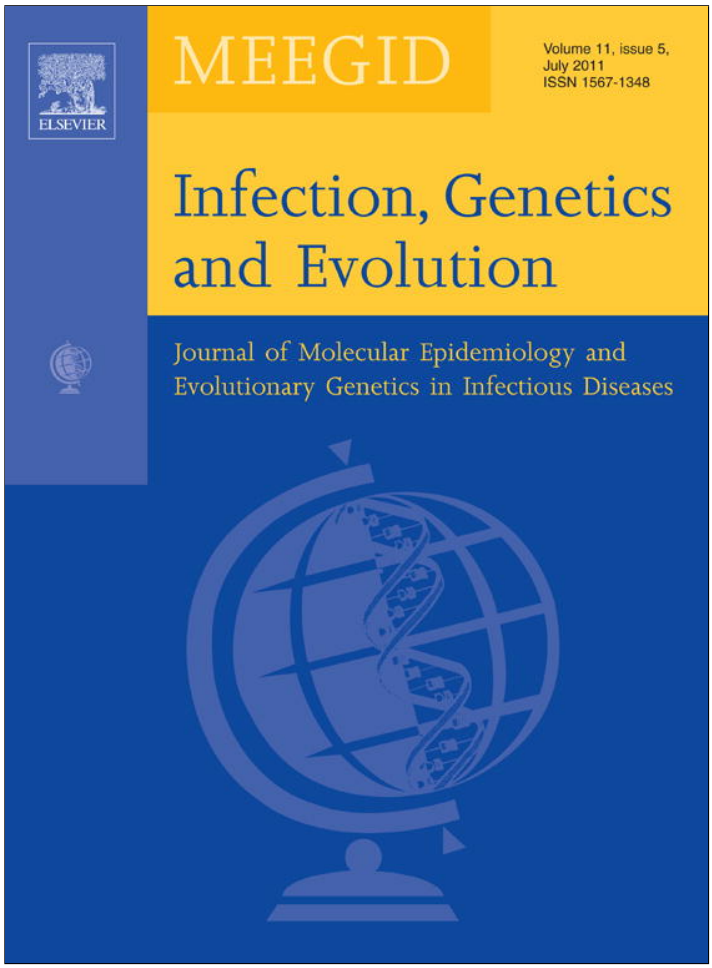

Infection, Genetics and Evolution 11 (2011) 955–959

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Resolving the question of trypanosome monophyly: A comparative genomics approach using whole genome data sets with low taxon sampling

Guy Leonard [1], Darren M. Soanes, Jamie R. Stevens *

Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter EX4 4QD, UK

ARTICLE INFO

Article history:
Received 7 January 2011
Received in revised form 10 March 2011
Accepted 11 March 2011
Available online 17 March 2011

Keywords:
Trypanosoma
Leishmania
Kinetoplastida
Monophyletic
Paraphyletic
Paralogous genes
Paralogue
Orthologue

ABSTRACT

Since the first attempts to classify the evolutionary history of trypanosomes, there have been conflicting reports regarding their true phylogenetic relationships and, in particular, their relationships with other vertebrate trypanosomatids, e.g. *Leishmania* sp., as well as with the many insect parasitising trypanosomatids. Perhaps the issue that has provided most debate is that concerning the monophyly (or otherwise) of genus *Trypanosoma* and, even with the advent of molecular methods, the findings of numerous studies have varied significantly depending on the gene sequences analysed, number of taxa included, choice of outgroup and phylogenetic methodology. While of arguably limited applied importance, resolution of the question as to whether or not trypanosomes are monophyletic is critical to accurate evaluation of competing, mutually exclusive evolutionary scenarios for these parasites, namely the 'vertebrate-first' or 'insect-first' hypotheses. Therefore, a new approach, which could overcome previous limitations was needed. At its most simple, the problem can be defined within the framework of a trifurcated tree with three hypothetical positions at which the root can be placed. Using BLASTp and whole-genome gene-by-gene phylogenetic analyses of *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania major* and *Naegleria gruberi*, we have identified 599 gene markers – putative homologues – that were shared between the genomes of these four taxa. Of these, 75 homologous gene families that demonstrate monophyly of the kinetoplastids were identified. We then used these data sets in combination with an additional outgroup, *Euglena gracilis*, coupled with large-scale gene concatenation and diverse phylogenetic techniques to investigate the relative branching order of *T. brucei*, *T. cruzi* and *L. major*. Our findings confirm the monophyly of genus *Trypanosoma* and demonstrate that <1% of the analysed gene markers shared between the genomes of *T. brucei*, *T. cruzi* and *L. major* reject the hypothesis that the trypanosomes form a monophyletic group.

