Current Biology 23, 1787–1793, September 23, 2013 ©2013 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.07.045

Report

Paratrypanosoma Is a Novel Early-Branching Trypanosomatid

Pavel Flegontov,[1,9] Jan Votýpka,[1,2,9] Tomáš Skalický,[1,3,9] Maria D. Logacheva,[4,5] Aleksey A. Penin,[4,6] Goro Tanifuji,[7] Naoko T. Onodera,[7] Alexey S. Kondrashov,[4,8] Petr Volf,[2] John M. Archibald,[7] and Julius Lukeš[1,3,*]

[1]Institute of Parasitology, Biology Centre, 37005 České Budějovice (Budweis), Czech Republic
[2]Department of Parasitology, Faculty of Science, Charles University, 12844 Prague, Czech Republic
[3]Faculty of Science, University of South Bohemia, 37005 České Budějovice (Budweis), Czech Republic
[4]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119991, Russia
[5]A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119991, Russia
[6]Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia
[7]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS B3H 4R2, Canada
[8]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

Summary

The kinetoplastids are a widespread and important group of single-celled eukaryotes, many of which are devastating parasites of animals, including humans [1–3]. We have discovered a new insect trypanosomatid in the gut of *Culex pipiens* mosquitoes. Glyceraldehyde-3-phosphate dehydrogenase- and SSU rRNA-based phylogenetic analyses show this parasite to constitute a distinct branch between the free-living *Bodo saltans* and the obligatory parasitic clades represented by the genus *Trypanosoma* and other trypanosomatids. From draft genome sequence data, we identified 114 protein genes shared among the new flagellate, 15 trypanosomatid species, *B. saltans*, and the heterolobosean *Naegleria gruberi*, as well as 129 protein genes shared with the basal kinetoplastid *Perkinsela* sp. Individual protein phylogenies together with analyses of concatenated alignments show that the new species, here named *Paratrypanosoma confusum* n. gen., n. sp., branches with very high support at the base of the family Trypanosomatidae. *P. confusum* thus represents a long-sought-after missing link between the ancestral free-living bodonids and the derived parasitic trypanosomatids. Further analysis of the *P. confusum* genome should provide insight into the emergence of parasitism in the medically important trypanosomatids.

Results

Isolation and Morphological and Ultrastructural Characterization

Out of 206 female mosquitoes (*Culex pipiens*) captured in Prague in June 2000, 25 were found to contain flagellates in their intestine. In most cases, these infections were confined to the midgut and stomodeal valve, characteristic for avian trypanosomes [4]. However, in the midgut and hindgut of several mosquitoes, different slowly moving flagellates, which were in one case successfully established as an axenic culture in SNB medium, were observed and were upon further study named *Paratrypanosoma confusum* n. gen., n. sp. (see below and "Taxonomic Summary").

The *P. confusum* culture is dominated by elongated promastigote-shaped cells, defined by the mutual position of the nucleus and kinetoplast DNA (kDNA) (Figure 1A). Occasionally, ovoid stages with morphology reminiscent of choanomastigotes occur (Figure 1B). Transmission electron microscopy (Figures 1C–1I) shows that the plasmalemma is underlain by a complete corset of subpellicular microtubules (Figures 1C and 1G). All kDNA is packed in a single dense disk, with minicircles stretched taut, located in the canonical position at the base of the single flagellum (Figures 1F and 1I). Considering the known correlation between the thickness of the disk and size of kDNA minicircles, their size in *P. confusum* is estimated to be 0.8 kb. The nucleus is usually located in or close to the center of the cell (Figures 1A, 1H, and 1I), and, as in other trypanosomatids, the kDNA division predates nuclear division (Figure 1H). A single long flagellum is supported by a prominent paraflagellar rod (Figure 1E), which is absent in the flagellar pocket (Figure 1D). Numerous vesicles reminiscent of acidocalcisomes and lipid bodies (Figures 1A–1C,1G–1I) are present throughout the cell.

SSU rRNA and gGAPDH Phylogeny

Nuclear small subunit (SSU) ribosomal RNA (rRNA) sequences were amplified from *P. confusum* DNA isolated from the axenic culture. In addition, identical or very similar partial SSU rRNA sequences have been amplified by nested PCR in monospecies pools of female *C. pipiens* and *C. modestus* mosquitoes trapped in southern Bohemia (Řežabinec) and southern Moravia (Mikulov), Czech Republic (data not shown). Short SSU rRNA fragments of 345 bp (accession numbers DQ813272–DQ813295) matching the *P. confusum* sequence with 95%–100% identity (data not shown) were previously amplified from *C. pipiens* and *C. tarsalis* mosquitoes collected in Colorado [5]. The *P. confusum* sequence and 219 nonidentical bodonid and trypanosomatid SSU rRNA sequences were aligned using SINA aligner [6] and were manually edited, resulting in 1,325 aligned characters (Figure S1A available online). The resulting maximum likelihood (ML) tree shows that *P. confusum* is clearly distinct from all known trypanosomatid clades: *Trypanosoma* [7], *Blastocrithidia-Leptomonas jaculum* [1, 8], *Herpetomonas* [9], *Phytomonas* [1], *Angomonas-Strigomonas* [10], *Sergeia*, *Leptomonas collosoma* [1, 8], and subfamily Leishmaniinae [11]. *P. confusum* is the most basal trypanosomatid branch with 98% bootstrap support (Figures 2 and S2A).

