

An evolutionary analysis of trypanosomatid GP63 proteases

Lina Ma · Kaifu Chen · Qingshu Meng · Qingyou Liu ·
Petrus Tang · Songnian Hu · Jun Yu

Received: 12 August 2010 / Accepted: 10 March 2011 / Published online: 19 April 2011
© Springer-Verlag 2011

Abstract The trypanosomatid GP63 proteases are known to be involved in parasite–host interaction and exhibit strong sequence and structural similarities to those of their hosts and insect vectors. Based on genome sequences of the three trypanosomatids, *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania* spp., we annotated all their GP63 proteases and divided highly duplicated *T. cruzi* GP63 proteases into four novel groups according to sequence features. In *Leishmania* spp., we studied the evolutionary dynamics of GP63 proteins and identified 57 amino acid sites that are under significant positive selections. These

sites may contribute to the functional variations of the GP63 proteases and provide clues for vaccine development.

Introduction

Three trypanosomatids—*Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*—have been sequenced so far (Berriman et al. 2005; El-Sayed et al. 2005b; Ivens et al. 2005). They are kinetoplastids according to the presence of a DNA-containing organelle known as kinetoplast in their single large mitochondrion. They are primitive eukaryotes (Sogin 1991) and often categorized into the super group Excavates (Adl et al. 2005; Simpson and Roger 2004). *T. brucei* and *T. cruzi* are the aetiological agents of African sleeping sickness (HAT) and Chagas' disease in South and Central America, respectively. *Leishmania* spp. causes a variety of diseases throughout the tropics and subtropics (Burri and Brun 2003). There are half a billion people, primarily in tropical and subtropical areas of the world, who are at the risk of contacting these diseases, and there are more than 20 million infections and 100,000 deaths per year caused by trypanosome infection (Stuart et al. 2008).

Although having diverged 200 to 500 million years ago (Douzery et al. 2004; Stevens et al. 2001; Overath et al. 2001), predating the emergence of mammals (O'Brien et al. 1999), the genomes of the trypanosomatid species, exhibits high synteny. A comparison of gene content and genome architecture of *T. brucei*, *T. cruzi*, and *L. major* revealed a conserved core proteome of ~6,200 genes that are largely syntenic polycistronic gene clusters (El-Sayed et al. 2005a) and gene families (Brenchley et al. 2007). As far as the *Leishmania* species are concerned, the orthologous chromosomes of *L. major*, *Leishmania braziliensis*, and *Leish-*

Electronic supplementary material The online version of this article (doi:10.1007/s00436-011-2348-x) contains supplementary material, which is available to authorized users.

L. Ma · K. Chen · Q. Meng · S. Hu (✉) · J. Yu (✉)
CAS Key Laboratory of Genome Sciences and Information,
Beijing Institute of Genomics, Chinese Academy of Sciences,
No.7 Beitucheng West Road, Chaoyang District, Beijing 100029,
People's Republic of China
e-mail: husn@big.ac.cn

J. Yu
e-mail: junyu@big.ac.cn

L. Ma · K. Chen · Q. Meng
Graduate University, Chinese Academy of Sciences,
Beijing 100049, People's Republic of China

P. Tang
Molecular Regulation and Bioinformatics Laboratory,
Chang Gung University,
Taoyuan 333, Taiwan

Q. Liu
Animal Reproduction Institute, Guangxi Key Laboratory
of Subtropical Bioresource Conservation and Utilization,
Guangxi University,
Nanning 530004, People's Republic of China

mania infantum are remarkably conserved in both gene content and order (Peacock et al. 2007), even though they diverged 40 to 80 million years ago (Fernandes et al. 1993).

For *Leishmania* spp., GP63 (also known as leishmanolysin) is considered to be a major virulence factor (Chaudhuri et al. 1989; Joshi et al. 2002; Yao et al. 2003). Unlike *Leishmania* spp., *T. brucei* employs a variant surface glycoprotein (VSG) as its main virulence factor and *T. cruzi* uses mucins and transsialidases/sialidases. Among the known virulence factors involved in host–parasite interaction, GP63 proteases are shared by the three trypanosomatids (Santos et al. 2006; Yao 2010). The *T. brucei* GP63 proteases are predominantly expressed in the bloodstream form, and the experimental evidence implicated that they may protect bloodstream trypanosomes against complement-mediated lysis by impairing a protein-processing function (LaCount et al. 2003; El-Sayed and Donelson 1997; Grandgenett et al. 2007). Neutralization assays have indicated that anti-GP63 serum has a significant inhibitive effect on *T. cruzi* infection (Kulkarni et al. 2009; Cuevas et al. 2003). Genome content and structure comparisons among *T. cruzi*, *T. brucei*, and *L. major* demonstrated that among the large-scale syntenic regions, gene divergence, acquisition and loss, and rearrangement have shaped each genome significantly (El-Sayed et al. 2005a). In order to further understand the virulence factor of these three parasites, we focused our analysis on their GP63 proteases, taking the advantage of the extremely homologous *Leishmania* spp. (Smith et al. 2007) and the knowledge about the parasitic trypanosomatid species (Lipoldova and Demant 2006). We also scrutinized sequence variations that may be involved in parasite–host interactions.

Materials and methods

Data sources and analysis tools

We downloaded 562 sequences annotated as “leishmanolysin” from Genbank, Refseq of NCBI, EMBL, DDBJ, and PDB, and Swiss-Prot on June 23 2008 and used HMMER (Krogh et al. 1994) version 2.3.2 (Linux), a freely distributable implementation of profile Hidden Markov Models software, to discover GP63 sequences from the Pfam database. After removal of short length sequences, we obtained 290 GP63 sequences from 63 species (more details are summarized in Online Resources 3). Information for all the species’ name and sequences’ accession can be referred to Online Resources 1 and 2, respectively.

