Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

**Article**

# Genome-Wide Survey and Evolutionary Analysis of Trypsin Proteases in Apicomplexan Parasites

Aylan Farid Arenas[#], Juan Felipe Osorio-Méndez[#], Andres Julian Gutierrez, and Jorge E. Gomez-Marin[*]

*Grupo de Parasitología Molecular (GEPAMOL), Centro de Investigaciones Biomédicas, Universidad del Quindío, Armenia, Colombia.*

Genomics Proteomics Bioinformatics 2010 Jun; 8(2): 103-112. DOI: 10.1016/S1672-0229(10)60011-3

## Abstract

Apicomplexa are an extremely diverse group of unicellular organisms that infect humans and other animals. Despite the great advances in combating infectious diseases over the past century, these parasites still have a tremendous social and economic burden on human societies, particularly in tropical and subtropical regions of the world. Proteases from apicomplexa have been characterized at the molecular and cellular levels, and central roles have been proposed for proteases in diverse processes. In this work, 16 new genes encoding for trypsin proteases are identified in 8 apicomplexan genomes by a genome-wide survey. Phylogenetic analysis suggests that these genes were gained through both intracellular gene transfer and vertical gene transfer. Identification, characterization and understanding of the evolutionary origin of protease-mediated processes are crucial to increase the knowledge and improve the strategies for the development of novel chemotherapeutic agents and vaccines.

**Key words**: Apicomplexa, trypsin, endosymbiotic gene transfer

## Introduction

Apicomplexa are a diverse group of protozoan parasites with an intracellular lifestyle. Members of this group include the causal agents of diseases such as toxoplasmosis (*Toxoplasma gondii*), malaria (*Plasmodium*), and cryptosporidiosis (*Cryptosporidium*) (*1, 2*). They invade their host cells through a parasite-mediated process in which several proteins secreted by three specialized organelles (named micronemes, rhoptries, and dense granules) mediate the attachment of the parasite to its host cell, modificate the host cell's physiology and mediate the formation of a parasitophorous vacuole where the parasite resides and replicates. Proteases belonging to different families are delivered by these secretory organelles and have been implicated in processes such as invasion and egress (*3*). For instance, subtilisin serine proteases have a role in the activation of other secretory proteins by the processing of immature proproteins, and rhomboid proteases cleave parasite adhesins to deliver the parasite into the luminal space of the parasitophorous vacuole (*4*). For this reason, proteases are considered potential targets for the development of new therapies against this group of parasites.

Trypsin is one of the largest serine protease sub-

[#] Equal contribution.
*Corresponding author.
E-mail: gepamol2@uniquindio.edu.co

families [named S1 according to MEROPS database classification of proteases (*5*)] in both prokaryotic and eukaryotic organisms and performs a wide range of functions. Chemical modification experiments and multiple sequence alignments have shown that several trypsins have three non-contiguous highly conserved amino acids, which are the active residues of this protease family: a histidine, an aspartic acid, and a serine flanked by the structurally important motif GD**S**GG (*6-8*). The majority of the proteases of this family enter the secretory pathway by an N-terminal signal peptide, where they can function as digestive proteins, regulate blood coagulation, act as growth cell factors, and activate mast cells (*9*). Despite that these groups of enzymes are found in eukaryotic as well as in prokaryotic organisms, there are no reports for genes encoding trypsin proteases in apicomplexan genomes (*4*). However, there is evidence suggesting that *Plasmodium falciparum* (the parasite responsible for the most serious human malaria) secretes trypsins; these trypsins are important in the invasion process because they apparently cleave surface proteins of the host erythrocytes (*10*).

Horizontal gene transfer (HGT) is a system that differs from the gene transfer in vertical inheritance of the DNA between parents and their offsprings, because in this process the genetic material is transferred between different evolutionary lineages. HGT is a common mechanism for the acquisition of genes encoding new functions in prokaryotes. In contrast, there are few cases of HGT involving eukaryotes in the literature. However, endosymbiotic relicts such as the chloroplast in plants, the mitochondria in eukaryotes, and the apicoplast in Apicomplexa are considered to be important sources of new genes because they transfer a part of their genetic material to the nuclear genome of the host organism by a process named intracellular gene transfer (IGT) (*11*).

Identification and understanding of the evolution of trypsin proteases in the apicomplexan lineage would help unravel their potential roles in various cellular processes, including their pathogenesis. In this work, 16 new genes encoding for trypsin proteases are identified in 8 apicomplexan genomes. Phylogenetic analysis suggests that these genes were gained through both IGT and vertical gene transfer.

# Results

## Identification of apicomplexan trypsin homolog sequences

Using the keyword "trypsin" on the Swiss-Prot/TrEMBL database through the ExPASy proteomics server, the prosite documentation PS00135 of serine proteases belonging to the trypsin family was obtained with a list of 5,743 sequences having the characteristic signature of this protein family. BLASTP searches were performed on the *T. gondii* (strain ME49) genome sequence by using a dataset of 50 bacterial trypsin proteases extracted from this list as queries. These searches retrieved three putative trypsin proteins (TGME49_062920, TGME49_077850, and TGME49_090840) that were used as queries to retrieve trypsins from a wide range of organisms including other apicomplexa. A hidden Markov model (HMM) profile from a multiple sequence alignment of trypsin domains extracted from the 134 BLASTP hits was used to perform a more sensitive search in 10 apicomplexan genomes. A total of 16 trypsin proteases were identified in 8 apicomplexan genomes for which the complete genomic data are available (**Table 1**). No hits were retrieved from *Cryptosporium* and *Babesia* genomes.