1. Introduction

The group of flagellate protozoa known as Kinetoplastida includes the genera *Trypanosoma* and *Leishmania,* both of which include vertebrate parasites of considerable medical and veterinary importance, in addition to others such as *Crithidia* and *Leptomonas*, which are parasites of arthropods (Lake et al., 1988; Podlipaev et al., 2004). The genus *Trypanosoma* includes: *Trypanosoma brucei*, an extracellular parasitic protist, which causes sleeping sickness in humans and a similar wasting disease known as Nagana in mammals (Hoare, 1972); *Trypanosoma cruzi*, an intracellular parasitic protist that causes Chagas disease in humans and also infects a range of mammals which can act as reservoirs of the human form of the disease (Hoare, 1972); and *Leishmania major*, an intracellular parasitic protist which causes the disease known as leishmaniasis in humans and animals – leishmaniasis can take a variety of forms, cutaneous, subcutaneous and visceral, and pathogenicity varies widely between hosts (Molyneux and Ashford, 1983).

Since the first attempts to classify the evolutionary history of these three important parasite genera, there have been conflicting reports regarding their true phylogenetic relationships, which have varied significantly depending on the gene sequences analysed, the number of taxa included, choice of outgroup and phylogenetic methodology employed (e.g. Alvarez et al., 1996; Hamilton et al., 2004; Hughes and Piontkivska, 2003; Lukeš et al., 1997; Maslov et al., 1996; Piontkivska and Hughes, 2005; Simpson et al., 2004, 2006; Stevens et al., 1999, 2001; Wright et al., 1999). The issue that appears to have provided most debate is that concerning the monophyly (or otherwise) of the trypanosomes

* Corresponding author at: Biosciences, College of Life and Environmental Sciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter EX4 4QD, UK. Tel.: +44 1392 723775; fax: +44 1392 723700.
E-mail addresses: guy.leonard@gmail.com (G. Leonard), D.M.Soanes@exeter.ac.uk (D.M. Soanes), J.R.Stevens@exeter.ac.uk (J.R. Stevens).
[1] Current address: Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.

1567-1348/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.03.005

(see Simpson et al., 2006 for an overview of this topic). While of arguably only limited applied importance, resolution of the question as to whether or not trypanosomes are monophyletic is critical to accurate evaluation of competing, mutually exclusive evolutionary scenarios for these parasites, namely the 'vertebrate-first' (Minchin, 1908; Wallace, 1966) or 'insect-first' (Baker, 1963, 1994; Hoare, 1972) hypotheses; see Hamilton and Stevens (2010) for an overview.

Fortunately, the problem can be defined relatively simply within the framework of a trifurcated tree – a topology with three branches – with three hypothetical positions at which the root can be placed (Fig. 1). Topology *X* describes the monophyly of trypanosomes, whereas topologies *Y* and *Z* describe the genus as paraphyletic. Of the studies cited above, early (typically less taxon-rich) rRNA-based studies have indicated paraphyly, while later studies have (depending on taxa included, outgroup and phylogenetic methodology) provided support for either outcome, some offering support for monophyly, with others suggesting paraphyly. Of those that provide support for paraphyly, most conform to topology Y, grouping *T. cruzi* with *Leishmania* sp. to the exclusion of *T. brucei*. Similarly, evidence from the genome content of *L. major* and *T. cruzi* (Berriman et al., 2005; El-Sayed et al., 2005; Ivens et al., 2005) – specifically protein families which are expanded in *T. brucei* compared to *L. major* and *T. cruzi*, and a large set of orthologues shared between *L. major* and *T. cruzi* – supports their grouping to the exclusion of *T. brucei*. Nevertheless, although gene content may support such a hypothesis, a number of other factors do not agree and, moreover, phylogenetic studies of a range of protein-coding genes (Hannaert et al., 1992, 1998; Hashimoto et al., 1995; Adjé et al., 1998; Simpson et al., 2002; Hamilton et al., 2004) have unequivocally supported monophyly of genus *Trypanosoma*.

Against this backdrop of conflicting topologies, improved phylogenetic methodologies and broader data sets have allowed for multiple nucleotide and protein alignments (2–9 genes) to be constructed (Hamilton et al., 2004, 2007; Simpson et al., 2002, 2004) and these analyses have indicated genus *Trypanosoma* to be monophyletic, grouping *T. brucei* and *T. cruzi* to the exclusion of *L. major* (Fig. 1, topology *X*). Nonetheless, inappropriate taxon sampling and the associated problems of compositional bias, hidden paralogy and horizontal gene transfer (HGT), continue to diminish support for relationships defined in phylogenetic analyses and have led to uncertainty in the reconstruction of the kinetoplastids' evolutionary history.

