Infection, Genetics and Evolution 33 (2015) 37–46

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

# Differences in inferred genome-wide signals of positive selection during the evolution of *Trypanosoma cruzi* and *Leishmania* spp. lineages: A result of disparities in host and tissue infection ranges?

Carlos A. Flores-López [a], Carlos A. Machado [b,*]

[a] Facultad de Ciencias, Universidad Autónoma de Baja California, Km. 103 Carretera Tijuana – Ensenada, Pedregal Playitas, 22860 Ensenada, Baja California, Mexico
[b] Department of Biology, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

*Article history:*
Received 1 November 2014
Received in revised form 19 March 2015
Accepted 9 April 2015
Available online 17 April 2015

*Keywords:*
*Trypanosoma cruzi*
*Leishmania* spp.
Positive selection
Evolutionary genomics
Chagas disease
Vaccine

ABSTRACT

*Trypanosoma cruzi* and *Leishmania* spp. are kinetoplastids responsible for Chagas disease and Leishmaniasis, neglected tropical diseases for which there are no effective methods of control. These two human pathogens differ widely in the range of mammal species they can infect, their cell/tissue tropism and cell invasion mechanisms. Whether such major biological differences have had any impact on genome-wide patterns of genetic diversification in both pathogens has not been explored. The recent genome sequencing projects of medically important species of *Leishmania* and *T. cruzi* lineages provide unique resources for performing comparative evolutionary analyses to address that question. We show that inferred genome-wide signals of positive selection are higher in *T. cruzi* proteins than in *Leishmania* spp. proteins. We report significant differences in the fraction of protein-coding genes showing evidence of positive selection in the two groups of parasites, and also report that the intensity of positive selection and the proportion of sites under selection are higher in *T. cruzi* than in *Leishmania* spp. The pattern is unlikely to be the result of confounding factors like differences in GC content, average gene length or differences in reproductive mode between the two taxa. We propose that the greater versatility of *T. cruzi* in its host range, cell tropism and cell invasion mechanisms may explain the observed differences between the two groups of parasites. Genes showing evidence of positive selection within each taxonomic group may be under diversifying selection to evade the immune system and thus, depending on their functions, could represent viable candidates for the development of drugs or vaccines for these neglected human diseases.



## 1. Introduction

Natural selection is the most important evolutionary force driving the diversification of all living organisms. Comparative and population genetic analyses of orthologous DNA sequences are routinely used for inferring the influence of natural selection in shaping levels and patterns of intra and interspecific nucleotide divergence and diversity in natural populations (Nielsen et al., 2005; Oleksyk et al., 2010; Fay, 2011). In comparative studies of orthologous protein-coding sequences, the action of natural selection during the divergence process is inferred from the value of the ratio of non-synonymous (dN) to synonymous (dS) substitutions (dN/dS or $\omega$). Proteins or sections of proteins under purifying (negative) selection (i.e. selectively constrained) have dN/dS <1, while proteins or sections of proteins that have experienced positive selection have dN/dS >1. However, as usually only a small proportion of codon sites are evolving under positive selection, averaging dN/dS over an entire protein is a very conservative approach for inferring natural selection. Consequently, methods that test for codon sites evolving under positive selection are more powerful and accurate (Yang and Swanson, 2002; Swanson et al., 2003).

Immune system elicitors or antigens from pathogens that evolve rapidly to avoid recognition from the host immune system constitute good examples of protein evolution driven by natural selection (Frank, 2002). With the availability of full genome sequences for multiple pathogen species, bioinformatics and evolutionary analyses focused on the use of dN/dS ratios have become powerful approaches for identifying proteins evolving under positive selection that may eventually become candidate genes for parasite control (Petersen et al., 2007; Soyer et al., 2009; Gu et al., 2011; Xu et al., 2011; Zhang et al., 2011; McCann et al., 2012).

* Corresponding author. Tel.: +1 (301) 405 9447.
*E-mail address:* machado@umd.edu (C.A. Machado).

http://dx.doi.org/10.1016/j.meegid.2015.04.008
1567-1348/

Chagas disease and Leishmaniasis are neglected tropical diseases for which vaccines are yet to be developed (Hotez et al., 2007; Bethony et al., 2011), making them excellent candidates for conducting *in silico* searches for protein-coding genes that can be targeted for vaccine or drug development. *Trypanosoma cruzi* is the etiological agent of Chagas Disease; it infects approximately 8–18 million people, and kills about 20 thousand people every year in Latin America (WHO, 2002; Rassi et al., 2010). Leishmaniasis, on the other hand, is caused by more than 20 species of the genus *Leishmania* and has a much wider geographic range, occurring in 88 different countries from Latin America, Asia, Africa and Europe. It is estimated that 12 million people are infected with *Leishmania*, with approximately 20,000–40,000 deaths per year (Desjeux, 2001; Alvar et al., 2012).

*Trypanosoma cruzi* and *Leishmania sp.* are protozoans that belong to the class Kinetoplastea, an early-diverged branch in the eukaryotic tree of life (Simpson et al., 2006). Members of this eukaryotic class have unique characteristics not found in other eukaryotes (Schmidt and Roberts, 2005) like polycistronic mRNA modification, uracil insertion modification of mRNA, and the presence of an enlarged mitochondria (i.e. kinetoplast) with a unique chromosomal architecture composed of a few mega circles and thousands of concatenated mini circles. Both parasites have evolved to be digenetic and have independently evolved intracellular evasion mechanisms within their mammal hosts (Sibley, 2011). Further, both groups of parasites are thought to have originated in South America (Yeo et al., 2005; Lukes et al., 2007; Zingales et al., 2012), even though they have now different geographic ranges and have adapted to very different insect vectors and mammal hosts. Sylvatic *Leishmania* has been found mostly in rodents, dogs, foxes and jackals, whereas *T. cruzi* has a much wider range of mammal hosts (>100 hosts), including opossums, armadillos, primates, raccoons, rodents, bats, dogs and humans (Schmidt and Roberts, 2005; Ready, 2013). However, one of the most dramatic differences between the two taxa is their contrasting cell/tissue tropism. *Leishmania* strictly invade macrophages and dendritic cells in the vertebrate host, although short-lived neutrophils are also targeted during the early invasion process (Liu and Uzonna, 2012). On the other hand, *T. cruzi* can invade a much wider array of cell types (e.g. myocytes, macrophages, cardiomyocytes, nerve cells, etc.) (Fernandes and Andrews, 2012; Moradin and Descoteaux, 2012), and in fact appears to be able to invade any type of cell *in vitro* (Manso-Alves and Arruda Mortara, 2009).

