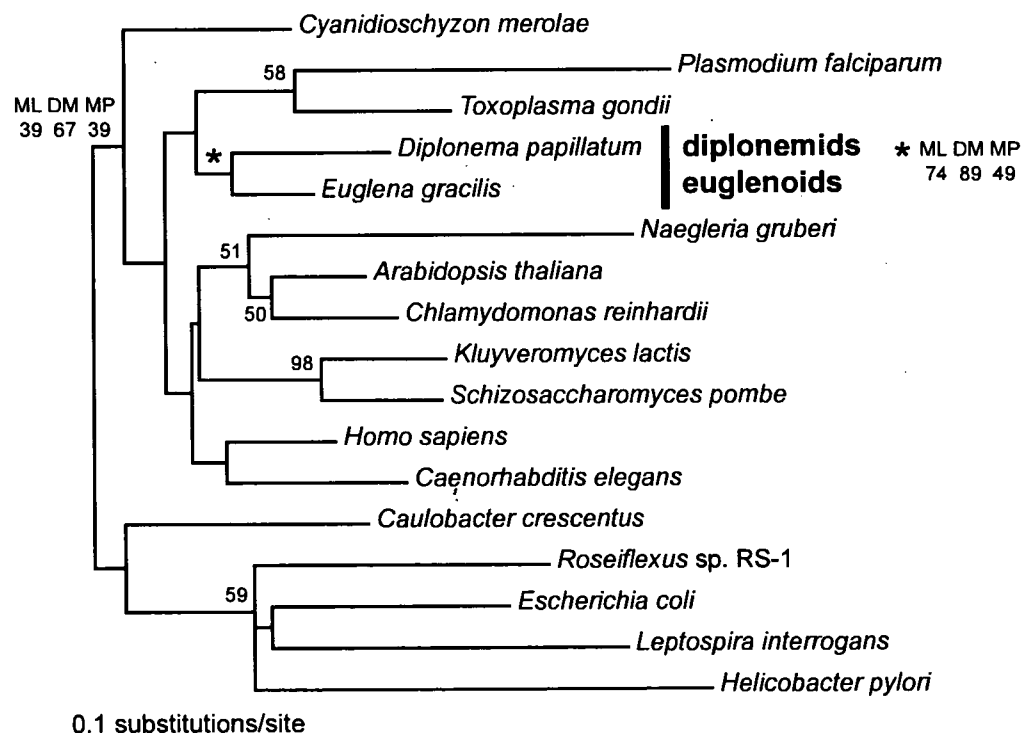


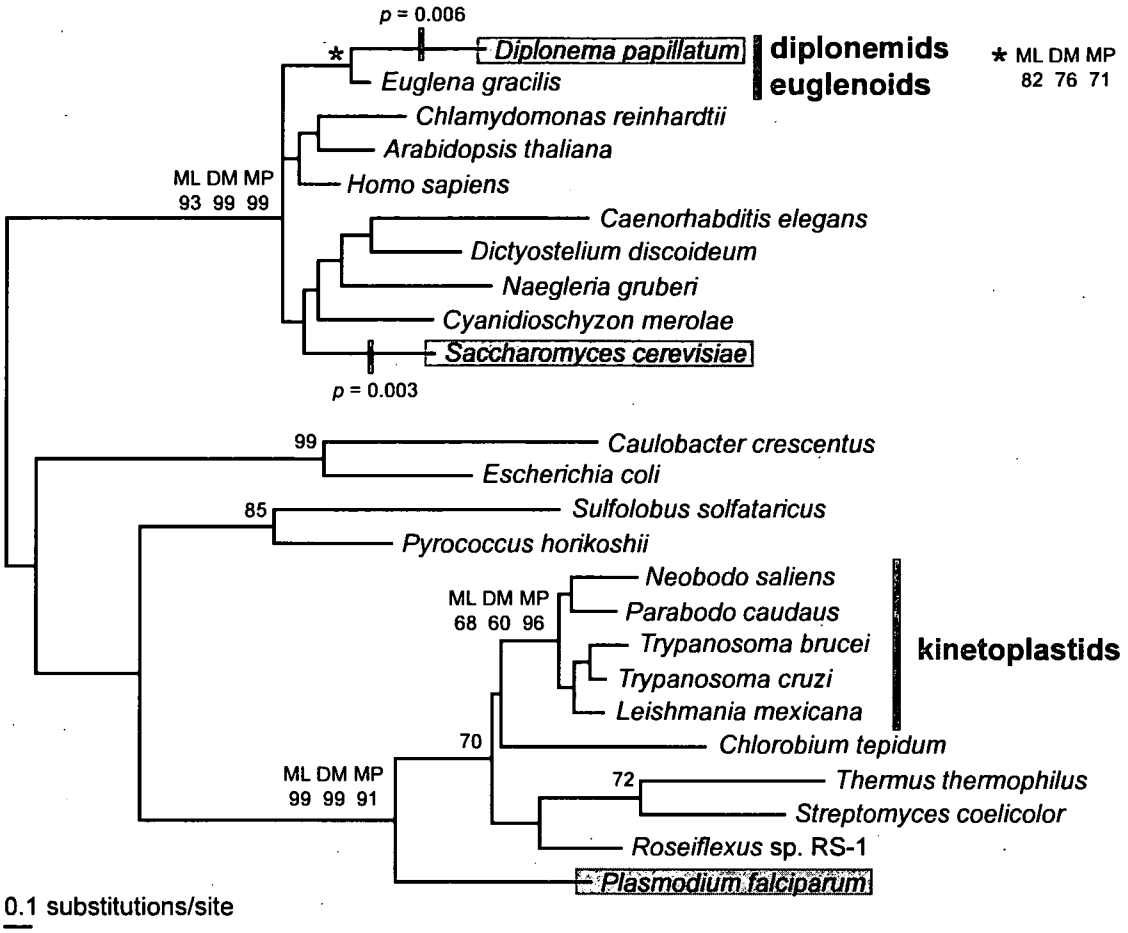
Makiuchi *et al.* Fig. 3

A)

Type of DHOD		FMN site	Orotate site
Family 2			
diplonemids	*Diplonema papillatum*	GFGFVEIGTVTPLPQP	YIVINVSSPNTPGLR
euglenoids	*Euglena gracilis*	GFGFVEIGTVTPLPQP	YLVINVSSPNTPGLR
Metazoa	*Homo sapiens*	GFGFVEIGSVTPKPQE	YLVVNVSSPNTAGLR
Viridiplantae	*Arabidopsis thaliana*	GFGFVEVGSVTPVPQE	YLVINVSSPNTAGLR
Fungi	*Schizosaccharomyces pombe*	GFSYLEIGSVTPKPQP	ILVINVSSPNTPGLR
Family 1A			
trypanosomatids	*Trypanosoma cruzi*	SSGALVSKSCTSAPRD	LLELNLSCPNVPGKP
bodonids	*Parabodo caudatus*	SSGSLITKSCTSAFRE	ILELNLSCPNVPGKP
	Neobodo saliens	ASGTLITKSCTAQQRD	LLELNLSCPNVPGKP
Fungi	*Saccharomyces cerevisiae*	KAGAFITKSATTLERE	ITELNLSCPNVPGKP
Firmicutes	*Lactococcus lactis* A	QAGAYITKSSTLEKRE	ITELNLSCPNVPGKP

B)




p = 0.006
Diplonema papillatum
Euglena gracilis
diplonemids
euglenoids
* ML DM MP
82 76 71
Chlamydomonas reinhardtii
Arabidopsis thaliana
Homo sapiens
ML DM MP
93 99 99
Caenorhabditis elegans
Dictyostelium discoideum
Naegleria gruberi
Cyanidioschyzon merolae
Saccharomyces cerevisiae
p = 0.003
99
Caulobacter crescentus
Escherichia coli
85
Sulfolobus solfataricus
Pyrococcus horikoshii
Neobodo saliens
ML DM MP
68 60 96
Parabodo caudaus
Trypanosoma brucei
Trypanosoma cruzi
Leishmania mexicana
kinetoplastids
70
Chlorobium tepidum
ML DM MP
99 99 91
72
Thermus thermophilus
Streptomyces coelicolor
Roseiflexus sp. RS-1
Plasmodium falciparum
0.1 substitutions/site