Therefore, a new approach, which could attempt to overcome or reduce the effects of these problems, was needed. Using whole genome datasets we undertook an analysis of protein-coding genes in order to resolve the branching relationships of *T. brucei*, *T. cruzi* and *L. major*. Using the bioinformatic methodology described by (Richards et al., 2009) to conduct whole-genome gene-by-gene phylogenetic analyses, we first identified reliable homologous gene families that demonstrated monophyly of the kinetoplastids within a broader eukaryotic phylogeny. We then used this data set in combination with an outgroup of *Naegleria gruberi* and/or *Euglena gracilis*, coupled with large-scale gene concatenation and diverse phylogenetic techniques to investigate the relative branching order of *T. brucei*, *T. cruzi* and *L. major*.

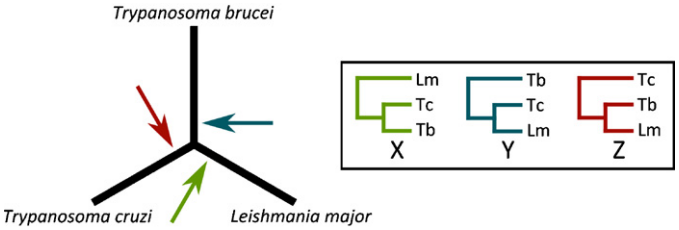

Fig. 1. A representation of a trifurcated tree topology for the kinetoplastids. The insert depicts three topologies, *X*, *Y* and *Z*, which show the branching order of the three kinetoplastid genera considered in this paper. Note that topologies *Y* and *Z* indicate paraphyly of the trypanosomes, whereas topology *X* depicts their monophyly.

2. Materials and methods

2.1. Pipelined genome-to-genome analysis of homologues

The methodology employed in this paper utilised an automatic tree-building pipeline (known as 'Darren's Orchard') described previously (Richards et al., 2009). Briefly, potential orthologues from across the predicted proteomes of *T. brucei*, *T. cruzi*, *L. major* and *N. gruberi* were subjected to a sequential, genome-to-genome BLASTp (Altschul et al., 1997) analysis against 795 other eukaryotic and prokaryotic genomes whose species were picked, as far as possible, to be representative of the whole tree of life.

Firstly, proteins from the four organisms were clustered using OrthoMCL (Li et al., 2003) to group together potential orthologues (*e*-value cut-off of $1e^{-20}$; inflation value 1.5). The pipeline was then used to create a phylogenetic tree for each of the 599 clusters identified with OrthoMCL, each cluster being screened against a database of 795 genomes, which comprised the mix of both eukaryotic and prokaryotic taxa utilised by Richards et al. (2009), together with all prokaryotic reference genomes available from NCBI at the time of analysis. The pipeline used BLASTp to select homologous proteins from genome databases, sequences were aligned using the program MUSCLE (Edgar, 2004), conserved regions from each alignment were sampled using GBLOCKS (Castresana, 2000), and phylogenetic trees were constructed using PhyML (Guindon and Gascuel, 2003) with a WAG + G + I substitution model (G + I parameters were estimated by PHYML). Further information on the pipeline used in this analysis can be found at: http://cogeme.ex.ac.uk/.

2.2. Outgroup choice

Outgroup choice was based on the most closely related sequenced genome available to the kinetoplastids at the time of analysis (Hampl et al., 2009), the genome of *N. gruberi* (Fritz-Laylin et al., 2010) available from the Joint Genome Institute website (http://genome.jgi-psf.org). *N. gruberi* was deemed most suitable for this set of analyses as it is contained within the super-phylum of the Discicristata (Cavalier-Smith, 1998) which are comprised of a set of unicellular protists and are so called as they contain mitochondria which possess disc-shaped cristae. Genus *Naegleria* is within the class Heterolobosea, the sister-group to Euglenozoa, which contains the Kinetoplastida.

2.3. BLASTp analysis and construction of concatenated data sets

The output of the BLASTp analyses showed the presence of 599 gene markers, putative homologues that were shared between the genomes of *T. brucei* (Tb), *T. cruzi* (Tc), *L. major* (Lm) and *N. gruberi* (Ng). These 599 maximum likelihood (ML) trees were assessed visually for the presence of kinetoplastid monophyly, which identified 75 reliable genes for analysis, which in turn showed support for three different possible tree topologies. These 75 genes were split into three data sets; data set *X* included all genes which recovered the topology of (Lm, (Tb, Tc)), data set *Y* included all genes which recovered the topology of (Tb, (Tc, Lm)) and data set *Z* included all genes which recovered the topology of (Tc, (Tb, Lm)). These topologies followed the three possible branching orders outlined in Fig. 1. Once the data sets were assembled, the protein

sequences from each tree were collated and concatenated together. A further data set was assembled; this comprised a concatenated alignment containing all of the genes present in data sets *X*, *Y* and *Z* and contained 36,278 individual sites.