There is a low representation of trypsin genes in different apicomplexan genomes compared with the higher number of copies of trypsin genes present in other eukaryotic organisms such as *Drosophila* and plants (*12, 13*). The highest representation of trypsins is in the *T. gondii* and *Neospora caninum* genomes, which share a set of four syntenic genes (Table 1). Two species of *Plasmodium*, *P. falciparum* and *P. vivax,* have two genes encoding trypsin proteases. In *P. knowlesi*, *P. yoelii*, and *Theileria*, just one copy of trypsin per genome sequence was found, respectively.

## Analysis of the trypsin sequences

A multiple sequence alignment of the 16 sequences homologous to the trypsin proteases found in the apicomplexan genomes (**Figure 1**) showed that 4 sequences (NC_LIV_051240, MAL8P1.98, PY01797, and XP_765845.1) have no histidine (H) of the trypsin catalytic triad and therefore are possibly proteolytically inactive proteins or pseudogenes. The

**Table 1 Main characteristics of the 16 trypsin genes identified in 8 apicomplexan genomes**

| Gene | Species | Accession number | ESTs | Upstream promoter regions | Exons | Gene size (bp) | ORF size (bp) | Genome localization |
|---|---|---|---|---|---|---|---|---|
| | | *Toxoplasma* (Type I) | | | | | | |
| 1 | Tg | TGME49_062920 | 48 | YES | 11 | 9,503 | 2,871 | VIIb |
| 2 | Tg | TGME49_077850 | 4 | YES | 11 | 7,933 | 2,235 | XII |
| 3 | Tg | TGME49_090840 | 9 | YES | 9 | 6,837 | 2,883 | IX |
| 4 | Tg | TGME49_118290 | 3 | YES | 8 | 5,474 | 972 | IV |
| | | *Toxoplasma* (Type II) | | | | | | |
| 1 | Tg | TGGT1_007070 | 48 | YES | 11 | 9,503 | 2,871 | VIIb |
| 2 | Tg | TGGT1_104800 | 4 | YES | 11 | 7,933 | 2,235 | XII |
| 3 | Tg | TGGT1_032310 | 9 | YES | 9 | 6,837 | 2,883 | IX |
| 4 | Tg | TGGT1_122960 | 3 | YES | 8 | 5,474 | 972 | IV |
| | | *Toxoplasma* (Type III) | | | | | | |
| 1 | Tg | TGVEG_071800 | 48 | YES | 11 | 9,503 | 2,871 | VIIb |
| 2 | Tg | TGVEG_028600 | 4 | YES | 10 | 7,933 | 2,124 | XII |
| 3 | Tg | TGVEG_084140 | 9 | YES | 9 | 6,837 | 2,880 | IX |
| 4 | Tg | TGVEG_009710 | 3 | YES | 7 | 5,474 | 1,089 | IV |
| | | *Plasmodium* | | | | | | |
| 5 | Pf | Pf_MAL8P1.126 | 1 | – | 5 | 3,139 | 2,613 | MAL 8 |
| 6 | Pf | Pf_MAL8P1.98 | 1 | – | 11 | 2,489 | 1,125 | MAL 8 |
| 7 | Pv | PVX_088155 | 0 | – | 5 | 2,913 | 2,430 | CM000442 |
| 8 | Pv | PVX_123160 | 5 | – | 14 | 5,097 | 2,112 | CM000455 |
| 9 | Pk | PKH_142640 | 0 | – | 15 | 5,470 | 2,112 | 14 |
| 10 | Py | PY01797 | 1 | – | 9 | 2,764 | 1,359 | MALPY00485 |
| | | *Neospora* | | | | | | |
| 11 | Nc | NC_LIV_090910 | 0 | – | 11 | 8,011 | 3,138 | VIIb |
| 12 | Nc | NC_LIV_145670 | 2 | – | 10 | 6,077 | 2,193 | XII |
| 13 | Nc | NC_LIV_113280 | 5 | – | 17 | 13,649 | 2,460 | IX |
| 14 | Nc | NC_LIV_051240 | 1 | – | 15 | 9,202 | 1,590 | IV |
| | | *Theileria* | | | | | | |
| 15 | Ta | XP_954412.1 | 1 | – | 9 | 2,108 | 1,731 | – |
| 16 | Tp | XP_765845.1 | 0 | – | – | – | – | – |

Note: The accession numbers, the number of EST alignments per gene, the prediction of upstream promoter regions [based on chip-chip data for *T. gondii* (*48*)], the exon counts, the gene and ORF sizes and the genome localization of the 16 putative trypsin protein sequences are indicated. Information of *T. gondii* and *N. caninum* trypsin genes is based on ToxoDB 5.1, and for *Plasmodium* is based on PlasmoDB 5.5. Analysis of *Theileria* genes were made by using sequences retrieved from the Welcome Trust Sanger Institute (http://www.sanger.ac.uk). Tg: *Toxoplasma gondii*; Pf: *Plasmodium falciparum*; Pv: *P. vivax*; Pk: *P. knowlesi*; Py: *P. yoelii*; Nc: *Neospora caninum*; Ta: *Theileria annulata*; Tp: *Theileria parva*.