Whether those major differences between *T. cruzi* and *Leishmania* spp. in mammal host diversity, cell/tissue tropism and, possibly, cell invasion mechanisms, have had any impact on patterns of genetic diversification genome-wide in both pathogens has yet to be explored. Here we present evolutionary analyses of protein-coding genes from five genomes of *T. cruzi* spp. and four genomes of *Leishmania* spp. Because *T. cruzi* has a broader host range and can invade a wider array of cell types than *Leishmania* spp. we hypothesize that *T. cruzi* proteins should show more evidence of adaptive evolution than *Leishmania* spp. proteins. Genes showing evidence of positive selection within each taxonomic group may be under diversifying selection to evade the immune system and thus, depending on their functions, could represent viable candidates for the development of methods of parasite control.

## 2. Materials and methods

### 2.1. Sequence data sets

We analyzed all available annotated genome sequences from *T. cruzi* and *Leishmania* spp. available in TriTrypDB release 7.0 (http://tritrypdb.org/tritrypdb/). As of January 2014 four *T. cruzi* genomes as well as the genome of the bat infecting species *T. cruzi marinkellei*, a close relative of *T. cruzi*, had been sequenced and annotated. The four *T. cruzi* genomes represent 3 of the 6 distinct typing units (DTUs) or lineages (TcI–TcVI) in which the genetic diversity of *T. cruzi* is currently divided (Zingales et al., 2009). They include the most divergent set of lineages of this taxon and thus provide a good representation of genetic divergence within *T. cruzi*. Because *T. cruzi* has a mainly clonal mode of reproduction (Tibayrenc et al., 1986; Tibayrenc and Ayala, 1988) *T. cruzi* DTUs do correspond to genetically isolated entities akin to species. Therefore, despite the difference in taxonomic nomenclature between the two groups we analyze here, the evolutionary divergence among *T. cruzi* DTUs is equivalent to the divergence among *Leishmania* species, making the comparison appropriate.

Two of the *T. cruzi* sequenced genomes (haplotypes Esmeraldo (TcII) and Non-Esmeraldo (TcIII)) were obtained during the sequencing of the genome strain of *T. cruzi* CL Brener (El-Sayed et al., 2005). This strain is a hybrid (Machado and Ayala, 2001, 2002; Brisse et al., 2003), the result of a hybridization event between two divergent lineages that took place ~0.4–0.8 million years ago (Flores-Lopez and Machado, 2011). In our analyses we only use sequences from the parental lineages (Esmeraldo and Non-Esmeraldo) that gave rise to the hybrid (El-Sayed et al., 2005), therefore avoiding the use of any sequences that could have recombined after the hybridization event. We also included two additional genome sequences from lineage TcI, *T. cruzi* Sylvio X10/1 (Franzen et al., 2011) and *T. cruzi* JRcl4 (unpublished but available in TriTrypDB), as well as the genome of the bat infecting *T. cruzi marinkellei* strain B7 (Franzen et al., 2011). The number of annotated protein coding genes in each sequenced genome is the following: Non-Esmeraldo (10,834), Esmeraldo (10,342), JRcl4 (7755), Sylvio (7456) and *T. cruzi marinkellei* (10,342).

The genetic diversity of *Leishmania* spp. has been divided into more than 30 species (Schmidt and Roberts, 2005). Five *Leishmania* species associated with humans have had their genomes sequenced: *L. major*, responsible for cutaneous leishmaniasis (CL) in the old world, *L. mexicana* and *L. braziliensis*, both of which cause CL in the new world, and *L. infantum* and *L. donovani* which cause visceral leishmaniasis (VL). *L. major* (strain Friedlin) was the first *Leishmania* species sequenced (Ivens et al., 2005), followed by *L. infantum* and *L. braziliensis* (Peacock et al., 2007). Annotated genome sequences of *L. mexicana* are not published but available in TriTrypDB. The annotated genome sequence of *L. donovani* is available in GeneDB (http://www.genedb.org/). However, this species has very low genetic divergence with *L. infantum* (~0.48% average nucleotide divergence) and for that reason it was not included in this study. The number of annotated protein coding genes in each of the 4 *Leishmania* genomes we included is the following: *L. major* (8408), *L. infantum* (8241), *L. braziliensis* (8357) and *L. mexicana* (8250).

### 2.2. Ortholog data sets

Reciprocal best hit blastx searches were conducted to find true orthologs within each taxonomic group. A blastx *E*-value of $10^{-5}$ was used as threshold for the orthologous search similarity criteria. The approach to identify orthologs using reciprocal best hit blastx searches is conservative due to its low false positive rate and medium false negative rate (Chen et al., 2007). Within each taxonomic group we conducted reciprocal best-hit blastx searches for all possible pairwise strain comparisons, and selected ortholog pairs that matched across all pairwise comparisons. This conservative approach filtered out almost all proteins that form part of the largest protein families found in *T. cruzi* (i.e. trans-sialidases, mucins, MASP, surface glycoprotein gp63 protease) due to large sequence similarities among protein members of these large gene families.

The final datasets of putative orthologs consisted of 3893 protein-coding genes in *T. cruzi* and 7439 protein-coding genes in *Leishmania* spp. *T. cruzi* strains were highly dissimilar in terms of the number of predicted protein-coding genes in each genome compared to *Leishmania* spp. The CL-Brener haplotypes had an average of 10,588 protein-coding genes per haplotype, similar to the 10,342 genes in *T. cruzi marinkellei*, while the TcI strains (Sylvio & JRcl4) had an average of 7605 protein-coding genes per strain. The first *T. cruzi* genome sequenced study noted that approximately 50% of the predicted protein-coding genes were members of few very large protein families (El-Sayed et al., 2005). The significantly smaller number of protein-coding genes predicted in the TcI strains is mostly due to differences in the copy number of these large protein families (mostly trans-sialidases, mucins, MASP and gp63s protein families) (Franzen et al., 2011). In contrast to *T. cruzi*, the number of predicted protein-coding genes in the genomes of *Leishmania* was very similar across species. The largest difference in number of protein-coding genes annotated among *Leishmania* species was 167, compared to approximately 3000 among *T. cruzi* strains. The average *Leishmania* genome contained 8314 predicted protein coding-genes, from which an ortholog data set of 7439 protein coding-genes was constructed.

