Genome Evolution and Phylogenomic Analysis of *Candidatus* Kinetoplastibacterium, the Betaproteobacterial Endosymbionts of *Strigomonas* and *Angomonas*

João M.P. Alves[1,3,*], Myrna G. Serrano[1], Flávia Maia da Silva[2], Logan J. Voegtly[1], Andrey V. Matveyev[1], Marta M.G. Teixeira[2], Erney P. Camargo[2], and Gregory A. Buck[1]

[1]Department of Microbiology and Immunology and the Center for the Study of Biological Complexity, Virginia Commonwealth University

[2]Department of Parasitology, ICB, University of São Paulo, São Paulo, Brazil

[3]Present address: São Paulo, Brazil

*Corresponding author: E-mail: alvesjmp@gmail.com.

Accepted: January 18, 2013

Data deposition: Finished genome sequences and annotations are available from GenBank under accession numbers CP003803, CP003804, CP003805, CP003806, and CP003807. Supplementary table S1, Supplementary Material online, also contains accession numbers to sequences used in the phylogenetic analyses.

Abstract

It has been long known that insect-infecting trypanosomatid flagellates from the genera *Angomonas* and *Strigomonas* harbor bacterial endosymbionts (*Candidatus* Kinetoplastibacterium or TPE [trypanosomatid proteobacterial endosymbiont]) that supplement the host metabolism. Based on previous analyses of other bacterial endosymbiont genomes from other lineages, a stereotypical path of genome evolution in such bacteria over the duration of their association with the eukaryotic host has been characterized. In this work, we sequence and analyze the genomes of five TPEs, perform their metabolic reconstruction, do an extensive phylogenomic analyses with all available Betaproteobacteria, and compare the TPEs with their nearest betaproteobacterial relatives. We also identify a number of housekeeping and central metabolism genes that seem to have undergone positive selection. Our genome structure analyses show total synteny among the five TPEs despite millions of years of divergence, and that this lineage follows the common path of genome evolution observed in other endosymbionts of diverse ancestries. As previously suggested by cell biology and biochemistry experiments, *Ca.* Kinetoplastibacterium spp. preferentially maintain those genes necessary for the biosynthesis of compounds needed by their hosts. We have also shown that metabolic and informational genes related to the cooperation with the host are overrepresented amongst genes shown to be under positive selection. Finally, our phylogenomic analysis shows that, while being in the Alcaligenaceae family of Betaproteobacteria, the closest relatives of these endosymbionts are not in the genus *Bordetella* as previously reported, but more likely in the *Taylorella* genus.

Key words: endosymbiont biology, phylogenomics, comparative genomics, Trypanosomatidae, selective pressure.

Introduction

The phenomenon of bacterial endosymbiosis has been increasingly shown to occur in a widespread range of eukaryotes ranging from protozoa to Metazoa. In Metazoa, insects (Kikuchi 2009) are most commonly implicated, but plants (Markmann and Parniske 2009), nematodes, Amoebozoa (Schmitz-Esser et al. 2010), and leeches (Perkins et al. 2005) have also been shown to harbor prokaryotic symbionts, which perform a variety of important roles for their hosts. Endosymbiotic associations characterized to date have exhibited a common pattern of genome evolution, the extent of which is directly proportional to the age of the association (Moya et al. 2008). Such common alterations in the endosymbiont genomes include marked genome size reduction, lowering of GC content, loss of transcriptional regulation genes (possibly due to the stability of the intracellular milieu), loss of DNA repair genes (possibly due to bottleneck effects given the small number of bacteria, frequently just one, in each host cell), extensive loss of biosynthetic metabolic capabilities (except those required by the host), reduction or loss of codon-usage bias, loss of mobile elements, reduction of the length of intergenic regions (frequently leading to a significant number of overlapping genes), and endosymbiotic gene transfer to the host nucleus (Moya et al. 2008; Nowack and Melkonian 2010).

The Kinetoplastida, one of the three classes of the phylum Euglenozoa, is comprised of a number of protozoan taxa of great ecological, medical, veterinary, and economical importance. The class is divided into the Trypanosomatidae, comprised only of highly successful parasitic species, and several bodonid clades, containing parasitic, commensal, or free-living species (Vickerman 1976; Campbell 1992; Cavalier-Smith 1993; Podlipaev 2001). Trypanosomatids are a highly successful taxon of parasitic organisms that are known to infect all classes of vertebrates, several classes of invertebrates, and many plants. Monoxenic parasites including *Angomonas*, *Crithidia*, *Blastocrithidia*, *Strigomonas*, *Herpetomonas,* and *Leptomonas* infect a wide range of insects (Vickerman 1976; Campbell 1992; Cavalier-Smith 1993; Podlipaev 2001). Newton and Horne (1957) first reported the presence of self-reproducing structures in the cytoplasm of *Strigomonas oncopelti*. These so-called bipolar bodies were the first endosymbionts described in trypanosomatids and were later shown to be of betaproteobacterial nature (Du et al. 1994; Umaki et al. 2009; Teixeira et al. 2011). Typically, there is only a single bacterium per trypanosomatid cell, and the host and parasite display a sophisticated interrelationship. Thus, the endosymbiont affects the morphology and ultrastructure of the host cells (Motta et al. 1997; de Souza and Motta 1999) and supplies the host with essential compounds including heme, nucleotides, and essential amino acids (de Souza and Motta 1999; Alves et al. 2011). Conversely, the endosymbionts are supplied with a stable environment and nutrients. Finally, the host nucleus and the bacterium have developed a form of communication that controls their cell cycles such that they replicate in synchrony (Motta et al. 2010).

Chang and coworkers (Chang and Trager 1974; Chang 1975a, 1975b; Chang et al. 1975) characterized the presence of endosymbionts in *S. culicis* and demonstrated the ability of this flagellate to synthesize heme. The latter ability was absent in flagellates artificially cured of their symbionts by chloramphenicol treatment, showing that the bacterium was responsible for this biosynthetic capability. The details of this biochemical collaboration have been recently published, showing that heme biosynthesis is performed partly by the endosymbiont and partly by host enzymes (which were laterally transferred from a Gammaproteobacteria), with neither organism capable of completing synthesis on its own (Alves et al. 2011). It is interesting to note that, while the flagellate can live in culture without the endosymbiont if the medium is supplemented with compounds the bacterium supplies, the opposite has not been possible, that is, the bacterial endosymbiont has not been successfully cultured outside its host to date (de Souza and Motta 1999).

Presently, six trypanosomatid species are known to harbor symbionts (Teixeira et al. 2011). They are referred to as SHTs (symbiont-harboring trypanosomatids) in contrast to regular trypanosomatids lacking symbionts. In this work, we sequenced and analyzed the entire genomes of the betaproteobacterial endosymbionts *Candidatus* Kinetoplastibacterium oncopeltii, *Ca.* K. blastocrithidii, *Ca.* K. galatii, *Ca.* K. desouzaii, and *Ca.* K. crithidii, which are referred to as TPEs (trypanosomatid proteobacterial endosymbionts) (Teixeira et al. 2011), and found in association with five SHTs: *S. oncopelti*, *S. culicis*, *S. galati*, *Angomonas desouzai*, and *A. deanei*, respectively. We have extensively annotated and analyzed the completed genome sequences of the five TPEs above, performing phylogenomic analyses with all sequenced betaproteobacterial genomes available to date (including two highly reduced and fast evolving genomes of endosymbionts from insects (McCutcheon and Moran 2010; Lopez-Madrigal et al. 2011), and three organisms each from the Alpha- and Gammaproteobacteria as outgroups. We also report the comparison of the TPE genomes with those of two of their relatives from the Alcaligenaceae family, that is, *Taylorella equigenitalis*, which is a pathogen of the equine genital tract (Hebert et al. 2011), and *Achromobacter xylosoxidans*, a free-living bacterium that is commonly found on human skin and gut, but can also be an opportunistic pathogen (Holmes et al. 1977; Strnad et al. 2011). These latter bacteria were chosen for being, respectively, the closest known relative of the TPEs and a free-living representative in the family.

Materials and Methods

Organisms and Growth Conditions

As the endosymbionts cannot be grown outside of their trypanosomatid host, we proceeded to shotgun sequence the mixture of host and bacterial DNA directly from the cultures, without any enrichment or separation of cells or DNA. The genomes of the following organisms and their endosymbionts were sequenced (numbers are culture identifiers from the Trypanosomatid Culture Collection of the University of São Paulo): *A. deanei* (TCC036E), *A. desouzai* (TCC079E), *S. culicis* (TCC012E), *S. galati* (TCC219), *S. oncopelti* (TCC290E). Organisms were grown in LIT media (Camargo 1964) supplemented with 2% fetal bovine serum.