We predicted glycosylphosphatidylinositol (GPI) modification site, using big-PI Predictor (http://mendel.imp.ac.at/sat/gpi/gpi_server.html). When sequences are predicted to

have GPI modification, *P* and *S* indicate higher and lower confidence, respectively. We display three-dimensional protein structures using the molecular graphics system PyMOL version 0.96 (<http://www.pymol.org/>).

Phylogeny

Since the most conserved portion of GP63 proteases across multiple species is about 100 aa in length (between amino acid position 295 and 393 of Leima90), we constructed the phylogenetic tree based on this segment of 290 GP63 sequences, using a maximum likelihood method implemented in the phylogeny inference package (version 3.67; Retief 2000). The robustness of the phylogenetic inferences was tested by bootstrapping, using 100 replications of the data based on the criterion of 50% majority-rule consensus.

Positive selection test

We used CodeML in the PAML package (Yang 1997) for the evolutionary analysis of the GP63 protease in *Leishmania* spp. There are three models in CodeML: branch, site, and branch–site. In particular, the branch model allows the ω ratio to vary among branches in the phylogenetic tree and is useful for detecting positive selection acting on particular lineages. The site model allows the ω ratio to vary among different sites, whereas the branch–site model allows the ω ratio to vary among both sites and lineages and attempts to detect positively selected sites along a limited number of sequences. We analyzed 17 *Leishmania* spp. GP63 sequences that have potential GPI structure to investigate if ω ratio varies among branches. After constructing free-ratio and one-ratio models, we used likelihood ratio test (LRT) to evaluate the opposing hypotheses. Since the null hypothesis that ω ratio varies among branches made better sense, we applied the branch–site model to find lineages (branches) and sites where positive selections are at work. We set the parameters as recommended (the Bayes empirical Bayes results of NSSites = 2 and NSSites = 8; Yang et al. 2005) for the site model analysis.

Results

A conserved motif implies functional role for GP63 proteases

We started our analysis by dividing GP63 proteases into four groups according to their taxonomy and pathogenicity: (1) non-pathogenic protists (such as *Tetrahymena thermophila* and *Paramecium tetraurelia*), (2) pathogenic protists (such as *Leishmania* spp., *T. cruzi*, and *T. brucei*), (3) multicellular animals, and (4) plants (Fig. 1a). Since the

pfam GP63 model is built with majority ruling, sequences (~500 aa) from group 2, group 3, and group 4 showed higher similarity, whereas sequences of group 1 have much shorter conserved length. Nevertheless, all sequences do share a 100-aa segment that follows a zinc-binding motif (HEXXH): between amino acid positions 340–430 of the pfam GP63 model and 295–393 of Leima90. The relative positions of this segment to the zinc-binding motif appeared group specific: 25 aa in pathogenic protists, 63 aa in animals, and 34 aa in plants; it is relatively variable in non-pathogenic protists: 30 to 50 aa. There are only a handful sequences (18 sequences: five from *T. thermophila*, one from *P. tetraurelia*, nine from *T. cruzi*, one from *Tetraodon nigroviridis*, one from *Rattus norvegicus*, and one from *Pan troglodytes*; see Online Resources 4) whose conserved segments are interrupted. We investigated the similarity matrix of the shared segment through pair-wise blast among the 290 sequences. The similarity score is significantly higher in *Leishmania* spp. and among multicellular eukaryotes (Fig. 1b).

We aligned the conserved segment of the 290 sequences and found 12 highly conserved sites (we numbered them sequentially; Online Resources 5). We suggest that His3 and Met5 are the two amino acid residues within the active

site for proteolysis function. In particular, His3 is the third histidine for zinc binding; Met5 is Met-turn-like, similar to other Metzincin class zinc proteases such as astacin and collagenase (Schlagenhauf et al. 1998). Among the remaining residues, cysteine may function in maintaining protease structure and the potential roles for the rest of the residues remain to be elucidated. Given the high degree of conservation and the presence of active sites, this shared segment is most likely to be a functional motif of GP63 protease in supporting catalytic reaction but not species-associated functions.

GP63 proteases of pathogenic protists bear more similarity to those of multicellular eukaryotes than non-pathogenic protists

We aligned the entire length of GP63 proteases from the pathogenic parasites and animals. In addition to the highly conserved segment, we found fifteen conserved Cys and four conserved Pro residues between multicellular eukaryotes and pathogenic protists (Fig. 2). The group 3 (animals) members are clustered into two clades: insects and others; the insect clade bears more similarity to that of the trypanosomatids than other animals (Online Resources 3). In addition, the

