GENE EXPANSION IN *TRICHOMONAS VAGINALIS*: A CASE STUDY ON TRANSMEMBRANE CYCLASES

JIKE CUI[1,2] TEMPLE F. SMITH[1]
jike@bu.edu tsmith@darwin.bu.edu

JOHN SAMUELSON[2]
jsamuels@bu.edu

[1] *Bioinformatics Program, Boston University, 44 Cummington St., Boston, MA, 02215*
[2] *Department of Molecular and Cell Biology, Boston University, 715 Albany St., Evans room 426, Boston, MA 02118*

The draft genome of *Trichomonas vaginalis* was recently published, but not much is known on why it has such a large genome. In part this size is due to many gene family expansions. For example we found over 100 members in the adenylyl cyclase family. About half are complete full length genes, and nearly half are initially confirmed to be pseudogenes, the remaining are either incomplete or the apparent result of assembly or sequencing problems. The family can be divided into two subgroups by sequence similarity. These can then be divided into functional and pseudo genes. Among all four of these sets the cyclase domain is very well conserved. We gave three possible hypotheses for that observation: a) Sequencing error or stop-codon read-through; b) Recency of duplication and mutation; c) The likelihood of functional pseudogene.

Keywords: *T. vaginalis*; cyclase; pseudogene; gene duplication.

1. Introduction

Trichomonas vaginalis is an anaerobic, parasitic flagellated protozoan. Infection with *T. vaginalis* is one of the most common sexually transmitted diseases for women with 8 million infections in North America and 180 million infections in the world each year [1-2]. The extracellular parasite resides in the urogenital tract of both sexes. *T. vaginalis* has a modified mitochondrion called the hydrogenosome, in which fermentation enzymes reside rather than enzymes of oxidative phosphorylation [3].

The recently published draft genome sequence of *T. vaginalis* by The Institute of Genomic Research (TIGR) reveals an abnormally large genome size of 160 Mb. ~60,000 protein-coding genes were identified, but only 65 were found to have introns. Two thirds of the genome consists of repeats and transposable elements [4].

Our examination of 12 animal pathogens, including fungi and protists, tells that their average genome size is ~20 Mb and ~6000 genes. These relatively small genomes provide efficiency in multiplication and infection. It is not clear why *T. vaginalis* possesses such a large genome, and how such massive gene expansion happened. The secretory pathway and signal transduction play important role in pathogenesis [28, 29]. Because cyclases are critical in eukaryotic signal transduction and have a unique structure discovered in our previous study, we propose a research project on a large family of putative transmembrane cyclase genes. And we hope the results can shed light on why and how the genome expansion happened.

2. Methods

2.1. *Identification of Cyclase Genes and Pseudogenes in T. vaginalis*

Cyclase domains were identified using NCBI Conservation Domain search tool [5]. The conserved cyclase domains and tblastn were used to search the *T. vaginalis* sequence scaffolds at the TIGR site using a cutoff E value of 1e-6 [6]. We identified 24 complete putative *T. vaginalis* transmembrane cyclases, which contained numerous transmembrane helices and a C-terminal cyclase. The length of these full-length transmembrane cyclases was ~1600 AAs (4,800 bps). In addition, we used blastx and those complete transmembrane cyclases to identify numerous other *T. vaginalis* genes encoding transmembrane cyclases, some of which were truncated or contained nonsense mutations (stop codons) or frame shifts.

2.2. *Initial Verification of Frameshift and Nonsense Mutation*

To verify each nonsense mutation, we took 30 AAs in the upstream and downstream of the stop site to make a query of 61 AAs. To verify frame shift, we took 100 bases of upstream and downstream of the end and start of each frame to make a query of 201 bases. Those queries were used to tblastn/blastn the trace from individual sequence reads of *T. vaginalis*. If there were 2 or more perfect matches, the mutation was considered real. This method actually found two assembly errors, which caused a frame shift.

2.3. *Softwares*

MUSCLE [7] and PIMA [8] are used for alignment. PAUP [9] and TREE-PUZZLE [10] were used for phylogenetic tree construction. PAML [13] was used for dN/dS analysis.