DNA Extraction and kDNA Depletion

Genomic DNA was extracted with the phenol/chloroform method (Ozaki and Czeko 1984). Steps were taken to minimize the amount of kinetoplast DNA present in the sequence, as previously described (Alves et al. 2011). Our results indicate that this enrichment protocol reduces the proportion of kDNA in the sample to less than approximately 5%, yielding good coverage of nuclear DNA, kDNA, and endosymbiont DNA (data not shown).

Genome Sequencing and Analysis, and Gene Calling and Annotation

Five micrograms of total kDNA-depleted genomic DNA was sequenced using standard pyrosequencing shotgun methodology according to Roche 454 protocols. Reads were assembled using Newbler software (Roche, version 2.3). Resulting contigs were initially segregated into endosymbiont- or host-derived sets by similarity of their sequences to currently available betaproteobacterial genomes. Bacterial chromosome sequence closing was performed by multiplex long-range polymerase chain reaction using primers designed based on contig ends, followed by capillary sequencing of the resulting products and manual gap closing. Genome atlases were drawn using DNAplotter (Carver et al. 2009). Dot plot comparisons were performed using the mummerplot program from MUMmer3 (Kurtz et al. 2004). Genes from TPEs were called using Glimmer (Delcher et al. 1999), and manually checked for completeness and adequate translation start and stop sites in comparison with previously characterized bacterial genes. Genes missed by Glimmer but identified by BLAST similarity searches were also added when shown to be complete. Ribosomal clusters were identified by similarity compared with their homologs from other Betaproteobacteria. Transfer RNAs were predicted using tRNAscan (Lowe and Eddy 1997) and predictions were refined using TFAM (Ardell and Andersson 2006) and its bacterial tRNA models and two of three previously described methodologies (Silva et al. 2006), namely the phenogram analysis and the comparison with the three precompiled fMet (X), Met (M), and Ile (I) tRNA profiles. For the phenogram method, the PHYLIP (Felsenstein 2005) package was used to calculate Jukes–Cantor distances and build the neighbor-joining tree. Prediction of cellular localization for proteins was done by SecretomeP (Bendtsen et al. 2005). Metabolic reconstruction and functional classification were performed by running ASGARD (Alves and Buck 2007), using as general databases both KEGG (Ogata et al. 1999) and UniRef100 (Suzek et al. 2007). Further Clusters of Orthologous Genes (COG) comparisons with other Alcaligenaceae bacteria were done using RPSBLAST (Altschul et al. 1997) against the COG database (Tatusov et al. 1997), and significant *P* values were calculated using *Ca.* K. blastocrithidii as the representative TPE. All searches were performed using an E value cutoff of 1E−6.

Orthologous Group Classification

Orthologous groups of Betaproteobacteria were determined using OrthoMCL (Li et al. 2003) version 2, using as input the predicted proteins from 119 organisms: the 5 TPEs, 108 other Betaproteobacteria, three Alphaproteobacteria, and three Gammaproteobacteria (supplementary table S1, Supplementary Material online). After the OrthoMCL run, we filtered the resulting groups to keep only those presenting at least 95% of the total number of organisms. For comparison of the five TPE genomes with each other, the same analysis was performed using only the five relevant protein sets.

Analysis of Natural Selection Pressure in *Ca.* Kinetoplastibacterium

Investigation of natural selection pressure was performed on all *Ca.* Kinetoplastibacterium orthologous groups where all five TPEs where represented, except EF-Tu. Amount and direction of selective pressure were calculated using the codeml program from PAML (Yang 2007), testing different models of codon evolution. We have investigated models allowing different d*N*/d*S* ratios (ω, nonsynonymous to synonymous substitution) for different codon positions along the gene. Site models used were M0, M1a, M2a, M7, and M8 (Yang et al. 2000; Wong et al. 2004), using the Bayes Empirical Bayes test with a probability of greater than 0.95 to identify positively selected codons (Yang et al. 2005).

Phylogenomic Inference

Phylogenetic analyses were performed in a two-step manner: First, a fast maximum likelihood phylogenetic inference of each orthologous group was performed to identify and remove inparalogs (sequences that appeared duplicated only inside a taxon) and misclassified orthologs—genes that were grouped in an orthologous group erroneously, due to partial similarity of sequence. Such sequences usually presented very long branch lengths or divided the phylogenetic tree in two parts that partially mirrored each other, and in such cases the smaller of the two subtrees was removed. Sequence alignments were redone after any removal of taxa. After filtering and cleanup, only groups containing at least 95% of the original number of taxa (i.e., 113) and no more than one gene per taxon were used in subsequent analyses. The second and final step of phylogenetic analysis was a maximum likelihood run using the supermatrix generated by concatenating the 233 protein alignments from the filtered orthologous groups described earler. Protein sequences were aligned by ClustalW2 (Larkin et al. 2007) using default parameters. Alignments were concatenated using FASconCAT v.1.0 (Kück and Meusemann 2010) after removal of ambiguously aligned columns by Gblocks (Castresana 2000) with default parameters except for the minimum length of a block (5 instead of 10 amino acids) and allowing gaps if present in less than half of the sequences in each column. Maximum likelihood phylogenetic inferences were performed by RAxML v.7.2.8 (Stamatakis 2006) on a Linux computer using 40 processor cores, under the WAG substitution model (Whelan and Goldman 2001), with four gamma-distributed heterogeneity rate categories and estimated empirical residue frequencies (model PROTGAMMAWAGF). One hundred different best tree searches were performed, and the tree with best likelihood found was kept. RAxML's rapid bootstrap was performed with 100 pseudoreplicates and support is only shown for branches with support of at least 50. The final tree was drawn and basically formatted by TreeGraph2 (Stöver and Müller 2010), with further cosmetic adjustments done using the Inkscape vector image editor (http://inkscape.org/, last accessed January 2013).

Results and Discussion

Overall Genomic Content

In this work, we sequence and characterize the genomes of five endosymbionts from trypanosomatid hosts. As previously suggested, the symbiotic association between ancestral trypanosomatids and Betaproteobacteria was probably the result of a single event that occurred in the late Cretaceous Period (Du et al. 1994; Teixeira et al. 2011), leading to tens of millions of years of symbiont and host co-evolution. Figure 1 presents the genome maps of *Ca.* K. crithidii and *Ca.* K. blastocrithidii, showing the typical bacterial features of GC skew and differential distribution of the protein-coding genes in the two chromosomal strands. As determined by comparing the 652 genes used for natural selection pressure analysis (discussed later), the genomes of the TPEs from *Strigomonas* hosts are quite similar and exhibit overall approximately 83% base sequence identity (with the highest similarity being ~85%, between *Ca.* K. oncopeltii and *Ca.* K. galatii). The genomes of TPEs from *Angomonas* species exhibit lower, approximately 73% base sequence identity. However, the base sequences of the *Strigomonas* TPEs exhibit only approximately 65% identity to those of the *Angomonas* TPEs. All these findings are concordant with these organisms' positioning and branch lengths in the phylogeny (fig. 2).

The genomes of the TPEs are markedly reduced in size compared with their closest nonendosymbiotic relatives, being between 810 and 833 kbp in length and presenting 693–742 protein-coding genes, 43 (or 44, in *Ca.* K. crithidii) tRNA genes, 9 ribosomal RNA genes and 4, 7, and 20 pseudogenes in *Ca.* K. galatii, *Ca.* K. blastocrithidii, and *Ca.* K. oncopeltii, respectively; only one pseudogene was identified in each of the endosymbionts from *Angomonas* (table 1). In comparison, the genome of *T. equigenitalis*, the closest known relative of *Ca.* Kinetoplastibacterium (fig. 2), is over twice the size and bears over twice as many protein-coding genes as the TPE genomes. However, the number of structural RNA genes in the endosymbionts and *T. equigenitalis* is similar (~40 tRNA genes and 3 rRNA clusters). The three CAT anti-codon tRNAs identified by tRNAscan could be confidently annotated as the initiator (fMet or X) tRNA, the elongator (Met or M), and a modified tRNA that is charged with isoleucine (Ile) instead of methionine, by using previously described methods (Silva et al. 2006). The presence of this modified tRNA is consistent with the presence in the genome of TilS (tRNA(Ile)-lysidine synthase), the enzyme responsible for the modification of a cytidine to lysidine. This modification is important because it allows the decoding of the AUA codon, for which no other tRNA was found in our predictions. These results were achieved unambiguously by the phenogram method, where the three different CAT tRNAs from the

