Protistology 18 (1): 72–81 (2024) | doi:10.21685/1680-0826-2024-18-1-7

ORIGINAL ARTICLE

The genome of the endosymbiont-harboring trypanosomatid *Kentomonas sorsogonicus*

Ana Luisa Elias Zavataro[1]*, Karolína Skýpalová[2]*, Percy Omar Tullume Vergara[1], Flavia Maia Silva[1], Anzhelika Butenko[2,3,4], Vyacheslav Yurchenko[2], Alexei Yu. Kostygov[2,5], João Marcelo Pereira Alves[1]**

[1] *Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil*
[2] *Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czechia*
[3] *Institute of Parasitology, Czech Academy of Sciences, České Budějovice, Czechia*
[4] *University of South Bohemia, Faculty of Science, České Budějovice, Czechia*
[5] *Zoological Institute of the Russian Academy of Sciences, St. Petersburg, Russia*
* These authors contributed equally.

| Submitted March 15, 2024 | Accepted March 27, 2024 |

Summary

The trypanosomatid subfamily Strigomonadinae, composed of the genera *Angomonas, Strigomonas*, and *Kentomonas*, is distinguished by the obligatory presence of endosymbiotic betaproteobacteria *Candidatus* Kinetoplastibacterium spp. This ancient and well-established symbiotic relationship features an intensive metabolic exchange and coordination of the cell division of the participants. In contrast to the extensively studied genera *Angomonas* and *Strigomonas*, little is known about *Kentomonas*, which has been described in 2014. Only one genome sequence of the bacterial endosymbiont (*Ca.* Ki. sorsogonicusi) is available. In this work, we report the high-quality genome sequence for the trypanosomatid *Kentomonas sorsogonicus*, obtained by a hybrid assembly of short Illumina and long PacBio read data. The assembly has a total length of ~34.8 Mb, with many scaffolds being as long as complete chromosomes in other trypanosomatid species. Our preliminary analysis demonstrates that the genome of this trypanosomatid is quite divergent, which significantly hampers functional annotation of its genes.

Key words: Trypanosomatidae, genomics, endosymbiosis, genome assembly

Introduction

The family Trypanosomatidae (Kinetoplastea, Euglenozoa) is composed of obligate parasites infecting a wide variety of organisms from ciliates to animals and plants (Kostygov et al., 2021). Their life cycles involve either one host (monoxenous species) or two hosts (dixenous species); in the latter case, the parasite alternates between vertebrate or plant and invertebrate (usually insect) hosts (Maslov et al.,

https://doi.org/10.21685/1680-0826-2024-18-1-7

**Corresponding author: João Marcelo Pereira Alves, Department of Parasitology, Institute for Biomedical Sciences, University of Sao Paulo, Av. Prof. Lineu Prestes, 1374, Sao Paulo 05508-000, SP, Brazil; jotajj@usp.br

2019). Although the monoxenous trypanosomatids have been investigated for well over a century, most attention has been given to their dixenous relatives, primarily those members of the genera *Trypanosoma* and *Leishmania* that have medical and veterinary impact. This imbalance was naturally reflected in significantly less attention paid to the genomics of monoxenous trypanosomatids. Recently, with the development of cheaper and faster sequencing technologies, the genomic diversity of the family Trypanosomatidae has been more thoroughly explored, although much remains to be done both in sequencing as well as in phenotypical and biological studies of these organisms (Yurchenko et al., 2021).

Amongst trypanosomatids, the subfamily Strigomonadinae is of special interest, as all its described representatives harbor betaproteobacterial endosymbionts. Each trypanosomatid cell has a single bacterium. This relationship appears to be long-standing and has been used as a model for studying the evolution of symbiosis and organelles for decades (de Souza and Motta, 1999). It is characterized by an intense exchange of metabolites, usually synthesized by the bacterium and provided to the trypanosomatid cell (Maslov et al., 2019). Their partnership is so close that it is not possible for the bacterium to survive outside of the host. This is corroborated by an observation that no endosymbiont-free Strigomonadinae has been found in nature thus far. The genetic basis of this long-studied collaboration has been recently elucidated for the members of the genera *Angomonas* and *Strigomonas* (and their corresponding endosymbionts of the genus *Candidatus* Kinetoplastibacterium) by genomic sequencing and analysis (Alves et al., 2011, 2013a, 2013b; Klein et al., 2013; Alves, 2017).

A third genus of Strigomonadinae, *Kentomonas*, has been described more recently (Votýpka et al., 2014) and is still understudied. The genome of its endosymbiont, *Ca.* Ki. sorsogonicusi, was sequenced (Silva et al., 2018), showing a significant reduction in size compared to that of bacteria from *Angomonas* and *Strigomonas* spp. (Alves et al., 2013b). Most strikingly, it was shown that the endosymbiont of *Kentomonas sorsogonicus* had lost the metabolic pathway for heme biosynthesis, which was considered a hallmark of the subfamily (Silva et al., 2018). This finding undermined the usage of the so-called hemin test (assessing auxotrophy for heme and related porphyrins) as a quick way of detecting endosymbiont presence. However, other aspects of the symbiotic relationship between *Kentomonas sorsogonicus* and *Ca.* Ki. sorsogonicusi have not been explored to a large extent due to the lack of the host genome.