### 2.3. Alignment and selection analyses

Sequences from each ortholog data set were translated to amino acids with translatorx3.pl (Abascal et al., 2010) and aligned with MUSCLE (Edgar, 2004). Aligned data sets were back translated into nucleotides for the selection analyses in PAML (Yang, 2007) (see below). Poorly aligned regions were removed using Gblocks (Castresana, 2000). Removing poorly aligned regions is a conservative approach that has recently been shown to outweigh the costs of removing true positively selected sites from the analyses (Privman et al., 2011). Thus, the true number of positively selected sites and/or proteins in our data set might actually be larger than what is presented here. Concatenated data sets of ~1.75 million base pairs of aligned protein-coding sequence for each group of taxa were used to reconstruct phylogenetic relationships among the *T. cruzi* or *Leishmania* genomes using Maximum likelihood in PhyML (Guindon et al., 2010) (Fig. 1). The DNA substitution models used in the phylogenetic reconstructions were selected for each taxonomic data set using jModeltest (Posada, 2008) (GTR + G for *T. cruzi* spp. and GTR + I for *Leishmania* spp.). The phylogenies were then input to PAML for conducting the positive selection analyses independently for each taxonomic dataset using site models that incorporate heterogeneity across sites in the estimates of the dN/dS ratio ($\omega$) and that use phylogenetic information. These tests do not rely on pairwise estimates of $\omega$ and have been shown (Dos Reis and Yang, 2013) to be less prone to biases caused by the observed dependency between genetic distance and $\omega$ in pairwise comparisons (Rocha et al., 2006; Wolf et al., 2009). Pairs of nested models, M7 (beta) versus M8 (beta & $\omega$) and M8 versus M8a (beta & $\omega = 1$), were compared using a likelihood ratio test (LRT). Significance of the LRT for M7 vs M8 was determined using 2 degrees of freedom. Since model M8a is not completely nested within M8, significance was determined by halving the $p$ value from a chi-square test with one degree of freedom as previously suggested (Yang, 2007). As we conducted multiple comparisons, the significance of each $p$-value was corrected using the false discovery rate method of Benjamini and Hochberg (Benjamini and Hochberg, 1995) with the p.adjust() function in *R*. To evaluate the effect of presumed differences in reproductive mode between the two taxa (*Leishmania* seems to undergo more sexual reproduction than *T. cruzi*), dN/dS values were compared for groups of genes with housekeeping and hypothetical functions to determine if there is evidence of relaxed purifying selection in *T. cruzi* (See Section 3).

### 2.4. Functional overrepresentation analyses

Functional annotations associated with *L. major* and the *T. cruzi* Non-Esmeraldo haplotype were used to determine if any molecular or biological functions were overrepresented among genes showing evidence of positive selection. Unfortunately, between 50% and 68% of the protein coding genes in *T. cruzi* and *L. major* have unknown functions based on genome annotations, and only 10% (~500 proteins) of *T. cruzi* proteins and 42% (3525 proteins) in *L. major* have a Gene Ontology term associated to them (www.genedb.org). Consequently, to conduct a more comprehensive functional over representation analysis, we additionally clustered all our protein data into Pfam clans (pfam.sanger.ac.uk/), a broader classification scheme that groups evolutionary related protein families into clans based on related structure, related function, and significant matching of the same sequence to databases of hidden Markov Models (HMMs) from different families and profile–profile comparisons (Finn et al., 2006). Clustering our data set into Pfam clans allowed us to include many more proteins into the functional overrepresentation analyses. Statistically overrepresented Pfam clans were identified with GeneMerge (Castillo-Davis and Hartl, 2003). All statistical overrepresentation tests were conducted with the protein lists predicted to be under positive selection from the M8 versus M8a model comparison.

### 2.5. Hypothetical clustering

Genes with unknown function (i.e. hypothetical genes) were grouped into protein clusters by conducting a self blastp search ($E$ value < $10^{-10}$) of all the hypothetical proteins from the Non-Esmeraldo genome or from the *L. major* genome. A numerical code (e.g. protein family 1, 2, 3, etc.) was given to all clusters, including proteins from clusters of size 1. GeneMerge was used to determine statistical overrepresentation of protein clusters that had evidence of sites under positive selection (based on the comparison of models M8 and M8a).

## 3. Results

### 3.1. More evidence of adaptive protein evolution in T. cruzi than in Leishmania spp.

Fig. 1 shows the phylogenetic relationships among the *Leishmania* spp. and *T. cruzi* genomes used in this study, reconstructed using maximum likelihood. The overall level of sequence divergence ($p$-distance, averaged across orthologous genes) among the *T. cruzi* genomes ranged between 0.015 and 0.086 (Table S1). The overall level of sequence divergence among *Leishmania* genomes ranged between 0.053 and 0.175 (Table S1). In *Leishmania* spp. our analyses identified 78 and 170 genes with evidence of positive selection using, respectively, the M8 vs M8a or M7 vs M8 tests of the codon site models, representing 1.0% or 2.3% of the 7439 orthologous gene data set from *Leishmania* (Table 1, $p < 0.01$, Tables S2–S3). In contrast, in *T. cruzi*, a total of 402 and 451 protein-coding genes presented evidence of positive selection using, respectively, the M8 vs M8a or M7 vs M8 tests of the codon site models, representing 10.33% or 11.59% of the 3893 *T. cruzi* genes used in this study (Table 1, $p < 0.01$, Tables S4–S5). The number of protein-coding genes showing evidence of positive selection is significantly higher in *T. cruzi* than in *Leishmania* spp. (M8 vs

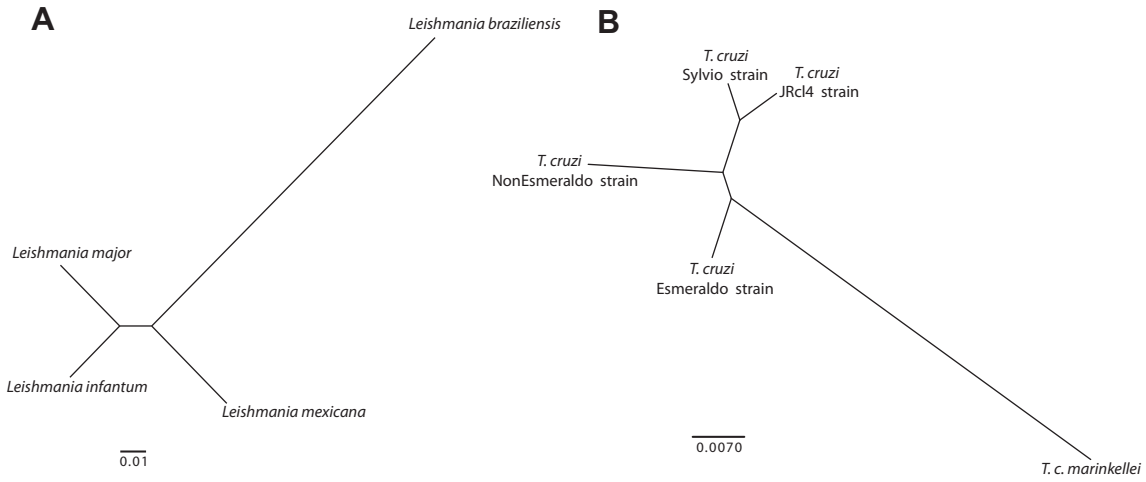


**Fig. 1.** Phylogenetic relationships of the *Leishmania* spp. or *Trypanosoma cruzi* genomes used in this study. The unrooted maximum likelihood phylogenies used in the PAML analyses are shown for (A) *Leishmania* spp. (B) *T. cruzi*. In each case, the phylogeny was reconstructed using a concatenated data set of 1.75 million base pairs of aligned coding sequence. The GTR substitution model was used in both analyses.