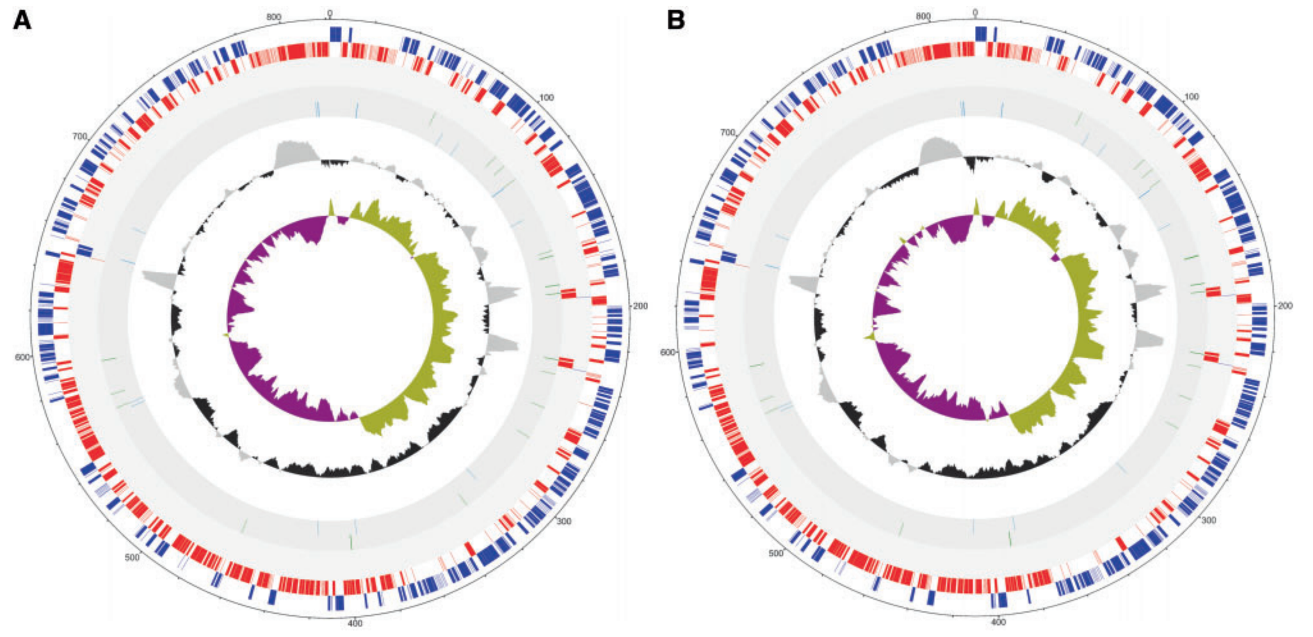

Fig. 1.—Genomic atlas of *Candidatus* Kinetoplastibacterium crithidii (*A*) and *Ca.* K. blastocrithidii (*B*). Tracks represent, from the outside to inside: genomic coordinates, in kbp; genes in the plus strand (defined as the strand where *dnaA* is located, blue); genes in the minus strand (red); rRNA genes in the plus strand; rRNA genes in the minus strand; tRNA genes in the plus strand; tRNA genes in the minus strand; GC content above or below average; GC skew.

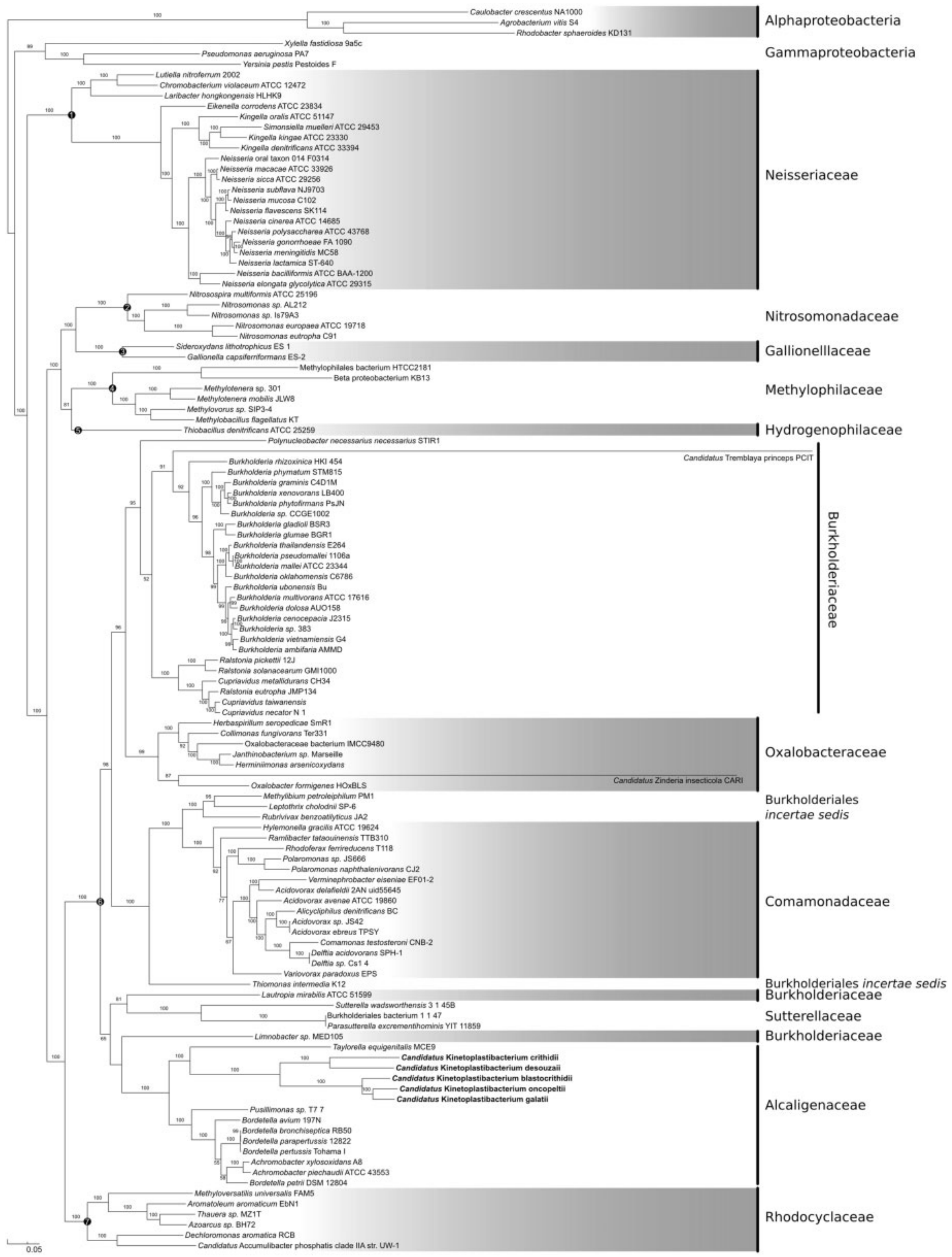
Caulobacter crescentus NA1000
Agrobacterium vitis S4
Rhodobacter sphaeroides KD131
Alphaproteobacteria
Xylella fastidiosa 9a5c
Pseudomonas aeruginosa PA7
Yersinia pestis Pestoides F
Gammaproteobacteria
Lutiella nitroferrum 2002
Chromobacterium violaceum ATCC 12472
Laribacter hongkongensis HLHK9
Eikenella corrodens ATCC 23834
Kingella oralis ATCC 51147
Simonsiella muelleri ATCC 29453
Kingella kingae ATCC 23330
Kingella denitrificans ATCC 33394
Neisseria oral taxon 014 F0314
Neisseria macacae ATCC 33926
Neisseria sicca ATCC 29256
Neisseria subflava NJ9703
Neisseria mucosa C102
Neisseria flavescens SK114
Neisseria cinerea ATCC 14685
Neisseria polysaccharea ATCC 43768
Neisseria gonorrhoeae FA 1090
Neisseria meningitidis MC58
Neisseria lactamica ST-640
Neisseria bacilliformis ATCC BAA-1200
Neisseria elongata glycolytica ATCC 29315
Neisseriaceae
Nitrosospira multiformis ATCC 25196
Nitrosomonas sp. AL212
Nitrosomonas sp. Is79A3
Nitrosomonas europaea ATCC 19718
Nitrosomonas eutropha C91
Nitrosomonadaceae
Sideroxydans lithotrophicus ES 1
Gallionella capsiferriformans ES-2
Gallionelllaceae
Methylophilales bacterium HTCC2181
Beta proteobacterium KB13
Methylotenera sp. 301
Methylotenera mobilis JLW8
Methylovorus sp. SIP3-4
Methylobacillus flagellatus KT
Methylophilaceae
Thiobacillus denitrificans ATCC 25259
Hydrogenophilaceae
Polynucleobacter necessarius necessarius STIR1
Candidatus Tremblaya princeps PCIT
Burkholderia rhizoxinica HKI 454
Burkholderia phymatum STM815
Burkholderia graminis C4D1M
Burkholderia xenovorans LB400
Burkholderia phytofirmans PsJN
Burkholderia sp. CCGE1002
Burkholderia gladioli BSR3
Burkholderia glumae BGR1
Burkholderia thailandensis E264
Burkholderia pseudomallei 1106a
Burkholderia mallei ATCC 23344
Burkholderia oklahomensis C6786
Burkholderia ubonensis Bu
Burkholderia multivorans ATCC 17616
Burkholderia dolosa AUO158
Burkholderia cenocepacia J2315
Burkholderia sp. 383
Burkholderia vietnamiensis G4
Burkholderia ambifaria AMMD
Ralstonia pickettii 12J
Ralstonia solanacearum GMI1000
Cupriavidus metallidurans CH34
Ralstonia eutropha JMP134
Cupriavidus taiwanensis
Cupriavidus necator N 1
Burkholderiaceae
Herbaspirillum seropedicae SmR1
Collimonas fungivorans Ter331
Oxalobacteraceae bacterium IMCC9480
Janthinobacterium sp. Marseille
Herminiimonas arsenicoxydans
Candidatus Zinderia insecticola CARI
Oxalobacter formigenes HOxBLS
Oxalobacteraceae
Methylibium petroleiphilum PM1
Leptothrix cholodnii SP-6
Rubrivivax benzoatilyticus JA2
Burkholderiales incertae sedis
Hylemonella gracilis ATCC 19624
Ramlibacter tataouinensis TTB310
Rhodoferax ferrireducens T118
Polaromonas sp. JS666
Polaromonas naphthalenivorans CJ2
Verminephrobacter eiseniae EF01-2
Acidovorax delafieldii 2AN uid55645
Acidovorax avenae ATCC 19860
Alicycliphilus denitrificans BC
Acidovorax sp. JS42
Acidovorax ebreus TPSY
Comamonas testosteroni CNB-2
Delftia acidovorans SPH-1
Delftia sp. Cs1 4
Variovorax paradoxus EPS
Comamonadaceae
Thiomonas intermedia K12
Burkholderiales incertae sedis
Lautropia mirabilis ATCC 51599
Burkholderiaceae
Sutterella wadsworthensis 3 1 45B
Burkholderiales bacterium 1 1 47
Parasutterella excrementihominis YIT 11859
Sutterellaceae
Limnobacter sp. MED105
Burkholderiaceae
Taylorella equigenitalis MCE9
Candidatus Kinetoplastibacterium crithidii
Candidatus Kinetoplastibacterium desouzaii
Candidatus Kinetoplastibacterium blastocrithidii
Candidatus Kinetoplastibacterium oncopeltii
Candidatus Kinetoplastibacterium galatii
Pusillimonas sp. T7 7
Bordetella avium 197N
Bordetella bronchiseptica RB50
Bordetella parapertussis 12822
Bordetella pertussis Tohama I
Achromobacter xylosoxidans A8
Achromobacter piechaudii ATCC 43553
Bordetella petrii DSM 12804
Alcaligenaceae
Methyloversatilis universalis FAM5
Aromatoleum aromaticum EbN1
Thauera sp. MZ1T
Azoarcus sp. BH72
Dechloromonas aromatica RCB
Candidatus Accumulibacter phosphatis clade IIA str. UW-1
Rhodocyclaceae
0.05