Next, the glycosomal glyceraldehyde-3-phosphate dehydrogenase (gGAPDH) gene was amplified and sequenced. An amino acid sequence alignment of 294 characters was constructed using 143 bodonid and trypanosomatid sequences (Figure S1B); phylogenetic model selection using Modelgenerator favored LG+Γ as the best model for this alignment. The

[9]These authors contributed equally to this work
*Correspondence: jula@paru.cas.cz

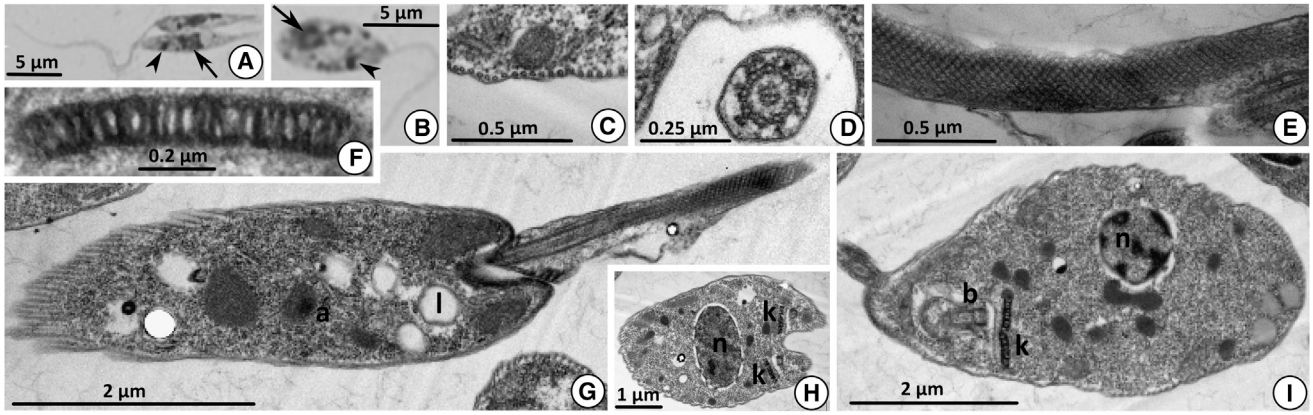

Figure 1. Morphology and Ultrastructure of *Paratrypanosoma confusum* n. sp.

(A and B) Light microscopy images. Giemsa staining of the predominant slender promastigotes (A) and infrequent oval-shaped promastigotes (B) reveals the position of the nucleus (arrow) and kinetoplast DNA (kDNA; arrowhead).

(C–I) Transmission electron microscopy images. k, kDNA; n, nucleus; b, basal body; l, lipid granule; a, acidocalcisome.

(C) The plasmalemma is supported by a corset of subpellicular microtubules.

(D) Section of the single flagellum within the flagellar pocket lacks the paraflagellar rod.

(E) An intricate meshwork of the prominent paraflagellar rod.

(F) Thin and wide kDNA disk composed of minicircles stretched taut.

(G) Longitudinally sectioned promastigote revealing full corset of subpellicular microtubules, external paraflagellar rod, and numerous lipid granules and putative acidocalcisomes.

(H) Division of the kDNA precedes nuclear division.

(I) Longitudinal section of an oval-shaped promastigote, revealing the relative position of the nucleus and the kDNA disk.

topology of the gGAPDH-based tree, which is, when compared with the SSU rRNA data set, underrepresented for the monoxenous lineages, further supports separation of *P. confusum* from the other trypanosomatid clades (bootstrap support 88%) (Figure S2B).

Phylogenomics

In order to further explore the possibility of *P. confusum* belonging to a novel trypanosomatid lineage, we used next-generation sequencing to produce a draft genome sequence. A paired-end Illumina library (insert size 330 ± 50 bp) was prepared from *P. confusum* total DNA and sequenced on a HiSeq 2000 instrument. Assembly of 40.1 million quality-filtered 101 bp reads gave scaffold N50 value of 11,534 bp. This assembly was used as a database from which to harvest protein genes for subsequent phylogenomic analyses.

Translated open reading frames (ORFs) >100 amino acids in length, from AUG to stop codons, were extracted from the *P. confusum* assembly. Simple ORF finding was used instead of more-sophisticated annotation methods because kinetoplastid genomes are essentially devoid of introns [2, 12]. Best reciprocal BLASTP hits at an E-value cutoff of 10^{-20} were found for *P. confusum* ORFs in annotated proteins or translated ORFs of 15 trypanosomatid species and the free-living bodonid *B. saltans* (Table S1, part A). Proteins from *Naegleria gruberi* (Excavata) (15,762 NCBI RefSeq entries) were used as outgroups. Proteins inferred from transcriptome data from the basal branching kinetoplastid *Perkinsela* sp. CCAP 1560/4, an endosymbiont of the amoebozoan *Neoparamoeba pemaquidensis* (G.T., P.F., N.T.O., J.L., and J.M.A., unpublished data), were included as outgroup sequences in some data sets.