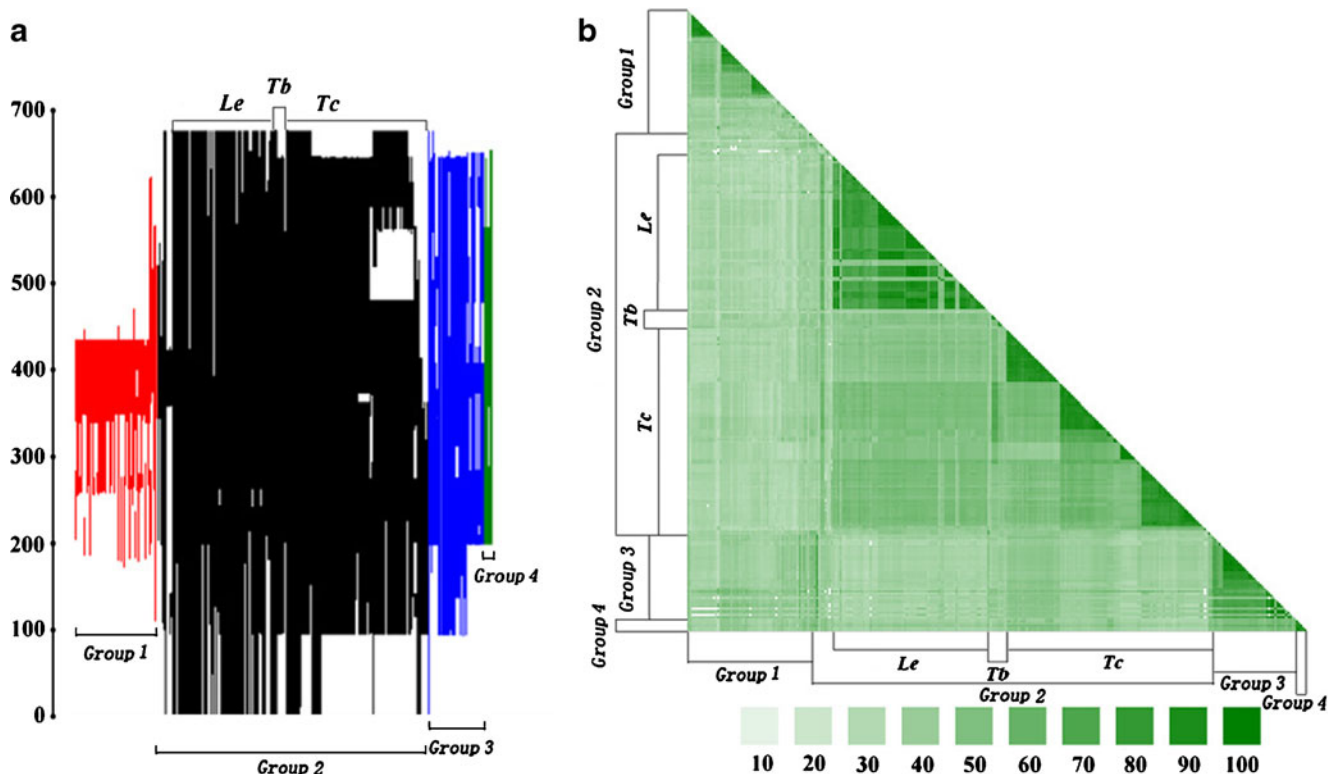


Fig. 1 The domain structure of GP63 proteases. **a** GP63 proteases were divided into four groups according to the taxonomy and pathogenicity: red group 1, black group 2, blue group 3, and green group 4. The y-axis indicates the length of the GP63 protease domain of the 294 sequences. There are an apparent motif absence of around

500 aa in *T. cruzi* and a less evident absence of around 350 aa. **b** Similarity matrix of the shared segment. Pair-wise blast was carried out among the 290 amino acid sequences, and the similarity score was colored. *Le* *Leishmania* spp., *Tb* *T. brucei*, *Tc* *T. cruzi*