Fig. 2.—Maximum likelihood supermatrix phylogeny (233 concatenated orthologs) of the Betaproteobacteria class, with Alpha- and Gammaproteobacteria as outgroups. For the Betaproteobacteria, families are indicated on the right of each clade. Numbers on branches represent bootstrap support (only support of 50 or greater shown). Black circles with white numbers mark the different betaproteobacterial orders: 1, Neisseriales; 2, Nitrosomonadales; 3, Gallionellales; 4, Methylophilales; 5, Hydrogenophilales; 6, Burkholderiales; 7, Rhodocyclales. The five endosymbionts sequenced in this work are represented in bold.

Table 1
Comparative Genome Statistics

	CKbla	CKcri	CKdes	CKgal	CKonc	Tequi	Axylo[a]
Genome length	820,029	821,932	833,125	822,140	810,172	1,695,860	7,359,146
Overall GC %	32.55	30.96	30.17	32.36	31.23	37.42	65.78
Protein-coding genes	723	730	742	726	693	1,556	6,815
% of genome in genes[b]	92.03	91.90	91.68	91.85	90.92	93.45	91.57
rRNA genes[c]	9	9	9	9	9	9	10
tRNA genes	43	44	43	43	43	38	60
Other noncoding RNAs	7	7	6	7	7	ND	ND
Pseudogenes	7	1	1	4	20	0	0
Average gene length[d]	1,004	1,010	1,007	1,012	1,017	1,008	986
Average intergenic length[b]	84	85	87	86	96	82	94

NOTE.—CKbla, *Candidatus* Kinetoplastibacterium blastocrithidii; CKcri, *Ca.* K. crithidii; CKdes, *Ca.* K. desouzaii; CKonc, *Ca.* K. oncopeltii; Tequi, *Taylorella equigenitalis*; Axylo, *Achromobacter xylosoxidans*; ND, not determined.

[a] *Achromobacter xylosoxidans* numbers include plasmid-located sequences.

[b] Includes protein-coding and RNA genes, and pseudogenes.

[c] Arranged in three identical clusters of the 16S, 23S, and 5S rRNA genes (except for *A. xylosoxidans*, where one of the clusters has a duplicated 5S gene).

[d] Only protein-coding genes.

TPEs cluster deep within each of the three sequence types (data not shown); comparison of the TPE sequences with the three profiles available from Silva et al. were not as definitive when it came to the distinction of the Ile and Met CAT tRNAs, although the distinction of fMet was very clear, as already observed by those authors for other organisms. Besides these three cases, all other tRNAscan predictions have been confirmed by matches to the corresponding bacterial TFAM profiles, which could also more easily distinguish the Ile and Met CAT tRNAs. Other noncoding RNAs were also found, that is, transfer-messenger, RNAse P class A, CspA thermoregulator, small signal recognition particle, SucA motif, and alpha operon ribosome-binding site RNA. The CspA thermoregulator RNA was found in two copies, associated with the corresponding CspA family beta-ribbon cold shock proteins (two different orthologs), except in *Ca.* K. desouzaii where only the second CspA thermoregulator was predicted. Given that this species also presents both CspA protein orthologs, it is likely that the prediction failed—some of the other CspA-associated RNAs had scores very close to the minimum cutoff for prediction.

The difference in genome size is much more marked in comparison with the free-living *A. xylosoxidans*, which presents a 7.36 Mbp genome with 6,815 protein-coding genes (both ~9 times larger than the respective endosymbiont numbers), 10 rRNA (distributed in three clusters), and 60 tRNA genes. In contrast to many other bacteria, for example, Xu et al. (2007), the tRNA genes of the TPEs are not concentrated in large clusters. The distribution of the tRNA genes in the TPE genomes is nearly identical, with only one additional tRNA for proline present in *Ca.* K. crithidii, located isolated and starting at position 328,161 (fig. 1). Average gene and intergenic region lengths (~1,000 and 82–96 bases, respectively) and amount of gene overlap are similar in all organisms analyzed, regardless of life-style. Almost all TPE protein-coding genes are single-copy. The exception—the *EF-Tu* gene—is duplicated in the TPE genomes. In *Ca.* K. galatii one of the copies (ST1E_0192) is present as a pseudogene, whereas both copies seem functional in the remaining four TPE genomes. It is not clear why the endosymbionts, in all other genomic aspects so reduced, would maintain two identical copies of this gene—although it would be tempting to speculate that these bacteria, which have also maintained the three ribosomal cluster copies customary in Betaproteobacteria and a high proportion of translation genes, still need to maintain high protein synthesis efficiency during at least part of their life cycle. The fact that one of them, the pseudogene in *Ca.* K. galatii, seems to be in the process of being lost suggests that this redundancy might not be essential.

Genome Organization and Synteny

Our genome structure analyses show that the genomes of these five *Ca.* Kinetoplastibacterium bacteria follow the path of genome evolution previously observed in other endosymbionts of diverse ancestries, and also exhibit total synteny. Synteny with *T. equigenitalis* is also relatively highly maintained (supplementary fig. S1, Supplementary Material online), with most blocks of matching sequence along the main diagonal lines. Interestingly, the dot-plot presents an overall X-shape, suggesting that bacteria in the Alcaligenaceae family follow the same pattern of inversions centered around the axis of replication—that is, the line connecting the origin and termination of replication in the genome map—as other bacteria previously studied (Tillier and Collins 2000; Silva et al. 2001). Comparison of *Ca.* Kinetoplastibacterium with *Pusillimonas* sp., a more distantly related Alcaligenaceae bacterium, by the same method yielded the same X-shaped dot-plot (not shown).

