Veterinary Parasitology 206 (2014) 267–276

Contents lists available at ScienceDirect

Veterinary Parasitology

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

Evidence for repeated gene duplications in *Tritrichomonas foetus* supported by EST analysis and comparison with the *Trichomonas vaginalis* genome

Jorge Oyhenart*, Javier D. Breccia

INCITAP – CONICET – Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa, Av. Uruguay 151, 6300 Santa Rosa, La Pampa, Argentina

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form
26 September 2014
Accepted 30 September 2014

Keywords:
Thrichomonosis
Expression
Evolution
Parasite
Protozoa
Venereal

ABSTRACT

Tritrichomonas foetus causes a venereal infection in cattle; the disease has mild or no clinical manifestation in bulls, while cows may present vaginitis, placentitis, pyometra and abortion in the more severe cases. *T. foetus* has one of the largest known genomes among trichomonads. However molecular data are fragmentary and have minimally contributed to the understanding of the biology and pathogenesis of this protozoan. In a search of new *T. foetus* genes, a detailed exploration was performed using recently available expressed sequences. Genes involved in the central carbon metabolism (phosphoenol pyruvate carboxykinase, glyceraldehyde-3-phosphate dehydrogenase, fructose-1,6-bisphosphate aldolase, thioredoxin peroxidase, alpha and beta chains of succinyl CoA synthetase, malate dehydrogenase, malate oxidoreductase and enolase) as well as in cell structure and motility (actin, α-tubulin and β-tubulin) were found duplicated and, in many cases, repeatedly duplicated. Homology analysis suggested that massive expansions might have occurred in the *T. foetus* genome in a similar way it was also predicted for *Trichomonas vaginalis*, while conservation assessment showed that duplications have been acquired after differentiation of the two species. Therefore, gene duplications might be common among these parasitic protozoans.

1. Introduction

Tritrichomonas foetus is an anaerobic parasitic protozoan known to cause a venereal disease in cattle. The microorganism may be found colonizing the bull's preputial cavity with mild or no clinical manifestation (Clark et al., 1974). Infection of the female may be the source of vaginitis, placentitis, uterine discharge, pyometra and abortion (Parsonson et al., 1976). The parasite is usually cleared from the cervical tract within 1–3 months while it may last for longer times in the male (Clark et al., 1974). *T. foetus* is worldwide distributed and the infection outcome is a significant impact in the herd productivity (Rae, 1989).

T. foetus cells are typically pear-shaped with three anterior and one recurrent or posterior flagellum. As most of the parabasalids, it is not known to form cysts. Endoflagellar or pseudocystic forms can be induced in culture by cold temperatures (Pereira-Neves and Benchimol, 2009). Pseudocysts would be present, and maybe occur more frequently than pear-shaped parasites, in infected bulls (Pereira-Neves et al., 2011). Flagellated and endoflagellar cells in contact with mammalian cells have also been described as acquiring amoeboid shape. Either form would be capable of promoting mammalian cell detachment and lysis (Pereira-Neves et al., 2012).

Molecular data are still fragmentary and have thus minimally contributed to the understanding of *T. foetus* biology

* Corresponding author. Tel.: +54 92954549591.
E-mail address: jorgeoyhenart@gmail.com (J. Oyhenart).

http://dx.doi.org/10.1016/j.vetpar.2014.09.024
0304-4017/

and pathogenesis. *T. foetus* has one of the largest observed genomes among trichomonads, at about 180 Mb (Zubáčová et al., 2008). It is distributed into 5 chromosomes that are thought to be stably inherited because no sexual stage has been described to date (Nadler and Honigberg, 1988; Tibayrenc et al., 1990; Yuh et al., 1997).

Ribosomal sequences are the most known sequences and the basis of *T. foetus* molecular diagnosis (Felleisen et al., 1998; Oyhenart et al., 2013). Other gene sequences have been described with the single purpose of undertaking taxonomic studies (Gerbod et al., 2004; Slapeta et al., 2012; Viscogliosi and Müller, 1998). The first *T. foetus* EST library was recently characterized (Huang et al., 2013). The overall data include about 2600 expressed genes, among which 45% appear to be novel sequences.

A single parabasalian genome has been sequenced to date. The *Trichomonas vaginalis* genome has approximately 170 Mb, a size comparable to the *T. foetus* genome (Zubáčová et al., 2008). Transposition elements would take account of a big proportion of the genome as about two thirds of the *T. vaginalis* genome is occupied with repeated elements of a single family (Pritham et al., 2007). Approximately 60,000 genes have been predicted in the *T. foetus* genome (Carlton et al., 2007). This number, 2–3 times higher than the human genome, is explained by repeated duplication of entire coding sequences.

Repeated genes occur in almost all organisms and are largely accepted as an important evolutionary mechanism (Ohno, 1982). The *T. vaginalis* genome seems to have retained multiple paralogous copies of a high amount of genes that could provide an opportunity to evolve in variable environmental conditions. It is not known if gene duplication is that common in *T. foetus* but previous efforts seem to indicate some genes would be present as different forms (Gerbod et al., 2004; Slapeta et al., 2012; Viscogliosi and Müller, 1998; Huang et al., 2013).

Expressed sequence tags (EST) are a popular and cost-effective means of initially cataloging many genes. DNA sequencing of randomly chosen clones from a cDNA library allow thousands of different transcripts to be identified. EST sequences can be assembled into consensus sequences or UniGene clusters that may in turn be compared to cDNA libraries obtained from different isolates. Such studies may help to identify single nucleotide polymorphisms (SNPs) and the variation within a species (Picoult-Newberg et al., 1999). Alternatively, the presence of SNPs or wrong sequence assemblies in the same library may provide evidence for heterozygosity, for the presence of homologous genes as well as for the existence of gene families.

Gene predictions from EST data are usually generated as consensus of automated pipeline results by employing comparative algorithms and data sources for gene and protein prediction. Comparative algorithms are inherently conservative, because of their reliance on gene and protein homology with other organisms, yielding predictions with high specificity but low sensitivity. Details in gene prediction thus must be obtained through more specific algorithms or by manual inspection and manipulation of sequence data.

In the search for new targets for diagnosis of *T. foetus* we undertook a detailed exploration of the Tf30924 cDNA library (Huang et al., 2013) and we found a high amount of genes may be repeatedly duplicated. We studied homologous *T. foetus* genes and compared them with orthologs in the *T. vaginalis* genome. We suggest that there would be striking resemblances between sequences in the genomes of *T. foetus* and *T. vaginalis*.

2. Methods

T. foetus cDNA sequences are available in the GenBank EST database as Tf30924 cDNA library Tritrichomonas foetus cDNA 5-, mRNA sequences (NCBI, 2014). This is a non-normalized 5′-end library obtained from the KV-1 strain (ATCC30924) which includes 4910 sequences with accession numbers from CX154307 to CX159216 (Huang et al., 2013).