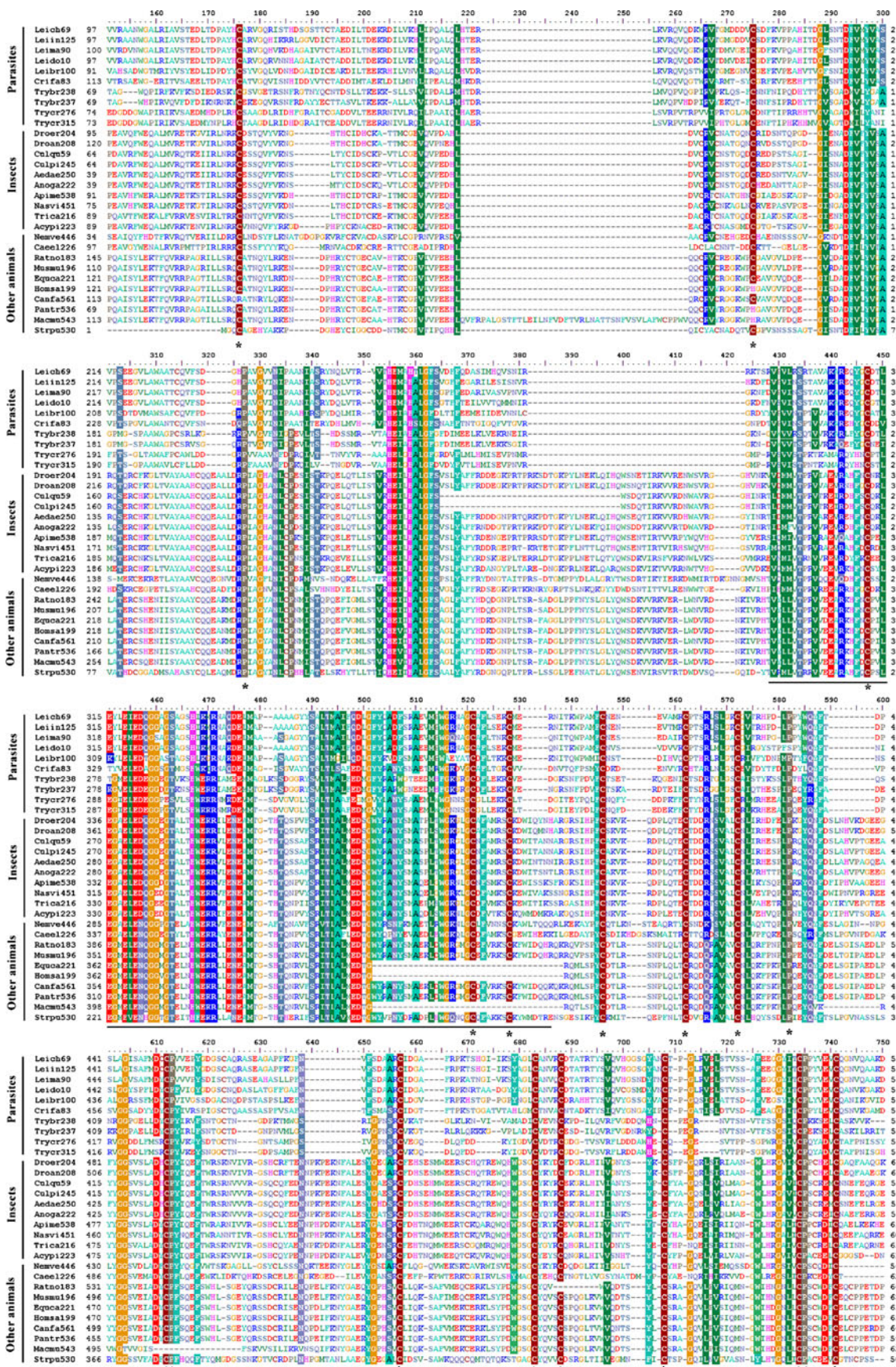


Fig. 2 Alignment of GP63 sequences between group 2 and group 3. We divided group 3 into two clades: insect and other animals. We selected ten representative sequences for each group and trimmed the portions with low similarity. We *color-coded* the sites that are conserved in more than 80% sequences, *underlined* the common segment, and *used asterisks* to indicate the 15 conserved Cys and four conserved Pro residues

presence or absence of GPI anchorage is also diverged among GP63 proteases; it is absent in group 1 and group 4, but in group 2, 11% and 5% of the proteases were predicted to have GPI at *P* and *S* levels, respectively. In group 3, the predictions are much higher, 21% and 38% at *P* and *S* levels, respectively. All the *P* level sequences belong to the insect clade.

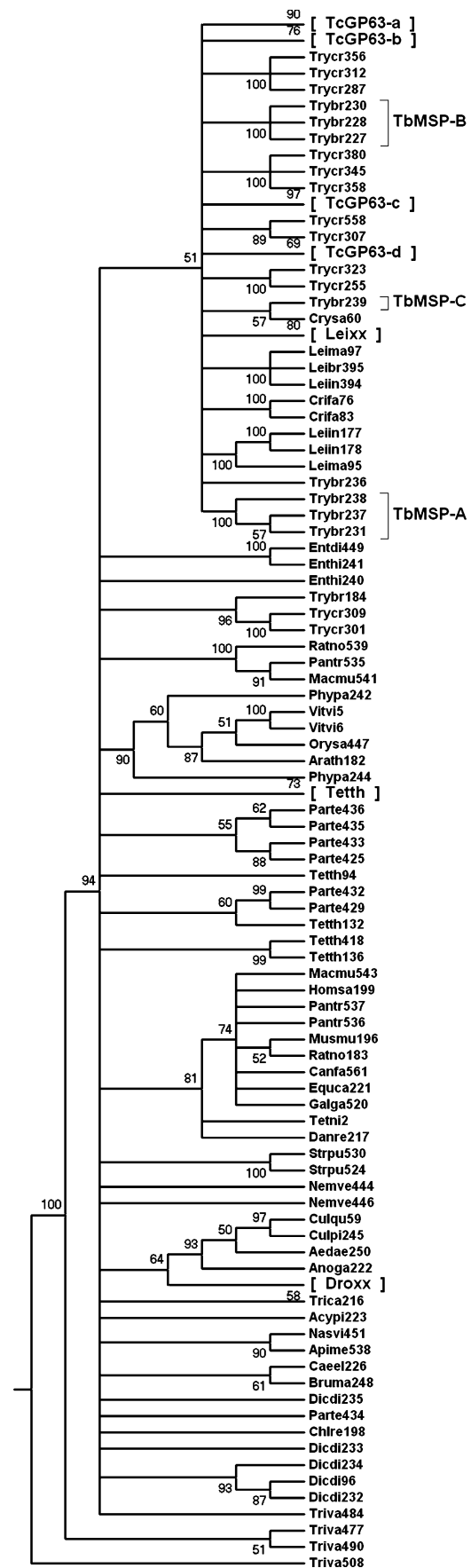
The *T. cruzi* GP63 proteases share a unique conserved domain

In *T. cruzi*, GP63 fell into four clades: TcGP63-a, TcGP63-b, TcGP63-c, and TcGP63-d (Fig. 3). The between-clade and within-clade amino acid similarities are 30% and 80%, respectively, as domain absence was frequently observed, involving both terminal and internal domains. Each clade has its signature deletions: 25 members of the TcGP63-a clade have N-terminal deletions, the 22 members of the TcGP63-b clade have deletions at both ends, eight members of the TcGP63-c clade have internal deletions of 10 aa (360–370 aa according to the pfam GP63 model), and 29 members of the TcGP63-d clade miss 80 aa (480–560 aa) when compared to the pfam model. We found that this missing 80 aa segment is highly conserved (data not shown).

Previous studies have suggested that TcGP63-b is involved in parasite infection. Three GP63 proteases in two groups were identified based on Northern blotting: Tcgp63-I, Tcgp63-II, and Tcgp63-III (Cuevas et al. 2003). Tcgp63-I was thought to be differentially expressed depending on the life cycle of *T. cruzi*, and the antibodies against Tcgp63-I were demonstrated to partially block the infection of Vero cells by trypomastigotes. Tcgp63-Ia and Tcgp63-Ib are GP63-1 and GP63-4, respectively, according to Grandgenett's classification (Grandgenett et al. 2000); both are grouped into TcGP63-b (Trycr355 and Trycr362) in our analysis.

In addition, among the 2,784 *T. cruzi* proteins (Atwood et al. 2005), we found that all GP63 proteases fell into one

Fig. 3 Maximum likelihood (ML) tree of 290 GP63 sequences from 63 species. *T. cruzi* has four main groups: TcGP63-a, TcGP63-b, TcGP63-c, and TcGP63-d. *T. brucei* has three groups: TbMSP-A, TbMSP-B, and TbMSP-C. Sequence is named with an abbreviation of the species name plus a serial number. For example, Leima90 is a GP63 sequence from the species *L. major* and 90 stands for the sequence series in our data. Clades with more than eight sequences of the same species were condensed. Names for all condensed branches were in the *parentheses*. The ML tree was constructed by using Phylip (also see Online Resource 6 for all data)



group: TcGP63-b. Therefore, TcGP63-b is also abundantly expressed in *T. cruzi*. However, the remaining three GP63 groups of *T. cruzi* (TcGP63-a, TcGP63-c, and TcGP63-d) have not yet been studied despite the fact that they share the highly conserved eighteen Cys and eight Pro residues (except a few members of TcGP63-d, which have lost the twelfth Cys residue). In addition, the predicted N-glycosylation sites seem to be specific for each group; the TcGP63-b members have two potential N-glycosylation sites, whereas TcGP63-a members have three (Cuevas et al. 2003).