**Table 1**
Differences in the number of proteins under positive selection between the two lineages. *T. cruzi* taxa set consists of Non-Esmeraldo and Esmeraldo haplotypes, Sylvio, Jrcl4 strains (all *T. cruzi cruzi* strains) and *T. cruzi marinkellei*. The *Leishmania* spp. taxa set consists of *L. braziliensis*, *L. mexicana*, *L. major* and *L. infantum*. The conserved section represents the number of proteins with no evidence of positive selection by the LRT of model M7 vs M8 and M8 vs M8a in PAML with $p < 0.01$ and $p < 0.05$.

| Taxonomic group | *T. cruzi* | | *Leishmania* spp. | |
|---|---|---|---|---|
| Evolutionary pressure | Conserved (%) | Positive selection (%) | Conserved (%) | Positive selection (%) |
| M7 vs M8 ($p < 0.01$) | 3442 (88.41%) *3680 (94.52%) | 451 (11.59%) *213 (5.48%) | 7269 (97.7%) *7430 (99.8%) | 170 (2.3%) *9 (0.12%) |
| M7 vs M8 ($p < 0.05$) | 3117 (80.06%) *3505 (90.03%) | 776 (19.94%) *388 (9.97%) | 7030 (94.5%) *7418 (99.7%) | 409 (5.5%) *21 (0.28%) |
| M8 vs M8a ($p < 0.01$) | 3491 (89.67%) *3716 (95.45%) | 402 (10.33%) *177 (4.55%) | 7361 (99%) *7431 (99.89%) | 78 (1.04%) *8 (0.1%) |
| M8 vs M8a ($p < 0.05$) | 3179 (81.65%) *3558 (91.4%) | 714 (18.34%) *335 (8.6%) | 7218 (97.02%) *7427 (99.84%) | 221 (2.97%) *12 (0.16%) |

* Numbers based on false discovery rate corrected *q*-values.

M8a: $\chi^2 = 542.29$, $p = 5.98 \times 10^{-120}$; M7 vs M8: $\chi^2 = 426.69$, $p = 8.53 \times 10^{-95}$) (Table 1, Fig. 2). Furthermore, the average dN/dS ratio in sites with dN/dS >1 in proteins showing significant evidence of positive selection (Fig. 3) and the proportion of codon sites under selection in those proteins (Fig. 4) are significantly higher in *T. cruzi* than in *Leishmania* spp. (Wilcoxon rank-sum test, $p < 0.0001$).

To examine the effect of removing highly divergent codons from the alignment we performed the *T. cruzi* ortholog alignments without alignment edition by Gblocks (Castresana, 2000), which removes highly divergent aligned regions. It is important to note that recent analyses (Privman et al., 2011) show that the benefits of using aligner filters in studies of this nature outweigh the costs of removing true positively selected sites from the analyses. Without the Gblocks alignment edition, the number of genes predicted to be under positive selection in *T. cruzi* spp. increases from 776 to 954 using the M7 vs M8 model comparison and from 714 to 892 using the M8 vs M8a comparison, confirming that our reported results are conservative. Similar results were observed in *Leishmania* spp., and thus the number of proteins predicted to be under positive selection in *T. cruzi* remained larger than those predicted in *Leishmania* spp. (not shown).

Low divergence at synonymous sites (dS) could generate false positives with high dN/dS particularly in pairwise sequence comparisons (Rocha et al., 2006; Wolf et al., 2009; Dos Reis and Yang, 2013). However, *T. cruzi* proteins under positive selection had a significantly higher divergence at synonymous sites compared to proteins under purifying selection (Wilcoxon rank-sum test, $p = 0.0144$), whereas the *Leishmania* spp. proteins under positive selection did not show any significant difference with those under purifying selection ($p = 0.38$). Furthermore, the distribution of dN/dS values for all proteins when comparing the most divergent lineages of *Leishmania* spp. (*L. braziliensis vs L. mexicana*) or *T. cruzi* (*T. cruzi Sylvio* strain vs *T. cruzi marinkellei*) is still significantly different between the two taxa (Wilcoxon rank-sum text, $p < 0.0001$) (Fig. 5). Therefore, it is unlikely that the results we report are caused by artifacts generated by low divergence at synonymous sites in the *T. cruzi* dataset.

Differences in GC content do not affect the performance of the codon models used (Zhai et al., 2012), and we saw no difference in GC content of genes with different evidence of selection (Table S7). The only sequence characteristic that has been shown to have an effect on the performance of the codon models used is gene length (Zhai et al., 2012), as there is more power to detect selection in longer genes (Anisimova et al., 2001). We in fact observed the same trend of significantly longer gene length in the set of positive selected genes and those under purifying selection both in *T. cruzi* (M7 vs M8: 2158 vs 1599, $p < 0.001$; M8 vs M8a: 2195 vs 1603, $p < 0.001$) and *Leishmania* spp. (M7 vs M8: 2280 vs 1998, $p = 0.015$; M8 vs M8a: 2458 vs 1999, $p = 0.024$) (Table S7). However, the protein coding genes in the *Leishmania* spp. data set were significantly longer than those of *T. cruzi* spp. ($p = 9.34 \times 10^{-13}$). Given that more selection was detected in the dataset with the shorter gene length, it is unlikely that this factor has influenced the overall results.