Sequence clusters containing all 17 kinetoplastid species/strains and an outgroup(s) were selected for further analysis. Clusters including sequences more than two times longer or shorter than the average for a given cluster were excluded, as were clusters including sequences with BLASTP hit length 1.5 times longer or shorter than average for the cluster and/or with an average identity in the BLASTP hit region <40% between an outgroup and the other species. The following six data sets were generated: (1) 114 sequence clusters with *N. gruberi* as an outgroup, including alignments with gaps if occurring in less than half of the sequences; (2) the same alignments as in (1) without gaps; (3) 129 sequence clusters with *Perkinsela* sp. as an outgroup; (4) the same alignments as in (3) without gaps; (5) 42 sequence clusters with both *N. gruberi* and *Perkinsela* sp.; and (6) the same alignments as in (5) without gaps (Tables S2 and S3; alignments are available upon request). Phylogenetic model selection with Modelgenerator favored the LG+Γ+F model for all six concatenated alignments, and ML trees were constructed with 1,000 bootstrap replicates using this model or the GTR+Γ model (Figure 3; Table S2, part A). In all six data sets, *P. confusum* branched between *B. saltans* and the genus *Trypanosoma* with 100% bootstrap support. The branching order of the other trypanosomatid species matched the expected pattern [1], with all nodes having 100% bootstrap support. Bayesian Monte Carlo Markov (MCM) chain analysis of the concatenated data sets, conducted with the Poisson+Γ+CAT or GTR+Γ+CAT models, showed the same branching position for *P. confusum*, with posterior probabilities ranging from 0.95 (gapped data set with *N. gruberi* and *Perkinsela* sp., Poisson+Γ+CAT model) to 1 (all data sets with GTR+Γ+CAT model; both data sets with *Perkinsela* sp., Poisson+Γ+CAT model) (Figure 3; Table S2, part A). Convergence of chains was estimated by comparison of bipartition frequencies in individual chains, discarding the first 2,000 cycles and taking each tree; maximum difference in frequencies among chains ranged from 0.11 (data set with *N. gruberi*, no gaps, Poisson+Γ+CAT model) to 0 (all data sets with GTR+Γ+CAT model;

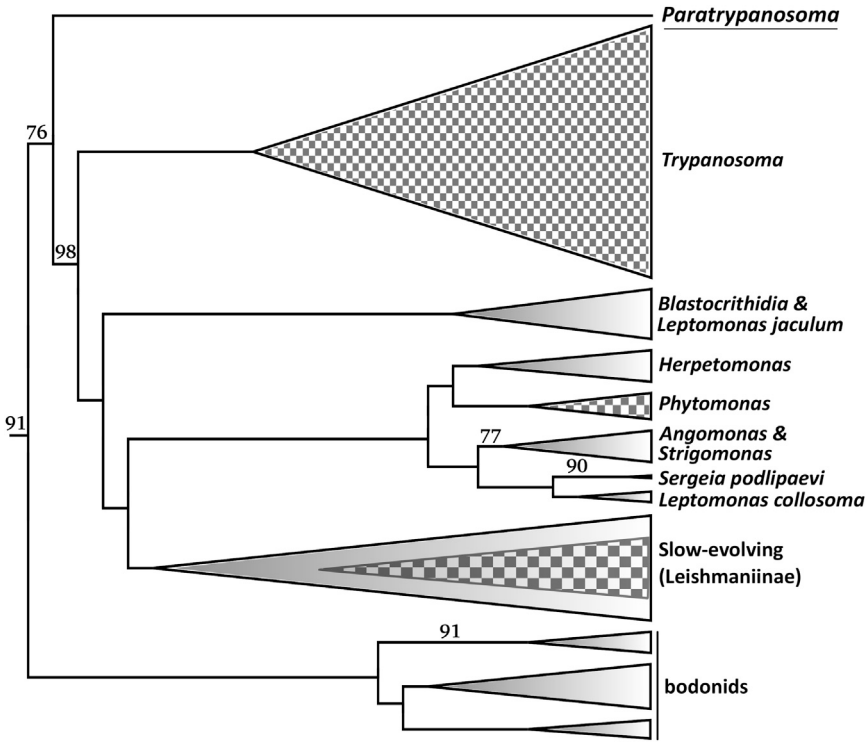

Figure 2. Maximum-Likelihood Phylogeny Based on the SSU rRNA Gene, Constructed under the GTR+Γ Model

Simplified representation; for the full tree, see Figure S2A. Clades containing dixenous species are highlighted with checkerboard shading. *P. confusum* is underlined. Bootstrap values >75% are displayed. See also Figures S1 and S2.

both data sets with *Perkinsela* sp., Poisson+Γ+CAT model). Groupings in conflict with the most probable topology corresponded to trees in which *P. confusum* branched specifically with *B. saltans* or with the outgroup (Table S2, part A). The GTR+Γ+CAT model performed better than the Poisson+Γ+CAT model according to cross-validation tests using data sets without gaps and resulted in perfect convergence in all data sets (Table S2, part A).

Removal of the BLASTP hit identity cutoff of 40% expanded the data set to 226 proteins when *N. gruberi* was used as an outgroup (Table S1, part B). Phylogenetic model selection with Modelgenerator favored the LG+Γ+F for the gapped and ungapped concatenated alignments, and ML trees for both alignments were constructed using this model (Table S2, part A). The results supported the position of *P. confusum* at the base of trypanosomatids in all bootstrap replicates. Bayesian analysis of this data set with the Poisson+Γ+CAT model was compromised by poor chain convergence, probably due to increased long-branch attraction (LBA) effects in a data set containing less conserved proteins (Table S2, part A). On the other hand, selection of more conserved proteins with a stricter BLASTP hit identity cutoff of 50% (data set with *N. gruberi*, 52 proteins; Table S1, part B) did not change the ML tree topology and support and decreased the frequency of conflicting bipartitions in the MCM chains under both the Poisson+Γ+CAT and the GTR+Γ+CAT models (Table S2, part A). The GTR+Γ+CAT had better fit than the Poisson+Γ+CAT model according to cross-validation tests with "*N. gruberi* 52" data sets (Table S2, part A).