2.4. Phylogenetic analyses

Each of the four concatenated alignments were then subjected to six different phylogenetic methodologies and alternative topology tests; two fast-ML topology tests (PhyML (Guindon and Gascuel, 2003) and RAxML (Stamatakis, 2006)) with both the LG model (Le and Gascuel, 2008; Le et al., 2008) and the best model optimised by MODELGENERATOR (Keane et al., 2006) (at the time of analysis the LG model was new and was not available in MODELGENERATOR), two approximate likelihood ratio tests (SH and X^2), a log-det approach with LDDist and finally a Bayesian analysis in the program MrBayes (Huelsenbeck and Ronquist, 2001).

2.5. Paralogue mirror-tree analysis

Amongst the 75 reliable genes/data sets identified for analysis, eight showed the presence of reciprocal rooting of paralogous genes within the kinetoplastids. These eight universal data sets were concatenated and subjected to the same methods and analysis as previously, in order to produce paralogous gene trees (Brown and Doolittle, 1995) or so called 'mirror-trees' (Pazos and Valencia, 2001), which each produce two distinct clades, where the root is inferred from the location of the branch connecting the pair of paralogous gene trees. Analyses were conducted twice, firstly with the two *Trypanosoma* genomes with that of *L. major* and, secondly, with the addition of *N. gruberi* as an outgroup.

2.6. Extended phylogenetic analysis with increased taxa and reduced gene sampling

To account for effects due to artefacts created by long-branch attraction (LBA) (Felsenstein, 1978; Philippe et al., 2005), three other closely related species were added to the phylogenetic analysis. These were *Leishmania braziliensis*, *Trypanosoma vivax* and *E. gracilis*. For *E. gracilis* (which is grouped within the Euglenozoa), data were obtained from expressed sequence tags (ESTs) from the taxonomically broad EST database (TBestDB) (TbestDB, 2010), which was then used as a second outgroup species; *Crithidia deanei* (Kinetoplastida), *Leptomonas seymouri* (Kinetoplastida), *Diplonema papilatum* (Euglenozoa) and *Bodo saltans* (Euglenozoa) were also assessed for their suitability as outgroups.

3. Results

3.1. Identification of potential homologues using OrthoMCL

Analysis of 52,411 proteins from the four organisms using OrthoMCL (Li et al., 2003) to group together potential orthologues (*e*-value cut-off of $1e^{-20}$; inflation value 1.5) produced 8888 clusters and 20,373 singletons. Of these, 599 clusters contained at least one protein from each species. Of these, 165 clusters had one protein from each species, the rest contained multi-gene families. The pipeline was then used to create a phylogenetic tree for each of the 599 clusters, sampling proteins from 795 taxa from across the tree of life.

3.2. Multi-gene phylogenetic analysis of kinetoplastids

Of the 599 trees resulting from the automated BLASTp analysis of the predicted functional proteins of the four organisms, 75 produced monophyly of the kinetoplastids. These 75 were sorted into three data sets, based upon which topology they presented (Fig. 1). Of these, 58 of the tree topologies recovered a grouping of *T. brucei* with *T. cruzi* to the exclusion of *L. major* (Lm, (Tb, Tc)); the two other possible topologies were recovered less frequently, with (Tb, (Tc, Lm)) occurring 8 times, and (Tc, (Tb, Lm)) 9 times. These latter 17 data sets, which did not support monophyly of the *Trypanosoma*, may therefore be the product of horizontal gene transfer, hidden paralogy, or artefacts of phylogenetic reconstruction; alternatively, these topologies may reflect the distinct gene histories of each of these 17 sets.

Data sets *X*, Y, *Z* and the full concatenated alignment were subjected to a range of phylogenetic topology tests and methodologies (RaxML/PhyML + LG/PhyML + G/SH/X^2/MrBayes), the results of which can be viewed in Fig. 2, where colour intensity of each data set varies depending on the level of support for each topology under a specific methodology/model. For example, data set *X* (green) has full support in all tests for the topology of (Lm, (Tb, Tc)) and no other topology is supported by this data set. Conversely, data set *Y* (blue/grey) shows differential support values across multiple topologies, with the majority of support being for topology *X*. Data set *Z* recovers near full support for the grouping of *T. brucei* and *T. cruzi*, with an insignificant posterior probability value under the RAxML maximum likelihood test for its own topology (Tc (Tb, Lm)). The complete concatenated data set (purple) recovered full support across all tests and methods for the grouping of *T. brucei* and *T. cruzi* together, to the exclusion of *L. major* (Lm (Tb, Tc)). Bayesian analysis of the full concatenation (of data sets *X*, *Y* and *Z*) produced complete, unequivocal support for a ((Ng, Lm),(Tb, Tc)) topology, with full unequivocal support for monophyly of the *Trypanosoma* in all six additional models assessed (7 in total, as per Fig. 2).