Sequences were clustered with a CD-Hit Suite algorithm (Huang et al., 2010) with an identity cut-off set at 0.9 or 0.95, by taking account of reverse-complementary strands during alignment. Clusters with high homology to known gene products and including 6 or more cDNAs were chosen for a first round of analysis. Clusters were reassembled with MUSCLE (Edgar, 2004) under a SeaView version 4.3.1 platform (Gouy et al., 2010) and gaps and single base insertions present in no more than one sequence were removed. Consensus sequences were generated and BLAST (Altschul et al., 1990) searches performed against the whole library. Every sequence was thus traced back to the original cluster inferred by CD-Hit. When new clusters were identified another round of cleaning, consensus generation and search against the database was performed.

Consensus sequences were then used for BLAST searches against non-related sequences with different parameters (word-size: 7 or 10, gap cost: 5-2 or 2-2 for existence-extension respectively, and no filtering for low-complexity regions). *T. vaginalis* nucleotide and protein sequences were then used for recovery of similar products through BLAST against non-related as well as to the *T. vaginalis* genome reference (http://www.ncbi.nlm.nih.gov/genome/258).

T. foetus and *T. vaginalis* sequences were aligned and gaps were removed or displaced in order to get properly aligned nucleotide and polypeptide sequences. Sequence distances were estimated under the SEAView platform with BioNJ (Gascuel, 1997) with 1000 bootstrap replications and without distance correction.

DNA distance matrices were obtained through Clustal 2.1 at the European Biotechnology Institute portal website (http://www.ebi.ac.uk). Protein sequence identities and similarities were calculated through the SIAS free service (http://imed.med.ucm.es/Tools/sias.html) with default parameters. The domain search for the inference of protein function or group assimilation was performed with Conserved Domain Architecture Retrieval Tool (Geer et al., 2002).

3. Results

EST analysis can help in revealing the presence of allelic forms, particularly for single nucleotide changes or polymorphisms (SNPs) in a diploid genome. Therefore, since

T. foetus is presumed to be a haploid organism SNPs would be missing. The examination of *T. foetus* strain 30924 cDNA library, showed 93 clusters containing 6–60 sequences without SNPs, arguing in favor of an organism with a unique, haploid genome. Actually, there is no rule indicating that a single base change would be originated from different alleles. However one or more base changes accumulated in several transcripts would be indicative of new copies of a single gene and the comparison with known sequences in a reference genome would help in the prediction of duplications. We found in the expression library the presence of two or more variants of homologous transcripts for the central pathways of energetic metabolism (fructose-1,6-bisphosphate aldolase, glyceraldehyde-3-phosphate dehydrogenase, phosphoenol pyruvate carboxykinase, malate dehydrogenase, malate oxidoreductase, succinyl CoA synthetase, thioredoxin peroxidase, enolase) as well as genes associated to cell structure and motility (actin, α-tubulin and β-tubulin).

3.1. Phosphoenol pyruvate carboxykinase (PEPCK)

The CD-Hit program (Supplementary File 1) ordered 49 putative PEPCK sequences in cluster 1. The alignment of 650 nucleotides showed a repetitive pattern of 12 nucleotide changes (with 2 amino acid changes) suggesting the presence of 2 different transcripts. Search by similarity with the 2 putative Tf-PEPCKs (Tf-pPEPCK1 and Tf-pPEPCK2) rendered other related sequences in the TfEST 30293 library. Two cDNAs in cluster 520 were identical to Tf-pPEPCK1, seven sequences in cluster 66 showed enough differences with Tf-pPEPCK1 and Tf-pPEPCK2 to be considered as issued from a different gene (Tf-pPEPCK3), and a fourth transcript (Tf-pPEPCK4) was deduced from 4 sequences in clusters 226 ($n = 3$) and 722 ($n = 1$).

Similarity search between putative Tf-PEPCK1-4 (Supplementary File 2) and *T. vaginalis* genome revealed five complete (TVAG_310250, TVAG_479540, TVAG_213710, TVAG_139300, TVAG_420390) and 2 interrupted (TVAG_314830, TVAG_434120) genes for such activity. Full coding regions for *T. vaginalis* PEPCKs are predicted around 1800 bp. and *T. foetus* sequences consistently covered the region spanning codons 8 through 205. *T. foetus* protein sequences were found 89–99% similar throughout this region and showed 81–84% similarity with *T. vaginalis* proteins (Supplementary File 3). An unrooted phylogenetic tree based on 198 amino acids of the putative Tf-PEPCK1-4 and 6 predicted proteins of *T. vaginalis* showed two different species specific branches (Fig. 1A). Such a clear division argues in favor of repeated duplication events occurring after the divergence of both parasites.

A 5′ terminal library construction has often cDNAs issued from 3′ termini. The C-terminal of 2 putative Tf-PEPCKs were found among sequences contained in cluster 36 ($n = 10$). They differed in 5 nucleotide positions along 580 bp. Overlapping sequences present in Clusters 41 ($n = 8$), 148 ($n = 4$), 254 ($n = 3$), 259 ($n = 3$) and 369 ($n = 2$) were used to fill a gap of more than 700 bp between 5′ and 3′ terminal ends. Tf-PEPCK1 and 2 complete sequences are >99% identical and have 82–85% identity with *T. vaginalis* paralogs.

3.2. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)

Three *T. foetus* sequences belonging to two GAPDH genes were previously reported (Viscogliosi and Müller, 1998, and Genbank U66072.1). They were named GAP1 and GAP2, having 76% identity at nucleotide level and 82% at polypeptide level. A missing codon helps to easily differentiate in GAP1 from GAP2. Sequences 100% identical to GAP1 were not found in the *T. foetus* EST-library. Clusters 2 and 22 had sequences predicted to encode the NAD-binding domain of 2 glyceraldehyde 3-phosphate dehydrogenases similar to GAP1 (pfam00044). Two thirds ($n = 37$) of the cDNAs (cluster 2 and 22) differed from the reference sequence of GAP 1 (AF022415) by 4 bp positions (C142T, A324G, C370T and A506C), 2 of these presumptively lead to amino acid changes (K121R and I182L, amino acid positions are based in the predicted complete sequence) (referred next as GAP1b). The remaining sequences ($n = 19$) showed 2 out of the 4 mentioned nucleotide substitutions (C370T and A506C) and one amino acid change (I182L) (referred next as GAP1c). The 3′-end, encoding the C-terminal or catalytic domain of GAPDH (pfam02800), was found in cluster 25 ($n = 13$). Ten cDNAs could be clearly distinguished from the reference sequence by the presence of 7 bp changes: A506C, A728G, C748T, T871A, C935G, C937T and T946C, three of them silent and the others leading to 3 amino acid substitutions (I182L, I256V and P325A). Three other sequences showed 6 out of the 7 cited nucleotide changes (A506C, A728G, C748T, T871A, C935G and T946C) and the same amino acid changes. The size of the protein allowed overlapping of 5′-end and 3′-end sequences. In comparison with GAP1, only one nucleotide change (A506C) was observed in the central region and it was found in both GAP1b and GAP1c. Regarding GAP2, only 6 sequences (cluster 86) of complementary DNAs matching exactly the 5′ termini (545 bases) were found in the library.