Posterior predictive analyses of mutational saturation under the Poisson+Γ+CAT model showed that the numbers of substitutions and homoplasies were not underestimated. As expected, their numbers per site, six and three, respectively, were the lowest for the most "conserved" data set (with *N. gruberi*; 52 proteins, no gaps) (Table S2, part B). Mutational saturation in this data set is on the same level as in the metazoan alignment used for inferring the placement of nematodes and platyhelminths under the Poisson+Γ+CAT model: 7.75 and 4.3 substitutions and homoplasies per site, respectively [13]. Under these conditions, the model predicted mutational saturation correctly and was therefore deemed to be resistant to LBA, as opposed to the WAG+Γ+F model [13].

Single-protein ML trees were constructed for the 52, 114, and 129 protein data sets with either *N. gruberi* or *Perkinsela* sp. as an outgroup. The topology in which *P. confusum* branches between *B. saltans* and the other trypanosomatids was the most frequently observed, consistent with the results of concatenated analyses. However, *P. confusum* forming a monophyletic group with *B. saltans* (located at different positions on the tree) was the second most frequent topology (Table S3). Other topologies in order of decreasing frequency were (1) *P. confusum* as the sister branch of the *Trypanosoma* clade only, (2) *P. confusum* branching before *B. saltans*, and (3) *P. confusum* as the sister branch of the subfamily Leishmaniinae and *Phytomonas* clade only (Table S3). The branching of *P. confusum* with *B. saltans* or deeper in the tree than *B. saltans* is most probably the result of LBA, and indeed such topologies are more frequent in single-protein trees derived from the mutationally saturated (Table S2, part B) data set of 226 proteins (Table S3). In contrast, the branching of *P. confusum* with the genus *Trypanosoma* is less obviously an artifact. While avian trypanosomes were also isolated from *Culex* mosquitoes [4, 5], *P. confusum* clearly lacks the trypomastigote morphology synapomorphic for the genus *Trypanosoma* [1]. We used topology tests to examine the possibility of a specific relationship between *P. confusum* and trypanosomes.

In the "*N. gruberi* 114," "*N. gruberi* 52," and "*Perkinsela* sp. 129" gapped and ungapped data sets, topologies within eight important clades on the tree were fixed: the outgroup, *B. saltans*, *P. confusum*, *Trypanosoma* spp., *Phytomonas serpens*, *Crithidia fasciculata* + *Leptomonas pyrrhocoris*, and finally *Leishmania* spp. + *Endotrypanum monterogeii*. All possible 10,395 topologies of the seven clades rooted with the outgroup were constructed for each data set, and per-site log likelihoods were calculated for all topologies under the LG+Γ+F or GTR+Γ phylogenetic models. In all cases, the approximately unbiased (AU) test did not support the grouping of *P. confusum* with *Trypanosoma* or with Leishmaniinae and *Phytomonas* at a p value cutoff of 10^{-4} (Table S4, part A).