As in other endosymbiotic associations (Moya et al. 2008; Nowack and Melkonian 2010), the genomes of the TPEs

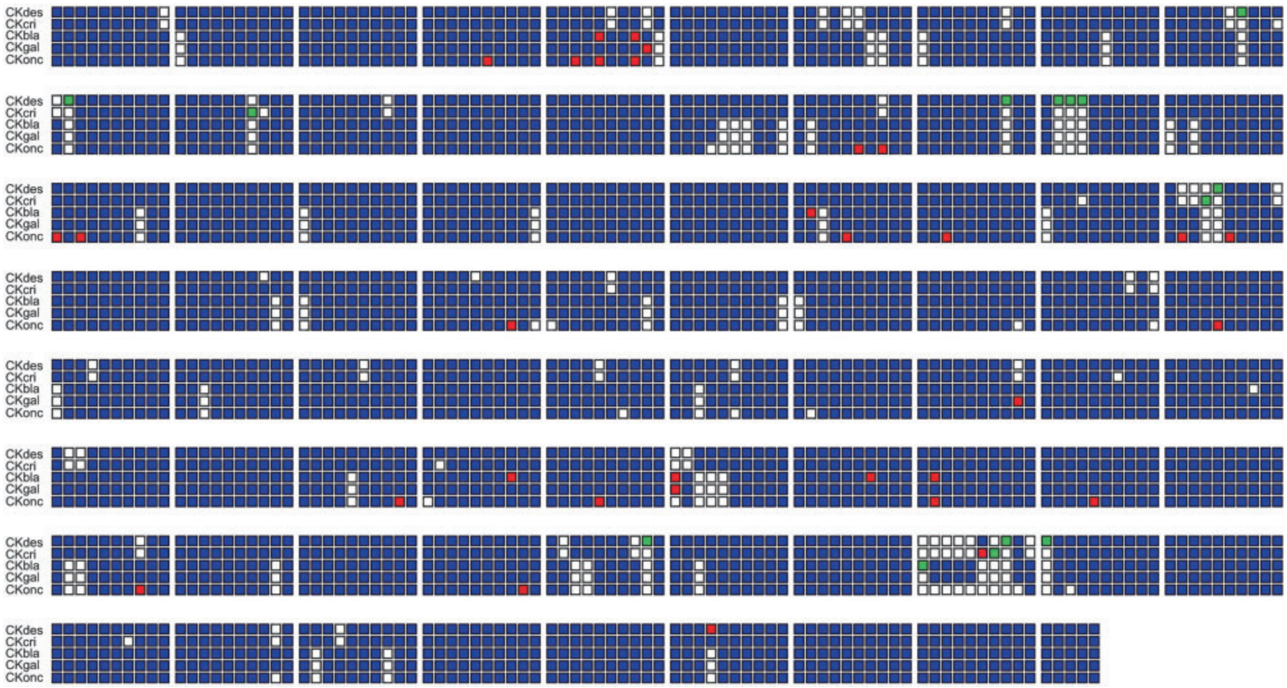

Fig. 3.—Orthologous group presence/absence display for the five *Candidatus* Kinetoplastibacterium genomes, in order of occurrence in the sequence after the *dnaA* gene. Boxes in the same column represent putative orthologs as identified by OrthoMCL. Organisms are denoted as: CKdes for *Ca.* K. desouzaii; CKcri for *Ca.* K. crithidii; CKbla for *Ca.* K. blastocrithidii; CKonc for *Ca.* K. oncopeltii; and CKgal for *Ca.* K. galatii. Box sizes are not proportional to gene lengths in the genome, and blocks are grouped in sets of 10 for aesthetic reasons only. Blue colour box denotes that ortholog is present in more than one organism; green, ortholog is present in one organism only; red, ortholog is present as a pseudogene; white, ortholog is absent.

apparently underwent a downward shift of their GC compositions to approximately 30% (table 1). In contrast, the genome of *Achromobacter* exhibits approximately 65.78% GC, whereas *T. equigenitalis*, the closest available relative of the TPEs, presents 37.42% GC. These values could suggest that the low GC content of the TPE genomes might not be completely associated with the process of becoming endosymbiotic, because their closest nonendosymbiotic relative has a relatively low GC value in comparison with other bacteria of the Alcaligenaceae family. All other Alcaligenaceae genomes currently available show GC contents ranging between 56.92% and 68.10%. Interestingly, like the TPEs, *T. equigenitalis* exhibits a markedly reduced genome size compared with the other Alcaligenaceae. Moreover, *T. equigenitalis* is an obligate intracellular pathogen and could be considered to be evolving toward an "endosymbiotic" life style. A more definitive conclusion on the significance of the variation in genome size and GC content among these bacteria will only be possible with more widespread sampling of taxa in the family.

Differences of Gene Content and Function among the TPEs

Although the endosymbionts present complete synteny and very similar overall characteristics, some differences in gene content, apparently due to differential gene loss, were observed. Figure 3 displays the presence or absence of protein-coding genes from the five TPE genomes aligned and in order from first (*dnaA*) to last gene. Interestingly, differences in gene content seem to appear in "hot-spots" of gene loss, interspersed with more stable blocks of genes. There also seems to be a trend of pseudogenes being located preferentially close to these "hot spots," suggesting that the presence of these pseudogenes is an early intermediate stage in the genome reduction process, as previously observed (Silva et al. 2001). Together, these observations suggest that functionally related genes from specific operons are being preferentially lost, as we have noticed in at least a couple of cases, for example, an operon related to ubiquinone biosynthesis, defective in the TPEs from *Angomonas*, and another related to thiamine biosynthesis, this time in process of loss from the genomes of TPEs from *Strigomonas* instead (not shown). This hypothesis is supported by the analysis of COG classifications for these "lost" genes. Although most COG categories are represented among the gene losses and pseudogenes, the distribution of these genes is not random ($P < 0.05$). Underrepresented (i.e., presenting less loss than expected by chance) categories include genes required for basic informational tasks and those related to the host-endosymbiont relationship; that is, "translation, ribosomal structure, and biogenesis" ($P = 2.7E{-}10$ using the two-tailed Fisher's exact test), "posttranslational modification, protein turnover, chaperones" ($P = 4.7E-06$), "DNA replication, recombination, and repair" ($P = 4.8E-05$), "energy production and conversion" ($P = 2.0E-04$), "amino acid transport and metabolism" ($P = 0.001$), and "nucleotide transport and metabolism" ($P = 0.02$). The finding that loss of genes from these categories is lower than expected suggests that most of the genes that could be lost from these categories have already been lost, and that many of the remaining ones might be needed to fulfill the needs of the endosymbionts or hosts. Overrepresented categories among gene losses and pseudogenes include, for the most part, genes of unknown or uncertain function, and genes related to the interface between the bacterium and its host; for example, "hypothetical protein" ($P = 9.5E-11$), "function unknown" ($P = 0.02$), "cell wall, membrane, and envelope biogenesis" ($P = 2.1E-08$), "coenzyme metabolism" ($P = 7.2E-06$), and "lipid metabolism" ($P = 0.02$) (note: although "hypothetical protein" is not a COG category, we have included such proteins in this analysis due to their abundance among genes lost in the TPEs).

The COG category "general function prediction only," which was not statistically overrepresented ($P = 0.40$), and the "hypothetical protein" class include a few genes predicted to encode membrane associated or transporter proteins, reinforcing the trend mentioned earlier of gene losses related to host–bacterium interface proteins. These observations argue that the genes of these functional categories are more dispensable by the TPEs. Cell wall- and lipid-related proteins, for example, might be dispensable given the protected environment inside the trypanosomatid host cell. Hypothetical proteins, on the other hand, could represent genes that were specific to the TPEs' former independent life, but which are uncharacterized given their obscure roles or limited taxonomic distribution, and that are now superfluous for their symbiotic intracellular life style.

Differences of Gene Content between the TPEs and Other Alcaligenaceae

Comparison of the presence of different functional categories as defined by COG has also shown interesting differences between the TPEs and their betaproteobacterial relatives (supplementary fig. S2, Supplementary Material online), demonstrating potential ancient losses that could have already occurred in all five TPEs, probably in their common ancestor—as a shorthand, we will be referring to the differences found as losses in the reduced TPE genomes, but it should be noted that, given the currently unsolvable lack of the relevant ancestral genomes, only a more thorough parsimony analysis of the gene presence patterns, and taking into account the underlying phylogeny, would be able to indicate the most likely reason for the differences with higher certainty. As previously observed in other endosymbionts (Moran et al. 2008), there is clear retention, compared with *Taylorella* and *Achromobacter*, of genes involved in replication ($P = 0.009$ and 1.2E−10, respectively) and translation ($P = 1.6E-09$ and $P = 1.2E-10$), as a proportion of whole genome gene content—although absolute numbers would seem to indicate otherwise given the large difference in genome size, specially between *Achromobacter* and the TPEs. In contrast, transcription-related genes are less represented in the TPE genomes than in *Achromobacter* ($P = 4.0E-10$). This finding might reflect the stability of the endosymbionts' intracellular milieu, which may decrease the requirement for regulatory molecules in relation to free-living organisms, which are much more likely to encounter differing environmental challenges. There was no statistically significant difference in the proportions of transcription-related genes between the TPEs and *T. equigenitalis*, which might reflect either shared ancestry of the reduction, or, more counter intuitively, that the pathogenic parasite *Taylorella* does not need extensive transcriptional regulatory capabilities in its intracellular niche in the equine genital tract. Other functional categories overrepresented in the TPEs in relation to both *Taylorella* and *Achromobacter* are "coenzyme transport and metabolism" ($P = 0.002$ and $P = 5.8E-11$) and "nucleotide transport and metabolism" ($P = 0.01$ and $P = 1.2E-10$), suggesting collaboration between the endosymbionts and their hosts in the synthesis of these compounds and, therefore, maintenance of necessary genes. Heme synthesis genes, for example, which were previously shown to be provided by the TPEs (Alves et al. 2011), are included in the "coenzyme" category. A more detailed analysis of the host genomes will be required to clarify whether a collaboration on the synthesis of nucleotides would benefit either, or both, of the two organisms. However, the "inorganic ion transport and metabolism" category is underrepresented in the TPEs as compared with the other two Alcaligenaceae examined ($P = 0.007$ and $P = 2.9E-06$), suggesting that the host could be supplying some of these metabolites.