Herein, we report the high-quality genome sequence for *K. sorsogonicus* MF-08, obtained by combining data from second- and third- generation sequencing technologies (by Illumina and PacBio, respectively). Our preliminary analysis demonstrates that this endosymbiont-bearing trypanosomatid has a very divergent genome, which significantly hampers functional annotation of its genes.

Methods

Genome sequencing and assembly

Organism cultivation, DNA extraction, as well as Illumina sequence data acquisition (100 bp paired-end reads) and processing were performed as described previously (Silva et al., 2018). Long-read sequencing was done commercially at DNALink (Seoul, Korea) using 5μg of genomic DNA and SMRTbell™ Template Prep Kit 1.0 (Pacific Biosciences, Manlo Park, USA) for library preparation. Sequencing was performed with the MagBead OneCellPerWell v1 protocol (insert size of ~20 kbp and movie time of 240 min).

Read quality was evaluated using FastQC v. 0.11.9 (Andrews, 2010). Jellyfish v. 2.3.0 was used to estimate the genome size based on k-mer distribution (Marçais and Kingsford, 2011). The hybrid assembly was performed by first assembling the PacBio data with Flye v. 2.9 b-1774 (Kolmogorov et al., 2019). The resulting fragments were then polished using the Illumina data with Pilon v. 1.24 (Walker et al., 2014) in three iterations and using parameters "--changes" and "--fix all". Scaffolds of the bacterial endosymbiont were identified by comparison with its complete genome (Silva et al., 2018) and removed before further analysis. Completeness of the genome assembly was evaluated using BUSCO v. 5.6.1 (Manni et al., 2021) with Euglenozoa_odb10 as reference database (130 genes). Only contigs with minimal length of 500 bp were used, and the results were visualized with ggplot2 v. 3.5.0 R package (Wickham, 2009). The genome assembly sequence produced in this work is available from GenBank under the accession number GCA_030347455.1.

Table 1. Genome assembly statistics for subfamily Strigomonadinae and reference trypanosomatids.

Species	Assembly size, Mb	N$_{50}$, bp	L$_{50}$	Longest contig, bp	GC, %	tRNA genes	ORFs	Average ORF length, bp	Part of genome in ORFs, %
Kentomonas sorsogonicus	34,81	430,516	26	1,139,634	55.9	103	12,238	1,419	49.89
Angomonas ambiguus	23,42	136,903	49	811,451	45	97	8,268	1,640	57.9
Angomonas deanei	20,98	774,942	10	1,502,655	49.9	59	10,365	1,239	61.23
Angomonas desouzai	24,25	5,727	1,063	64,124	49	67	11,037	1,253	57.03
Strigomonas culicis	23,59	2,337	3,086	158,194	54.3	45	12,083	1,131	50.16
Strigomonas galati	27,24	6,697	1,079	55,781	50	77	10,289	1,523	57.56
Strigomonas oncopelti	24,96	4,543	1,539	37,394	55	119	10,187	1,505	61.44
Wallacemonas collosoma	25,69	167,338	46	600,384	57	74	8,903	1,830	63.44
Sergeia podlipaevi	26,88	45,748	173	241,952	48.5	95	8,722	1,79	58.08
Leishmania major	32,86	1,091,540	11	2,682,151	59.5	84	8,424	1,921	49.26
Trypanosoma brucei	22,15	2,224,448	4	2,825,021	47.5	81	9,788	1,542	68.16
Paratrypanosoma confusum	27,55	220,321	34	1,420,330	61.5	60	8,659	1,842	57.91

Functional annotation

Genes were predicted with AUGUSTUS v. 3.3.3 (Stanke et al., 2006), and the respective proteins were functionally annotated using HMMER v. 3.3.2 with Pfam v. 36.0 as a database and e-value threshold of 1e-5 (Potter et al., 2018; Mistry et al., 2021). The best hits found in database for each protein were selected for protein annotation using sequence and profile searches against public databases of sequences, protein domains, and orthologous groups. In addition, transfer RNA genes were predicted with tRNAscan v.2.0.9 (Chan and Lowe, 2019) in default settings in all genomes, for which publicly data of tRNA were unavailable (Table 1).

The predicted proteins of *K. sorsogonicus* MF-08 were included in a dataset also comprising: i) all other described Strigomonadinae (*Angomonas* and *Strigomonas* spp.), ii) members of the two genera most closely related to this subfamily (*Sergeia podlipaevi* and *Wallacemonas collosoma*); and reference species (*Leishmania major, Trypanosoma brucei* and *Paratrypanosoma confusum*) (Fig. 1, Suppl. Table S1). The sequences of all these species were submitted to eggNOG-mapper v. 2.1.12 (Cantalapiedra et al., 2021) with eggNOG 5 as a reference database (Huerta-Cepas et al., 2019). The target taxon was set to Eukaryota and the transfer of both experimental and electronic annotations were allowed. Assigned categories of clusters of orthologous groups (COGs) were visualized using ggplot2. Proteins with no orthologs in the eggNOG database were added to the category “no COG assigned”.