There are 6 complete GAPDH sequences (TVAG_347410, TVAG_412780, TVAG_366380, TVAG_446910, TVAG_475220 and TVAG_476100) in the *T. vaginalis* draft genome. Some of them have been confirmed by cloning (Markos et al., 1993; Viscogliosi and Müller, 1998; Carlton et al., 2007). Similarity analysis suggested 5 *T. vaginalis* sequences are >98% identical (>98% similarity) while the sixth, GAP2 (TVAG_476100), would be more distant with a mean identity of 95% (97% similarity). Phylogenetic reconstructions of *T. foetus* and *T. vaginalis* GAP sequences were based in amino acids 14–336 with distance-based, maximum-parsimony and maximum-likelihood methods rendered 2 robust branches with an early diverging GAP2 gene, and several GAP1-like proteins in both species (Fig. 1B). Five insertion/deletion codons in the N-terminal region and one in the C-terminus greatly help in distinguishing between *T. foetus* and *T. vaginalis* putative GAPs. While a single codon insertion/deletion around position 100–117 allows the rapid recognition between GAP1 and GAP2 in both species. These findings support that GAP1

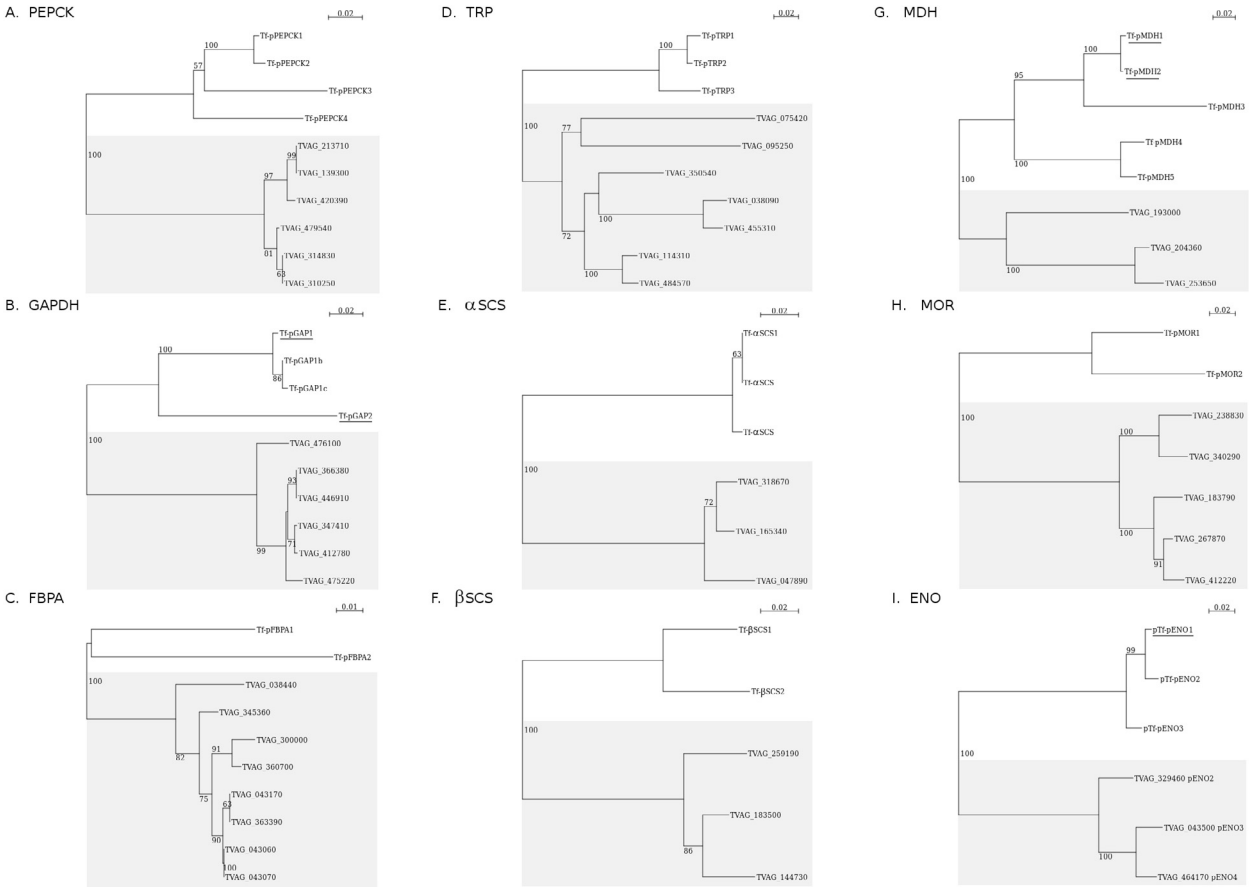

Fig. 1. Gene duplications in metabolic genes of *T. foetus* and *T. vaginalis*. Unrooted maximum-likelihood trees inferred from phosphoenol pyruvate carboxykinase (A, PEPCK), glyceraldehyde-3-phosphate dehydrogenase (B, GAPDH), thioredoxin peroxidase (D, TRP), α chain of succinyl CoA synthetase (E, αSCS) or beta (F, βSCS) subunits, malate dehydrogenase (G, MDH), malate oxidoreductase (H, MOR) and enolase (I, ENO) sequences from *T. foetus* and *T. vaginalis*. The fructose-1,6-bisphosphate aldolase tree (C, FBPA) was rooted based on species specific sequences. Nucleotide sequences are available in the GenBank database and in supplementary files as described in the text. Bayesian posterior probabilities are given as percentages near the individual nodes. Nodes with values of <50% are not shown. Sequences already described in *T. foetus* are underlined and *T. vaginalis* sequences are shaded in gray. Scale bars indicate substitutions per base pair.

and GAP2 diverged in a common ancestor, while GAP1 duplications took place after speciation.

3.3. Fructose-bisphosphate aldolase (FBPA)

Three clusters (3, 161 and 184) contained 42 sequences that could be confidently aligned into a transcript encoding for a putative class-II FBPA (pfam01116). Two other clusters (10 and 61) grouped 28 sequences for a similar product. Putative proteins are 85% identical (91% similar) and were named Tf-pFBPA1 and Tf-pFBPA2. BLAST search suggested 8 FBPA similar coding regions would be present in the *T. vaginalis* genome (TVAG_043060, TVAG_043070, TVAG_043170, TVAG_363390, TVAG_038440, TVAG_300000, TVAG_360700 and TVAG_345360). These sequences are highly conserved with identities superior to 95% (>97% similarity). *T. vaginalis* sequences coding for putative FBPAs were more distant from Tf-pFBPA2 (84–86% identity, 91–92% similarity) than from Tf-pFBPA1 (87–89% identity, 97–100% similarity). A phylogenetic reconstruction was performed on the basis of 324 predicted amino acids and results suggested that Tf-pFBPA1 and Tf-pFBPA2 may have diverged soon after speciation (Fig. 1C).