Another interesting observation is the lower representation of "poorly characterized" genes (e.g., not matching any COG category) in the TPE genomes compared with the other Alcaligenaceae. This observation likely reflects the loss of nonessential genes, or of genes specific to independent life that have not yet been studied. The proportions of genes not classified in any COG are strikingly lower in the TPEs than in either *Taylorella* ($P = 1.8E-10$) or *Achromobacter* ($P = 3E-10$). The category "lipid transport and metabolism" is overrepresented in the endosymbionts compared with *Taylorella* but not *Achromobacter* ($P = 0.03$ and $P = 0.16$, respectively). The significance of this lower amount of *Taylorella* lipid-related genes, compared with the other two organisms, is unclear. Compared with *Achromobacter*, the number of other underrepresented categories in the TPEs is higher, and includes "secondary metabolites biosynthesis, transport and catabolism" (3.7E−05), "cell motility" ($P = 1.3E-04$), "function unknown" ($P = 0.002$), "amino acid transport and metabolism" ($P = 0.05$), and "signal transduction mechanisms"

($P = 0.03$). It is not surprising that a primarily free-living bacterium would have greater need for these functional gene categories than the endosymbionts, which do not require systems for motility, and have reduced needs for signal transduction for reasons similar to those discussed earlier for transcription regulation. Moreover, the relative overabundance of proteins of unknown function could be explained by the same logic that explains the overrepresentation of hypothetical proteins. Also, without collaboration from a host cell to supply amino acids, *Achromobacter* must be genetically competent to synthesize more of these compounds than parasites or endosymbionts.

Several categories of genes are overrepresented in the TPEs and *Taylorella* in relation to *Achromobacter*, including "posttranslational modification, protein turnover, chaperone" ($P = 5.8\text{E}{-}05$), "cell wall/membrane biogenesis" ($P = 4.0\text{E}{-}4$), and "cell cycle control, cell division, chromosome partitioning" ($P = 6.4\text{E}{-}04$). Neither of these categories is statistically different between TPEs and *Taylorella* ($P = 0.23$, 0.93, and 0.59, respectively), suggesting that the TPEs are close to having the minimal complement of genes from these categories, although some may still be lost (discussed earlier in the comparative analyses of the gene losses and pseudogenes in the five TPE genomes).

Recombination and Repair Genes

Recombination and repair-related genes are also of interest in endosymbionts, given that the extreme organizational stability of their genomes is commonly thought to be due to the loss of recombination genes and mobile elements (Silva et al. 2003). However, the *Ca.* Kinetoplastibacterium genomes contain all genes for the RecOR pathway of homologous recombination involving single-strand DNA breaks: *recAGJOR*, *ssb*, and *ruvABC*. The exact same arrangement is present in the genomes of *Taylorella* and *Achromobacter*, both of which also lack *recF* and thus also apparently use the RecOR version of the homologous recombination pathway (Sakai and Cox 2009). As recBCD are missing in all of these genomes, homologous recombination involving double-strand breaks does not seem possible in any of these organisms. On the other hand, two bacterial genes for nonhomologous end-joining (Ku protein and DNA ligase D1) are absent from both the TPEs and *Taylorella* but present in *Achromobacter* and many other betaproteobacterial genera currently present in KEGG. Given the isolation of the TPEs inside of the host precluding contact with other bacterial DNA, lack of *recBCD* is probably dispensable, and the presence of the RecOR pathway in the endosymbionts is maintained to perform essential functions of DNA damage repair. We speculate that the RecOR homologous recombination system could potentially mediate chromosomal inversions, and therefore explain the diversity in cluster arrangements in the Alcaligenaceae, described later.

Ribosomal Gene Cluster Organization

The ribosomal gene clusters in the endosymbiont genomes exhibit the customary betaproteobacterial organization (supplementary fig. S3, Supplementary Material online). There are three ribosomal clusters in all of the endosymbionts, as in other known Alcaligenaceae—although *Pusillimonas* has only two. With the exception of *A. xylosoxidans*, which bears a duplicated 5S rRNA gene in one of the clusters, the genes in the rDNA clusters from all bacteria included in table 1 are arranged in exactly the same in order: 16S-Ile-Ala-23S-5S, where Ile and Ala represent isoleucine and alanine tRNA genes, respectively. However, the orientations of the individual clusters in relation to each other vary. In *T. equigenitalis* and *Bordetella avium*, all of the clusters are in the plus strand (defined here as being the strand of the *dnaA* gene), as are all genes in each cluster. In contrast, in *T. asinigenitalis*, *A. xylosoxidans*, and other *Bordetella* species, two clusters are encoded in one strand (sometimes the plus strand, other times the minus strand), and the third is encoded from the other strand.

The organization of the rDNA gene clusters of these members of the Alcaligenaceae varies surprisingly, and often is not conserved across closely related taxa; compare the *Bordetella* species shown in supplementary figure S3, Supplementary Material online. These observations argue that the ribosomal region of these organisms is structurally pliable and subject to rearrangement, potentially by the RecOR system.

Phylogenomic Analysis of the Betaproteobacteria

We performed maximum likelihood analysis of 233 concatenated proteins from the 5 endosymbionts, 109 Betaproteobacteria from 7 different orders and 11 different families (fig. 2), and 3 Alpha- and 3 Gammaproteobacteria as outgroups. Organisms were selected from GenBank based on the availability of both genome sequences and predicted protein sequences. Genes from the 119 organisms were selected utilizing a "relaxed single-copy" approach, where inparalogs were initially allowed (all but one were removed from the final analyses) and genes were required to be present in only 95% of the organisms. This approach permitted us to use many more genes and organisms than a strict "single-copy in all organisms" approach would permit (not shown).

After removal of ambiguous positions from the alignment, 71,747 columns were analyzed, with only approximately 0.68% and 1.94% of the matrix composed of gaps or missing data, respectively. The resulting phylogenetic tree (fig. 2) shows, as previously suggested by more restricted phylogenetic analyses of a few genes (Alves et al. 2011), that the five TPEs are divided in two clades reflecting the taxonomy and phylogeny of their host species, trypanosomatids from the *Angomonas* and *Strigomonas* genera. Also as previously shown (Du et al. 1994; Umaki et al. 2009; Teixeira et al. 2011), the TPEs are within the family Alcaligenaceae but, in contrast to previous analyses, they group with the *Taylorella* instead of the *Bordetella*. This discrepancy is due to the previous lack of sequences intermediary between *Bordetella* and the TPEs; that is, *Taylorella* and *Pusillimonas*, which have had their genomes sequenced only very recently (Cao et al. 2011; Hebert et al. 2011). In general, bootstrap support values are very high (more than 95, and most often 100) across the whole tree, showing strong support of the phylogeny by the data. Bootstrap values for the branches among the TPEs are all 100, as are the ones for the branch-linking *Taylorella* and the TPEs, and the branch grouping all the Alcaligenaceae.

Only one low support value (of 47) is present, for the branch-positioning *Limnobacter* (classified as a Burkholderiaceae) as sister group of the Alcaligenaceae. The only other disagreement between our phylogeny and the previous taxonomic classification of the Betaproteobacteria here analyzed is the positioning of *Lautropia*, classified as a Burkholderiaceae, as a sister group of the Sutterellaceae with reasonably high bootstrap support (value of 81). Bootstrap support for other orders and families of the Betaproteobacteria is very high and groups are nearly all monophyletic, with few exceptions. The basal family, amongst those with sequenced genomes, seems to be the Neisseriaceae, as shown by rooting the tree with the Alpha- and Gammaproteobacteria.