Analysis of orthologs

Inference of groups of orthologous proteins (OGs) was performed on the same dataset as above using OrthoFinder v. 2.5.5 with BLAST as a sequence search program and other parameters at their default values (Emms and Kelly, 2019). Shared and species-specific OGs were visualized with the R package UpSetR v. 1.4.0 (Lex et al., 2014).

To assess protein sequence divergence for the species in the dataset, we analyzed all 2,528 OGs containing a single protein per species. The sequence pairwise identities within each orthogroup were computed using a custom Python script employing “align.globalxx” function from the BioPython package (Cock et al., 2009). The distributions of obtained values for each species were visualized as boxplots in Microsoft Excel.

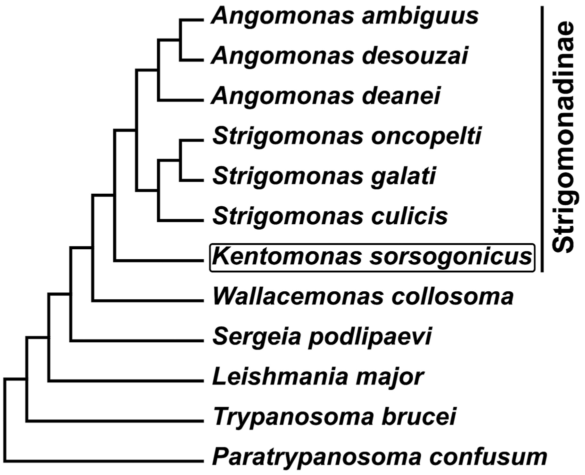

Fig. 1. Schematic phylogenetic tree showing relationships between the species included in the dataset used in this work. The tree is based on the previously published phylogenomic reconstructions (Silva et al., 2018; Kostygov et al., 2024).

Results and discussion

The assembly of ~146 thousand PacBio reads (~1.2 Gb) after polishing by 45.7 million Illumina reads (~4.6 Gb) resulted in 308 scaffolds, of which three belonged to the endosymbiont and were removed from the dataset. The final *K. sorsogonicus* genome assembly has a length of ~34.8 Mb, with the largest scaffold being ~1.1 Mb long. This is close to an independent estimate of genome size using k-mer distribution analysis, which yielded 38.41 Mb using all Illumina reads (without removal of those for the endosymbiont, whose genome size is 0.74 Mb). The resulting assembly is not only highly contiguous (as judged by the N50 value of ~431 Kb), but also nearly complete as judged by the BUSCO scores, with only four missing and five fragmented single copy orthologs out of 130 in the BUSCO dataset (Fig. 2). Quality-wise, it is similar to the only chromosome-level genome assembly of Strigomonadinae – that of *Angomonas deanei* ATCC PRA-265 – with two missing and five fragmented genes. Of note, the reference trypanosomatid genome for *Leishmania major* has zero missing or fragmented genes (Ivens et al., 2005). Meanwhile, in *K. sorsogonicus* one of the BUSCO orthologs is duplicated, which was not detected in *A. deanei* or *L. major* (Fig. 2).

Based on available assemblies, we compared genome characteristics of *K. sorsogonicus*, other members of Strigomonadinae, the representatives of two closest outgroup genera (*Wallacemonas* and *Sergeia*), as well as three reference species: *L. major, Trypanosoma brucei*, and *Paratrypanosoma confusum* (Table 1). The assembled genome of *K. sorsogonicus* is the largest among all Strigomonadinae studied so far and in our dataset is similar in size only to that of the quite distant relative – *L. major* (Table 1). Although the genome assembly of *S. culicis* is ~7.5 Mb smaller, it contains about the same number of open reading frames (ORFs) as *Kentomonas* (both species demonstrate the highest values among all compared species). However, this large number of genes in the case of *S. culicis* can be artifactual, since this is the most fragmented assembly in the dataset as judged by N50 and L50 values (Table 1) and the largest count of fragmented BUSCOs (Fig. 2). The average ORF length in K. sorsogonicus (1,419 bp) falls within the range of the values for other Strigomonadinae (1,131–1,640 bp), but is shorter than in the considered outgroups (1,505–1,921, Table 1). The percentage of the *Kentomonas* genome representing protein-coding genes (49.89%) is rather small, and only *S. culicis* and *L. major* have comparable values (50.16 and 49.26%, respectively). Our preliminary analysis suggests that this low proportion in *K. sorsogonicus* can be due to the expansion of repetitive elements, which occupy 21% of the genome as compared to less than 15% in *A. deanei* (data not shown). With respect to the number of tRNA genes, Strigomonadinae show the widest range among all considered trypanosomatids, with the minimal and maximal values observed in the members of the genus *Strigomonas* – *S. culicis* and *S. oncopelti* (45 and 119, respectively). Of note, the second highest number belongs to *K. sorsogonicus* (Table 1). Apparently, all the listed differences contribute to the observed genome size variation in Strigomonadinae.