3.4. Thioredoxin peroxidase (TRP)

Complete coding sequences for three TRPs (Typical 2-Cys Peroxiredoxin (PRX) family [cd03015]) from *T. foetus* were found and named Tf-pTRP1 (clusters 7 and 120, $n = 29$), Tf-pTRP2 (cluster 11, $n = 21$) and Tf-pTRP3 (cluster 27, $n = 12$). A fourth potential gene represented by a single transcript (CX156366) was not considered in the analysis. The translation starting site rarely appeared in the sequences, the predicted protein from the alignments with *T vaginalis* was around 194 amino acids. A BLAST search showed seven complete genes for putative TRPs in *T. vaginalis* (TVAG_114310, TVAG_484570, TVAG_350540, TVAG_075420, TVAG_038090, TVAG_455310, TVAG_095250), another one lacking the 3′-end extremity (TVAG_528900) and other with both ends missing (TVAG_528900). Predicted Tf-pTRP1-3

polypeptides are 94–99% identical (96–100% similar) and have 66–77% identity (73–84% similarity) with *T. vaginalis* putative TRPs. Phylogenetic analysis showed a monophyletic group with duplications occurring after divergence of both parasites (Fig. 1D).

3.5. α-Chain of succinyl CoA synthetase (αSCS)

Sequences presumptively coding for the α-chain of succinyl CoA synthetases were found in cluster 35. The cDNAs covered the first 2/3 of putative αSCSs homologous of *T. vaginalis*, including the CoA binding domain (pfam02629) and CoA-ligase domain (pfam00549). Three groups ($n = 5$, 3 and 2 cDNAs) containing 27 changes along 600 bp were found in the same cluster. Alignment of *T. foetus* and *T. vaginalis* αSCS sequences showed that approximately 4–6 codons would still be missing at 5′-end of the three open reading frames. Two sequences in cluster 335 overlapped with the first group (Tf-αSCS1) while the second and third groups still incomplete at 3′-end. Two amino acid changes would distinguish Tf-αSCS2 from Tf-αSCS1 and 3. Base changes that allowed distinction of the three putative αSCS genes occurred mostly at wobble codon bases.

For *T. vaginalis* αSCSs, also called adhesin protein 33 or AP-33, three similar genes with 12 amino acid replacements mostly at the N-terminus were found (TVAG_047890, TVAG_318670 and TVAG_165340). Similarity analysis showed those sequences are 97–99% conserved (96–98% identity). Phylogeny reconstruction based on the alignment of the first 198 amino acids of *T. foetus* and *T. vaginalis* proteins showed an equilibrated unrooted tree with parallel duplication-divergence events (Fig. 1E). Identity between *T. foetus* and *T. vaginalis* sequences was 76–77% (80–82% similarity) and the speciation seems to have been accompanied by an early insertion of 2 codons in the *T. foetus* N-terminal region (between amino acids 10 and 11 of *T. vaginalis* αSCSs).

3.6. β-Chain of succinyl CoA synthetase (βSCS)

Transcripts presumptively encoding the β subunit of succinyl CoA synthetase were also found in the library. The coding sequence for a putative βSCS1 was obtained from clusters 53 ($n = 8$), 280 ($n = 3$) and 594 ($n = 2$). A similar sequence, putative βSCS2, was assembled from cDNAs grouped into clusters 130 ($n = 4$) and 157 ($n = 4$). Other Tf-βSCS sequences are possibly present in the genome of *T. foetus* but could not be clearly detached from reading errors (e.g. CX156316). Putative βSCS 1 and 2 showed 92% identity (95% similarity) over the 401 deduced amino acid positions. There are three related βSCS sequences in *T. vaginalis* (TVAG_259190, TVAG_144730 and TVAG_183500), also called adhesin proteins AP51-1, 2 and 3, which are 76–78% identical to *T. foetus* polypeptides (83–85% similar). Phylogenetic trees of βSCS sequences rendered 2 robust branches with recently evolved isoforms (Fig. 1F).

3.7. Malate dehydrogenase (MDH)

Two sequences were previously described as putative malate dehydrogenases, MDH1 and MDH2, in *T. foetus* (AF307994 and AF307995). Identical transcripts to those genes were not found in the library, nevertheless the product of clusters 18 ($n = 17$), 109 ($n = 5$) and 324 ($n = 2$) overlap into a messenger putatively encoding for a polypeptide (Tf-pMDH3) 85% identical (88% similar) to both MDH1 and 2. Another sequence, Tf-pMDH4, merged sequences in clusters 34 ($n = 10$), 634 ($n = 2$) and 1240 ($n = 1$). It appeared to be more distant with identities of 80, 81 and 67% for Tf-pMDH1, Tf-pMDH2 and Tf-pMDH3 respectively. A fifth transcript was deduced from cluster 34 ($n = 1$) and cluster 279 ($n = 3$). Some errors can still be present in Tf-pMDH5, particularly in the 5′-end, but several nucleotide changes and an uninterrupted polypeptide justified the assignment. Tf-pMDH4 and Tf-pMDH5 proteins are 96% identical (98% similar) and they seem more related to Tf-pMDH1 and Tf-pMDH2 than to Tf-pMDH3. BLAST search showed three complete (TVAG_193000, TVAG_204360 and TVAG_253650) and two incomplete (TVAG_196230 and TVAG_196240) genes in *T. vaginalis*. Sequence similarity analysis of *T. vaginalis* putative proteins showed TVAG_193000 is 84% similar (74% identity) to TVAG_204360 and TVAG_253650 with the last two being more closely related (96% identity, 97% similarity). Interspecies identities are between 63 and 69% and phylogenetic analysis argues in favor of a common origin followed by duplications after speciation (Fig. 1G).

3.8. Malate oxidoreductase (MOR)

Clusters 72 ($n = 7$), 217 ($n = 3$), 227 ($n = 3$) and 869 ($n = 1$) rendered a contig (Tf-pMOR1) containing 2 domains (pfam00390, pfam03949) found in malate oxidoreductases. A highly similar sequence (Tf-pMOR2) to the N-terminal domain (amino acids 1–236) was gathered from clusters 8 ($n = 23$), 84 ($n = 5$) and 1035 ($n = 1$). On the other hand, five complete (TVAG_183790, TVAG_238830, TVAG_412220, TVAG_340290, TVAG_267870) and two incomplete (TVAG_416100, TVAG_068130) genes with high similarity to Tf-pMORs were located in the *T. vaginalis* genome Their products are predicted to be hydrogenosomal malate dehydrogenases and evidence supports a possible role in cell adhesion. Alternative names are: adhesin protein 65 or AP65 (O'Brien et al., 1996). Identity of the putative proteins Tf-pMORs 1 and 2 was 85% (91% similar) and both sequences were 62–67% identical (73–78% similar) to TVAG predicted proteins. This gene family seems to have evolved late after speciation (Fig. 1H).