We did not define new GP63 family members in *T. brucei*, where nine non-redundant sequences are grouped into three as reported previously (LaCount et al. 2003). One striking feature of the *T. brucei* GP63 groups is that they are scattered into the *Leishmania* spp. groups and the *T. cruzi* groups. It suggests that the GP63 genes (paralogs) were duplicated and started to diverge from each other before their speciation events. TbMSP-A (Trybr231, Trybr237, and Trybr238 in this study) has a unique extended C-terminal region rich in serines and glutamates. TbMSP-B (Trybr228, Trybr227, and Trybr230) is constitutively expressed in both the procyclic and the bloodstream stages, whereas TbMSP-A and TbMSP-C (Trybr239) only express at the bloodstream stage. Experimental evidence showed that depletion of TbMSP-B inhibits the release of VSG and thus increases the steady level of cell-associated VSG (LaCount et al. 2003).

Positive selection of *Leishmania* spp. GP63 sites

Leishmania spp. infection is known to lead to different sicknesses: cutaneous leishmaniasis (CL), mucocutaneous leishmaniasis (ML), and visceral leishmaniasis (VL). CL heals without any treatment, ML causes disfiguring, and VL is fatal if left untreated. Using the tools in the PAML, we examined the sequence variations of *Leishmania* GP63 proteases to evaluate sites of positive selection as well as their functional implications. To focus on a set of limited number and relevant sequences, we selected those predicted to be GPI-anchored based on big-PI algorithm because previous evidence suggested that most GP63 are anchored to the plasma membranes through a GPI anchor (Bordier et al. 1986). We obtained 17 sequences from eight species (*L. major*, *Leishmania amazonensis*, *L. infantum*, *Leishmania chagasi*, *Leishmania donovani*, *Leishmania guyanensis*, *Leishmania panamensis*, and *L. braziliensis*), and a segment from Val100 to Asn577 according to Leima90 was used for our analysis (Schlagenhauf et al. 1998).

We used Consurf (Landau et al. 2005) to investigate how variable sites have distributed on the tertiary structure of the GP63 sequences. Our results showed that the zinc-binding surface is quite variable and its protuberant positions, also highly variable, seem to be advantageous for interacting with host proteins (Fig. 4b), which agree with the non-

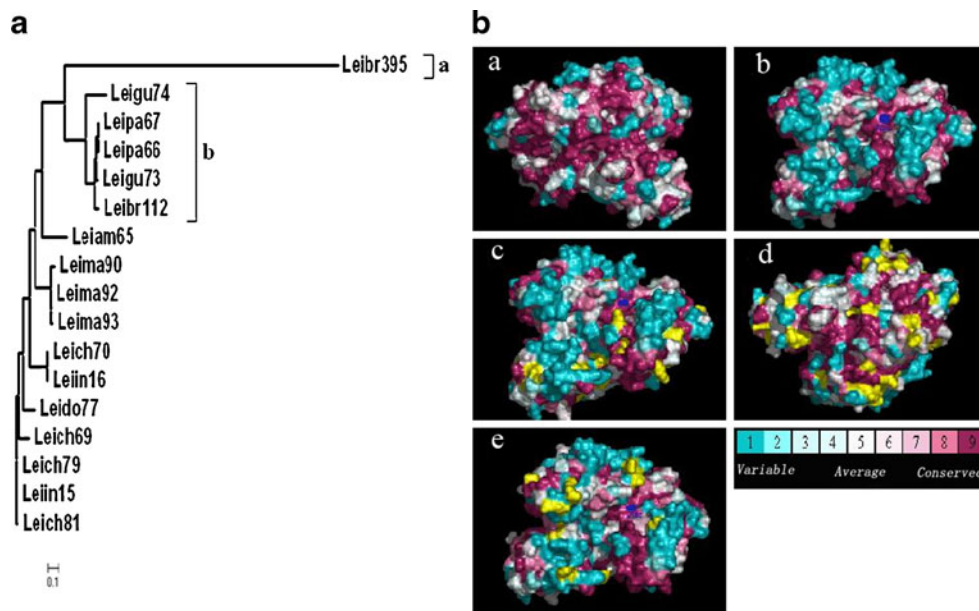


Fig. 4 Positive selection analysis of the *Leishmania* spp. GP63 proteases. We used a segment (Val100 to Asn577 according to Leima90) of the 17 mature sequences. **a** Neighbor-joining (NJ) tree was built for a Leibr395 and b the branch of Leigu73, Leigu74, Leipa66, Leipa67, and Leibr112. **b** Positively selected sites were mapped to the tertiary structure of the GP63 protease: a and d, the back side of the zinc-binding surface; b, c, and e, the zinc-binding

surface. The zinc atom is shown as a blue pellet. In a and b, different colors were used to represent the conservation condition for each site. In c, d, and e, the yellow-coded sites are positively selected ($P > 95\%$). b and c shows the positive selected sites of branch-site model. e shows the positive selected sites based on the site model. All the pictures were processed by using PyMOL

Table 1 LRT for positive selection of *Leishmania* spp.