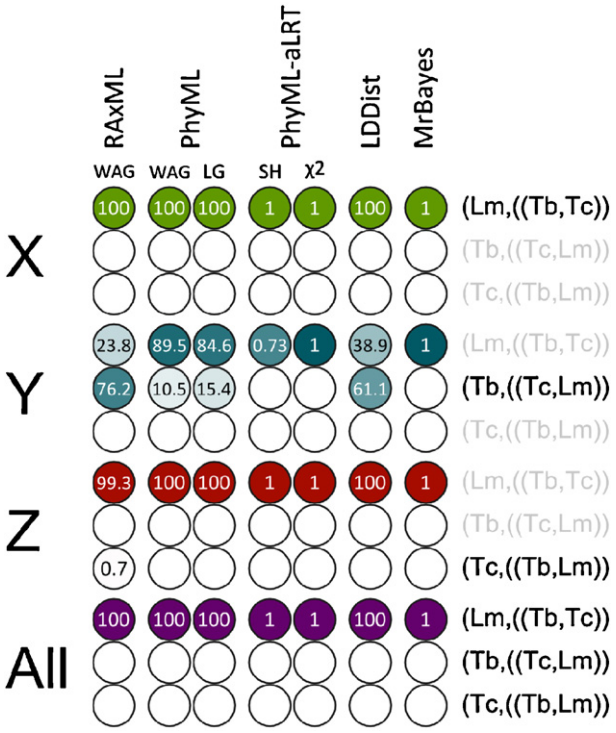

Fig. 2. Schema outlining the support values returned for each topology for the three data sets (*X*, *Y* and *Z*) and the total concatenated data set (All). Main topologies for each data set are shown in black, non-predominant topologies are shown in grey. [Note, concatenated data sets for *Y* and Z, although derived from individual analyses that support a *Y* or *Z* topology (Fig. 1), once concatenated together (All) strongly support topology *X*. Such a finding suggests the source of error here is probably phylogenetic artefact and not HGT or hidden paralogy].

The conflicting support present in data set *Y* (Fig. 2) is a result of the phylogenetic analyses performed on the eight conflicting sets of gene markers constituting this data set and the associated concatenated alignment. Consequently, to better understand the varied phylogenetic signal present in data set Y, separate Bayesian analyses were performed for each of the eight gene markers, together with assessment of additional models (RaxML/PhyML + LG/PhyML + G/SH/X^2/Mr Bayes) (Supplementary Data 1). Of the resulting eight trees, three recovered monophyly of *Trypanosoma* ((Ng, Lm),(Tb, Tc)), albeit with varying levels of support depending on the model used, four recovered ((Ng, Tb),(Tc, Lm)), and one recovered ((Ng, Tc),(Tb, Lm)); support values across methods and trees especially the five trees which did not recover a ((Ng, Lm),(Tb, Tc)) topology – were variable and often weak, with the ((Ng, Tc),(Tb, Lm)) topology in particular receiving uniformly very weak support with all methods/models. The low support values for topology ((Ng, Tb),(Tc, Lm)) is particularly surprising given that they represent the topology of the data set (*Y*). Thus, overall, the mixed support for the different topologies in data set *Y* is indicative of the conflicting signal within this data set.

3.3. Paralogue mirror-tree analysis

Amongst the 75 data sets which produced monophyly of the kinetoplastids, eight data sets showed the presence of reciprocal rooting of paralogous genes within the kinetoplastids. Analysis of paralogous gene trees (Brown and Doolittle, 1995; Pazos and Valencia, 2001) was conducted twice, firstly with the three trypanosomatid genomes (i.e., the two *Trypanosoma* genomes, plus that of *L. major*) and, secondly, with the addition of *N. gruberi* as an outgroup. In both cases the same topology was recovered with full, almost unequivocal support across all tests and methods for the grouping of *T. brucei* and *T. cruzi* together, to the exclusion of *L. major* (Fig. 3).