3.9. Enolases (ENO)

Enolase proteins carry over different tasks in the cells. They are predicted to have distant or even different origins. A single enolase gene was previously found in *T. foetus* (Gerbod et al., 2004). Complementary DNAs putatively encoding this enolase (Tf-pENO1) were not found in the expression library but three similar sequences were identified. A consensus sequence obtained from

clusters 43 ($n=9$), 31 ($n=11$), 52 ($n=8$), 288 ($n=3$), 373 ($n=2$), 610 ($n=2$), 657 ($n=2$), 1561 ($n=1$) and 2286 ($n=1$) showed 99% similarity (through 334 amino acids) with Tf-pENO1. Another consensus, obtained from clusters 44 ($n=9$), 31 ($n=2$), 549 ($n=2$) and 1527 ($n=1$) showed 98% identity with Tf-pENO1. A fourth protein more distantly related, was found in cluster 137 ($n=4$) and it was only 70% similar with the former group. The search for similar enolases sequences in the *T. vaginalis* rendered seven complete genes (TVAG_148010, TVAG_329460, TVAG_043500, TVAG_464170, TVAG_263740, TVAG_487600, TVAG_282090). Three of these sequences (TVAG_329460, TVAG_043500, TVAG_464170) are known as putative enolases 2, 3 and 4 and were found more alike to Tf-pENO 1, 2 and 3. An unrooted tree built up from the putative enolases showed two species specific branches with similar splitting patterns suggesting a parallel way of proteins acquisition (Fig. 1I).

3.10. Actin

There are not actin sequences reported for *T. foetus*. Considering that is still in discussion if *T. foetus* and *T. suis* are different microorganisms or not, we include in the study an actin sequence (acc. AB468092) recently reported by Noda et al. (2012). It lacks about 220 bases at the 5′-end and 21 bases 3′-end of the entire open reading frame. The first half of the sequence showed 86% identity with reads grouped in cluster 4 and the tail portion showed 93% identity with cDNAs clustered under number 46. *T. foetus* sequences in cluster 4 (5′-end, $n=29$) showed a striking feature of repeated changes falling in wobble bases giving unchanged protein products. Therefore, this group would be composed by 7 kinds of transcripts raised from a similar number of genes. On the other hand, a BLAST search for cluster 46 (3′-end, $n=9$) showed clusters 202 ($n=3$), 267 ($n=3$), 360 ($n=2$) and 836 ($n=2$) contained putative actin encoding transcripts. Due to the poor overlapping among these 3′-end sequences and those grouped in cluster 4 was not possible obtain consensus sequences representing the whole transcripts.

Local alignment search against *T. vaginalis* sequences showed 10 complete and 3 interrupted coding regions for actin proteins. The nucleotide sequences also showed frequent changes falling in wobble codon bases (48 nt changes in 1100 bp), and just 2 changes in the same codon that raise with three different polypeptides modified in one amino acid (Q, S or K at position 308). A reconstruction of the distances separating the known *T. foetus* and *T. vaginalis* actin sequences showed unrooted trees like those exemplified in Fig. 2A and B. Where nucleotide and polypeptide sequences seem to have evolved in a similar way, with repeated duplication events. The selection pressure is evident by the high number of synonymous substitutions and the virtual absence of non-synonymous changes. The *T. suis* actin sequence (TF-pActin1) showed 2 amino acids replacements, F141L and V213I. Homologies with *T. foetus* and *T. vaginalis* sequences suggest they could be sequencing errors.

3.11. α-Tubulin

Two *T. foetus* α-tubulin gene sequences were previously cloned and are available under accession numbers AY277784 and AY277785 (Gerbod et al., 2004). These sequences would encode amino acids 26–406 (with reference in *T. vaginalis* paralogs) from α-tubulin (cd02186) proteins predicted to have 452 amino acids. They show 90% identity at nucleotide level and 99% identity at protein level. Clustering of similar cDNAs from the *T. foetus* library showed a single group (5′-end, cluster 9) of 21 members with high similarity. These sequences showed repeated changes regularly spaced every third base suggesting they may be transcribed from different genes. Sequences with high identity to the 3′-end of such genes were found in different clusters (105, 150, 151, 1032, 1640 and 2231). Aligned sequences showed only partial overlapping but repeated changes located in wobble bases were evident. Overall sequence alignment for α-tubulin is self confident until codon 218.

A search for similar coding sequences in the *T. vaginalis* genome showed 7 complete (TVAG_467840, TVAG_360870, TVAG_206890, TVAG_196270, TVAG_359090, TVAG_312330, TVAG_448390) and 3 interrupted α-tubulin genes (TVAG_519620, TVAG_345420, TVAG_523980). These genes showed regular nucleotide changes at synonymous positions and identities of 97–99% at nucleotide level and 100% at protein. *T. foetus* and *T. vaginalis* α-tubulin genes showed 85–87% identity and proteins 93–95% identity. Phylogenetic trees built from nucleotide and protein sequences reproduce the picture obtained with actin sequences and arguments in favor of a high selection pressure acting on gene duplications encoding a structural protein (Fig. 2C and D).

3.12. β-Tubulin (β-Tub)

T. foetus sequences for β-tubulin 1 and β-tubulin 2 (cd02187) were previously described and can be found in the GeneBank database under accession numbers AY277786 and AY277787 (Gerbod et al., 2004). These are partial transcripts that would encode amino acids 17 to 381 (with reference in *T. vaginalis* paralogs) of proteins expected to be 447 amino acids long. Sequences presumptively encoding for the N-terminal region of β-tubulin were found in clusters 21 ($n=16$), 30 ($n=11$), 526 ($n=2$), 200 ($n=2$) and 744 ($n=1$). Inspection of alignments showed frequent synonymous changes that helped in prediction of 7 different genes. For the already known sequences of β-tubulin 1 and β-tubulin 2 were found 1 and 5 transcripts respectively. Identity among β-tubulin nucleotide sequences was 97–99% and no difference was found at protein level. Blast search led to 9 similar sequences in the *T. vaginalis* genome (TVAG_073810, TVAG_008680, TVAG_456920, TVAG_062880, TVAG_034440, TVAG_525430, TVAG_148390, TVAG_338530, TVAG_200200) being the last two are more distant. Identity analysis of codons 15 to 185 showed that *T. foetus* and *T. vaginalis* sequences are 86–89% identical at the nucleotide level while they would produce proteins which are more than 97% identical.