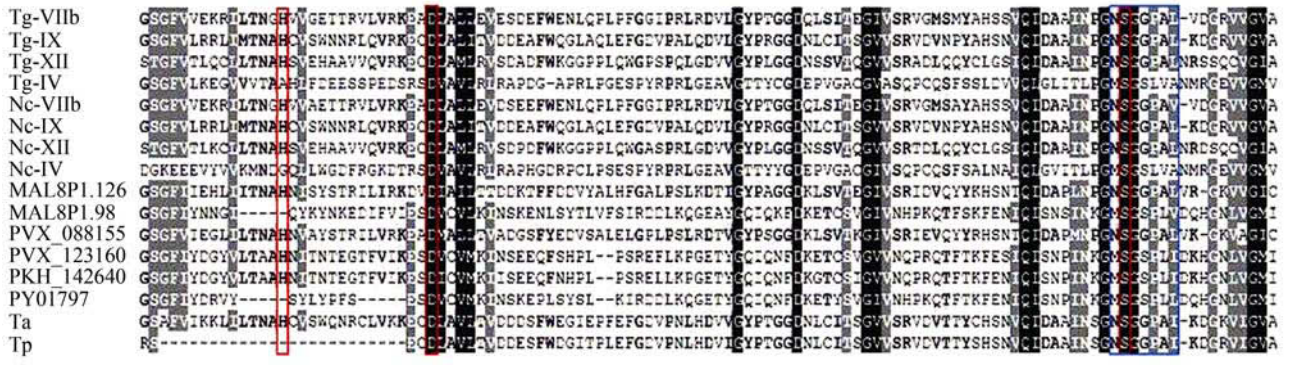


**Figure 1** Multiple alignment of the 16 trypsin domain sequences identified in 8 apicomplexan genomes. Amino acids belonging to the trypsin catalytic triad are indicated by red boxes. The characteristic trypsin motif GNSGGPAL is indicated by the blue box. Alignment was edited by using GeneDoc program through deleting gap columns. Conserved positions are shaded as black and semi-conserved positions are shaded as grey.

other 12 putative trypsins are probably active proteases encoded by transcriptionally active genes because they have all catalytic residues including the conserved motif GN**S**GGPAL flanking the catalytic serine. Three facts provide strong evidence for the expression of trypsins in apicomplexa (Table 1): (1) No premature stop codons interrupting the open reading frames encoding trypsin amino acid sequences were identified; (2) There are ESTs that support the transcription of most genes; (3) Chip-chip data from *T. gondii* predict upstream transcriptional promoter regions for all *T. gondii* trypsin genes.

When the exon-intron structure of the *T. gondii* trypsin genes is compared among representative strains of the three *T. gondii* clonal lineages (strains ME49, GT1, and VEG), only the trypsin genes located on the chromosomes VIIb and IX have the same exon counts. However, when the comparison is made between *T. gondii* and *N. caninum* genes, only the exon-intron structure of the gene located on chromosome VIIb is interspecifically conserved (Table 1). It is important to note that despite the high synteny level of *T. gondii* and *N. caninum* genomes, the *T. gondii* genes located on chromosomes IX and IV had lost 7 to 8 exons with respect to their *N. caninum* orthologs. Interspecific comparison of exon-intron structure of *Plasmodium* trypsin genes suggests that *P. falciparum* and *P. vivax* have two pairs of ortholog trypsin genes based on their similar exon-intron structures. Only one trypsin gene per genome sequence was found in *P. knowlesi*, *P. yoelii*, and *Theileria*.

Protein subcellular prediction and the presence of coexisting domains are important characteristics to infer the potential role of a protein. Prediction for secretion and for N-terminal transit peptides to the mitochondria and the chloroplast showed that six apicomplexan trypsins are predicted as secretory proteins, four are predicted as mitochondrial proteins, and three have a putative N-terminal transit peptide to the chloroplast (**Table 2**). BLASTP searches by using *T. gondii* trypsin sequences as queries in the GenBank and in the protease database MEROPS retrieved trypsins from bacteria, plants and metazoans. The majority of these sequences are also predicted as both secretory and organellar proteases. Analysis of accessory domains present in trypsins from this wide phylogenetic spectrum showed that most trypsins are commonly present with other protein domains, generally one or two copies of PDZ repeats. The 16 apicomplexan trypsins have 4 domain architectures: 6 without accessory domains, 7 with a C-terminal PDZ repeat, 1 with an N-terminal CS domain and a C-terminal PDZ repeat, and 2 with a C-terminal PDZ repeat and an MMR_HSR1 domain (**Table 3**). The