3.4. Phylogenetic analysis with increased taxa and reduced gene sampling

The 58 gene families identified that comprised data set *X* (i.e., produced topology *X*, Fig. 1) were subjected to additional BLASTp searches against *C. deanei*, *L. seymouri*, *D. papilatum* and *B. saltans*. However, the majority of gene families either did not return sufficient reliable BLASTp hits from the available EST databases, or the ESTs were simply unavailable; this resulted in a somewhat reduced data set. Further sequences were recovered from two other kinetoplastids: *T. vivax* and *L. braziliensis*, both from GeneDB (Hertz-Fowler and Hall, 2004). The genomes of *Leishmania mexicana* and *Trypanosoma congolense* were not included as they were considered too closely related to *L. major* and *T. brucei*.

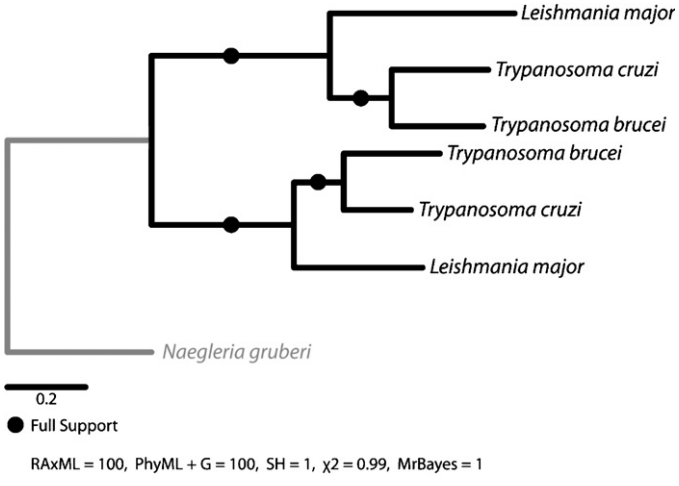

Fig. 3. Bayesian analysis (MrBayes) generated from eight reciprocally rooted paralogous data sets, indicating full support for the monophyletic grouping of *T. brucei* and *T. cruzi*. *N. gruberi* (grey) was included in the secondary analysis; the same topologies and almost identical support values were recovered.

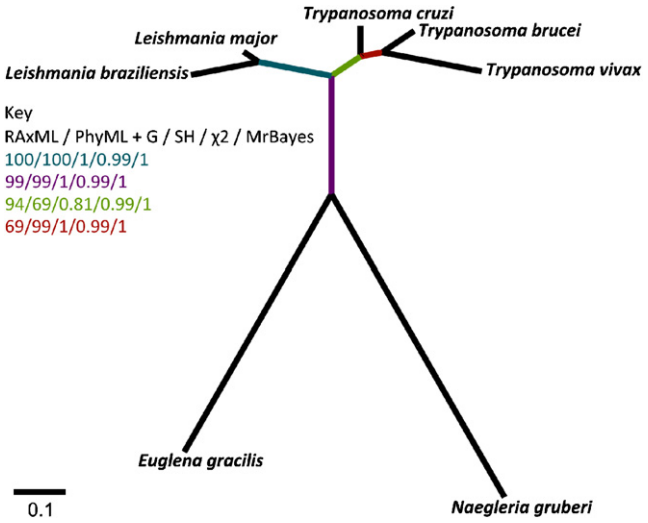

Fig. 4. Bayesian analysis (MrBayes) including data from three additional taxa, *L. braziliensis*, *T. vivax* and *E. gracilis*. Topology support values in the key match corresponding internal branches of the tree. The green branch defines monophyly of *Trypanosoma* to the exclusion of both the *Leishmania* species.

Despite a reduction in gene family number and sites available for phylogenetic reconstruction, the additional analysis, which included three other closely related species, *L. braziliensis*, *T. vivax* and *E. gracilis*, demonstrated strong support for monophyly of genus *Trypanosoma* (Fig. 4).

4. Discussion

This paper attempts to resolve the contentious nature of the monophyly (Stevens et al., 1999, 2001; Hamilton et al., 2004; Simpson et al., 2006) or paraphyly (Hughes and Piontkivska, 2003; Piontkivska and Hughes, 2005) hypotheses for the phylogeny of genus *Trypanosoma*. We have used a range of robust phylogenetic techniques adapted for analysis of few taxa, but utilising a 'whole genome'-based approach. The analyses used multi-gene concatenated alignments and selected the most appropriate outgroup taxa available at the time. This strategy demonstrated near-total support for the monophyly of *T. brucei* and *T. cruzi*, and has resolved the branching order of key kinetoplastid taxa. Additionally, this work outlines a new strategy for resolving phylogenies of whole genome data sets with low taxon sampling (3–4 genomes).