We investigated the functional diversity of proteins of *K. sorsogonicus* and compared it to that of other Strigomonadinae and trypanosomatids outside of this subfamily by annotating them with eggNOG mapper and clustering according to the nomenclature used by this program (Fig. 3). A profile of functional categories revealed to be quite similar for the majority of the considered species, with similar proportions of the functional categories and the largest of them being "posttranslational modification, protein turnover, chaperones", "translation, ribosomal structure and biogenesis", and "signal transduction mechanisms" (Fig. 3). The two exceptions were *T. brucei*, with a considerably

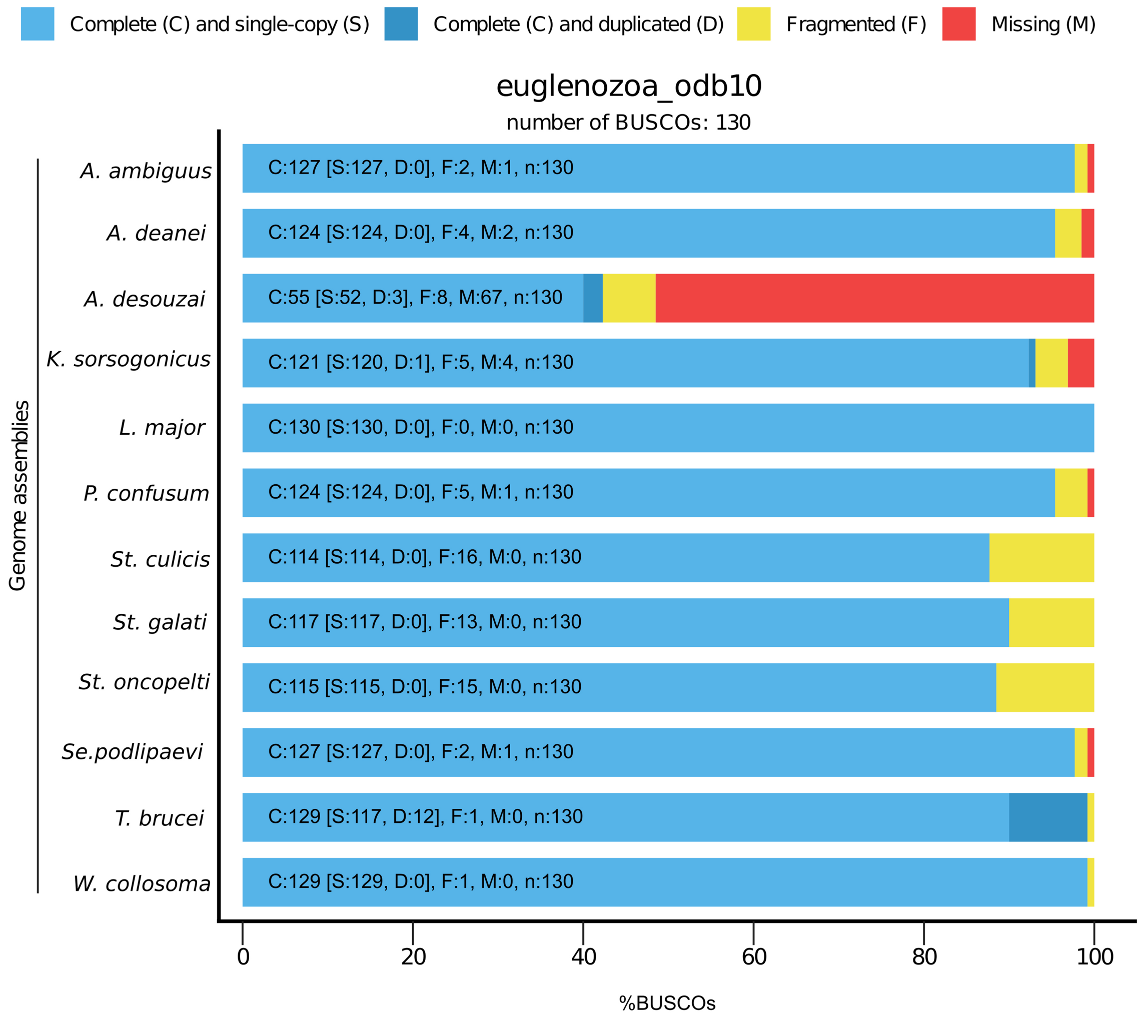

Fig. 2. Genome assembly completeness assessed by the presence of Benchmarking Universal Single-Copy Orthologs (BUSCOs) from the Euglenozoa_odb10 database.

increased category "inorganic ion transport and metabolism" and *K. sorsogonicus*, possessing as much as 59.5% of proteins not assigned to any category by the eggNOG mapper, whereas other species had only 34.2 – 45.8% of such proteins (Fig. 3). Such a high proportion of the unidentified proteins in *K. sorsogonicus* led to the shrinkage of most other categories making profile comparison between this and other species unreliable.