**Table 2 Prediction of secretion and subcellular localization of *T. gondii* trypsin homologous sequences**

| Prokaryotic trypsins | Secretory | No secretory | | |
|---|---|---|---|---|
| Cyanobacteria (10) | 8 | 2 | | |
| Clostridia (8) | 7 | 1 | | |
| Alpha-proteobacteria (7) | 6 | 1 | | |
| Beta-proteobacteria (11) | 11 | 0 | | |
| Delta-proteobacteria (6) | 5 | 1 | | |
| Gamma-proteobacteria (4) | 4 | 0 | | |
| Proteobacteria (1) | 1 | 0 | | |
| Bacilli (10) | 10 | 0 | | |
| Spirochetes (3) | 1 | 2 | | |
| Chlamydiae (1) | 1 | 0 | | |
| Actinobacteria (3) | 1 | 2 | | |
| Deinococcus (4) | 4 | 0 | | |
| Thermotogae (2) | 1 | 1 | | |
| Dictyoglomi (2) | 0 | 2 | | |
| Chlorobi (4) | 4 | 0 | | |
| Deferribacteres (1) | 1 | 0 | | |
| Eukaryotic trypsins | Secretory | Chloroplast | Mitochondria | Other |
| Apicomplexa (16) | 6 | 3 | 4 | 3 |
| Plants (3) | 0 | 1 | 1 | 1 |
| Metazoa (38) | 12 | - | 13 | 8 |

Note: The number of retrieved sequences for each taxonomic group is indicated in parenthesis. Prediction of secretion for prokaryotic sequences is based on either SignalP 3.0 (*46*) or SecretomeP 2.0 (*47*) and for eukaryotic sequences is based on either SignalP 3.0 (*46*) or TargetP 1.1 (*45*). Transit peptides for mitochondria and chloroplast were predicted by TargetP 1.1 (*45*).

**Table 3 Domain architecture of trypsins**

| Domain architecture | Archaea | Bacteria | Apicomplexa | Other eukaryotes |
|---|---|---|---|---|
| Tryp | 0 | 3 | 6 | 3 |
| Tryp - PDZ | 3 | 42 | 7 | 22 |
| Tryp - PDZ - PDZ | 0 | 32 | 0 | 0 |
| CS - Tryp - PDZ | 0 | 0 | 1 | 0 |
| KAZAL - Tryp - PDZ | 0 | 0 | 0 | 2 |
| IB - KAZAL - Tryp - PDZ | 0 | 0 | 0 | 13 |
| 2-Hacid_dh_C - Tryp | 0 | 0 | 0 | 1 |
| Tryp - PDZ - MMR_HSR1 | 0 | 0 | 2 | 0 |
| Domain architecture | Secretory | Mitochondria | Chloroplast | Other |
| Tryp | 2 | 3 | 2 | 5 |
| Tryp - PDZ | 39 | 10 | 1 | 24 |
| Tryp - PDZ - PDZ | 30 | 0 | 0 | 2 |
| CS - Tryp - PDZ | 0 | 0 | 0 | 1 |
| KAZAL - Tryp - PDZ | 0 | 2 | 0 | 0 |
| IB - KAZAL - Tryp - PDZ | 12 | 0 | 0 | 1 |
| 2-Hacid_dh_C - Tryp | 0 | 0 | 0 | 1 |
| Tryp - PDZ - MMR_HSR1 | 0 | 2 | 0 | 0 |

Note: Protein domains were defined as in the SMART database (*40*). The classification of the sequences is based on the NCBI taxonomy database. Prediction of secretion for prokaryotic sequences is based on either SignalP 3.0 (*46*) or SecretomeP 2.0 (*47*) and for eukaryotic sequences is based on either SignalP 3.0 (*46*) or TargetP 1.1 (*45*). Transit peptides for mitochondria and chloroplast were predicted by TargetP 1.1 (*45*).

last two domain architectures are specific to the apicomplexan trypsins. A possible relationship between protein localization and its function was also analyzed by counting the predicted subcellular localization for all the domain architectures. Trypsins without additional domains are equally predicted as secretory, organellar or with other protein localization. In contrast, trypsins with the accessory domains PDZ, IB, KAZAL and MMR_HSR1 are mainly predicted as secretory or mitochondrial proteins (Table 3).

### Phylogenetic analysis

High similarity BLAST scores showed that the apicomplexan trypsin genes have high homology with cyanobacterial trypsins, especially with *Arabidopsis thaliana* and *Oryza sativa* DegP proteases. These results suggest that apicomplexan trypsins were gained by HGT from bacterial trypsins or by IGT from an endosymbiotic relict. In order to test whether apicomplexan trypsins were gained by HGT or IGT, we used different phylogenetic methods. The topologies of the trees showed two well supported clusters by bootstrap values containing apicomplexan trypsin sequences, indicating two different evolutionary origins for these genes (**Figure 2**). The first group of apicomplexan trypsins possibly evolved from IGT from a green or red algae endosymbiotic relict (possibly the apicoplast or the mithocondrion), because apicomplexan sequences are clustered with plant, metazoan and cyanobacterial trypsins. In this cluster, a gene duplication event in the common ancestor of the closely related species *T. gondii* and *N. caninum* was identified. The second cluster of apicomplexan trypsins seems to be evolved before the above-mentioned IGT, because these sequences do not have a close evolutionary relationship with sequences of any other lineages.