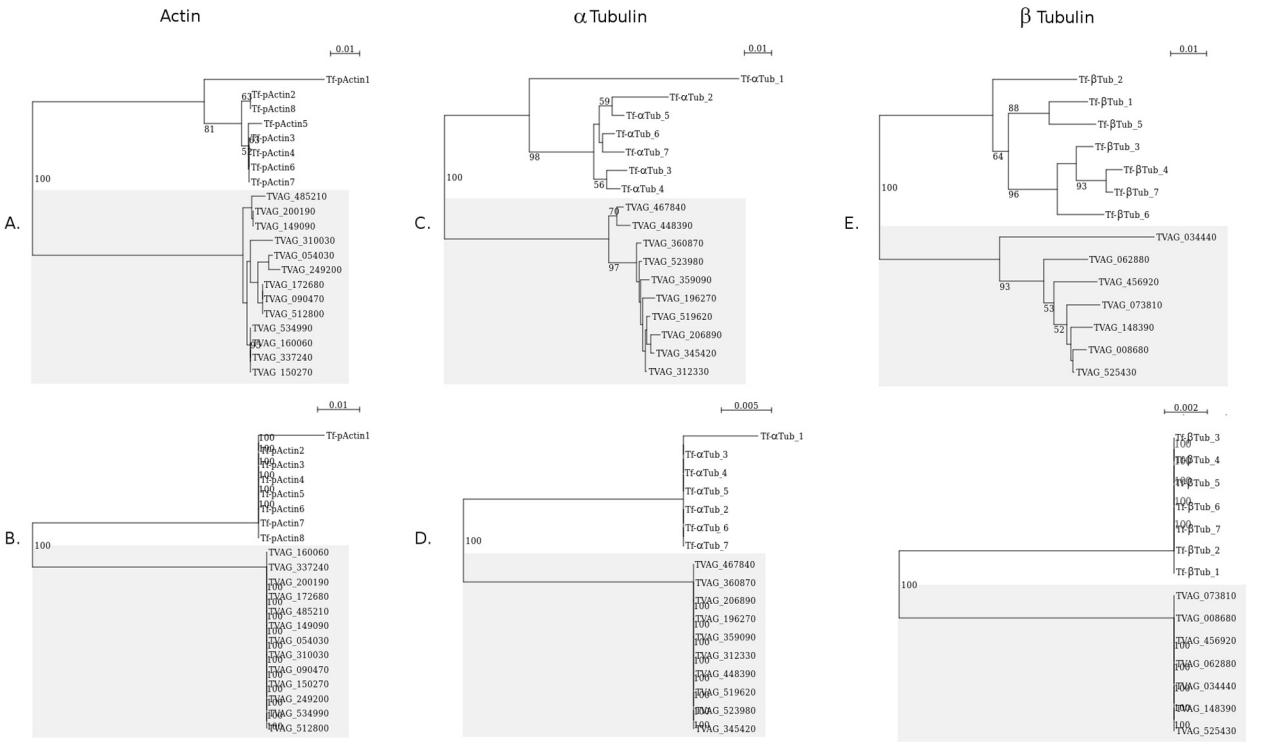

Fig. 2. Gene duplications in structural genes of *T. foetus* and *T. vaginalis*. Unrooted maximum-likelihood trees inferred from nucleotide and amino acid sequences from Actin (A and B, respectively), αTubulin (C and D) and βTubulin (E and F), similar sequences from *T. foetus* and *T. vaginalis*. Nucleotide sequences are available in the GenBank database and in supplementary files. Bayesian posterior probabilities are given as percentages near the individual nodes. Nodes with values of <50% are not shown. Sequences already described in *T. foetus* are underlined and *T. vaginalis* sequences are shaded in gray. Scale bars indicate substitutions per base pair.

Distances between *T. foetus* and *T. vaginalis* β-tubulins are depicted through phylogenetic trees in Fig. 2E and F.

3.13. Summary of gene duplications

Computational clustering methods employed here and in the original description of the EST cDNA library (Huang et al., 2013) achieved similar results predicting the absence of the transcripts GAP1, pMDH1, pMDH2, α-tubulin 1 and α-tubulin 2. Both methods have difficulties to discriminate between alike transcripts. This fact imposed the need of visual inspection and sequence manipulation, in order to judge if they bear sequencing errors or if they should be assigned to a different source gene.

T. foetus pTRP4, pMDH3, pMDH5, Actin5, Actin7, Actin8, α-tubulin6, α-tubulin7 and bTub7 paralogs were thus predicted after detailed visual analysis and repeated sequence comparison. Table 1 shows the *T. foetus* duplicated genes and summarizes the described findings.

4. Discussion

Computational clustering techniques are invaluable tools for the study of expression data. Transcripts can be quickly inferred, ordered and potential functions listed. Nevertheless the partitions returned by clustering algorithms are compendious and must be validated (Handl et al., 2005). The first *T. foetus* cDNA library obtained from the strain 30394 was a great contribution toward the understanding of its biology and pathogenesis (Huang et al., 2013). Herein, the assessment of the library was performed targeting to partially assembled sequences, unknown genes and variations that indicated presence of gene duplications. It revealed that several genes involved in energetic metabolism, cell structure and motility would effectively be duplicated. Moreover several genes were found repeatedly duplicated. In a process of gene duplication, natural selection is intended to maintain the original function of a preexisting gene while a copy could loose the function or acquire a new one (reviewed in Magadum et al., 2013). Otherwise duplicated genes could produce relatively larger doses of an unchanged protein product. A different function or specificity, or the reinforcement of the ancient function seems to be common routes followed by metabolic and structural gene duplications here described.

Changes in metabolic gene paralogs seem to be randomly produced in *T foetus*, while product divergence was found proportionally distributed along the phylogenetic trees. This observation suggested that gene preservation and neo-functionalization were outstanding processes for this microorganism. In contrast, the changes in the structural paralogs argue in favor of a high selection pressure hampering the expression of a different product. Actin, α-tubulin and β-tubulin nucleotide sequences showed enough differences to complicate the estimation of time elapsed since the separation of *T. foetus* or *T. vaginalis*

Table 1
Duplicated genes of *T. foetus*. Gene copies were numbered according to the order in which they were discovered. The relationship between each predicted gene and the clusters automatically obtained in this work (CD-HIT clusters) or consensus sequences from Supplementary File in Huang et al. (PHRED-PHRAP contigs) are shown.