The relatively low number of clustered proteins in *K. sorsogonicus* may be a consequence of significant sequence divergence within this species, affecting the functional annotation procedure. This assumption is substantiated by a long branch leading to *K. sorsogonicus* on a phylogenomic tree (Kostygov et al., 2024), as well as our estimates of pairwise sequence identities between proteins encoded by single-copy orthologous genes of the trypanosomatids in our dataset (Fig. 4). The median value of these pairwise identities among *K. sorsogonicus* proteins and those of other species is 35.6%. This is, on average, by 6% lower than for other Strigomonadinae and is comparable only to the values for the evolutionary most divergent species in the dataset – *T. brucei* and *P. confusum* (35.6 and 34.5%, respectively; Fig. 4).

Out of approximately 117,000 proteins used for OG interference, around 108,900 (93%) were assigned to 10,204 groups containing at least two proteins (Suppl. Table S2). Of these, 4,119 OGs (likely corresponding to housekeeping genes that had been inherited from the last common ancestor of the family) were shared by all analyzed species (Fig. 5). The 12,238 proteins of *K. sorsogonicus*

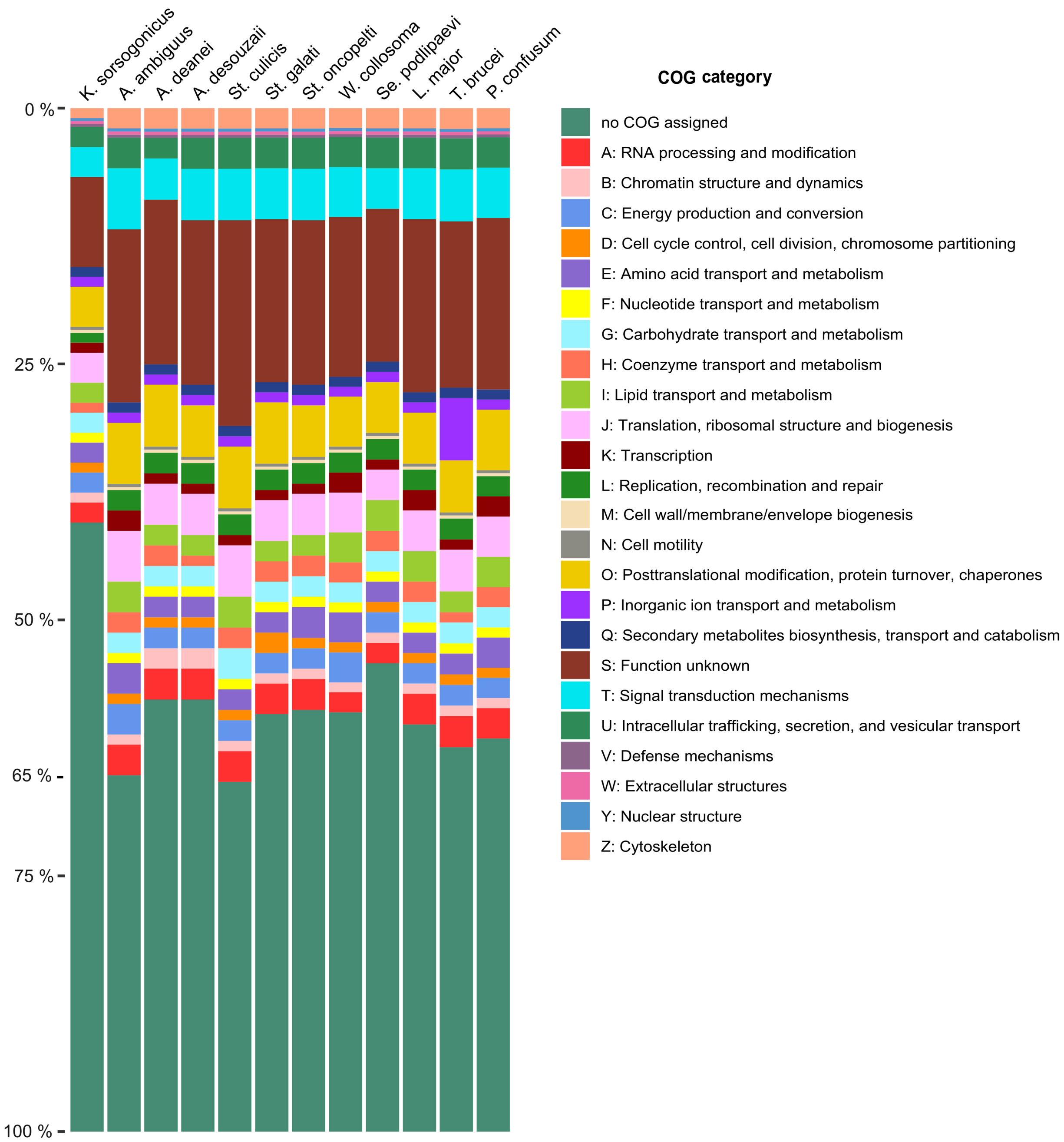

Fig. 3. Clusters of orthologous groups and functional classification of trypanosomatid proteins according to eggNOG nomenclature. Plot showing the percentage proportion (X-axis) of proteins falling into each category for each species (Y-axis).