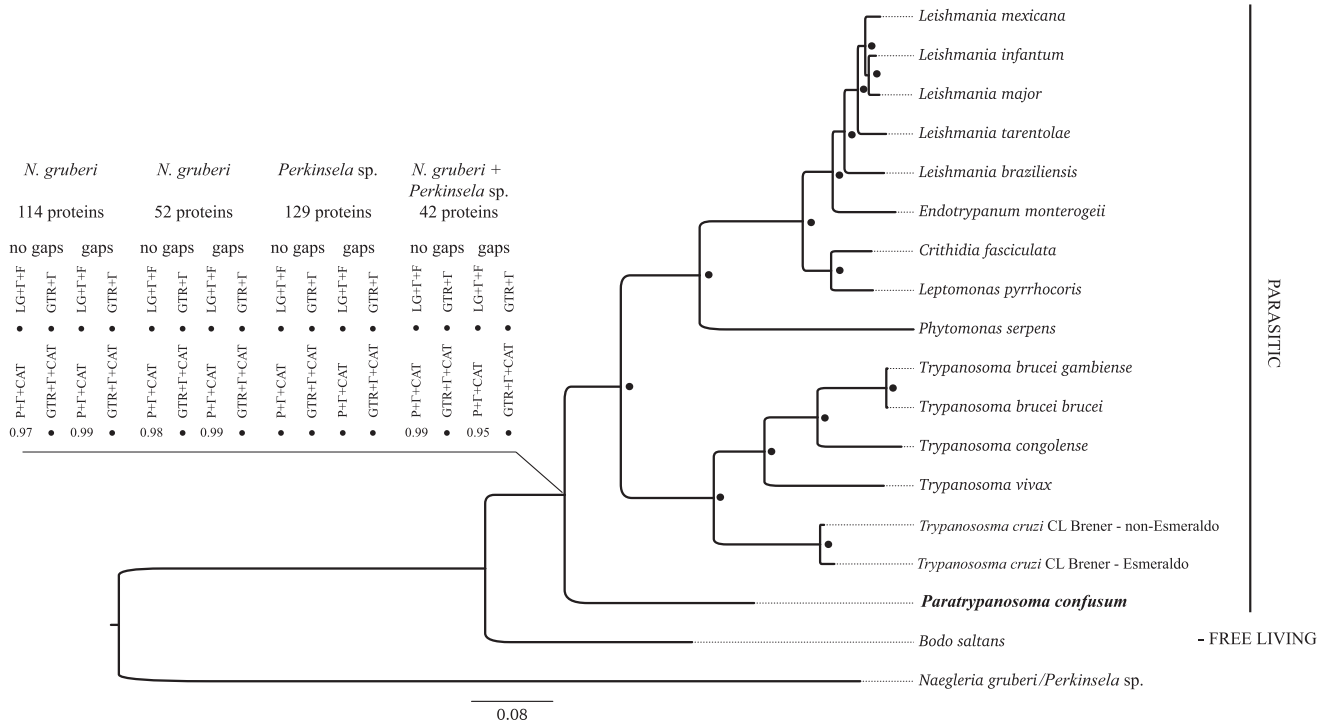

Figure 3. Maximum-Likelihood Phylogenetic Tree Based on Concatenated Protein Alignments of 18 Species

Nodes having 100% bootstrap support or posterior probabilities of 1.0 in all data sets and under all phylogenetic models tested are marked by black circles. Support for the position of *P. confusum* (in bold) is shown in a separate table. The scale bar indicates the inferred number of amino acid substitutions per site. See also Tables S1 and S2.

For most data sets, none of the 10,394 alternative topologies were supported by the AU test at a p value cutoff of 0.05. Exceptions included only topologies strongly conflicting previous data, such as the branching of *P. confusum* basal to *B. saltans* and the branching of *L. major* basal to *C. fasciculata* and *P. serpens* (Table S4, part B).

Taxonomic Summary

The taxonomic summary of *Paratrypanosoma confusum* n. gen., n. sp., is as follows:

- Class Kinetoplastea Honigberg, 1963 emend. Vickerman, 1976
- Subclass Metakinetoplastina Vickerman, 2004
- Order Trypanosomatida Kent, 1880 stat. nov. Hollande, 1952
- Family Trypanosomatidae Doflein, 1951

Paratrypanosomatinae n. subfam. Votýpka and Lukeš 2013

The newly described subfamily belongs to the obligatory parasitic uniflagellate family Trypanosomatidae, with kDNA arranged in a single compact disk at the base of the flagellum. Diagnosis is phylogenetically defined by branching at the base of all trypanosomatids, according to SSU rRNA and multiple protein-coding genes. The type genus is *Paratrypanosoma* n. gen. Votýpka and Lukeš 2013.

Paratrypanosoma confusum n. sp. Votýpka and Lukeš 2013

The dominant morphotype observed in the axenic culture is an elongated promastigote, 9.8 ± 2.1 (7.2–16.2) μm long and 2.0 ± 0.3 (1.5–2.9) μm wide, with 12.2 ± 4.8 (7.0–39.9) μm long flagellum (n = 50). The nucleus and the kDNA are situated in the anterior end of the cell. The distance between the anterior end and the kDNA and the nucleus is 2.3 ± 0.3 (1.5–2.9) μm and 4.6 ± 0.7 (3.3–6.2) μm, respectively (n = 50). The thickness of the kinetoplast is 116.4 ± 11.7 (94.4–155.5) nm (n = 50). Short oval promastigotes of varying sizes were rare.

Type host and locality: intestine of mosquito female *C. pipiens* (Diptera: Nematocera: Culicidae) captured on June 28, 2000 in the vicinity of Prague-Prosek (50°6′45.81″N, 14°29′14.53″E).
Additional hosts and localities: female mosquitoes *Culex pipiens* and *C. modestus* in Řežabinec (49°15′09″N, 14°05′32″E) and Mikulov (48°46′30″N, 16°43′30″E), Czech Republic.
Type material: the designated hapantotype is cryopreserved as an axenic culture of *P. confusum* (isolate CUL13) deposited in the slide collection at Charles University, Prague.
Etymology: the species name was given to reflect the misleading morphology.
Gene sequences: The GenBank accession numbers are KC534633–KC534828.

Discussion

Using morphology and phylogenomics, we have described a new kinetoplastid, *Paratrypanosoma confusum*, which constitutes the most basal trypanosomatid lineage branching between the free-living *B. saltans* and the parasitic *Trypanosoma* spp. and other trypanosomatids. Kinetoplastid

flagellates (Kinetoplastea, Euglenozoa) are ubiquitous single-celled eukaryotes best known as pathogens of humans and other animals, responsible for African sleeping sickness, Chagas disease, leishmaniases, and other diseases. They are traditionally split into the bodonids, which are comprised of biflagellate free-living, commensalic, or parasitic members, and the obligatory parasitic trypanosomatids, which are equipped with a single flagellum [2, 14]. Bodonids and trypanosomatids also share some unusual molecular features, such as packaging of kDNA, RNA editing, polycistronic transcription, highly modified base J, and massive *trans*-splicing [12, 14, 15]. Extensive phylogenetic analyses of about a dozen bodonid and more than a hundred trypanosomatid species have shown that the latter group is monophyletic, whereas bodonids are clearly paraphyletic [14, 15].