3. Results and Discussion

3.1. *Number and Topology of Cyclase Genes in T. vaginalis*

Cyclases play very important roles in eukaryotic signal transduction. The second messenger cAMP (adenosine 3’,5’-cyclic monophosphate) is synthesized by an adenylyl cyclase (AC) from ATP, and cGMP (guanosine 3’,5’-cyclic monophosphate) is made by a guanylyl cyclase (GC) from GTP [14]. Cyclase normally is not a big protein family, for example, we only found one cyclase in yeast. But we discovered that there are over 100 copies of the putative transmembrane cyclase genes in *T. vaginalis*. Unpublished data from our lab suggest the cyclase domains of these *T. vaginalis* transmembrane cyclases are adenylyl cyclase, and unpublished mRNA data suggest that these cyclases are expressed constitutively, that is, they are expressed at the same time.

Each full length cyclase has 12 to 16 transmembrane helixes (TMH) and a cyclase domain at the C-terminal. Some of them are truncated at the N terminal but all have at least one transmembrane helix, which means there is no cytosolic cyclase in *T. vaginalis*.

3.2. *How Did the Duplication Happen?*

Gene duplication may occur in homologous recombination, retrotransposition event, or duplication of an entire chromosome [15]. To find out more details about cyclase gene families' duplication event, we asked the following questions:

- Do the cyclase genes locate in subtelomere?
- Did the cyclase genes duplicate with other repetitive sequences?
- Were flanking sequences of the cyclase genes also duplicated?
- Were the cyclase genes duplicated by retrotransposition?

Genes in subtelomere usually have multiple copies. The cyclase family has ~100 copies, so it is reasonable to speculate that it might be in subtelomere. However Fig. 1 obviously does not support that. On average 83.7% of genes on a scaffold do not have a homolog in that scaffold, with SD of 15.7%. Although the percentage could be greatly affected by the length of scaffold, in scaffolds with more than 80 genes, the two numbers are 73.2% and 5.9%. It demonstrates that those scaffolds are not in subtelomere.

Sometimes genes could be carried by repeat elements and got duplicated when repeats move and copy themselves across the genome. We studied transposons, microsatellites, and virus-like repeat families in the upstream and downstream 5000 bases of each cyclase gene. While many repeat families were found, only one pair of genes has the same repeat at the same location. It seems cyclase genes were not duplicated with repeats.

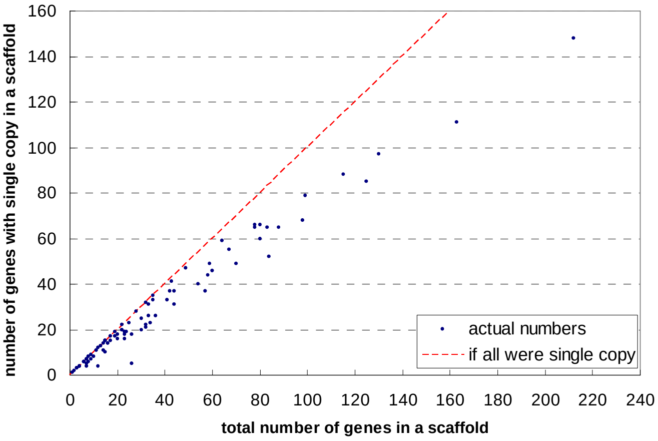

Fig. 1. Total number of genes and number of single copy genes in each of the 90 scaffolds which contains cyclase genes.

If a gene is duplicated in recombination event, its flanking sequence will probably be duplicated too. We searched ORFs in the 5000 bases region on two sides of each cyclase gene, but did not find any synteny, except some very repetitive proteins, like DNA polymerase and BspA-like surface antigen.

If cyclase genes were not duplicated with flanking sequence, they probably were copied with retrotransposons. Such a gene could have a poly-A present in its close downstream if the duplication happens very recently. And as discovered in *Entamoeba histolytica*, it may have a 9-20 bases repeat at the beginning and end of the copied region [16]. However no such evidence has been found so far among even the closest cyclase genes.

3.3. *Pseudogenes in Cyclase Genes of T. vaginalis*

Among the discovered cyclase genes, about half are complete genes ranging from 1449 to 1681 AAs. A dozen are either truncated genes that range from 340 to 1322 AAs, or incomplete genes due to the sequencing or assembly problems. The rest, which is nearly half, are pseudogenes that contain frameshifts and nonsense mutations. Fig. 2 illustrates the transmembrane cyclases in different life stages.