Model	2ΔL	df	P
Branch model 1 ^a	94.28814	30	<0.01
Branch model 2	28.556642	28	>0.01
Branch–site model	19.985484		<0.01
Site model			
Model 1–Model 2	95.640432	2	<0.01
Model 7–Model 8	115.15132	2	<0.01

LRT likelihood ratio test

^a Branch model 1 was tested for the 17 sequences that include Leibr395. Branch model 2 was tested for the 16 sequences that exclude Leibr395

synonymous variations that occur externally and around the active site (Alvarez-Valin et al. 2000).

In the test for positive selections, we put forward two opposite hypotheses based on two different models. One assumes that all 17 sequences have the same ω ratio, based on the one-ratio model. The other assumes that some sequences may have different ω ratios, based on the free-ratio model. The LRT results are shown in Table 1. The lnL values for the one-ratio and free-ratio models are $-7,806.594318$ and $-7,759.450251$, respectively. Therefore, free-ratio model was considered better in this case. Given that df was 30, providing a value of 94.288134 for $2\Delta L$ and the P value less than 0.01, our test suggested that some sequences may evolve at a different ratio. Since Leibr395 represents a more divergent lineage as compared to the remaining sequences (Fig. 4a), we supposed that Leibr395 has a different evolutionary pace in the branch–site model and our tests accepted the hypothesis. We thus defined 45 variation sites (105, 119, 123, 127, 153, **154**, 166, 176, 181, 182, 206, **220**, 235, 246, 248, 255, 256, 259, 267, **301**, 302, 305, 307, 309, **314**, 318, 332, 351, 392, 401, 468, 483, 484, 486, 496, 506, 532, 536, 537, 542, 547, 555, 570, 575, and **576**, which are positioned according to the mature sequence of Leima90. Positions in bold have a $P > 99\%$, while the remaining are $99\% > P > 95\%$. These P values are different from the P value in LRT and represent the possibility for a site to be positively selected) that are supposed to contribute to the different ω ratios. After having localized all positions on the tertiary structure, we noticed that most of the positions are at the back of the zinc-binding surface (these sites are at the front if you look from the opposite angle; Fig. 4b). We also examined the three species (*L. braziliensis*, *L. guyanensis*, and *L. panamensis*) in lineage b, which have longer branches when compared to the rest and were reported to have a close phylogenetic status and all induce mucocutaneous leishmaniasis (Osorio et al. 1998; Di Lella et al. 2006), but free-ratio test (16 sequences without Leibr395) showed insignificant positive selection.

We used the site model to analyze the remaining 16 sequences and found 12 positively selected sites, **136**, **275**, **278**, **396**, **424**, **426**, **429**, **431**, **432**, **469**, **474**, and **502** (the positions in bold have a $P > 99\%$ and the rest are $99\% > P > 95$). The ω ratios of these sites are all greater than one. They were again mapped to the surrounding area of the zinc-binding site in the tertiary structure and mostly in protuberant positions (Fig. 4b). Although the two models, the branch–site and the site models, give rise to non-overlapping positively selected positions and have different locations on the tertiary structure, the nonrandom distributions suggested that these variations or sites may have functional implications in parasite–host interactions.

To determine whether the positively selected sites correlate with any specific leishmaniasis, we extracted the positively selected positions defined by the site model from each sequence to form a positive selection pattern for the sequences. We also divided the species into three groups according to different types of leishmaniasis: CL (*L. major* and *Leishmania aethiopica*), CL and ML (*L. braziliensis*, *L. panamensis*, and *L. guyanensis*), and CL and VL (*L. chagasi*, *L. infantum*, and *L. donovani*). The results are quite significant, where each group not only has their own distinct patterns but also more than one sub-patterns (Table 2).

Table 2 Positively selected sites of *Leishmania* spp.

	Sequences	Positively selected sites ^a
CL&ML	Leipa66	A T G N G I T T S R S D
	Leipa67	A T G N G I T T S R S D
	Leigu73	A T G N G I T T S R S D
	Leibr112	A V G N G I T T S R S D
	Leigu74	A T G N K A T T P D P D
CL&VL	Leich70	L T R D A S T Q S N M V
	Leiin126	L T R D A S T Q S N M V
	Leich79	L G R R K G R P D R G I
	Leiin125	L G R R K G R P D R G I
	Leich81	L G R R K G R P D R G Q
	Leich69	D D S R K G R P D R G I
Leido77	D V R D V G S I P R G V	
CL	Leima92	A P R Q A G R P G R H I
	Leima93	A P R Q A G R P G R H I
	Leima90	A P R Q A G R P G R H I
	Leiam65	V T G N M D P Q A S D L

a: The characters in bold have $p > 99\%$ and the rest are $99\% > p > 95\%$. Different patterns in each group are shaded in blue

Discussion

In this study, we found that GP63 proteases evolved differently among trypanosomatids, especially in *T. cruzi*, whose GP63 proteases were duplicated extensively and brought in several novel GP63 domains as compared to those of *T. brucei* and most *Leishmania* spp. In addition, we found that the trypanosomatid GP63 protease domain bears more similarity to that of its hosts than to that of their non-pathogenic counterparts. Compared to the vectors (insects) and hosts (human and rodents), the parasitic GP63 proteases are more similar to those of their vectors. These results suggest that GP63 proteases may indeed play significant roles in host–parasite interactions.