Differences in reproductive mode (asexual vs sexual) can be problematic for interpreting differences in patterns of adaptive evolution. Theory suggests that asexual lineages should undergo a relaxation of purifying selection with the consequence of an

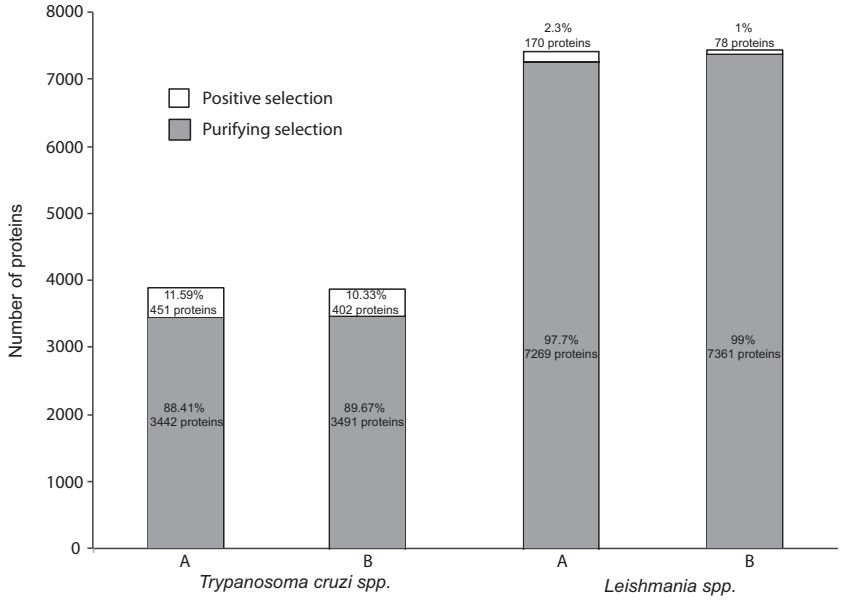


**Fig. 2.** Difference in the number of proteins showing evidence of positive selection within each taxa. (A) Data for codon models M7 versus M8 ($p < 0.01$). (B) Data for codon models M8 versus M8a ($p < 0.01$). The observed difference is highly significant (M8 vs M8a: $\chi^2 = 693.33$, $p = 8.43 \times 10^{-153}$; M7 vs M8: $\chi^2 = 503.86$, $p = 1.37 \times 10^{-111}$).

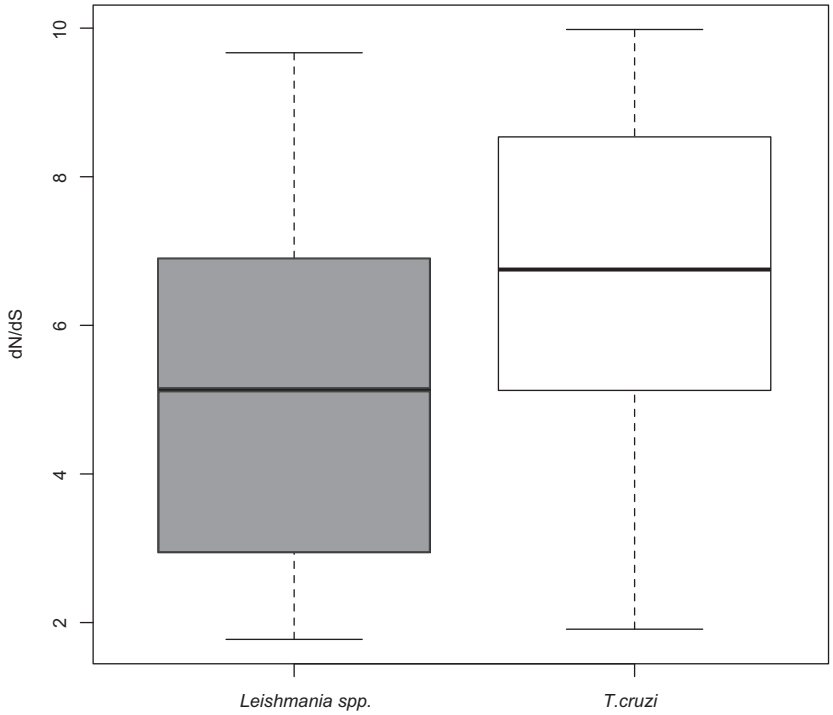


**Fig. 3.** dN/dS for sites predicted to be evolving under positive selection. Boxplot of dN/dS values in sites evolving under positive selection in all the proteins showing evidence of adaptive evolution using the M8 vs M8a model tests ($p < 0.05$). The two distributions are significantly different (Wilcoxon rank-sum test, $p < 0.0001$). dN/dS values higher than 10 were removed from the analysis (differences between species remained highly significant without the removal of outliers, $p < 0.0001$). Boxes encompass the lower and upper quartiles, with the internal line representing the median and whiskers extending to the 2.5th and 97.5th percentiles.

overall increase of Ka/Ks ratios relative to sexually reproducing lineages due to the faster accumulation of deleterious mutations expected in smaller populations (Charlesworth and Wright, 2001; Glemin, 2007), although empirical results have not always been consistent with that prediction (Paland and Lynch, 2006; Barraclough et al., 2007; Henry et al., 2012; Ollivier et al., 2012). It is widely accepted that *T. cruzi* reproduces mostly asexually (Tibayrenc et al., 1986; Tibayrenc and Ayala, 2002) although there is strong evidence that rare sexual recombination events have occurred (Machado and Ayala, 2001; Brisse et al., 2003; Flores-Lopez and Machado, 2011), a finding consistent with the observed capacity of this parasite to undergo genetic exchange in the lab (Gaunt et al., 2003). Although there has been more controversy about the reproductive mode of *Leishmania* species (Tibayrenc and Ayala, 2013), recent studies on *L. donovani/infantum* (Rogers et al., 2014), *L. major* (Akopyants et al., 2009; Inbar et al., 2013)

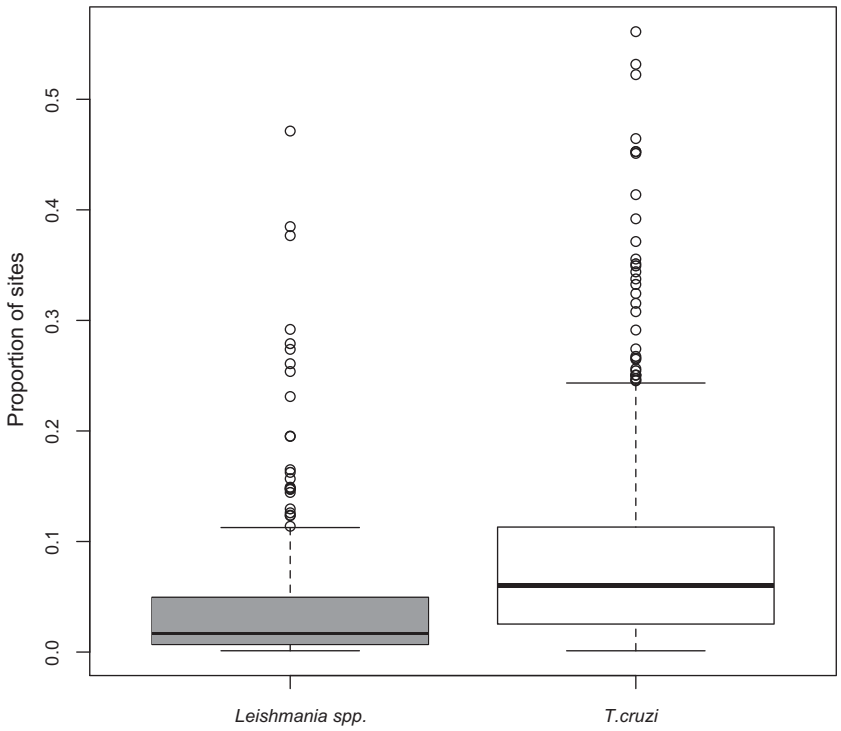


**Fig. 4.** Proportion of sites predicted to be under positive selection for all proteins showing evidence of adaptive evolution under codon models M8 vs M8a ($p < 0.05$). The two distributions are significantly different (Wilcoxon rank-sum test, $p < 0.0001$). Boxes encompass the lower and upper quartiles, with the internal line representing the median and whiskers extending to the 2.5th and 97.5th percentiles.