## Discussion

In most eukaryotic taxonomic groups, serine protease superfamily has undergone repeated cycles of duplication and divergence to perform a wide spectrum of physiological activities (*14, 15*). Multiple copies of trypsin genes have been found in animals (*16-23*) and plants. For instance, clusters of tandemly repeated trypsin genes are present in both human and *Drosophila melanogaster* genome sequences (*16, 24*). In

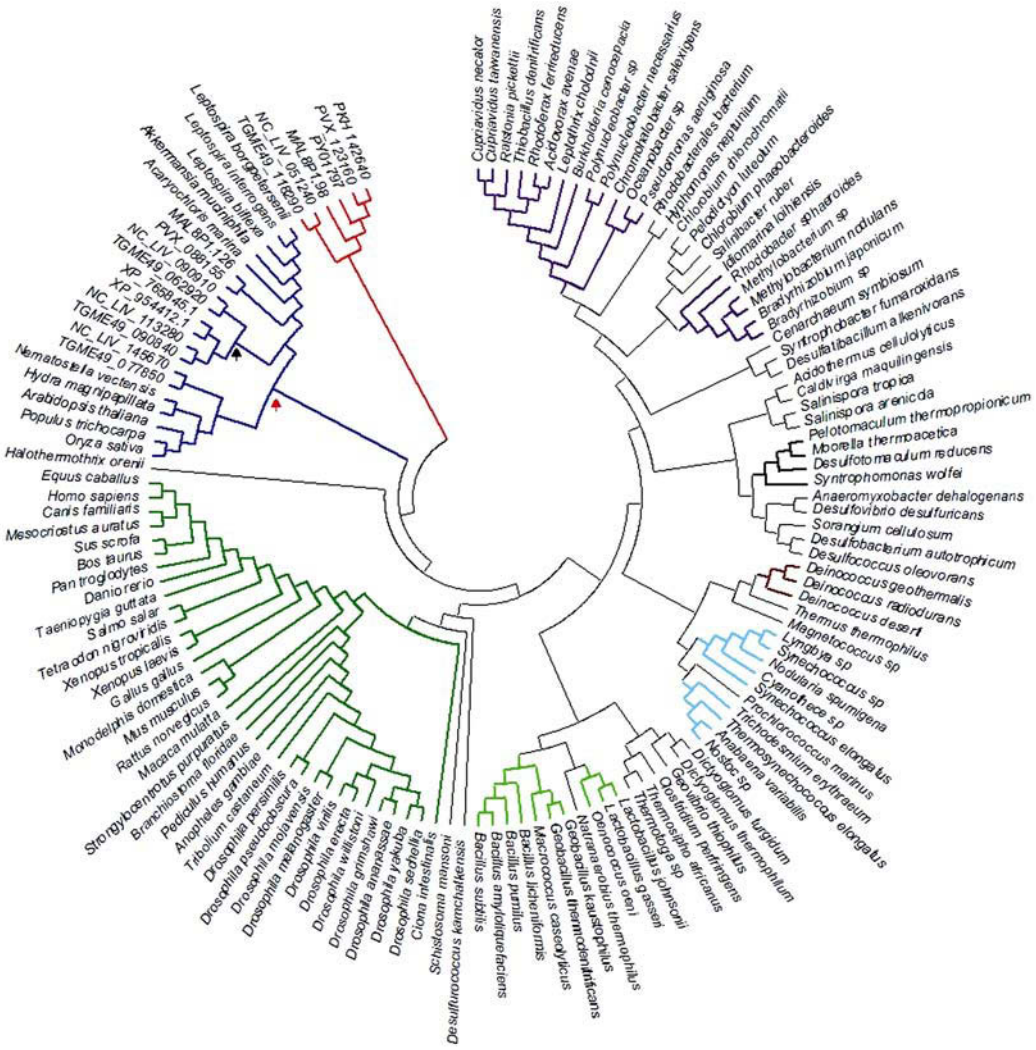


**Figure 2** Bootstrap consensus tree inferred using the neighbor-joining method. Confidence values were assessed from 5,000 replicates. All positions containing gaps and missing data were eliminated from the dataset. Phylogenetic analyses were conducted in MEGA4 software (*44*). Branches supported by significant bootstrap values (>50) are highlighted by using the following color code. Light blue: Cyanobacteria; Brown: Deinococcus; Grey: Chloribium; Violet: Proteobacteria; Green: Metazoa; Light green: Bacilli; Black: Clostridia; Blue: Apicomplexan trypsin from IGT (Cluster 1 in the text); Red: Other apicomplexan trypsins (Cluster 2 in the text). The black arrow head indicates a possible segmental duplication event, and the red arrow head indicates the IGT event from an endosymbiont to the nuclear genome of an apicomplexan ancestor. The apicomplexan trypsins accession numbers (Table 1) are used as operational taxonomic units.

plants, there are reports of several trypsin genes located in different chromosomes (*13*). These examples suggest that both local duplication events by unequal crossing-over and segmental duplication have contributed to the expansion of this gene family in the eukaryotic genomes. Apicomplexan ancestral trypsin genes may have undergone a single segmental duplication event after a possible IGT in the *Toxoplasma-Neospora* more recent common ancestor (Figure 2). Therefore, both local and segmental duplications may have no significant contribution to the expansion of trypsins in the apicomplexan parasites.