Gene	CD-Hit clusters[a]	PHRED-PHRAP contigs[b]	References	GenBank
Tf-pPEPCK1	1, 520	0777 and 0782		
Tf-pPEPCK2	1	Similar to 0729		
Tf-pPEPCK3	66	Similar to 0698		
Tf-pPEPCK4	226, 722	Similar to 0594		
Tf-GAP1	No	No	Viscogliosi and Müller, 1998	AF022415.1
Tf-GAP1b	2, 22	TfEST_0774		
Tf-GAP1c	2, 22	TfEST_0786, 0667		
Tf-GAP2	86	5-end in 0146, 0586	Viscogliosi and Müller, 1998	AF022416.1, U66072.1
Tf-pFBPA1	3, 161, 184	0507, 0521, 0737 and 0769		
Tf-pFBPA2	10, 61	0789		
Tf-pTRP1	7, 120	0641, 0707, 0791		
Tf-pTRP2	11	0683, part of 0761		
Tf-pTRP3	27	0739		
Tf-pTRP4	541	No		
Tf-αSCS1	35	0433		
Tf-αSCS2	35	0295		
Tf-αSCS3	35	0684		
Tf-βSCS1	53, 280, 594	Part in 0742		
Tf-βSCS2	130, 157	Part in 0614		
Tf-pMDH1	No	No	Lee, Moore, Koszul and Müller	AF307994.1
Tf-pMDH2	No	No	Lee, Moore and Koszul	AF307995.1
Tf-pMDH3	18, 109, 324	No		
Tf-pMDH4	34, 634, 1240	0744		
Tf-pMDH5	34, 279	No		
Tf-pMOR1	72, 217, 227, 869	0752		
Tf-pMOR2	8, 84, 1035	0773		
Tf-pENO1	No	No	Gerbod et al., 2004	AY277773.1
Tf-pENO2	31, 43, 52, 288, 373, 610, 657, 1561, 2286	0778		
Tf-pENO3	31, 44, 549, 1527	0715 and 0200		
Tf-pENO4	137	0553		
Tf-Actin1	No	0755	Noda et al., 2012	AB468092[c]
Tf-Actin2	4	0736		
Tf-Actin3	4	0654		
Tf-Actin4	4	0639		
Tf-Actin5	4	No		
Tf-Actin6	4	663		
Tf-Actin7	4	No		
Tf-Actin8	4	No		
Tf-αTub1	No	No	Gerbod et al., 2004	AY277784
Tf-αTub2	No	No	Gerbod et al., 2004	AY277785
Tf-αTub3	9	0721 and part 0741		
Tf-αTub4	9	0679		
Tf-αTub5	9	0608		
Tf-αTub6	9	No		
Tf-αTub7	9	No		
Tf-βTub_1	21, 30, 200	No	Gerbod et al., 2004	AY277786
Tf-βTub_2	21	Similar to 0754	Gerbod et al., 2004	AY277787
Tf-βTub_3	21, 30, 200	Similar to 0758, 0284 and 0391		
Tf-βTub_4	21	Similar to 0494		
Tf-βTub_5	21	Similar to 0604		
Tf-βTub_6	21	Similar to 0747		
Tf-βTub_7	526	No		

[a] Only clusters with 5′ sequences are described.
[b] Huang et al. (2013).
[c] Partial sequence from *T. suis*.

paralogs. Structural gene duplications may for instance have occurred before speciation and probably subjected to gene conversion processes. In this sense, previous findings suggest that structural genes may have been expanding in other parabasalians (Noda et al., 2012). It would be interesting to know the mechanisms directing such processes since the control of each parasite affection might lie in this knowledge.

The number of metabolic gene duplications in *T. foetus* seems reduced when compared to the number of genes predicted in the *T. vaginalis* genome. Nevertheless several EST projects show that not all the *T. vaginalis* paralogs are simultaneously expressed (http://trichdb.org). Recent evidence supports that some copies of paralogous genes in the *T. vaginalis* genome are expressed under certain culture conditions while others stay silenced (Horváthová et al., 2012). If similar regulation mechanisms exist in *T. foetus*, it is reasonable to think that we have found few paralogs genes while other paralogs would be expressed under different conditions. In contrast with the findings concerning metabolic genes, *T. foetus* was found to express a high number of structural genes, nearly as many as those observed in the *T. vaginalis* genome. For instance, the appealing character in these species, as in many parabasalians, is the sophisticated motility system. Adding new copies of genes involved in the self-propelled motion might have represented a selective advantage by increasing gene dosage. This observation is in agreement with the three structural paralogs genes studied here and it was not found in such extend for metabolic genes as it was pointed out above. The identity of the protein product translated from each gene family strongly argues in favor of a high balancing selection and the need of maximal production.

An organism with capacity to live in the bulls and in the cows reproductive system is expected to have taken advantage of processes that wide protein diversity. Gene duplication is predicted to take place with a decrease in fitness, particularly if recombination is a frequent event (Ohta, 1987; Moitra and Dean, 2011). However despite the finding of genes apparently related to meiosis (Malik et al., 2008) there is no evidence in the scientific literature supporting *T. vaginalis*, *T. foetus* or other parabasalians undergo recombination or sexual division. Therefore, repeated duplications may thus accumulate without perceptible decrease in fitness. These parasites may have been endowed with a large battery of very similar genes that give raise to a quasi heterozygous state. The sequencing of the *T. foetus* genome, cDNA libraries in different culture conditions and further studies into the genome evolution will probably shed light onto survival and pathogenic mechanisms of these parasitic protozoans.

Acknowledgements

This study was supported by the ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica), CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) and the UNLPam (Universidad Nacional de La Pampa). We extend our thanks to Marie Oudot.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.vetpar.2014.09.024.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Carlton, J.M., Hirt, R.P., Silva, J.C., Delcher, A.L., Schatz, M., Zhao, Q., Wortman, J.R., Bidwell, S.L., Alsmark, U.C.M., Besteiro, S., Sicheritz-Ponten, T., Noel, C.J., Dacks, J.B., Foster, P.G., Simillion, C., Van de Peer, Y., Miranda-Saavedra, D., Barton, G.J., Westrop, G.D., Müller, S., Dessi, D., Fiori, P.L., Ren, Q., Paulsen, I., Zhang, H., Bastida-Corcuera, F.D., Simoes-Barbosa, A., Brown, M.T., Hayes, R.D., Mukherjee, M., Okumura, C.Y., Schneider, R., Smith, A.J., Vanacova, S., Villalvazo, M., Haas, B.J., Pertea, M., Feldblyum, T.V., Utterback, T.R., Shu, C.-L., Osoegawa, K., de Jong, P.J., Hrdy, I., Horvathova, L., Zubacova, Z., Dolezal, P., Malik, S.-B., Logsdon Jr., J.M., Henze, K., Gupta, A., Wang, C.C., Dunne, R.L., Upcroft, J.A., Upcroft, P., White, O., Salzberg, S.L., Tang, P., Chiu, C.-H., Lee, Y.-S., Embley, T.M., Coombs, G.H., Mottram, J.C., Tachezy, J., Fraser-Liggett, C.M., Johnson, P.J., 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science 315, 207–212.

