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