A protein domain could be defined as an independent evolutionary unit that may either occur in single domain proteins or with other units in multi-domain proteins. A domain might have an independent function or possibly will contribute to the function of a protein in cooperation with other domains (*25*, *26*). Single and multiple PDZ domains in trypsin proteases have been found in plants, bacteria,

yeast and mammals (*27, 28*). PDZ is one of the most common protein-protein interaction domains and plays a major role in the assembly of the multimeric protein complexes. This functional role of PDZ domains is facilitated by their ability to recognize and bind C-terminal short specific motifs of the target proteins to internal peptide sequences, which enables them to recognize and bind to diverse ligands. They may have modulated the function and localization of associated proteins such as the trypsins by mediating the protease binding to its substrate (*29-31*).

With respect to the origin of trypsin domains with C-terminal PDZ repeats, the identification of this architecture in bacteria, metazoans, plants and apicomplexa indicates either that it arose prior to the divergence of bacteria and eukaryotes or that HGT led to the spread of this architecture through both prokaryotic and eukaryotic lineages. Despite of the similarity in domain architectures of many trypsin proteases, there are diverse subcellular localizations and unique domain combinations for some prokaryotic and eukaryotic taxonomic groups. For instance, two PDZ domain repetitions are present in various bacterial groups but not in the analyzed eukaryotic sequences. Similarly, there are combinations of the eukaryotic trypsins with other protein domains that are not present in the prokaryotic trypsins. For instance, KAZAL and MMR_HSR1 domains are present just in metazoan and in apicomplexan trypsins, respectively. Trypsins are also predicted as secretory, mitochondrial or chloroplastic proteins. These diverse domain architectures and localizations suggest a wide range of roles for trypsin proteases in both eukaryotes and prokaryotes.

Apicomplexan parasites possess a non-photosynthetic plastid of secondary endosymbiotic origin called the apicoplast (*32*). The evolutionary origin of the apicoplast is still being debated, with either a green or red alga positioned as a donor lineage (*33, 34*). The apparently common occurrence of plastid secondary endosymbiosis throughout the eukaryotic organisms provides widespread potential for IGT to the eukaryotic nuclear genome (*35*). It is reported that apicomplexan genes are closely related with plant, algae, and bacterial genes. These unexpected relationships suggest that IGT events from the endosymbiots that originated the apicoplast and the mithocondrion were important sources of new genes in the apicomplexan ancestor (*11, 36*).

Significant BLAST similarity scores of sequences of one lineage with distantly related sequences are an important clue to suggest HGT or IGT. However, BLAST-based similarity scores have received particular criticism when used as the sole criterion for assessing these kinds of phenomena (*37, 38*). The relationships found in the first cluster in the phylogenetic trees suggest that an HGT event from cyanobacteria or an IGT event from a green algae secondary endosymbiont (the group that originated both the apicoplast in apicomplexa and the chloroplast in plants) transferred trypsin genes into the nucleus of an apicomplexa ancestor. In the second cluster, possibly these apicomplexa trypsin genes were inherited from an ancient eukaryotic organism. The lack of genes belonging to the first cluster of trypsins in *Cryptosporium* and *Babesia* genomes supports the IGT event for these genes because these parasites have secondarily lost the apicoplast. The absence of genes of the second cluster of trypsins in these organisms could be explained by the secondary loss of these genes.

HGT and IGT may change the way to combat diseases caused by protozoan parasites. The identification of genes possibly acquired by IGT such as the trypsin genes in apicomplexa would increase the attention of this phenomenon as source for gaining new proteins that may acquire new functions in invasion and pathogenesis processes. For this reason, IGT and HGT should be taken into account in building new therapeutic targets. Additionally, new evidence of this event in these parasites should generate a different view about several aspects of parasite biology such as their gene structure, genetic variability, pathology and epidemiology.

## Materials and Methods

### Identification of trypsin proteins in apicomplexa

To search for sequences of well characterized trypsin proteases, the keyword "trypsin" was used as query on Swiss-Prot/TrEMBL database through the ExPASy proteomics server (http://www.expasy.ch/sprot/). 50 bacterial sequences of this dataset were then used as

queries for BLASTP (E-value threshold = 1e-10) in the ToxoDB 5.1 database (http://www.toxodb.org) to retrieve homologous trypsin sequences in the genome of the well characterized apicomplexan parasite *T. gondii*. In order to discard false positives (*T. gondii* protein sequences with significant statistic similarity but without the conserved amino acids typical of this family), we searched for domains and functional motifs in Smart and Prosite databases (*39*, *40*).