Accepted for publication in *Protist*, 4 Feb 2008

Evolutionary analysis of synteny and gene fusion for pyrimidine biosynthetic enzymes in Euglenozoa: An extraordinary gap between kinetoplastids and diplonemids

Takashi Makiuchi [a], Takeshi Annoura [a], Tetsuo Hashimoto [b], Eri Murata [a], Takashi Aoki [a], and Takeshi Nara [a,1]

[a] Department of Molecular and Cellular Parasitology, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan
[b] Institute of Biological Sciences, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-8572, Japan

Running title: Evolution of synteny and gene fusion in Euglenozoa

[1]Corresponding author;
fax +81 3 5800 0476
e-mail tnara@med.juntendo.ac.jp (T. Nara).

Supplementary Materials and Methods

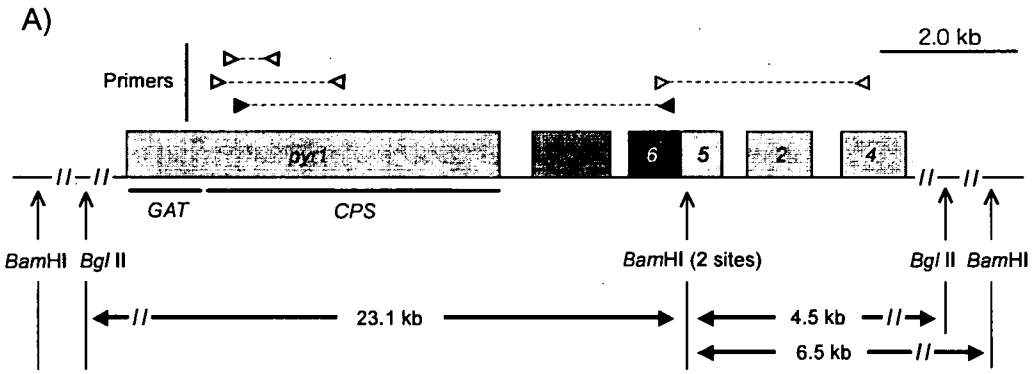
Southern blot analysis: To prepare the DNA probe for each *pyr* gene of *Parabodo caudatus*, PCR amplification was performed using a DIG probe synthesis kit (Roche Diagnostics), the *pyr*-specific primer set, and *P. caudatus* genomic DNA. The primers were designed to give an average size of 300 bp for each *pyr* gene, and the resulting DNA probes corresponded to the following positions: *pyr1*, nt 1877-2181 in the ORF; *pyr3*, nt 115-531; *pyr6/pyr5*, nt 555-819; *pyr2*, nt 74-610; and *pyr4*, nt 8-336. Three μg of *P. caudatus* genomic DNA was digested with *Bgl*II, *Bam*HI, or both overnight at 37°C, separated on 0.8 % agarose gels, and transferred to positively-charged nylon membranes (Roche Diagnostics). Hybridization with each DNA probe and colorimetric detection of the signals were performed under the conditions recommended by the manufacturer.