We also identified different positive selection patterns for the three types of leishmaniasis. It was reported that patients cured from visceral leishmaniasis or cutaneous leishmaniasis respond to GP63 proteases differently: the former induces Th2-like response and the latter shows Th1-like response (Kurtzhals et al. 1994; Kemp et al. 1994). Furthermore, four *Leishmania* promastigote species, *L. donovani*, *L. amazonensis*, *L. infantum*, and *L. major*, exhibit different sensitivities to human complement (Dominguez et al. 2002). In each type of symptoms, we obtained unique patterns that can be subdivided into a limited number of sub-patterns. In fact, *Leishmania* species that induce the same disease symptoms also differ in immunoreactions (McMahon-Pratt and Alexander 2004; Wilson et al. 2005). The three species *L. chagasi*, *L. infantum*, and *L. donovani* all induce VL and CL, but only *L. chagasi* leads to common CL lesions. Dermal lesions at the inoculation site of *L. infantum* resemble lesions caused by dermatotropic species. *L. donovani* is unique in producing disseminated post-kala-azar dermal leishmaniasis, which occurs in a percentage of cured VL cases (Launois et al. 2008). In addition, even different strains of the same species can manifest different disease sequelae (Sharma et al. 2005). For instance, two *L. major* strains, MRHP/SU/59 Neals and WHOM/IR/-/173, not only differ in virulence but also respond differently to the same vaccine. Moreover, there are possibilities that disease diversity may be contributed by different host immune systems. For instance, in a mouse model, resistance and susceptibility to infection were found to be correlated with the development of CD4⁺ Th1 and CD4⁺ Th2 responses, respectively (Reiner and Locksley 1995). As a final note, GP63 protease expression levels may also result in different disease symptoms as *Leishmania* spp. display variable GP63 expression pattern even at the same stages (Yao et al. 2007).

Current therapies for trypanosomatid-associated diseases are costly, often poorly tolerated, and not always effective. Therefore, the development of alternative therapies, including vaccines, is of essence in the prevention of trypanosomatid

infection. For *Leishmania*, the pathogenic mechanisms have been thoroughly elucidated and the biggest obstacle now for vaccine development is the variability of antigens among different *Leishmania* species (Spitzer et al. 1999). Our analysis provides a fresh look at the conserved and variable sequences, which may lead to new designs for the vaccines against the pathogenic trypanosomatids.

Acknowledgments We are very grateful to Drs. Yingfeng Luo and Dan Qi for their useful suggestions and comments during the study and Mr. Joe Yu for editing the manuscript. The work is supported by a grant from National Science and Technology Key Project (2008ZX1004-013), the Ministry of Science and Technology of the People's Republic of China.

References

- Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MF (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* 52:399–451
- Alvarez-Valin F, Tort JF, Bernardi G (2000) Nonrandom spatial distribution of synonymous substitutions in the gp63 gene from *Leishmania*. *Genetics* 155:1683–1692
- Atwood JA III, Weatherly DB, Minning TA, Bundy B, Cavola C, Opperdoes FR, Orlando R, Tarleton RL (2005) The *Trypanosoma cruzi* proteome. *Science* 309:473–476
- Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Bohme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou U, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivans A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309:416–422
- Bordier C, Etges RJ, Ward J, Turner MJ, Cardoso de Almeida ML (1986) *Leishmania* and *Trypanosoma* surface glycoproteins have a common glycopospholipid membrane anchor. *Proc Natl Acad Sci USA* 83:5988–5991
- Brenchley R, Tariq H, McElhinney H, Szoor B, Huxley-Jones J, Stevens R, Matthews K, Taberner L (2007) The tritryp phosphatome: analysis of the protein phosphatase catalytic domains. *BMC Genomics* 8:434

- Burri C, Brun R (2003) Human African trypanosomiasis. In: Cook GC, Zumla A (eds) *Manson's Tropical Diseases*, 21st edn. WB Saunders, London, pp 1303–1323
- Chaudhuri G, Chaudhuri M, Pan A, Chang KP (1989) Surface acid proteinase (gp63) of *Leishmania mexicana*. A metalloenzyme capable of protecting liposome-encapsulated proteins from phagolysosomal degradation by macrophages. *J Biol Chem* 264:7483–7489
- Cuevas IC, Cazzulo JJ, Sanchez DO (2003) Gp63 homologues in *Trypanosoma cruzi*: surface antigens with metalloprotease activity and a possible role in host cell infection. *Infect Immun* 71:5739–5749
- Di Lella F, Vincenti V, Zennaro D, Afeltra A, Baldi A, Giordano D, Pasanisi E, Bacciu A, Bacciu S, Di Lella G (2006) Mucocutaneous leishmaniasis: report of a case with massive involvement of nasal, pharyngeal and laryngeal mucosa. *Int J Oral Maxillofac Surg* 35:870–872
- Dominguez M, Moreno I, Lopez-Trascasa M, Torano A (2002) Complement interaction with trypanosomatid promastigotes in normal human serum. *J Exp Med* 195:451–459
- Douzery EJ, Snell EA, Baptiste E, Delsuc F, Philippe H (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci USA* 101:15386–15391
- El-Sayed NM, Donelson JE (1997) African trypanosomes have differentially expressed genes encoding homologues of the *Leishmania* gp63 surface protease. *J Biol Chem* 272:26742–26748
- El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renaud H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N (2005a) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404–409
- El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B (2005b) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309:409–415
- Fernandes AP, Nelson K, Beverley SM (1993) Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism. *Proc Natl Acad Sci USA* 90:11608–11612
- Grandgenett PM, Coughlin BC, Kirchhoff LV, Donelson JE (2000) Differential expression of gp63 genes in *Trypanosoma cruzi*. *Mol Biochem Parasitol* 110:409–415
- Grandgenett PM, Otsu K, Wilson HR, Wilson ME, Donelson JE (2007) A function for a specific zinc metalloprotease of African trypanosomes. *PLoS Pathog* 3:1432–1445
- Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley SM, Bianchetti G, Borzym K, Bothe G, Bruschi CV, Collins M, Cadag E, Ciarlioni L, Clayton C, Coulson RM, Cronin A, Cruz AK, Davies RM, De Gaudenzi J, Dobson DE, Duesterhoeft A, Fazelina G, Fosker N, Frasch AC, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram JC, Muller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'Neil S, Pentony M, Pohl TM, Price C, Purnelle B, Quail MA, Rabinowitz E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz JC, Rutter S, Saunders D, Schafer M, Schein J, Schwartz DC, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G, Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith DF, Blackwell JM, Stuart KD, Barrell B, Myler PJ (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309:436–442
- Joshi PB, Kelly BL, Kanhawi S, Sacks DL, McMaster WR (2002) Targeted gene deletion in *Leishmania major* identifies leishmanolysin (gp63) as a virulence factor. *Mol Biochem Parasitol* 120:33–40
- Kemp M, Hey AS, Kurtzhals JA, Christensen CB, Gaafar A, Mustafa MD, Kordofani AA, Ismail A, Kharazmi A, Theander TG (1994) Dichotomy of the human T cell response to *Leishmania* antigens. I. Th1-like response to *Leishmania major* promastigote antigens in individuals recovered from cutaneous leishmaniasis. *Clin Exp Immunol* 96:410–415
- Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden markov models in computational biology. Applications to protein modeling. *J Mol Biol* 235:1501–1531
- Kulkarni MM, Olson CL, Engman DM, McGwire BS (2009) *Trypanosoma cruzi* gp63 proteins undergo stage-specific differential posttranslational modification and are important for host cell infection. *Infect Immun* 77:2193–2200
- Kurtzhals JA, Hey AS, Jardim A, Kemp M, Schaefer KU, Odera EO, Christensen CB, Githure JI, Olafson RW, Theander TG et al (1994) Dichotomy of the human T cell response to *Leishmania* antigens. II. Absent or Th2-like response to gp63 and Th1-like response to lipophosphoglycan-associated protein in cells from cured visceral leishmaniasis patients. *Clin Exp Immunol* 96:416–421
- LaCount DJ, Gruszynski AE, Grandgenett PM, Bangs JD, Donelson JE (2003) Expression and function of the *Trypanosoma brucei* major surface protease (gp63) genes. *J Biol Chem* 278:24658–24664
- Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Bental N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33:W299–W302
- Launois P, Tacchini-Cottier F, Kiény MP (2008) Cutaneous leishmaniasis: progress towards a vaccine. *Expert Rev Vaccin* 7:1277–1287
- Lipoldova M, Demant P (2006) Genetic susceptibility to infectious disease: lessons from mouse models of leishmaniasis. *Nat Rev Genet* 7:294–305
- McMahon-Pratt D, Alexander J (2004) Does the *Leishmania major* paradigm of pathogenesis and protection hold for new world cutaneous leishmaniasis or the visceral disease? *Immunol Rev* 201:206–224
- O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA (1999) The promise of comparative genomics in mammals. *Science* 286(458–462):479–481