For finding trypsin sequences of a wide range of organisms for further analysis, BLASTP (*41*) searches were conducted in MEROPS pepunit database (*5*) by using the three putative trypsin protein sequences retrieved from ToxoDB 5.1 as queries. BLASTP searches were also performed in EuPathDB 2.0 (http://eupathdb.org/eupathdb/) and in the NCBI non-redundant protein database (http://www.ncbi.nlm.nih.gov/) to retrieve trypsin sequences from gene predictions of apicomplexan genomes and from metazoan organisms, respectively. Searches in MEROPS and EuPathDB databases were conducted setting an E-value threshold of 1e-15. An E-value threshold of 1e-05 was used for retrieving sequences in NCBI non-redundant protein database. Except for the apicomplexa, only the best BLAST hits per species were taken into account for further analysis.

More sensitive detection of trypsin sequences in apicomplexan genomes was conducted with HMMER 2.3.2 package (*42*). Searches were performed by using the hmmsearch program with an HMM profile constructed with the hmmbuild program from a multiple sequence alignment of trypsin domains extracted from the retrieved protein sequences with BLASTP program. Trypsin domains were defined as in the multiple sequence alignment section. Searches were performed on protein sequence predictions of Toxo DB 5.1 (for *T. gondii* and *N. caninum*), CryptoDB 4.0 (http://cryptodb.org/cryptodb/) (for *Cryptosporidium*), PlasmoDB 5.5 (http://plasmodb.org/plasmo/) (for *Plasmodium*) and EuPathDB 2.0 (for *T. annulata* and *T. parva*). *Babesia bovis* protein sequences were retrieved from the NCBI non-redundant protein database.

### Multiple sequence alignment

Trypsin homolog sequences retrieved from BLASTP and HMMER searches were multiply aligned by using default parameters of ClustalW2 (http://www.ebi.ac.uk/clustalw/) (*43*). Alignments were inspected by eye and manually edited with GeneDoc program (http://www.nrbsc.org/gfx/genedoc/index.html). Trypsin domains for each sequence were defined from this multiple alignment as the longest possible continuous amino acid sequence homologous in all proteins. Trypsin domain sequences were then multiply aligned for performing the phylogenetic analysis.

### Phylogenetic analysis

The phylogenetic analysis was performed from a multiple alignment of trypsin domain sequences of all BLASTP and HMMER hits. Neighbor joining, minimum evolution, and maximum parsimony bootstrap consensus trees were constructed. The confidence of the trees was assessed by performing 5,000 bootstrap replicates. The evolutionary analysis was conducted in MEGA4 software (*44*).

### Protein sequence analysis

The prediction of both protein subcellular localization and Pfam domains for the retrieved sequences was conducted by using the following bioinformatic approach: First, protein secretion and transit peptides for the mitochondria and chloroplast in eukaryotic sequences were identified in TargetP 1.1 server (*45*) setting a specificity value of 0.90. Then, N-terminal signal peptides of bacterial and eukaryotic sequences were predicted with SignalP 3.0 server (*46*) by using the default thresholds of Smean and Sprob scores. Next, non-classical secretion of bacterial proteins was predicted with SecretomeP 2.0 server (*47*) by using the default threshold of SecP score. Finally, Pfam domains present in both prokaryotic and eukaryotic sequences were searched in the SMART database (*39*).

## Acknowledgements

JFOM is a recipient of a young researcher award from Colciencias (Colombia).

## Authors' contributions

AJG, AFA, JFOM, JEGM conceived and designed the study, and analyzed the data. AFA and JFOM performed the bioinformatic analysis. JFOM and AFA wrote the paper. All authors read and approved the final manuscript.

## Competing interests

The authors have declared that no competing interests exist.

# References


1 Hyde, J.E. 2005. Drug-resistant malaria. *Trends Parasitol.* 21: 494-498.

2 Plowe, C.V. 2005. Antimalarial drug resistance in Africa: strategies for monitoring and deterrence. *Curr. Top Microbiol. Immunol.* 295: 55-79.

3 Carruthers, V. and Boothroyd, J.C. 2007. Pulling together: an integrated model of *Toxoplasma* cell invasion. *Curr. Opin. Microbiol.* 10: 83-89.

4 Kim, K. 2004. Role of proteases in host cell invasion by *Toxoplasma gondii* and other Apicomplexa. *Acta Trop.* 91: 69-81.

5 Rawlings, N.D., *et al.* 2008. MEROPS: the peptidase database. *Nucleic Acids Res.* 36: D320-325.

6 Hartley, B.S. 1964. Amino-acid sequence of bovine chymotrypsinogen-A. *Nature* 201: 1284-1287.

7 Dixon, G.H., *et al.* 1956. Peptides combined with 14C-diisopropyl phosphoryl following degradation of 14C-DIP-trypsin with alpha-chymotrypsin. *Biochim. Biophys. Acta* 19: 193-195.

8 Shaw, E., *et al.* 1965. Evidence for an active-center histidine in trypsin through use of a specific reagent, 1-chloro-3-tosylamido-7-amino-2-heptanone, the chloromethyl ketone derived from Nα-tosyl-L-lysine. *Biochemistry* 4: 2219-2224.