A few previously unclassified (according to NCBI's Taxonomy database) bacterial species present in our tree can now be confidently placed in families. Thus, the Burkholderiales *incertae sedis Methylibium*, *Leptothrix*, and *Rubrivivax* cluster together, and as sister group of the Family Comamonadaceae with bootstrap support of 100 and divergence level consistent with its potential placement in this family. The next sister group of these bacteria, *Thiomonas*, is also currently unclassified at the family level; its divergence level from any other bacterial group included in our phylogeny is similar to that observed among the known families included, and might warrant the creation of a new family. *Ca.* Accumulibacter and the extremely reduced endosymbiont *Ca.* Tremblaya, both also previously unclassified at the family level, have been placed by our phylogenomic analysis in the Rhodocyclaceae and Burkholderiaceae families, respectively, with very high bootstrap support values of 100 and 95.

Two branches in the phylogeny were very long, that is, the branches for the highly reduced endosymbionts *Ca.* Tremblaya and the *Ca.* Zinderia. It is well known that long branches can cause problems for phylogenetic inference, by leading to the positively misleading long-branch attraction phenomenon (Felsenstein 1978), which can be a problem for any phylogenetic method if it is extremely enough. This would lead to unnatural groups being highly supported by bootstrap, whereas sometimes lowering bootstrap support values in other parts of the tree. With that in mind, we have also run the betaproteobacterial phylogeny described without including these two fast-evolving lineages to check for any changes in tree topology and bootstrap support. Except for the lack of the two species, the rest of the resulting tree (not shown) has identical topology to the one presented in figure 2, and bootstrap values are very similar, therefore indicating that the presence of the two long branches is not biasing the rest of the phylogeny in any appreciable way.

Analysis of Natural Selection Pressure

We have analyzed all 652 single-copy genes that were present in all 5 TPEs to scan for potential instances of positive selection in these bacteria. The nature of selective pressure is determined by calculating the d*N*/d*S* (nonsynonymous to synonymous substitution) ratio (ω), where $\omega < 1$, $\omega = 1$, and $\omega > 1$ mean negative, neutral, or positive evolution, respectively (Yang and Bielawski 2000). We tested a number of different models accounting for some modes of evolution (see Materials and Methods). As expected (Yang et al. 2000), tests of model M0 (allowing for only one ω category over time and across codon sites in genes) identified no genes with any signal of positive selection, with ω ranging from 0.00168 to 0.17109, thus indicating strong purifying evolution overall in TPE genes. However, it is common that a few codon sites in a gene could be subjected to positive selection (e.g., epitopes and extra-cellular residues) among a mostly conserved sequence, leading to a very low average ω for the gene as a whole. Thus, it is necessary to test sequence evolution with models (M2a and M8, here) that allow the occurrence of ω values above 1 on different codon sites along the genes, and compare those with their corresponding null models, which do not allow for $\omega > 1$ (M1a and M7, respectively). Using such models, we have identified several genes that contain one or more codons that seem to have undergone positive selection among the orthologs tested (table 2). As expected (Anisimova et al. 2001), the M1a–M2a test is more conservative and identified only two genes containing one codon each undergoing positive selection, whereas the M7–M8 test identified 20 genes containing codons with at least 0.95 probability of being under positive selection (one gene had one codon with better than 0.99 probability). All of these genes presented just one codon undergoing positive selection, except for one gene (ATP-dependent RNA helicase RhlE) with two such codons.

Categories of genes undergoing positive selection often include those related to membrane or extracellular proteins, proteins related to defense and environmental adaptation, or proteins involved in host–pathogen interaction (Aris-Brosou 2005; Anisimova et al. 2007). Less common, however, are reports of less obvious housekeeping genes undergoing positive evolution, although the significance of these findings is usually unexplored (Xu et al. 2011). In this work, we have seen many such genes amongst the candidates for positive selection (table 2), including many genes belonging to COG categories dealing with "amino acid transport and metabolism" (four genes—or five, if the methyltransferase is included),

Table 2

Genes Under Positive Selection in *Ca.* Kinetoplastibacterium

Annotation	COG Category Class[a]
Shikimate kinase	**Amino acid transport and metabolism**
Glutamine amidotransferase[b,c]	**Amino acid transport and metabolism**
3-Isopropylmalate dehydrogenase small subunit	**Amino acid transport and metabolism**
Leucyl aminopeptidase	**Amino acid transport and metabolism**
Small-conductance mechanosensitive ion channel protein (MscS)	Cell envelope biogenesis, outer membrane
Lipoprotein NlpD	Cell envelope biogenesis, outer membrane
Uroporphyrin-III C-methyltransferase (hemX)	Coenzyme metabolism
Chromosomal replication initiator protein (dnaA)	**DNA replication, recombination, and repair**
DNA polymerase III subunit delta	**DNA replication, recombination, and repair**
TatD DNAse family protein	**DNA replication, recombination, and repair**
ATP-dependent RNA helicase (rhlE)[d]	**DNA replication, recombination, and repair**
F-type H+-transporting ATPase subunit gamma	Energy production and conversion
F-type H+-transporting ATPase subunit epsilon	Energy production and conversion
Methyltransferase	General function prediction only
Fe/S cluster insertion protein ErpA[c]	Posttranslational modification, protein turnover, and chaperones
Predicted ATPase of the DUF815 family and AAA + superfamily	Posttranslational modification, protein turnover, and chaperones
Small subunit ribosomal protein S14	Translation, ribosomal structure, and biogenesis
Large subunit ribosomal protein L2	Translation, ribosomal structure, and biogenesis
Large subunit ribosomal protein L9	Translation, ribosomal structure, and biogenesis
Glutaminyl-tRNA synthetase	Translation, ribosomal structure, and biogenesis

[a] COG category classes in bold typeface are overrepresented (see text) in relation to the genome; those in normal typeface have representation not significantly different from the genome.

[b] Confidence of the positive selection inference, as calculated by PAML, is greater than 0.99 (all others are between 0.95 and 0.99).

[c] Genes identified both by M1a–M2a and M7–M8 tests.

[d] Two amino acids were detected as being under positive selection (all other had one).

"translation, ribosomal structure and biogenesis" (four genes), and "DNA replication, recombination, and repair" (four genes). Of these categories, only the amino acid- and DNA replication-related ones are overrepresented or nearly so, at the 0.05 level, in relation to the whole genome (Fisher's exact test *P* values of 0.0249 and 0.0867, respectively). Out of the 20 genes identified as possibly having undergone positive selection, only two (lipoprotein NlpD and small-conductance mechanosensitive ion channel protein MscS) belong to a category more obviously associated with this mode of evolution, namely "cell envelope biogenesis, outer membrane," because such proteins face the exterior of the cell and might participate more directly in interactions with hosts or, in case of free-living organisms, the environment. Although the delta and epsilon subunits of the F-type H+-transporting ATPase (ATP synthase) participate in a complex of proteins that is membrane-associated, these two specific subunits are part of the F1-unit, located on the cytoplasmic side. A previous report (Xu et al. 2011) on the selective pressure in genes from the parasitic bacterium *Actinobacillus pneumoniae* also found, in addition to the regularly found membrane-associated and virulence-related genes, several housekeeping genes presenting evidence for positive selection. Interestingly, some of the genes were in common with our observations (e.g., isopropylmalate dehydrogenase small subunit and mechanosensitive ion channel) or from the same pathways or processes (helicases, heme biosynthesis, DNA replication proteins, genes related to amino acid biosynthesis, and translation). In the endosymbionts of trypanosomatids, we hypothesize that the categories of genes found to be under positive selection probably reflect their interactions with their hosts, which might have led to accelerated evolution in processes related to compounds made by the TPEs for the host (amino acids, heme) and to DNA replication (helicase, polymerase, replication initiator)—synchronization of replication between the two organisms could be responsible for positive evolution in this category of genes.

Conclusions

As observed in other endosymbiont genomes previously analyzed, despite significant sequence divergence, genome rearrangement has not occurred in millions of years of separation between the TPEs. This is probably associated with the loss of some, but not all, of the DNA recombination genes. More specifically, nonhomologous end-joining genes are absent, whereas homologous recombination genes for the RecOR pathway are present.

The observed genome size difference between the TPEs and their nearest nonendosymbiont relatives is clearly due to loss of genes instead of gene or intergenic region size reduction, which are both essentially identical among all Alcaligenaceae analyzed here. Gene loss may occur by accumulation of substitutions that turn the gene into a pseudogene, followed by eventual sequence loss during replication. The categories of genes lost also match the needs of the endosymbiont and its host, with genes needed for the collaboration being preferentially retained in the TPEs. Overall, the vast majority of the genes conserved in these five endosymbionts are identical, implying that much of the gene loss occurred in a common progenitor.

Our results also show that the genomic data presented herein strongly support previous biochemical observations concerning the contribution of the bacterial symbionts to their trypanosomatid hosts. One example is the previously reported collaboration of the SHTs and TPEs in the synthesis of the essential compound heme, with the TPE genomes supplying most, but not all, of the genes necessary (Alves et al. 2011)—the trypanosomatid still performs the final three steps of the synthesis. One other such collaboration currently under analysis involves the amino acid metabolism pathways (in preparation) and, accordingly, a high number of amino acid synthesis genes have been retained in the TPE genomes. A more closely focused analysis of both endosymbiont and host genomes will ascertain which, if any, of the other preferentially retained gene categories are kept due to the host–endosymbiont relationship and not just to the specific needs of the intracellular bacterium.