Clark, B.L., Parsonson, I.M., Dufty, J.H., 1974. Experimental infection of bulls with *Tritrichomonas foetus*. Aust. Vet. J. 50, 189–191.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Felleisen, R.S., Lambelet, N., Bachmann, P., Nicolet, J., Müller, N., Gottstein, B., 1998. Detection of *Tritrichomonas foetus* by PCR and DNA enzyme immunoassay based on rRNA gene unit sequences. J. Clin. Microbiol. 36, 513–519.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14, 685–695.

Geer, L.Y., Domrachev, M., Lipman, D.J., Bryant, S.H., 2002. CDART: Protein homology by domain architecture. Genome Res. 12 (10), 1619–1623, http://dx.doi.org/10.1101/gr.278202.

Gerbod, D., Sanders, E., Moriya, S., Noël, C., Takasu, H., Fast, N.M., Delgado-Viscogliosi, P., Ohkuma, M., Kudo, T., Capron, M., Palmer, J.D., Keeling, P.J., Viscogliosi, E., 2004. Molecular phylogenies of Parabasalia inferred from four protein genes and comparison with rRNA trees. Mol. Phylogenet. Evol. 31, 572–580.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224.

Handl, J., Knowles, J., Kell, D.B., 2005. Computational cluster validation in post-genomic data analysis. Bioinformatics 21, 3201–3212, http://dx.doi.org/10.1093/bioinformatics/bti517.

Horváthová, L., Šafaríková, L., Basler, M., Hrdy, I., Campo, N.B., Shin, J.-W., Huang, K.-Y., Huang, P.J., Lin, R., Tang, P., Tachezy, J., 2012. Transcriptomic identification of iron-regulated and iron-independent gene copies within the heavily duplicated Trichomonas vaginalis genome. Genome. Biol. Evol. 4 (10), 1017–1029, http://dx.doi.org/10.1093/gbe/evs078.

Huang, K.-Y., Shin, J.-W., Huang, P.-J., Ku, F.-M., Lin, W.-C., Lin, R., Hsu, W.-M., Tang, P., 2013. Functional profiling of the *Tritrichomonas foetus* transcriptome and proteome. Mol. Biochem. Parasitol. 187, 60–71.

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., Ravikesavan, R., 2013. Gene duplication as a major force in evolution. J. Genet. 92 (1), 155–161.

Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M., Logsdon, J.M., 2008. An Expanded inventory of conserved meiotic genes provides evidence for sex in Trichomonas vaginalis. PloS One 3 (8), e2879, http://dx.doi.org/10.1371/journal.pone.0002879.

Moitra, K., Dean, M., 2011. Evolution of ABC transporters by gene duplication and their role in human disease. Biol. Chem. 392, 29–37.

Nadler, S.A., Honigberg, B.M., 1988. Genetic differentiation and biochemical polymorphism among trichomonads. J. Parasitol. 74, 797–804.

NCBI Resource Coordinators, 2014. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 42 (D1), D7–D17, http://dx.doi.org/10.1093/nar/gkt1146.

Noda, S., Mantini, C., Meloni, D., Inoue, J.-I., Kitade, O., Viscogliosi, E., Ohkuma, M., 2012. Molecular phylogeny and evolution of parabasalia with improved taxon sampling and new protein markers of actin and elongation factor-1α. PloS One 7 (1), e29938, http://dx.doi.org/10.1371/journal.pone.0029938.

O'Brien, J.L., Lauriano, C.M., Alderete, J.F., 1996. Molecular characterization of a third malic enzyme-like AP65 adhesin gene of *Trichomonas vaginalis*. Microb. Pathog. 20, 335–349.

Ohno, S., 1982. Evolution is condemned to rely upon variations of the same theme: the one ancestral sequence for genes and spacers. Perspect. Biol. Med. 25, 559–572.

Ohta, T., 1987. Simulating evolution by gene duplication. Genetics 115 (1), 207–213.

Oyhenart, J., Martínez, F., Ramírez, R., Fort, M., Breccia, J.D., 2013. Loop mediated isothermal amplification of 5.8S rDNA for specific detection of *Tritrichomonas foetus*. Vet. Parasitol. 193, 59–65, http://dx.doi.org/10.1016/j.vetpar.2012.11.034.

Parsonson, I.M., Clark, B.L., Dufty, J.H., 1976. Early pathogenesis and pathology of *Tritrichomonas foetus* infection in virgin heifers. J. Comp. Pathol. 86, 59–66.

Pereira-Neves, A., Benchimol, M., 2009. *Tritrichomonas foetus*: budding from multinucleated pseudocysts. Protist 160, 536–551.

Pereira-Neves, A., Campero, C.M., Martínez, A., Benchimol, M., 2011. Identification of *Tritrichomonas foetus* pseudocysts in fresh preputial secretion samples from bulls. Vet. Parasitol. 175, 1–8.

Pereira-Neves, A., Nascimento, L.F., Benchimol, M., 2012. Cytotoxic effects exerted by *Tritrichomonas foetus* pseudocysts. Protist 163, 529–543.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M., 1999. Mining SNPs from EST databases. Genome Res. 9, 167–174.

Pritham, E.J., Putliwala, T., Feschotte, C., 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390, 3–17.

Rae, D.O., 1989. Impact of trichomoniasis on the cow-calf producer's profitability. J. Am. Vet. Med. Assoc. 194, 771–775.

Slapeta, J., Müller, N., Stack, C.M., Walker, G., Lew-Tabor, A., Tachezy, J., Frey, C.F., 2012. Comparative analysis of *Tritrichomonas foetus* (Riedmüller, 1928) cat genotype, *T. foetus* (Riedmüller, 1928) cattle genotype and *Tritrichomonas suis* (Davaine, 1875) at 10 DNA loci. Int. J. Parasitol. 42, 1143–1149.

Markos, A., Miretsky, A., Müller, M., 1993. A glyceraldehyde-3-phosphate dehydrogenase with eubacterial features in the amitochondriate eukaryote, *Trichomonas vaginalis*. J. Mol. Evol. 37, 631–643.

Tibayrenc, M., Kjellberg, F., Ayala, F.J., 1990. A clonal theory of parasitic protozoa: the population structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and Trypanosoma and their medical and taxonomical consequences. Proc. Natl. Acad. Sci. U.S.A. 87, 2414–2418.

Viscogliosi, E., Müller, M., 1998. Phylogenetic relationships of the glycolytic enzyme, glyceraldehyde-3-phosphate dehydrogenase, from parabasalid flagellates. J. Mol. Evol. 47, 190–199.

Yuh, Y.S., Liu, J.Y., Shaio, M.F., 1997. Chromosome number of *Trichomonas vaginalis*. J. Parasitol. 83, 551–553.

Zubáčová, Z., Cimbůrek, Z., Tachezy, J., 2008. Comparative analysis of trichomonad genome sizes and karyotypes. Mol. Biochem. Parasitol. 161, 49–54.