Of course, conclusions reached using a low-taxon approach are inevitably limited and are, to a large extent, dependent on the idiosyncrasies of the taxa included. Thus, while this study sheds light on the nature of trypanosome monophyly, a more taxon rich data set will be required to fully elucidate the evolutionary history of the kinetoplastids. Additionally, we are aware of the potential confounding effects of long-branch attraction that use of *N. gruberi* as an outgroup may have given rise to; certainly, in subsequent studies, a more closely related outgroup should help to alleviate this problem, while also allowing a potentially greater number of potential clusters to be evaluated by OrthoMCL analysis.

Nevertheless, rejection of monophyly for genus *Trypanosoma* in the current data set appears to be very weak; indeed, of the initial 599 gene families only 17 (Fig. 2, data sets *Y* and *Z*) appeared to suggest alternative topologies. Moreover, further analysis of the

genes in data set *Z* returned a phylogeny supporting monophyly of the trypanosomes, along with four of the eight genes in data set *Y* (Fig. 2; Supplementary Data 1). This suggested that only five genes (from data set *Y*), or <1% of the original 599 gene markers shared between the genomes of *T. brucei*, *T. cruzi*, *L. major* and *N. gruberi*, reject the hypothesis that the trypanosomes form a monophyletic group. Thus, our findings provide strong support for the monophyly of genus *Trypanosoma*.

Additionally, the approach outlined here, i.e., the use of an automated tree topology building/assessment pipeline (Richards et al., 2009) and reciprocally rooted paralogous gene trees (mirrortrees), has been completed for all gene families within an entire genome and is proposed as a new systematic method for investigating phylogenetic conflicts. Specifically, these methods can be adapted for addressing phylogenetic conflict hypotheses amongst trifurcated branching relationships, where taxon sampling is limited to a few whole genomes and where selection of distantly related taxa as an outgroup is the only viable option.

Acknowledgments

We are grateful to the University of Exeter for providing funding for this work. We thank Dr T. A. Richards, Exeter (currently Natural History Museum, London), for help and advice on this research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2011.03.005.

References

Adjé, C.A., Opperdoes, F.R., Michels, P.A.M., 1998. Molecular analysis of phosphoglycerate kinase in *Trypanoplasma borreli* and the evolution of this enzyme in Kinetoplastida. Gene 217, 91–99.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Alvarez, F., Cortinas, M.N., Musto, H., 1996. The analysis of protein coding genes suggests monophyly of *Trypanosoma*. Mol. Phylogenet. Evol. 5, 333–343.

Baker, J.R., 1963. Speculations on the evolution of the family Trypanosomatidae Doflein 1901. Exp. Parasitol. 13, 219–233.

Baker, J.R., 1994. The origins of parasitism in the protists. Intl. J. Parasitol. 24, 1131–1137.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., et al., 2005. The genome of the African trypanosome *Trypanosoma brucei*. Science 309, 416–422.

Brown, J.R., Doolittle, W.F., 1995. Root of the universal tree of life based on ancient aminoacyl-transfer RNA synthetase gene duplications. Proc. Natl. Acad. Sci. U.S.A. 92, 2441–2445.

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

Cavalier-Smith, T., 1998. A revised six-kingdom system of life. Biol. Rev. Camb. Philos. Soc. 73, 203–266.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.-N., et al., 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309, 409–415.

Felsenstein, J., 1978. Cases in which parsimony or compatibility will be positively misleading. Syst. Zool. 27, 401–410.

Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., et al., 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140, 631–642.

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

Hampl, V., Hug, L., Leigh, J.W., Dacks, J.B., Lang, B.F., et al., 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc. Natl. Acad. Sci. U.S.A. 106, 3859–3864.

Hannaert, V., Blaauw, M., Kohl, L., Allert, S., Opperdoes, F.R., Michels, P.A., 1992. Molecular analysis of the cytosolic and glycosomal glyceraldehyde-3-phosphate dehydrogenase in *Leishmania mexicana*. Mol. Biochem. Parasitol. 55, 115–126.

Hannaert, V., Opperdoes, F.R., Michels, P.A.M., 1998. Comparison and evolutionary analysis of the glycosomal glyceraldehyde-3-phosphate dehydrogenase from different Kinetoplastida. J. Mol. Evol. 47, 728–738.

Hashimoto, T., Nakamura, Y., Kamaishi, T., Adachi, J., Nakamura, F., Okamoto, K., Hasegawa, M., 1995. Phylogenetic place of kinetoplastid protozoa inferred from a protein phylogeny of elongation factor 1α. Mol. Biochem. Parasitol. 70, 181–185.

Hamilton, P.B., Stevens, J.R., Gaunt, M.W., Gidley, J., Gibson, W.C., 2004. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Intl. J. Parasitol. 34, 1393–1404.