Our phylogenetic analysis conclusively shows the positioning of the TPEs in the Alcaligenaceae family, as the sister group of *Taylorella* instead of *Bordetella*, as previously reported in analyses that employed less extensive taxonomic sampling. We can also confidently place the few previously unclassified Betaproteobacteria from our analyses in known families, with high bootstrap support. The only genus with uncertain placement in our analyses is *Limnobacter* which, while securely grouped in the order Burkholderiales, was not strongly associated with any of its families. Analyses of different genes, not constrained by the necessity of their presence in the TPEs as was the case here, will be required to investigate the taxonomic classification of *Limnobacter*.

And finally, we have also identified a number of genes possibly under positive selective pressure in the TPE genomes. Contrary to most previous similar studies on selective pressure in microorganisms, these were mostly intracellular and metabolic genes; normally, genes in the interface with the environment (or host) are the ones most often subjected to such pressure. We suggest that the genes identified in the TPEs as being under positive evolutionary pressure may reflect the collaboration between host and endosymbiont in metabolism, and also the necessity for synchronization in their life cycles.

Supplementary Material

Supplementary table S1 and figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Marta Campaner and Carmen C. Takata (USP), Vladimir Lee, Stephanie L. Hendricks, and Yingping Wang (VCU) for technical support, and Carlisle G. Childress Jr. and J. Michael Davis (VCU Center for High-Performance Computing). Sequencing was performed by the Nucleic Acids Research Facilities at VCU. This work was supported by the National Science Foundation grant NSF DEB-0830056 to G.A.B. and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) grant to M.M.G.T. and E.P.C.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Alves JMP, Buck GA. 2007. Automated system for gene annotation and metabolic pathway reconstruction using general sequence databases. Chem Biodivers. 4:2593–2602.

Alves JMP, et al. 2011. Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. PLoS One 6:e23518.

Anisimova M, Bielawski J, Dunn K, Yang Z. 2007. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. BMC Evol Biol. 7: 154.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 18:1585–1592.

Ardell DH, Andersson SGE. 2006. TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. Nucleic Acids Res. 34:893–904.

Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol. 22:200–209.

Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. 2005. Non-classical protein secretion in bacteria. BMC Microbiol. 5:58.

Camargo EP. 1964. Growth and differentiation in *Trypanosoma cruzi*. I. Origin of metacyclic trypanosomes in liquid media. Rev Inst Med Trop Sao Paulo. 6:93–100.

Campbell DA. 1992. *Bodo caudatus* medRNA and 5S rRNA genes: tandem arrangement and phylogenetic analyses. Biochem Biophys Res Commun. 182:1053–1058.

Cao B, et al. 2011. Complete genome sequence of *Pusillimonas* sp. T7-7, a cold-tolerant diesel oil-degrading bacterium isolated from the Bohai Sea in China. J Bacteriol. 193:4021–4022.

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. 2009. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25: 119–120.

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. 1993. Kingdom protozoa and its 18 phyla. Microbiol Rev. 57:953–994.

Chang KP. 1975a. Haematophagous insect and haemoflagellate as hosts for prokaryotic endosymbionts. Symp Soc Exp Biol. 407–428.

Chang KP. 1975b. Reduced growth of *Blastocrithidia culicis* and *Crithidia oncopelti* freed of intracellular symbiotes by chloramphenicol. J Protozool. 22:271–276.

Chang KP, Chang CS, Sassa S. 1975. Heme biosynthesis in bacterium-protozoon symbioses: enzymic defects in host hemoflagellates and complemental role of their intracellular symbiotes. Proc Natl Acad Sci U S A. 72:2979–2983.

Chang KP, Trager W. 1974. Nutritional significance of symbiotic bacteria in two species of hemoflagellates. Science 183:531–532.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27: 4636–4641.

de Souza W, Motta MC. 1999. Endosymbiosis in protozoa of the Trypanosomatidae family. FEMS Microbiol Lett. 173:1–8.

Du Y, Maslov DA, Chang KP. 1994. Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. Proc Natl Acad Sci U S A. 91:8437–8441.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.69 [Distributed by the author]. Seattle (WA): Department of Genetics, University of Washington.

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

Hebert L, et al. 2011. Genome sequence of *Taylorella equigenitalis* MCE9, the causative agent of contagious equine metritis. J Bacteriol. 193: 1785.

Holmes B, Snell JJ, Lapage SP. 1977. Strains of *Achromobacter xylosoxidans* from clinical material. J Clin Pathol. 30:595–601.

Kikuchi Y. 2009. Endosymbiotic bacteria in insects: their diversity and culturability. Microbes Environ. 24:195–204.

Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. Mol Phylogenet Evol. 56:1115–1118.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Lopez-Madrigal S, Latorre A, Porcar M, Moya A, Gil R. 2011. Complete genome sequence of "*Candidatus Tremblaya princeps*" strain PCVAL, an intriguing translational machine below the living-cell status. J Bacteriol. 193:5587–5588.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955–964.

Markmann K, Parniske M. 2009. Evolution of root endosymbiosis with bacteria: how novel are nodules? Trends Plant Sci. 14:77–86.

McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol. 2:708–718.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 42:165–190.

Motta MC, et al. 1997. Ultrastructural and biochemical analysis of the relationship of *Crithidia deanei* with its endosymbiont. Eur J Cell Biol. 72:370–377.

Motta MCM, et al. 2010. The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus. PLoS One 5: e12415.

Moya A, Peretó J, Gil R, Latorre A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat Rev Genet. 9: 218–229.

Newton BA, Horne RW. 1957. Intracellular structures in *Strigomonas oncopelti*. I. Cytoplasmic structures containing ribonucleoprotein. Exp Cell Res. 13:563–574.

Nowack ECM, Melkonian M. 2010. Endosymbiotic associations within protists. Philos Trans R Soc Lond B Biol Sci. 365:699–712.

Ogata H, et al. 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27:29–34.

Ozaki LS, Czeko YMT. 1984. Genomic DNA cloning and related techniques. In: Morel CM, editor. Genes and antigens of parasites. A laboratory manual. Rio de Janeiro (Brazil): Fundação Oswaldo Cruz. p. 165–185.

Perkins SL, Budinoff RB, Siddall ME. 2005. New gammaproteobacteria associated with blood-feeding leeches and a broad phylogenetic analysis of leech endosymbionts. Appl Environ Microbiol. 71:5219–5224.

Podlipaev S. 2001. The more insect trypanosomatids under study-the more diverse Trypanosomatidae appears. Int J Parasitol. 31:648–652.

Sakai A, Cox MM. 2009. RecFOR and RecOR as distinct RecA loading pathways. J Biol Chem. 284:3264–3272.

Schmitz-Esser S, et al. 2010. The genome of the amoeba symbiont "*Candidatus Amoebophilus asiaticus*" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. J Bacteriol. 192:1045–1057.

Silva FJ, Belda E, Talens SE. 2006. Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. Nucleic Acids Res. 34: 6015–6022.

Silva FJ, Latorre A, Moya A. 2001. Genome size reduction through multiple events of gene disintegration in Buchnera APS. Trends Genet. 17: 615–618.

Silva FJ, Latorre A, Moya A. 2003. Why are the genomes of endosymbiotic bacteria so stable? Trends Genet. 19:176–180.

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

Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:7.

Strnad H, et al. 2011. Complete genome sequence of the haloaromatic acid-degrading bacterium *Achromobacter xylosoxidans* A8. J Bacteriol. 193:791–792.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

Teixeira MMG, et al. 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.

Tillier ERM, Collins RA. 2000. Genome rearrangement by replication-directed translocation. Nat Genet. 26:195–197.

Umaki A, et al. 2009. Complete genome of the endosymbiont from *Crithidia deanei*, Proceedings XIII International Congress of Protistology/XXV Annual Meeting of the Brazilian Society of Protozoology/XXXVI Annual Meeting on Basic Research in Chagas Disease; Armação de Búzios, Brazil. p.152.

Vickerman K. 1976. The diversity of the kinetoplastid flagellates. In: Lumsden WHR, Evans DA, editors. Biology of the Kinetoplastida. London: Academic Press. p. 1–34.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041–1051.

Xu P, et al. 2007. Genome of the opportunistic pathogen *Streptococcus sanguinis*. J Bacteriol. 189:3166–3175.

Xu Z, Chen H, Zhou R. 2011. Genome-wide evidence for positive selection and recombination in *Actinobacillus pleuropneumoniae*. BMC Evol Biol. 11:203.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.

Associate editor: Dan Graur