were clustered into 8,766 OGs, of which 2,112 (24%) were not present in other species (Suppl. Table S2, Fig. 5). These are the largest number and the largest proportion of inferred species-specific OGs among all species in the dataset, substantially bigger even compared to the earliest-branching trypanosomatid – *P. confusum* (1,258 out of 7,834 OGs, 16%). A functional annotation of these uniquely present orthogroups even with a fairly low e-value threshold of 1e-5 was successful only for 274 of them (13%, Suppl. Table S3), further corroborating our hypothesis of extreme sequence divergence. Such results additionally point to the exceptional divergence of gene/protein sequences in *K. sorsogonicus* and do not allow relying on the results of the inference of uniquely absent or present

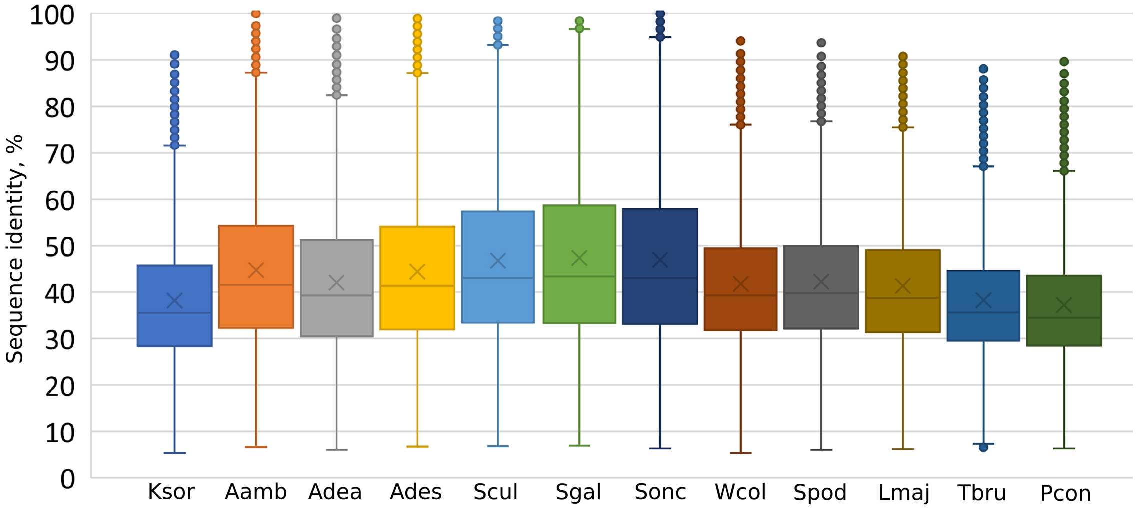

Fig. 4. Distributions of pairwise sequence identities for proteins from single-copy orthologous groups. The horizontal line and cross within a box represent the median and mean values, respectively. Circles correspond to outliers (values outside the interquartile range (IQR) ± 1.5 IQR). Species *abbreviations*: Aamb – *Angomonas ambiguus*, Adea – *A. deanei*, Ades – *A. desouzai*, Ksor – *Kentomonas sorsogonicus*, Lmaj – *Leishmania major*, Pcon – *Paratrypanosoma confusum*, Scul – *Strigomonas culicis*; Sgal – *St. galati*, Sonc – *St. oncopelti*; Spod – *Sergeia podlipaevi*; Tbru – *Trypanosoma brucei*, Wcol – *Wallacemonas collosoma*.

proteins in this species (Suppl. Table S3). However, this does not preclude the analysis (although it can be non-exhaustive) of the whole subfamily Strigomonadinae, or its *Angomonas* + *Strigomonas* clades, specifically.

The whole subfamily shared unique 14 OGs (Suppl. Fig. S1). Besides hypothetical proteins, those were predominantly enzymes involved in amino acid metabolism and horizontally transferred from bacteria (Alves et al., 2013a). In addition, there was one subfamily-specific amastin and three enzymes with broadly specified functions: a (di)oxygenase, a hydrolase and a adenylate/guanylate cyclase (Suppl. Table S4). Among the uniquely absent (i.e. lost) 47 OGs, there were paraflagellar rod components, chaperones, and single proteins with various functions (transcription factor, a receptor for a host hormone, apoptosis regulator, etc.) (Suppl. Table S4).

The set of 48 OGs that we inferred as uniquely present in *Angomonas* and *Strigomonas* (responsible for cellular metabolism, transport, signaling and membrane dynamics; Fig. 5, Suppl. Table S5) is problematic for the reasons that have been mentioned above (i.e. they are probably artifactually absent in *Kentomonas*). More reliable were the results concerning the proteins, which are absent (i.e. lost) from *Angomonas* and *Strigomonas*. There were 18 such OGs involved in amino acid transport, protein folding, nucleocytoplasmic transport, signal transduction, or nicotinamide metabolism (Suppl. Table S5). Interestingly, there was also one more paraflagellar component, evidencing that this structure in strigomonadines was reduced gradually, with a more advanced state in the two crown genera (Suppl. Table S5).

Thus, our data demonstrate high overall divergence of *Kentomonas* genome, which represents an essential challenge for its analysis. However, our preliminary analysis sheds some light on the evolution of Strigomonadinae by revealing stepwise changes in the gene repertoire.