The origin of the extremely successful trypanosomatid life style, which combines a vertebrate (usually warm-blooded) host with an invertebrate (usually insect) vector, has been debated for more than a century [16, 17]. The insect-early scenario is now generally favored [1], since phylogenies constructed from multiple nuclear-encoded proteins suggest that the dixenous (two-host) genera *Leishmania* and *Phytomonas* are nested within clades that otherwise consist of monoxenous (single-host) insect trypanosomatids [1, 8, 11, 18]. Recent molecular surveys uncovered a major hidden diversity of insect trypanosomatids, greatly exceeding that of the dixenous genera [8–11, 18, 19]; globally, more than 10% of all dipterans, fleas, and hemipterans may be infected [1].

Hence, the most likely scenario for the evolution of dixenous parasitism postulates that an ancestor of *Leishmania* parasitizing a blood-sucking insect was injected into a vertebrate host during blood feeding and established itself in that niche. This course of events is supported by the discovery of an amber-trapped phlebotomine sand fly that was massively infected by flagellates virtually indistinguishable from the extant *Leishmania*; the insect's intestinal tract also contained nucleated red blood cells, likely originating from a "dinosaur" [20]. The protist was dated to ~220 million years ago, indicating that the establishment of the dixenous life cycle may be a fairly ancient event [15, 20]. Phylogenetic position of *Phytomonas* favors a similar scenario, in which flagellates established themselves in plants only after being transmitted to them by infected sap-sucking insects [1].

The third group known to have adopted a dixenous life style is the emblematic genus *Trypanosoma*, which thrives in a wide variety of hosts, ranging from deep-sea fish and desert reptiles to birds and mammals, including humans [2, 3, 7]. Trypanosomes have been extensively studied since Bruce's discovery of sleeping sickness [21], and their diversity is fairly well known [3, 7]. With the advent of molecular techniques, it was shown that the genus *Trypanosoma* constitutes the most basal trypanosomatid branch, the monophyly of which withstood phylogenetic scrutiny, yet sometimes its early-branching position could not be resolved with confidence [1–3, 7, 22]. However, since the time of Léger and Minchin [16, 17], the search for monoxenous ancestors of *Trypanosoma* has been ongoing, which would illuminate the evolution of the *Trypanosoma* life cycle and emergence of its extremely successful parasitic strategy.

P. confusum is the first flagellate to fit this bill. Due to its origin from female mosquitoes, its monoxenous status may be questioned, but since it was repeatedly encountered in mosquitoes in the Czech Republic (this work) and US [5], yet was never found in much-better-studied vertebrates, we consider its transmission to the latter hosts to be highly unlikely. Our extensive phylogenetic analyses argue against a specific sister relationship between *P. confusum* and *Trypanosoma* spp. and other trypanosomatids with genome sequences available. The free-living biflagellate *B. saltans*, which is radically different from parasitic species in morphology and biology, represents the closest known outgroup to *P. confusum* [12, 14, 15]. Thus, detailed genome sequence analysis of *P. confusum* should provide crucial information about the switch to a parasitic life style followed by the uniquely successful expansion of trypanosomatid flagellates.

P. confusum promastigote morphology, which occurs in various trypanosomatid clades, may be considered ancestral for the group as a whole, while the trypomastigote and epimastigote morphologies represent a synapomorphy of the genus *Trypanosoma*. Significantly, the main ultrastructural characters shared by all trypanosomatids are already present in *P. confusum*.

Since all SSU rRNA sequences obtained for *P. confusum* on two continents (this work; [5]) are virtually identical, strongly indicating their monospecific character, *Paratrypanosoma* may be an example of a cosmopolitan yet species-poor clade. Most surveys targeting non-*Leishmania* and non-*Trypanosoma* flagellates involved only two insect orders, Heteroptera and Diptera, with very few reports available from other groups, such as Hymenoptera [23] and Siphonaptera (J.V. and J.L., unpublished data; [1]). To what extent this reflects the actual distribution of parasites in insects, or simply investigator bias, remains unclear. In any case, the finding of the most deeply diverging branch in a dipteran host suggests that association of trypanosomatids with this insect order may be an ancestral state, with its spread to other insects, plants and vertebrates occurring secondarily.

Experimental Procedures

Collection of Insects, Cultivation, Microscopy, and DNA Isolation

Mosquito females were collected by miniature Centers for Disease Control (CDC) traps when attacking sparrow-hawk (*Accipiter nisus*) nestlings in Prague [4] or by dry ice- or animal-baited CDC traps in several wetland regions in southern Bohemia and Moravia. Mosquitoes were dissected under a stereomicroscope, their alimentary tracts were examined using light microscopy, and axenic culture was established as described previously [4]. The cultures were kept at 23°C, and cells were processed for light and transmission electron microscopy as described elsewhere [11].

Total DNA was isolated from axenic *P. confusum* culture or from ethanol-preserved environmental samples of mosquito females grouped in monospecific pools using the High Pure PCR Template Preparation Kit (Roche) according to the manufacturer's manual. SSU rRNA and glycosomal GAPDH genes were amplified using primers S762 (5′-GACTTTTGCTTCC TCTA[A/T]TG-3′) and S763 (5′-CATATGCTTGTTTCAAGGAC-3′) and M200 (5′-ATGGCTCC[G/A/C][G/A/C]TCAA[G/A]GT[A/T]GG[A/C]AT-3′) and M201 (5′-TA[G/T]CCCCACTCGTT[G/A]TC[G/A]TACCA-3′), respectively. Upon gel purification with the Gel Extraction Kit (Roche), the amplicons were directly sequenced.