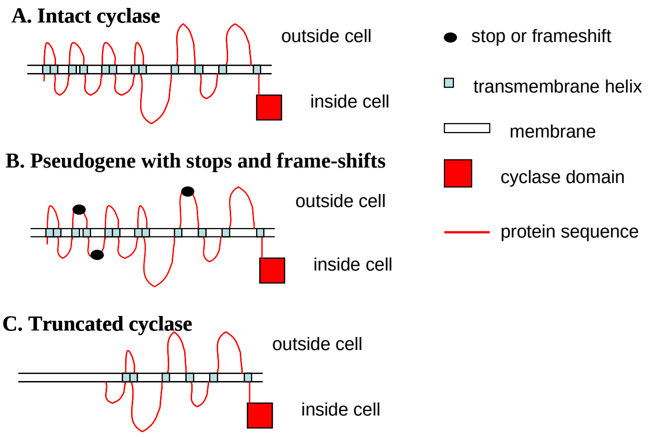

Fig. 2. Intact, pseudo, and truncated transmembrane cyclase of *T. vaginalis*.

If we refer the region before the cyclase domain as transmembrane region, an absolute majority of frameshifts and nonsense mutations happens in the transmembrane region, and very few are in the cyclase domain.

3.4. *Are Those Pseudogenes Really Pseudo?*

Alignment of the cyclases reveals that there are lots of variations in the transmembrane region, but the cyclase domain is highly conserved not only for those full length complete genes, but also in those pseudogenes. Phylogenetic analysis on the amino acid and nucleotide sequences shows that there is not a separation between pseudogenes and complete genes, and instead, they group together. Fig. 3 illustrates that grouping.

The above observation is unusual. Normally a pseudogene loses its function and is free from any selection constraint. The pseudogene then has a faster mutation clock than a functional conserved domain, and so they are not likely to group together. To study in more detail, the mutation on each codon position of the cyclase domain in 41 pairs of close full length functional genes, and 49 pairs of close functional and pseudogenes were compared. Table 1 lists the result.

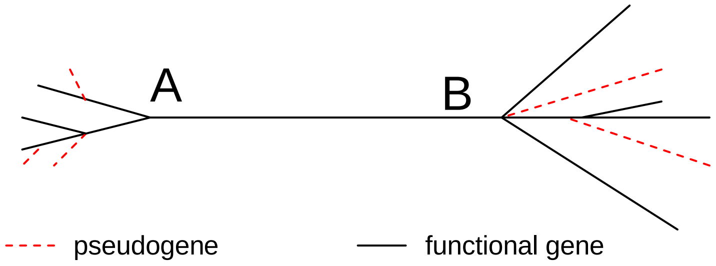

Fig. 3. A cartoon illustration of the phylogenetic grouping of the cyclase domain amino acid sequences in transmembrane cyclases using maximum likelihood.

Table 1. Number and percentage of mutation at each codon position

Pairs	statistics	1st position		2nd position		3rd position		sum
49 pairs of complete genes vs pseudogenes	average	35	20.5%	20	11.1%	99	67.1%	154
	StDev	17	6.0%	14	5.6%	24	11.2%	49
41 pairs of complete genes vs complete genes	average	42	22.1%	24	12.2%	106	64.2%	172
	StDev	23	5.7%	16	5.2%	20	10.4%	56

Surprisingly, the pseudogenes behave exactly like the functional genes. Most mutations happen at the third (wobble) codon position which is often silent, and the first codon position which is silent sometimes too. Further analysis on selection constraint using PAML program [13] gave the dN/dS ratio of 0.0436 for the functional genes, and 0.0438 for pseudogenes, suggesting that mutations in the cyclase domains of both complete genes and pseudogenes are highly selected against.

We offer the following hypotheses for the above observation,

1. Those frameshifts and nonsense mutations may be sequencing error, or they maybe real but somehow *T. vaginalis* has a mechanism to read through stop codon in translation, therefore these pseudogenes can still be expressed and be functional.
2. Those pseudogenes are real, but the frameshifts and nonsense mutations happened very recently, therefore there has not been ample time for mutations to happen in a larger scale yet.
3. Those pseudogenes cannot be expressed; however they are not treated as junk and are still being conserved. And they might serve some function in certain events.