Hamilton, P.B., Gibson, W.C., Stevens, J.R., 2007. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. Mol. Phylogen. Evol. 44, 15–25.

Hamilton, P.B., Stevens, J.R., 2010. Classification and phylogeny of *Trypanosoma cruzi*. In: Telleria, J., Tibayrenc, M. (Eds.), American Trypanosomiasis: Chagas Disease, One Hundred Years of Research. Elsevier, Chapter 13, pp. 321–338.

Hertz-Fowler, C., Hall, N., 2004. Parasite genome databases and web-based resources. Methods Mol. Biol. 270, 45–74.

Hoare, C.A., 1972. The Trypanosomes of Mammals. Blackwell, Oxford, p. 749.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Hughes, A.L., Piontkivska, H., 2003. Phylogeny of Trypanosomatidae and Bodonidae (Kinetoplastida) based on 18S rRNA: evidence for paraphyly of *Trypanosoma* and six other genera. Mol. Biol. Evol. 20, 644–652.

Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., et al., 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science 309, 436–442.

Keane, T., Creevey, C., Pentony, M., Naughton, T., McInerney, J., 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. 6, 29.

Lake, J.A., de la Cruz, V.F., Ferreira, P.C., Morel, C., Simpson, L., 1988. Evolution of parasitism: kinetoplastid protozoan history reconstructed from mitochondrial rRNA gene sequences. Proc. Natl. Acad. Sci. U.S.A. 85, 4779–4783.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320.

Le, S.Q., Lartillot, N., Gascuel, O., 2008. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. B: Biol. Sci. 363, 3965–3976.

Li, L., Stoeckert Jr., C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.

Lukeš, J., Jirků, M., Doležel, D., Kral'ová, I., Hollar, L., Maslov, D.A., 1997. Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic. J. Mol. Evol. 44, 521–527.

Maslov, D.A., Lukes, J., Jirku, M., Simpson, L., 1996. Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: implications for the evolution of parasitism in the trypanosomatid protozoa. Mol. Biochem. Parasitol. 75, 197–205.

Minchin, E.A., 1908. Investigations on the development of trypanosomes in tsetse flies and other Diptera. Q. J. Microsc. Sci. 52, 159–260.

Molyneux, D.H., Ashford, R.W., 1983. The Biology of *Trypanosoma* and *Leishmania*, Parasites of Man and Domestic Animals. Taylor and Francis, London, p. 294.

Pazos, F., Valencia, A., 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng. 14, 609–614.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., Delsuc, F., 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5, 50.

Piontkivska, H., Hughes, A.L., 2005. Environmental kinetoplastid-like 18S rRNA sequences and phylogenetic relationships among Trypanosomatidae: paraphyly of the genus Trypanosoma. Mol. Biochem. Parasitol. 144, 94–99.

Podlipaev, S.A., Sturm, N.R., Fiala, I., Fernanades, O., Westenberger, S.J., Dollet, M., Campbell, D.A., Lukes, J., 2004. Diversity of insect trypanosomatids assessed from the spliced leader RNA and 5S rRNA genes and intergenic regions. J. Eukaryot. Microbiol. 51, 283–290.

Richards, T.A., Soanes, D.M., Foster, P.G., Leonard, G., Thornton, C.R., Talbot, N.J., 2009. Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell 21, 1897–1911.

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

Simpson, A.G.B., Gill, E.E., Callahan, H.A., Litaker, R.W., Roger, A.J., 2004. Early evolution within Kinetoplastids (Euglenozoa), and the late emergence of trypanosomatids. Protist 155, 407–422.

Simpson, A.G.B., Stevens, J.R., Lukes, J., 2006. The evolution and diversity of kinetoplastid flagellates. Trends Parasitol. 22, 168–174.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Stevens, J.R., Noyes, H.A., Dover, G.A., Gibson, W.C., 1999. The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. Parasitology 118, 107–116.

Stevens, J.R., Noyes, H.A., Schofield, C.J., Gibson, W., 2001. The molecular evolution of Trypanosomatidae. Adv. Parasitology 48, 1–56.

TbestDB, 2010. Taxonomically Broad EST Database. , http://amoebidia.bcm.umontreal.ca/pepdb/.

Wallace, F.G., 1966. The trypanosomatid parasites of insects and arachnids. Exp. Parasitol. 18, 124–193.

Wright, A.-D.G., Li, S., Feng, S., Martin, D.S., Lynn, D.H., 1999. Phylogenetic position of the kinetoplastids, *Cryptobia bullocki*, *Cryptobia catostomi*, and *Cryptobia salmositica* and monophyly of the genus *Trypanosoma* inferred from small subunit ribosomal RNA sequences. Mol. Biochem. Parasitol. 99, 69–76.