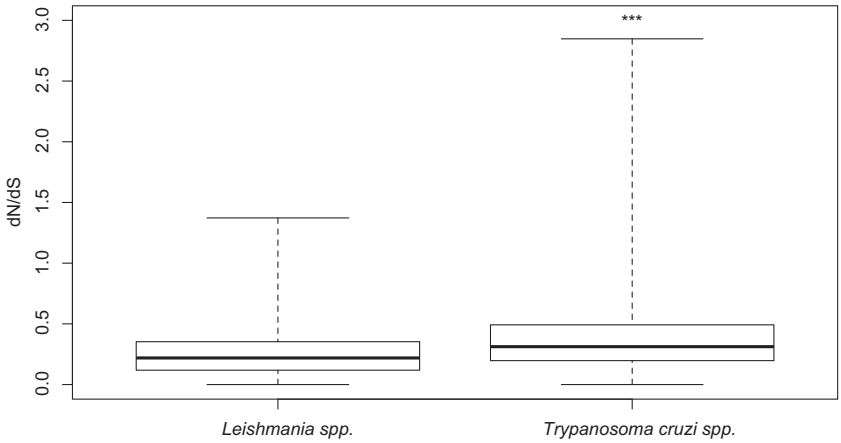


**Fig. 5.** Comparison of dN/dS values for all protein coding genes. dN/dS values were estimated between the two most divergent lineages of *Leishmania* spp. (*L. braziliensis vs L. mexicana*) or *T. cruzi* (*T. cruzi* Sylvio vs *T. cruzi marinkellei*), and dN/dS values higher than 3 were removed from the analyses. The two distributions are significantly different between taxa ($p = 3.25 \times 10^{-62}$). If outliers with dN/dS >3 are not removed the difference between species remains highly significant ($p < 0.0001$). Boxes encompass the lower and upper quartiles, with the internal line representing the median and whiskers extending to the 2.5th and 97.5th percentiles.

and *L. braziliensis* (Rougeron et al., 2009) suggest that *Leishmania* alternate clonal propagation with sexual reproduction. Given that *Leishmania* seems to undergo more sexual reproduction than *T. cruzi* (acknowledging that we have no information for *T. cruzi marinkellei*) we hypothesize that if our reported results are due to reduced levels of genome-wide purifying selection rather than increased adaptive evolution in *T. cruzi* we should observe evidence of relaxed purifying selection in housekeeping genes in this taxon. For each taxon we compared dN/dS values from 47 genes that had the same housekeeping functions in both taxa (e.g. DNA and RNA polymerases, Ribosomal proteins, histones, cytochromes, kinases, helicases) with dN/dS values of approximately 300 genes encoding hypothetical proteins (Table S6). The latter set of genes are expected to have lower selective constraints than housekeeping genes and were chosen at random from the large suite of annotated hypothetical genes from each taxon. The dN/dS values of housekeeping genes were significantly lower than hypothetical genes for both *T. cruzi* spp. ($p = 3.07 \times 10^{-25}$; housekeeping genes mean dN/dS = 0.223, hypothetical genes mean dN/dS = 0.748) and *Leishmania* spp. ($p = 8.5 \times 10^{-15}$; housekeeping genes mean dN/dS = 0.167, hypothetical genes mean dN/dS = 0.387) (Fig. 6). The difference in mean dN/dS values between housekeeping and hypothetical genes is thus much greater in *T. cruzi* spp. than in *Leishmania* spp. suggesting that purifying selection is still playing a very important role in *T. cruzi* protein evolution. That conclusion is reinforced by the additional observation that while the

difference between both taxa in dN/dS values for housekeeping genes is barely significant ($p = 0.03$) it is highly significant for hypothetical genes ($p = 1.78 \times 10^{-57}$). The results are also consistent with a previous detailed population genetic study that also found evidence of purifying selection in two housekeeping genes of *T. cruzi* (Machado and Ayala, 2002).

### 3.2. Functional overrepresentation of genes under positive selection

In *Leishmania* spp., 4 functional categories were statistically overrepresented in the genes showing evidence of positive selection ($p < 0.01$) (Table 2). The most significant was glutathione peroxidase, followed by ATP binding cassette, iron superoxide dismutase and cysteine peptidase. Interestingly, proteins with those functions have been shown to play some role in the evolution of drug resistance (ATP binding cassette) (Purkait et al., 2012), in the interaction with the host immune system (Cysteine peptidase) (Mottram et al., 2004; Alexander and Bryson, 2005), or have been proposed as vaccine candidates for Leishmaniasis (iron superoxidase dismutase) (Daifalla et al., 2012) (see Section 4).

In *T. cruzi*, there were only 2 over represented functions among those showing evidence of positive selection ($p < 0.01$): genes with hypothetical functions (discussed below) and genes with mucin function (Table 2). The latter is an important finding given the important role played by mucins in protecting the parasite from both vector and mammal host defense mechanisms and their role in guaranteeing an anchorage point during the parasite invasion process (Buscaglia et al., 2006; De Pablos and Osuna, 2012).

### 3.3. Hypothetical genes under positive selection

Although the ortholog dataset of *T. cruzi* had a smaller number of hypothetical proteins than *Leishmania* spp. (2796 vs 5054), a significantly higher proportion of hypothetical proteins were predicted to be under positive selection in *T. cruzi* (M8 vs M8a: 572 vs 140, $\chi^2 = 547.07$, $p = 5.46 \times 10^{-121}$; M7 vs M8: 585 vs 242, $\chi^2 = 388.02$, $p = 2.23 \times 10^{-86}$). In an attempt to increase the power of the analysis of proteins with unknown function, all those proteins were clustered based on sequence similarity. Of the 2796 hypothetical proteins within the *T. cruzi* ortholog data set, 461 clusters were formed after the sequence similarity cluster criteria ($E$ value $< 10^{-10}$). In the *Leishmania* spp. ortholog data set 830 clusters were formed. These results show, as expected, that most hypothetical proteins are not part of large protein families, reducing the power of functional overrepresentation analyses. Of the clusters of hypothetical proteins recovered, only 3 clusters in *T. cruzi* and 2 clusters in *Leishmania* spp. were overrepresented among those under positive selection (Table 3).