3.5. *Evaluation of Hypothesis One and Two*

To check the sequencing error and the likelihood of *T. vaginalis* reading through nonsense mutations, PCR and sequencing of some of the critical mutation sites will be conducted using the G3 strain that is used in the TIGR sequencing project. Although most of the frameshifts and nonsense mutations have been verified by the sequencing trace file, there is not an absolute certainty until the experiment proves that. Stop-codon read-through in mRNA translation is rare, but it has been reported in yeast, human, and *E. coli* [17-20]. We will test its likelihood in *T. vaginalis* too.

Hypothesis two is in agreemenet with TIGR's conslusion that the genome expansion is recent due to the low polymorphism within high repeat protein families, and evidence of repeat expansion after *T. vaginalis* and *T. tenax* diverged [4]. However it can not explain a) Why very few frameshifts and nonsense mutations happened in the cyclase domain. b) The cyclase domains of A have much greater conservation than those of B, as shown in Fig. 2, while their TM regions have similar degree of sequence variation. Actually the TM region of group A has on average more frameshifts and stops than that of B. So the recency of the duplication in cyclase family does not seem to explain everything here.

3.6. *Pseudogenes in the Whole Genome -- Evaluation of Hypothesis Three*

Most pseudogenes come from duplicated genes because its pseudogenization is less likely to be deleterious than a singleton. Human has about 30,000 genes with 38% of duplicated genes [21], and 12,000 pseudogenes have been identified [22]. TIGR predicted that there are about 60,000 genes in *T. vaginalis* but did not mention pseudogenes. We speculate that a significant portion of the 60,000 genes might be pseudogenes. All of the pseudo and incomplete cyclase genes are included in TIGR predicted genes which start after the last frameshift or nonsense mutation. Although the amount of pseudogenes in other large gene families can not be estimated until a similar survey is conducted, our search of nonsense mutation in the whole genome has found about 3000 pseudogenes with nonsense mutation, and there could be similar or greater number of pseudogenes with frameshift. Large number of pseudogenes are present in the family of ankyrin repeat proteins, hypothetical protein, conserved hypothetical protein, adenylate cyclase, vsaA, surface antigen BspA, ANK-repeat protein, CG1651-PD-related, Dentin sialophosphoprotein precursor, ABC transporter protein, kinases, major facilitator superfamily protein, leucine rich repeat family protein, and Transmembrane amino acid transporter protein. Many of those families are involved in secretary pathway and signal transduction system, which play important role in pathogenesis.

Hartl suggests that the rate of deleting junk DNA decides genome size, and a low rate would accumulate many pseudogenes, longer introns, and intergenic regions in a genome [23-25]. However it is different in *T. vaginalis*, where there might be many pseudogenes but very little introns. Is it possible that those pseudogenes are not real junk and do serve a purpose in certain situation? That is not impossible! It was discovered that pseudogenes could act as the supplier of certain variable region to immunoglobulin gene in chicken [26], and that some pseudogenes conserve their sequences and can still be revived and functional in cow and human [27], but the mechanism is not totally understood yet.

4. Conclusion and Future Work

T. vaginalis has an enormously large genome, and many protein families underwent massive duplication. We proposed a research project to study transmembrane cyclase in *T. vaginalis*, a very important protein in cell signal transduction. Our initial results show that there are over 100 copies of genes in this family, about half are full length genes, and nearly half are pseudogenes that are initially confirmed by the sequencing trace files. The family can be roughly separated into two groups by sequence similarity based on their cyclase domains, which is very well conserved in both complete genes and surprisingly pseudogenes too.

The conservation of cyclase domain in pseudogenes is not expected. We suggested three possible reasons: a) sequencing error and stop-codon read-through which will be tested by our experiments; b) recency of duplication and mutation. It is likely to be true but can not explain some discrepancies between the TM region and cyclase domain, and between the group A and group B. c) The likelihood of functional pseudogene which is possible but does not have any evidence until we see more experimental support.

We also tried to look into how cyclase genes were duplicated. We found that they are not in subtelomere, and there is no evidence to support that they might be copied with repeats, retrotransposons, or flanking regions.

We hope that after a larger survey on other duplicated protein families and having more experimental data on the pseudogenes, we could shed light on why *T. vaginalis* possesses such a huge genome, how genes are duplicated, the quantity of its pseudogenes, and their evolution histories.

References

[1] http://www.cdc.gov/ncidod/dpd/parasites/trichomonas/factsht_trichomonas.htm

[2] Hook, E., Trichomonas vaginalis--no longer a minor STD, *Sex. Transm. Dis.*, 26(7):388-389, 1999.