High-Throughput DNA Sequencing and Sequence Assembly

To construct the libraries for whole-genome sequencing, DNA was processed as described in the TruSeq DNA Sample Preparation Guide (Illumina). The library with length of 330 ± 50 bp (according to analysis on Agilent 2100 Bioanalyzer) was selected for sequencing. The library was quantified using fluorimetry with Qubit (Invitrogen) and real-time PCR and diluted up to final concentration of 8 pM. The diluted library was clustered on a paired-end flowcell (TruSeq PE Cluster Kit v3) using a cBot instrument and sequenced using a HiSeq2000 sequencer with the TruSeq SBS Kit v3-HS with a read length of 101 bp from each end. The following bases/reads were removed at the filtering stage using PRINSEQ: 3′ tails with Phred

quality values (QVs) <3, homopolymeric 3′ tails >10 nt, reads with length <75 nt, reads with average base QV <30, and reads with more than one undetermined base. Genome assembly was performed using Velvet 1.2.03 with a range of assembly parameters tested: odd k-mer values from 25 to 49; k-mer coverage cutoffs 3, 6, or 9; and expected k-mer coverage from 60 to 10,000. The best scaffold N50 value of 11,534 bp was obtained with a k-mer value of 27, coverage cutoff 6, and expected coverage 10,000.

Phylogenetic Analyses

The SSU rRNA alignment was constructed using the SINA aligner website [6] and edited manually; the ML tree for SSU rRNAs was constructed with RAxML 7.2.8 using the GTR+Γ model, with 1,000 bootstrap replicates. For the gGAPDH alignment, the ML tree was constructed using the LG+Γ model and 1,000 bootstrap replicates. Protein alignments were made using MUSCLE 3.8.31 with maximum number of iterations set to 128. Site selection was performed with Gblocks 0.91b [24] using the following settings: minimum number of sequences for a conserved position, ten; minimum number of sequences for a flanking position, 12; maximum number of contiguous nonconserved positions, eight; minimum length of a block, ten; and allowed gap positions, in half of sequences or none. Selection of appropriate model parameters was done with Modelgenerator [25] using four Γ rate categories. Single-protein ML trees were constructed using RAxML 7.2.8, the LG+Γ model and 100 bootstrap replicates in all cases. For multiprotein concatenated alignments, the ML trees were constructed using different phylogenetic models and 1,000 bootstrap replicates (identical random seeds were used in all analyses), and Bayesian phylogenetic trees were constructed using PhyloBayes 3.3f, with eight chains run for 10,000 cycles. Posterior predictive tests of substitutions and homoplasies were performed using PhyloBayes 3.3f. Cross-validation of the Poisson+Γ+CAT model versus the reference GTR+Γ+CAT model was performed using the PhyloBayes package as follows: ten replicates of each data set were split into a learning set (nine-tenths of the initial data set size) and a test set (one-tenth of the initial data set size); each model was run with each replicated learning set for 10,000 cycles under the topology estimated by the model itself on the full data set; and cross-validated log-likelihood scores of each test set were computed, taking each tree after discarding first 2,000 trees, and combined. For the purpose of topology testing, lists of all possible topologies were generated using Tresolve (Barrel-o-Monkeys package, http://rogerlab.biochemistryandmolecularbiology.dal.ca/Software/Software.htm), and per-site log likelihoods for the topologies were calculated using RAxML 7.2.8. Bootstrap replicates of per-site log likelihoods were made using CONSEL 0.1i with the "-b 10" option producing ten data sets of scales 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, and 1.4, with 100,000 replicates in each. AU test p values and their SEs were calculated using CONSEL 0.1i [26].

Accession Numbers

The GenBank accession numbers KC534504–KC534632 for the transcriptomic data reported in this paper were obtained by sequencing *Perkinsela* sp. strain CCAP1560/4. The GenBank accession numbers KC534633–KC534828 for the genomic data reported in this paper were obtained by sequencing *Paratrypanosoma confusum*. The GenBank accession numbers KC543586–KC543700 for the genomic data reported in this paper were obtained by sequencing *Leptomonas pyrrhocoris* strain H10.

Supplemental Information

Supplemental Information includes two figures and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.07.045.

Acknowledgments

The help of Eva Suková with photodocumentation, Aleš Horák with phylogenetic analyses, and Milena Svobodová and Jana Rádrová with isolation and PCR detection of flagellates in mosquito hosts is appreciated. This work was supported by the Czech Grant Agency (P305/11/2179, 206/09/H026) and a Praemium Academiae award to J.L., who is also a Fellow of the Canadian Institute for Advanced Research (CIFAR). Research in the Archibald Lab was supported by an operating grant from the Canadian Institutes of Health Research (CIHR). J.M.A. holds a CIHR New Investigator Award and is a CIFAR Fellow. M.D.L., A.A.P., A.S.K. were supported by Russian Ministry of Education and Science (G34.31.0008). The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the program LM2010005, is appreciated. The access to the CERIT-SC computing and storage facilities provided under the program Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, regulation number CZ. 1.05/3.2.00/08.0144, is acknowledged.