**Table 2**
Functional overrepresentation of genes showing evidence of positive selection.

| Predicted function | *N* | *n* | *p* | Gene names |
|---|---|---|---|---|
| *Trypanosoma cruzi* | | | | |
| Hypothetical | 2792 | 585 | 0.0072 | See Supp. Mat. |
| Mucin associated surface protein (MASP) | 3 | 3 | 0.0079 | Tc00.1047053511839.20<br>Tc00.1047053506815.20<br>Tc00.1047053507071.180 |
| *Leishmania* spp. | | | | |
| Glutathione peroxidase & synthetase | 2 | 2 | 0.00043037 | LmjF.36.3010<br>LmjF.14.0910 |
| ATP-binding cassette protein subfamily A, D & G | 34 | 5 | 0.00062779 | LmjF.11.1270<br>LmjF.11.1290<br>LmjF.27.0970<br>LmjF.33.1860<br>LmjF.06.0090 |
| Iron superoxide dismutase | 5 | 2 | 0.0041293 | LmjF.32.1820<br>LmjF.32.1830 |
| Cysteine peptidase | 7 | 2 | 0.0084367 | LmjF.29.0820<br>LmjF.19.1420 |

*Note: N*: number of proteins of this function in ortholog data set. *n*: number of proteins of this function under positive selection under model M8 vs M8a. *p*: statistical significance estimated from GeneMerge. Gene codes are the gene codes for Non-Esmeraldo (*T. cruzi*) and *Leishmania major* found in Tritrypdb.org.

## 4. Discussion

We show that we can infer more evidence of positive selection in *T. cruzi* proteins than in *Leishmania* spp. proteins. Because *T. cruzi* has a mainly clonal mode of reproduction (Tibayrenc et al., 1986; Tibayrenc and Ayala, 1988) *T. cruzi* DTUs do correspond to genetically isolated entities akin to species. Thus, the evolutionary divergence among *T. cruzi* DTUs is equivalent to the divergence among *Leishmania* species, making the comparison appropriate. First, we report that a significantly larger fraction of protein-coding genes show evidence of positive selection in *T. cruzi* than in *Leishmania* spp. (Table 1, Fig. 2). Furthermore, we report that the average dN/dS of sites under positive selection and the proportion of sites under positive selection in genes showing a signal of positive selection are significantly higher in *T. cruzi* than in *Leishmania*

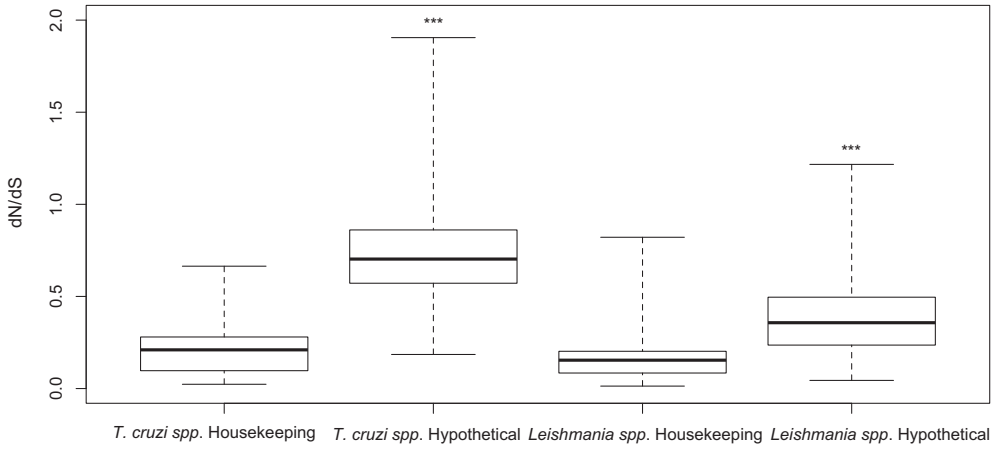


**Fig. 6.** Comparison between dN/dS values of housekeeping and hypothetical genes in *T. cruzi* spp. and *Leishmania* spp. dN/dS values were extracted from the output files of model M8 (see methods) and represent the average dN/dS of the estimated parameter from every branch of the tree. *** Significant difference between the average dN/dS values of housekeeping genes and hypothetical genes in *T. cruzi* spp. ($p = 3.07 \times 10^{-25}$) or *Leishmania* spp. ($p = 8.5 \times 10^{-15}$).

**Table 3**
Statistically overrepresented hypothetical protein clusters under positive selection in *T. cruzi* and *Leishmania* spp.

| Protein cluster | Contributing proteins | Population fraction | Study fraction | *p*-Value | *p*-Value (BF)* | dN/dS | % Sites dN/dS > 1 | Ov. expr | Syntenic | % Identity within cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| *T. c.* 1 | Tc00.1047053506559.20 | 2/7079 | 2/479 | 0.0045696 | 0.39299 | 25.44 | 0.014 | Am | Chr34 | 27% |
| | Tc00.1047053509039.10 | | | | | 12.64 | 0.034 | Epi | Chr32 | |
| *T. c.* 2 | Tc00.1047053507031.130 | 3/7079 | 2/479 | 0.013093 | 1 | 6.12 | 0.063 | NA | Chr40 | 39.25% |
| | Tc00.1047053509569.140 | | | | | 10.10 | 0.092 | NA | Chr12 | |
| *T. c.* 3 | Tc00.1047053505999.170 | 9/7079 | 3/479 | 0.019016 | 1 | 30.08 | 0.005 | NA | Chr9 | 48.25% |
| | Tc00.1047053508299.30 | | | | | 24.15 | 0.116 | NA | Chr38 | |
| | Tc00.1047053511287.4 | | | | | 21.38 | 0.023 | NA | Chr40 | |
| *L.* spp. 1 | LmjF.09.1020 | 3/6031 | 2/102 | 0.000840 | 0.08404 | 10.84 | 0.00557 | | Chr9 | 33% |
| | LmjF.32.0510 | | | | | 5.91 | 0.03476 | | Chr32 | |
| *L.* spp. 2 | LmjF.29.1500 | 17/6031 | 2/102 | 0.032657 | 1 | 6.37 | 0.01263 | | Chr29 | 32.25% |
| | LmjF.11.0670 | | | | | 392.16 | 0.00450 | | Chr11 | |

*Note:* dN/dS: Average dN/dS for sites with dN/dS > 1 from Model8 in PAML. % sites dN/dS > 1: estimated from the M8 model implemented in PAML. Ov. Expr. NA: Data Non available. Am: Over-expressed in amastigote stage. Epi: over expressed in epimastigote stage.
* Bonferroni corrected *p*-values.

spp. (Fig. 3). What makes our results more startling is the fact that the majority of surface protein families of *T. cruzi*, which make up almost 50% of the genes in its genome were unavoidably filtered out of our gene data set due to the strict orthologous gene criteria that was used. Given the immune related function of those protein families our results are indicative that the number of proteins that have been under positive selection in *T. cruzi* should be larger than the number reported here.