[3] Upcroft, P. and Upcroft, J., Drug targets and mechanisms of resistance in the anaerobic protozoa, *Clin. Microbiol. Rev.*, 14(1):150-164, 2001.

[4] Carlton J.M., *et al.*, Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis, *Science*, 315(5809):207-212, 2007.

[5] http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

[6] http://www.tigr.org/tdb/e2k1/tvg/

[7] Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, 32(5):1792-1797, 2004.

[8] Smith, R.F. and Smith, T.F., Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling, *Protein Eng.*, 5(1):35-41, 1992.

[9] Paup: Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer Associates, Sunderland, Massachusetts, 2002.

[10] Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A.., TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing, Bioinformatics., 18(3):502-504, 2002.

[11] Sawyer, S., Statistical tests for detecting gene conversion, *Mol. Biol. Evol.*, 6(5):526-538, 1989.

[12] Martin, D.P., Williamson, C., and Posada, D., RDP2: recombination detection and analysis from sequence alignments, *Bioinformatics*, 21(2):260-262, 2005.

[13] http://abacus.gene.ucl.ac.uk/software/paml.html

[14] Roelofs, J., Smith, J. L., and Van Haastert, P. J. M., cGMP signalling: different ways to create a pathway, *TRENDS Genet.,* 19(3):132-134, 2003.

[15] Zhang, J., Evolution by gene duplication: an update, *Trends in Ecology & Evolutio,* 18(6):292-298, 2003.

[16] Van Dellen, K., Field, J., Wang, Z., Loftus, B., and Samuelson, J., LINEs and SINE-like elements of the protist Entamoeba histolytica, *Gene*., 297(1-2):229-239, 2002.

[17] Williams, I., Richardson, J., Starkey, A., and Stansfield, I., Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae, *Nucleic Acids Res.*, 32(22):6605-6616, 2004.

[18] Namy, O., Duchateau-Nguyen, G., Hatin, I., Hermann-Le Denmat, S., Termier, M., and Rousset, J.P., Identification of stop codon readthrough genes in Saccharomyces cerevisiae, *Nucleic Acids Res.*, 31(9):2289-2296, 2003.

[19] Lai, C.H., Chun, H.H., Nahas, S.A., Mitui, M., Gamo, K.M., Du, L., and Gatti, R.A., Correction of ATM gene function by aminoglycoside-induced read-through of premature termination codons, *Proc. Natl. Acad. Sci. USA*, 101(44):15676-15681, 2004.

[20] Engelberg-Kulka, H. and Schoulaker-Schwarz, R., Stop is not the end: physiological implications of translational readthrough, *J. Theor. Biol.*, 131(4):477-485, 1988.

[21] Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A., Evolutionary analyses of the human genome, *Nature*, 409(6822):847-849, 2001.

[22] Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M., Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome, *Genome Res.*, 13(12):2541-2558, 2003.

[23] Hartl, D.L. and Wirth, D.F., Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of Plasmodium falciparum, *Mol. Biochem. Parasitol.*, 150(1):10-24, 2006.

[24] Petrov, D.A. and Hartl, D.L., High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups, *Mol. Biol. Evol.*, 15(3):293-302, 1998.

[25] Lozovskaya, E.R., Nurminsky, D.I., Petrov, D.A., and Hartl, D.L., Genome size as a mutation-selection-drift process, *Genes. Genet. Syst.*, 74(5):201-207, 1999.

[26] Ota, T. and Nei, M., Evolution of immunoglobulin VH pseudogenes in chickens, *Mol. Biol. Evol.*, 12(1):94–102, 1995.

[27] Kleineidam, R.G., Jekel, P.A., Beintema, J.J., and Situmorang, P., Seminal-type ribonuclease genes in ruminants, sequence conservation without protein expression?, *Gene.*, 231(1-2):147-153, 1999.

[28] Lopez, L.B., Braga, M.B., Lopez, J.O., Arroyo, R., and Costa e Silva Filho, F., Strategies by which some pathogenic trichomonads integrate diverse signals in the decision-making process, *An. Acad. Bras. Cienc.*, 72(2):173-186, 2000.

[29] Kucknoor, A.S., Mundodi, V., and Alderete, J.F., The Proteins Secreted by Trichomonas vaginalis and Vaginal Epithelial Cell Response to Secreted and Episomally-Expressed AP65, *Cell. Microbiol.*, 2007.