Acknowledgments

This work was supported by the grant #2013/14622-3 from São Paulo Research Foundation (FAPESP) and Czech Grant Agency grant 23-07695S. The infrastructure used in this project was partially funded by the EU's Operational Program "Just Transition" LERCO CZ.10.03.01/00/22_003/0000003.

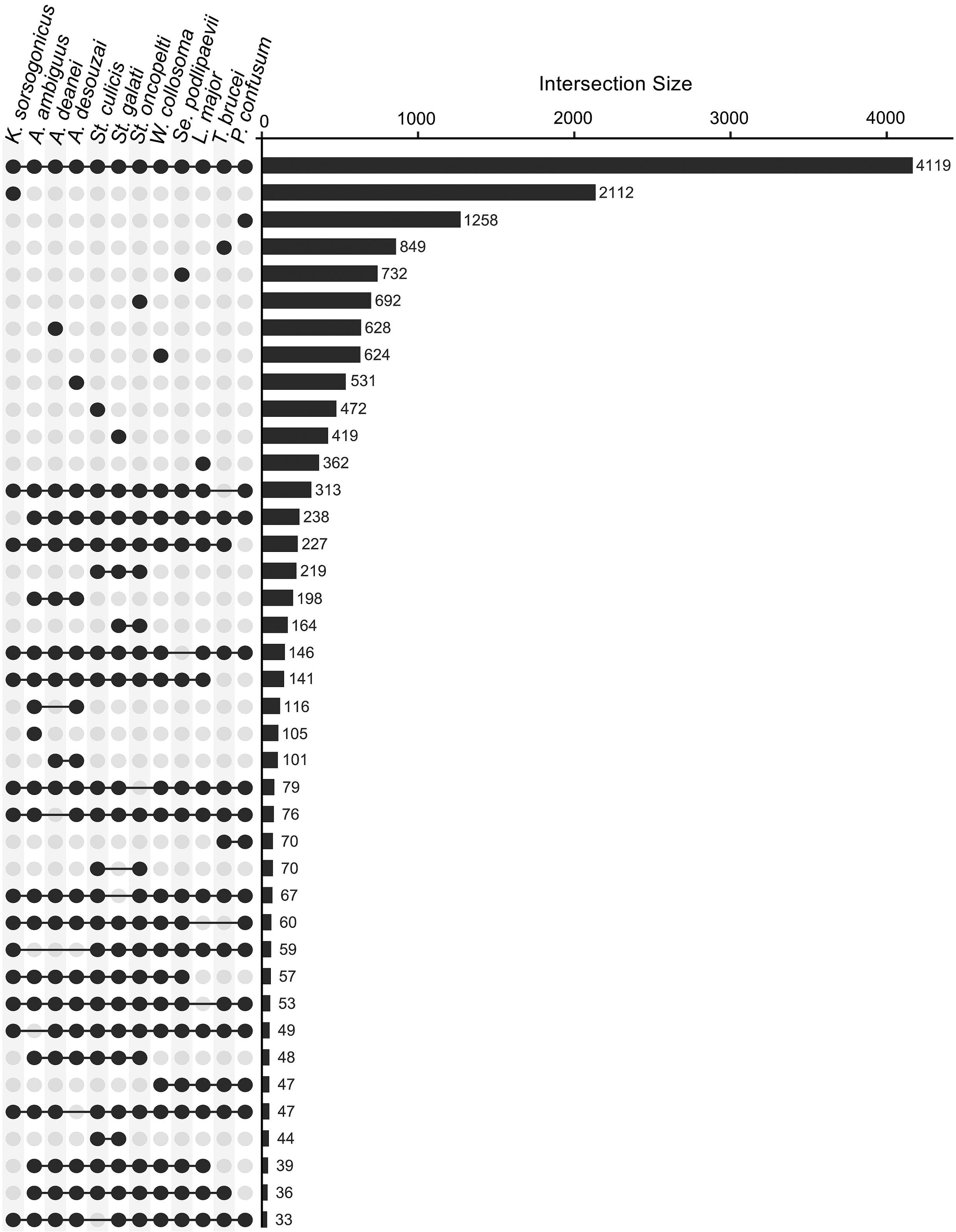

Fig. 5. Orthologous groups shared among 12 trypanosomatid species visualized using UpSetR software. Plot shows the number of orthologous groups (Y-axis) and species composition for each intersection (X-axis). Only 40 largest intersections are shown.

References

Alves J.M., Klein C.C., da Silva F.M., Costa-Martins A.G. et al. 2013a. Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. BMC Evol. Biol. 13: 190. https://doi.org/10.1186/1471-2148-13-190

Alves J.M., Serrano M.G., Maia da Silva F., Voegtly L.J. et al. 2013b. Genome evolution and phylogenomic analysis of *Candidatus Kinetoplastibacterium*, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. Genome Biol Evol. 5 (2): 338–350. https://doi.org/10.1093/gbe/evt012

Alves J.M., Voegtly L., Matveyev A.V., Lara A.M. et al. 2011. Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. PLOS ONE. 6 (8): e23518. https://doi.org/10.1371/journal.pone.0023518