Received: January 15, 2013
Revised: June 17, 2013
Accepted: July 12, 2013
Published: September 5, 2013

References

1. Maslov, D.A., Votýpka, J., Yurchenko, V., and Lukeš, J. (2013). Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. Trends Parasitol. *29*, 43–52.
2. Simpson, A.G., Stevens, J.R., and Lukeš, J. (2006). The evolution and diversity of kinetoplastid flagellates. Trends Parasitol. *22*, 168–174.
3. Adams, E.R., Hamilton, P.B., and Gibson, W.C. (2010). African trypanosomes: celebrating diversity. Trends Parasitol. *26*, 324–328.
4. Votýpka, J., Szabová, J., Rádrová, J., Zídková, L., and Svobodová, M. (2012). *Trypanosoma culicavium* sp. nov., an avian trypanosome transmitted by *Culex* mosquitoes. Int. J. Syst. Evol. Microbiol. *62*, 745–754.
5. Van Dyken, M., Bolling, B.G., Moore, C.G., Blair, C.D., Beaty, B.J., Black, W.C., 4th, and Foy, B.D. (2006). Molecular evidence for trypanosomatids in *Culex* mosquitoes collected during a West Nile virus survey. Int. J. Parasitol. *36*, 1015–1023.
6. Pruesse, E., Peplies, J., and Glöckner, F.O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics *28*, 1823–1829.
7. Leonard, G., Soanes, D.M., and Stevens, J.R. (2011). Resolving the question of trypanosome monophyly: a comparative genomics approach using whole genome data sets with low taxon sampling. Infect. Genet. Evol. *11*, 955–959.
8. Votýpka, J., Klepetková, H., Jirků, M., Kment, P., and Lukeš, J. (2012). Phylogenetic relationships of trypanosomatids parasitising true bugs (Insecta: Heteroptera) in sub-Saharan Africa. Int. J. Parasitol. *42*, 489–500.
9. Borghesan, T.C., Ferreira, R.C., Takata, C.S., Campaner, M., Borda, C.C., Paiva, F., Milder, R.V., Teixeira, M.M., and Camargo, E.P. (2013). Molecular phylogenetic redefinition of *Herpetomonas* (Kinetoplastea, Trypanosomatidae), a genus of insect parasites associated with flies. Protist *164*, 129–152.
10. Teixeira, M.M., Borghesan, T.C., Ferreira, R.C., Santos, M.A., Takata, C.S., Campaner, M., Nunes, V.L., Milder, R.V., de Souza, W., and Camargo, E.P. (2011). Phylogenetic validation of the genera *Angomonas* and *Strigomonas* of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts. Protist *162*, 503–524.
11. Jirků, M., Yurchenko, V.Y., Lukeš, J., and Maslov, D.A. (2012). New species of insect trypanosomatids from Costa Rica and the proposal for a new subfamily within the Trypanosomatidae. J. Eukaryot. Microbiol. *59*, 537–547.
12. Jackson, A.P., Quail, M.A., and Berriman, M. (2008). Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). BMC Genomics *9*, 594.
13. Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. *7*(*Suppl 1*), S4.
14. Moreira, D., López-García, P., and Vickerman, K. (2004). An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. Int. J. Syst. Evol. Microbiol. *54*, 1861–1875.
15. Deschamps, P., Lara, E., Marande, W., López-García, P., Ekelund, F., and Moreira, D. (2011). Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. Mol. Biol. Evol. *28*, 53–58.
16. Léger, L. (1904). Sur les affinités de l'Herpetomonas subulata et la phylogénie des trypanosomes. Comptes Rendus des Séances de la Société de Biologie *57*, 615–617.
17. Minchin, E.A. (1908). Investigations on the development of trypanosomes in the tsetse flies and other Diptera. Q. J. Microsc. Sci. *52*, 159–260.

18. Votýpka, J., Maslov, D.A., Yurchenko, V., Jirků, M., Kment, P., Lun, Z.R., and Lukeš, J. (2010). Probing into the diversity of trypanosomatid flagellates parasitizing insect hosts in South-West China reveals both endemism and global dispersal. Mol. Phylogenet. Evol. *54*, 243–253.
19. Týč, J., Votýpka, J., Klepetková, H., Šuláková, H., Jirků, M., and Lukeš, J. (2013). Growing diversity of trypanosomatid parasites of flies (Diptera: Brachycera): frequent cosmopolitism and moderate host specificity. Mol. Phylogenet. Evol. *69*, 255–264.
20. Poinar, G., Jr., and Poinar, R. (2004). *Paleoleishmania proterus* n. gen., n. sp., (Trypanosomatidae: Kinetoplastida) from Cretaceous Burmese amber. Protist *155*, 305–310.
21. Bruce, D. (1895). Preliminary Report on the Tsetse Fly Disease or Nagana, in Zululand (London: Durban, Bennett & Davis).
22. Hughes, A.L., and 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.
23. Runckel, C., Flenniken, M.L., Engel, J.C., Ruby, J.G., Ganem, D., Andino, R., and DeRisi, J.L. (2011). Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, *Nosema*, and *Crithidia*. PLoS ONE *6*, e20656.
24. Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.
25. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., and Mclnerney, J.O. (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.
26. Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. *51*, 492–508.