9 Baptista, A.M., *et al.* 1998. The origin of trypsin: evidence for multiple gene duplications in trypsins. *J. Mol. Evol.* 47: 353-362.

10 Blackman, M.J. 2000. Proteases involved in erythrocyte invasion by the malaria parasite: function and potential as chemotherapeutic targets. *Curr. Drug Targets* 1: 59-83.

11 Huang, J., *et al.* 2004. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int. J. Parasitol.* 34: 265-274.

12 Wang, S., *et al.* 1999. Concerted evolution within a trypsin gene cluster in *Drosophila*. *Mol. Biol. Evol.* 16: 1117-1124.

13 Tripathi, L.P. and Sowdhamini, R. 2006. Cross genome comparisons of serine proteases in *Arabidopsis* and rice. *BMC Genomics* 7: 200.

14 Neurath, H. 1984. Evolution of proteolytic enzymes. *Science* 224: 350-357.

15 Neurath, H. 1989. The diversity of proteolytic enzymes. In *Proteolytic Enzymes: A Practical Approach* (eds. Beynon, R.J. and Bond, J.S.), pp. 1-13. IRL Press, Oxford, UK.

16 Davis, C.A., *et al.* 1985. A gene family in *Drosophila melanogaster* coding for trypsin-like enzymes. *Nucleic Acids Res.* 13: 6605-6619.

17 Muller, H.M., *et al.* 1993. Members of a trypsin gene family in *Anopheles gambiae* are induced in the gut by blood meal. *EMBO J.* 12: 2891-2900.

18 Casu, R.E., *et al.* 1994. Isolation of a trypsin-like serine protease gene family from the sheep blowfly *Lucilia cuprina*. *Insect Mol. Biol.* 3: 159-170.

19 Emi, M., *et al.* 1986. Cloning, characterization and nucleotide sequence of two cDNAs encoding human pancreatic trypsinogens. *Gene* 41: 305-310.

20 Stevenson, B.J., *et al.* 1986. Sequence organization and transcriptional regulation of the mouse elastase II and trypsin genes. *Nucleic Acids Res.* 14: 8307-8330.

21 Shi, Y.B. and Brown, D.D. 1990. Developmental and thyroid hormone-dependent regulation of pancreatic genes in *Xenopus laevis*. *Genes Dev.* 4: 1107-1113.

22 Wang, K., *et al.* 1995. Isolation and characterization of the chicken trypsinogen gene family. *Biochem. J.* 307: 471-479.

23 Roach, J.C., *et al.* 1997. The molecular evolution of vertebrate trypsinogens. *J. Mol. Evol.* 45: 640-652.

24 Rowen, L., *et al.* 1996. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* 272: 1755-1762.

25 Ponting, C.P. and Russell, R.R. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* 31: 45-71.

26 Vogel, C., *et al.* 2004. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* 14: 208-216.

27 Pallen, M.J., *et al.* 1997. The HtrA family of serine proteases. *Mol. Microbiol.* 26: 209-221.

28 Ponting, C.P. 1997. Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Sci.* 6: 464-468.

29 Spiers, A., *et al.* 2002. PDZ domains facilitate binding of high temperature requirement protease A (HtrA) and tail-specific protease (Tsp) to heterologous substrates through recognition of the small stable RNA A (ssrA)-encoded peptide. *J. Biol. Chem.* 277: 39443-39449.

30 Murwantoko, M.Y., *et al.* 2004. Binding of proteins to the PDZ domain regulates proteolytic activity of HtrA1 serine protease. *Biochem. J.* 381: 895-904.

31 Wilken, C., *et al.* 2004. Crystal structure of the DegS

stress sensor: how a PDZ domain recognizes misfolded protein and activates a protease. *Cell* 117: 483-494.
32 McFadden, G.I., *et al.* 1996. Plastid in human parasites. *Nature* 381: 482.
33 Fast, N.M., *et al.* 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol. Biol. Evol.* 18: 418-426.
34 Köhler, S., *et al.* 1997. A plastid of probable green algal origin in Apicomplexan parasites. *Science* 275: 1485-1489.
35 Delwiche, C.F. 1999. Tracing the thread of plastid diversity through the tapestry of life. *Am. Nat.* 154: S164-S177.
36 Keeling, P.J. and Palmer, J.D. 2001. Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. *Proc. Natl. Acad. Sci. USA* 98: 10745-10750.
37 Salzberg, S.L., *et al.* 2001. Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292: 1903-1906.
38 Stanhope, M.J. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411: 940-944.
39 Schultz, J., *et al.* 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* 95: 5857-5864.
40 Hofmann, K., *et al.* 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27: 215-219.
41 Altschul. S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
42 Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755-763.
43 Thompson, J., *et al.* 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
44 Tamura, K., *et al.* 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.
45 Emanuelsson, O., *et al.* 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300: 1005-1016.
46 Bendtsen, J.D., *et al.* 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340: 783-795.
47 Bendtsen, J.D. 2005. Non-classical protein secretion in bacteria. *BMC Microbiol.* 5: 58.
48 Gissot, M. 2007. Epigenomic modifications predict active promoters and gene structure in *Toxoplasma gondii*. *PLoS Pathog.* 3: e77.