- Osorio LE, Castillo CM, Ochoa MT (1998) Mucosal leishmaniasis due to *Leishmania (viannia) panamensis* in Colombia: clinical characteristics. *Am J Trop Med Hyg* 59:49–52
- Overath P, Haag J, Lischke A, O'HUigin C (2001) The surface structure of trypanosomes in relation to their molecular phylogeny. *Int J Parasitol* 31:468–471
- Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream MA, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, Faulconbridge A, Jeffares D, Depledge DP, Oyola SO, Hilley JD, Brito LO, Tosi LR, Barrell B, Cruz AK, Mottram JC, Smith DF, Berriman M (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet* 39:839–847
- Reiner SL, Locksley RM (1995) The regulation of immunity to *Leishmania major*. *Annu Rev Immunol* 13:151–177
- Retief JD (2000) Phylogenetic analysis using phylip. *Methods Mol Biol* 132:243–258
- Santos AL, Branquinha MH, D'Avila-Levy CM (2006) The ubiquitous gp63-like metalloprotease from lower trypanosomatids: in the search for a function. *An Acad Bras Cienc* 78:687–714
- Schlagenhauf E, Etes R, Metcalf P (1998) The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63). *Structure* 6:1035–1046
- Sharma NL, Mahajan VK, Kanga A, Sood A, Katoch VM, Mauricio I, Singh CD, Parwan UC, Sharma VK, Sharma RC (2005) Localized cutaneous leishmaniasis due to *Leishmania donovani* and *Leishmania tropica*: preliminary findings of the study of 161 new cases from a new endemic focus in Himachal Pradesh, India. *Am J Trop Med Hyg* 72:819–824
- Simpson AG, Roger AJ (2004) The real 'kingdoms' of eukaryotes. *Curr Biol* 14:R693–R696
- Smith DF, Peacock CS, Cruz AK (2007) Comparative genomics: from genotype to disease phenotype in the *Leishmaniases*. *Int J Parasitol* 37:1173–1186
- Sogin ML (1991) Early evolution and the origin of eukaryotes. *Curr Opin Genet Dev* 1:457–463
- Spitzer N, Jardim A, Lippert D, Olafson RW (1999) Long-term protection of mice against *Leishmania major* with a synthetic peptide vaccine. *Vaccine* 17:1298–1300
- Stevens JR, Noyes HA, Schofield CJ, Gibson W (2001) The molecular evolution of *Trypanosomatidae*. *Adv Parasitol* 48:1–56
- Stuart K, Brun R, Croft S, Fairlamb A, Gurtler RE, McKerrow J, Reed S, Tarleton R (2008) Kinetoplastids: related protozoan pathogens, different diseases. *J Clin Invest* 118:1301–1310
- Wilson ME, Jeronimo SM, Pearson RD (2005) Immunopathogenesis of infection with the visceralizing *Leishmania* species. *Microb Pathog* 38:147–160
- Yang Z (1997) Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556
- Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118
- Yao C, Donelson JE, Wilson ME (2003) The major surface protease (MSP or GP63) of *Leishmania* sp. biosynthesis, regulation of expression, and function. *Mol Biochem Parasitol* 132:1–16
- Yao C, Donelson JE, Wilson ME (2007) Internal and surface-localized major surface proteases of *Leishmania* spp. and their differential release from promastigotes. *Eukaryot Cell* 6:1905–1912
- Yao C (2010) Major surface protease of trypanosomatids: one size fits all? *Infect Immun* 78:22–31