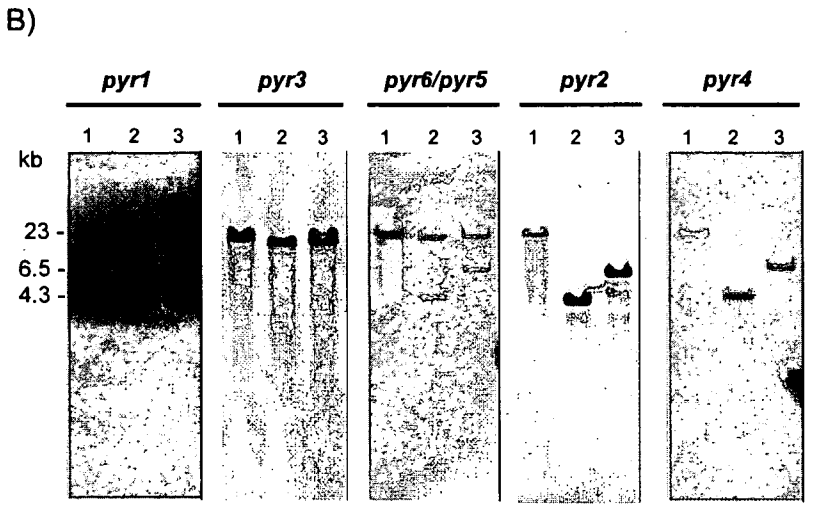
Legends for Supplementary Figures

Supplementary Fig. S1. The *pyr* gene cluster in the bodonid, *Parabodo caudatus*. (A) Schematic diagram of the *P. caudatus pyr* gene cluster, in which *pyr1, pyr3, pyr6/pyr5, pyr2,* and *pyr4* are juxtaposed in this order. Positions of the primers and the corresponding PCR products are indicated by arrowheads and dashed lines, respectively. There are two closely located *Bam*HI sites within *pyr6/pyr5*. (B) Southern blot analysis of the *pyr* gene cluster in *P. caudatus*. *P. caudatus* genomic DNA was digested with *Bgl*II (lane 1), *Bgl*II and *Bam*HI (lane 2), or *Bam*HI (lane 3) and hybridized with each DNA probe. A *Bgl*II fragment of > 23 kb hybridized with all *pyr* genes. Double digestion allows separation of a ≈ 23 kb fragment containing *pyr1, pyr3,* and the 5'-half of *pyr6/pyr5* and a 4.5 kb fragment, containing the 3'-half of *pyr6/pyr5, pyr2,* and *pyr4*. Digestion with *Bam*HI alone yields a >23 kb longer fragment (> 23 kb) containing *pyr1, pyr3,* and the 5'-half of *pyr6/pyr5*, and a 6.5 kb fragment containing the 3'-half of *pyr6/pyr5, pyr2,* and *pyr4*. Note that two bands are observed with the *pyr6*/pyr5 probe in the *Bgl*II/*Bam*HI and *Bam*HI restricted DNA, because two *Bam*HI sites are located within the probe.

Supplementary Fig. S2. Monophyly of kinetoplastids in the phylogenetic tree of the CPS domain of pyr1. The consensus tree of the maximum likelihood (ML) method with 100 replicates is shown. The tree was inferred by the JTT model taking across-site rate heterogeneity into consideration. The α-value of the Γ shape parameter used in the analysis was 0.64852. Bootstrap proportion (BP) values are attached to the internal branches. Branches with less than 50% BP support are unmarked. BP values for a kinetoplastid clade are given by ML, distance matrix (DM) and maximum parsimony (MP) methods. The length of each branch is proportional to the estimated number of substitutions. Analysis was performed using 828 unambiguously aligned amino acid sites with 26 taxa, corresponding to residues 378-471, 479-564, 570-636, 656-706, 709-711, 713-756, 758-768, 805-838, 844-868, 880-902, 905-916, 924-996, 999-1103, 1106-1118, 1121-1140, 1142-1173, 1178-1194, 1197-1237, 1252-1299 and 1400-1428 of the *P. caudatus* sequence.

Supplementary Fig. S3. Monophyly of kinetoplastids in the phylogenetic tree of pyr3. The consensus tree of the maximum likelihood (ML) method with 100 replicates is shown. The tree was inferred by the JTT model taking across-site rate heterogeneity into consideration. The α-value of the Γ shape parameter used in the analysis was 0.88172. Bootstrap proportion (BP) values are attached to the internal branches. Branches with less than 50% BP support are unmarked. BP values for a kinetoplastid clade are marked by an asterisk and given by ML, distance matrix (DM) and maximum parsimony (MP) methods. The length of each branch is proportional to the estimated number of substitutions. Analysis involved 174 unambiguously aligned amino acid sites with 30 taxa, corresponding to residues 3-16, 41-48, 50-52, 70-73, 76-79, 102-118, 138-148, 180-201, 214-222, 235-249, 252-257, 260-276, 281-300, 319-337 and 366-370 of the *P. caudatus* sequence.


A)
2.0 kb
Primers
pyr1
6
5
2
4
GAT
CPS
BamHI
Bgl II
BamHI (2 sites)
Bgl II
BamHI
23.1 kb
4.5 kb
6.5 kb


B)
pyr1
pyr3
pyr6/pyr5
pyr2
pyr4
1 2 3
kb
23 -
6.5 -
4.3 -