Alves J.M.P. 2017. Amino acid biosynthesis in endosymbiont-harbouring Trypanosomatidae. In: The handbook of microbial metabolism of amino acids. (Ed. D'Mello J.P.F.). CAB International, Oxfordshire, UK, pp. 371–383.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Cantalapiedra C.P., Hernández-Plaza A., Letunic I., Bork P. and Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. 38 (12): 5825–5829. https://doi.org/10.1093/molbev/msab293

Chan P.P. and Lowe T.M. 2019. tRNAscan-SE: Searching for tRNA genes in genomic sequences. Methods Mol. Biol. 1962: 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Cock P.J., Antao T., Chang J.T., Chapman B.A. et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25 (11): 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

de Souza W. and Motta M.C. 1999. Endosymbiosis in protozoa of the Trypanosomatidae family. FEMS Microbiol Lett. 173 (1): 1–8. https://doi.org/10.1111/j.1574-6968.1999.tb13477.x

Emms D.M. and Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome. Biol. 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y

Huerta-Cepas J., Szklarczyk D., Heller D., Hernandez-Plaza A. et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47 (D1): D309–D314. https://doi.org/10.1093/nar/gky1085

Ivens A.C., Peacock C.S., Worthey E.A., Murphy L. et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science. 309 (5733): 436–442. https://doi.org/10.1126/science.1112680

Klein C.C., Alves J.M., Serrano M.G., Buck G.A. et al. 2013. Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. PLOS ONE. 8 (11): e79786. https://doi.org/10.1371/journal.pone.0079786

Kolmogorov M., Yuan J., Lin Y. and Pevzner P.A. 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37 (5): 540–546. https://doi.org/10.1038/s41587-019-0072-8

Kostygov A.Y., Albanaz A.T.S., Butenko A., Gerasimov E.S. et al. 2024. Phylogenetic framework to explore trait evolution in Trypanosomatidae. Trends Parasitol. 40 (2): 96–99. https://doi.org/10.1016/j.pt.2023.11.009

Kostygov A.Y., Karnkowska A., Votýpka J., Tashyreva D. et al. 2021. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. Open Biol. 11 (3): 200407. https://doi.org/10.1098/rsob.200407

Lex A., Gehlenborg N., Strobelt H., Vuillemot R. and Pfister H. 2014. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph. 20 (12): 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Manni M., Berkeley M.R., Seppey M., Simro F.A. and Zdobnov E.M. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38 (10): 4647–4654. https://doi.org/10.1093/molbev/msab199

Marçais G. and Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27 (6): 764–770. https://doi.org/10.1093/bioinformatics/btr011

Maslov D.A., Opperdoes F.R., Kostygov A.Y., Hashimi H. et al. 2019. Recent advances in trypanosomatid research: genome organization, expression, metabolism, taxonomy and evolution. Parasitology. 146 (1): 1–27. https://doi.org/10.1017/S0031182018000951

Mistry J., Chuguransky S., Williams L., Qureshi M. et al. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49 (D1): D412–D419. https://doi.org/10.1093/nar/gkaa913

Potter S.C., Luciani A., Eddy S.R., Park Y. et al. 2018. HMMER web server: 2018 update. Nucleic Acids Res. 46 (W1): W200–W204. https://doi.org/10.1093/nar/gky448

Silva F.M., Kostygov A.Y., Spodareva V.V., Butenko A. et al. 2018. The reduced genome of *Candidatus* Kinetoplastibacterium sorsogonicusi, the endosymbiont of *Kentomonas sorsogonicus* (Trypanosomatidae): loss of the haem-synthesis pathway. Parasitology. 145 (10): 1287–1293. https://doi.org/10.1017/S003118201800046X

Stanke M., Keller O., Gunduz I., Hayes A. et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34 (Web Server issue): W435–439. https://doi.org/10.1093/nar/gkl200

Votýpka J., Kostygov A.Y., Kraeva N., Grybchuk-Ieremenko A. et al. 2014. *Kentomonas* gen. n., a new genus of endosymbiont-containing trypanosomatids of Strigomonadinae subfam. n. Protist. 165 (6): 825–838. https://doi.org/10.1016/j.protis.2014.09.002

Walker B.J., Abeel T., Shea T., Priest M. et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE. 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963

Wickham H. 2009. ggplot2: Elegant graphics for data analysis. Springer, New York.

Yurchenko V., Butenko A. and Kostygov A.Y. 2021. Genomics of Trypanosomatidae: where we stand and what needs to be done? Pathogens. 10 (9): 1124. https://doi.org/10.3390/pathogens10091124

Supplementary materials

Fig. S1. Orthologous groups shared among 12 trypanosomatid species visualized using UpSetR software. Plot shows the species composition for each intersection and the corresponding number of orthologous groups (X and Y axes, respectively). The 70 largest intersections are shown.

Table S1. Trypanosomatid species included into the dataset used in this work for comparative analyses.

Table S2. Protein families inferred by the analysis of orthologous groups.

Table S3. Functional annotation of OGs uniquely present and absent in *K. sorsogonicus*.

Table S4. Functional annotation of OGs uniquely present or absent in Strigomonadinae.

Table S5. Functional annotation of OGs uniquely present or absent in *Angomonas* and *Strigomonas* spp.