The reported results are not artifacts of the analyses. Removing or not removing highly divergent regions using GBlocks does not affect the overall results. Further, low levels of sequence divergence at silent sites have not influenced the inference of adaptive evolution. Although the overall level of divergence was smaller among *T. cruzi* strains (*p*-distance: 0.015–0.086) than among *Leishmania* species (*p*-distance: 0.053–0.175), a similar comparative genomic study in primates observed the opposite trend that we report: significantly smaller number of proteins were predicted to be under positive selection with the less divergent sample of primate species (George et al., 2011). Moreover, we show that the pattern is not the result of possible differences in reproductive mode between the two taxa, or differences in GC content or gene length.

We propose that the larger number of hosts mammal species, the larger number of target cells and tissues it can infect, and the more diverse intracellular invasive developmental stages of *T. cruzi* are the underlying reasons behind the observed difference in the number of proteins inferred to have experienced adaptive evolution in the two taxa. Those major differences in the biology of the two groups of parasites should influence the level of interaction between the parasite, surface receptors of host target cells and/or the immune system of their different hosts, and can therefore influence patterns of protein evolution in immune elicitors of the parasite and in proteins involved in cell invasion. Species from the genus *Leishmania* are known to almost exclusively depend on macrophages and dendritic cells for their intracellular survival mechanism (Liu and Uzonna, 2012) within their relatively small number of mammal hosts (Schmidt and Roberts, 2005; Ready, 2013). Although there are some lizard-infecting species of *Leishmania* that live in the lumen of the cloacae, the intestine or in the bloodstream of lizards and do not infect macrophages, none of these species were included in this study and there has been a debate among taxonomists about the placement of these species within the genus *Leishmania* or the subgenus Sauroleishmania (Noyes et al., 1998). On the other hand *T. cruzi* has the ability to invade any type of cell in humans during the initial phases of infection, even though the parasite does tend to have tropism toward muscle and nerve cells (Schmidt and Roberts, 2005). The pathology of *T. cruzi* in its estimated 180 mammal reservoir species is not well known (WHO, 2002; Noireau et al., 2009). It is unlikely that the capacity to invade multiple cell types is a characteristic unique to the interaction between *T. cruzi* and humans given the recent age of this interaction and the vast number and diversity of mammal species *T. cruzi* is known to infect. The wider range of host cells and host species that *T. cruzi* can infect, combined with its more complex life cycle, exposes the parasite to more diverse intra- and extracellular environments and has thus exposed a larger number of its proteins to selective pressures during its evolution compared to *Leishmania*. Proteins directly exposed to the immune system as well as proteins that directly interact with surface receptors of the wide range of cell types *T. cruzi* can invade should have been exposed to different levels of selection. Therefore, we hypothesize that the higher versatility of *T. cruzi* is the most likely reason for the significant differences in the fraction of proteins under positive selection (Fig. 2) and the more intense levels of selection in those proteins between the two taxa (Fig. 3). We stress the fact that selection pressures have been exerted mostly by non-human hosts of these parasites given the short time of association of these parasites with humans.

The availability of annotated genome sequences from human pathogens has made feasible the application of *in silico* methods to identify proteins with immunogenic properties that could become candidates for parasite control. Recent studies have provided solid evidence that candidates for vaccine development can be identified by surveying parasite's genomes for proteins in which few amino acid sites have experienced high rates of amino acid substitution (consistent with the action of positive selection) while the rest of the protein is under strong purifying selection (Suzuki, 2004; Gu et al., 2011; McCann et al., 2012). The rationale behind this idea is that proteins with those characteristics have regions that are rapidly evolving because of their recognition by the host's immune system, but also have conserved regions under strong negative selection that may have a critical role in the biology of the pathogen. Effective vaccines or drugs with long-term effectiveness would target those conserved regions rather than the rapidly changing regions of those proteins (Burton et al., 2012), although there is some evidence that polymorphic regions under balancing selection can also be very effective immune elicitors (Osier et al., 2007; Weedall and Conway, 2010).

Our study is the first to use comparative evolutionary analyses for generating a preliminary list of potentially useful immunogenic proteins in Leishmaniasis and Chagas disease (Tables S2–S5). In *Leishmania* spp. the most interesting overrepresented function with genes under positive selection was iron superoxidase dismutase, since one gene from this protein family was recently proposed as a vaccine candidate for Leishmaniasis due to the protective role it induced in mice (Daifalla et al., 2012). That result further reinforces the point that evolutionary approaches can play an

important role in detecting immunogenic molecules. In addition, we found four additional overrepresented functions, some of which have been shown to play some role in the evolution of drug resistance (ATP binding cassette) or in the interaction with the host immune system (Cysteine peptidase). ATP binding cassette was the second most significant overrepresented function. Members of this large protein family are known to be involved in the development of amphotericin-B drug resistance in *L. donovani* (Purkait et al., 2012). This antifungal compound (amphotericin-B) is the main drug therapy employed by the WHO to treat visceral leishmaniasis. It would be of particular interest to determine if the sites predicted to be under positive selection in these proteins are directly involved in the development of amphotericin-B drug resistance. Further, Cysteine peptidases are known to play a very important role in the manipulation of the host's immune response in *L. mexicana* by controlling the T-helper cell response, which has been shown to determine the fate of the infection (Mottram et al., 2004; Alexander and Bryson, 2005). Interestingly the phylogenetic branches leading to *L. mexicana* in both ortholog data sets appear to be under positive selection (data not shown).

In *T. cruzi* we found that Mucin proteins were overrepresented among those showing signals of positive selection. However, while the heavy glycosylation and hypervariability of Mucin proteins does not favor their consideration as logical drug/vaccine candidates for *T. cruzi*, other characteristics like their cell membrane location and role in the mechanisms of intracellular invasion (Buscaglia et al., 2006) suggests the contrary. Unfortunately there are still many proteins of this parasite with unknown function, and our analyses suggest that many of these unknown proteins have experienced adaptive evolution. However, without a biological function associated to these proteins and without knowing whether they are cell surface proteins it is difficult to evaluate their utility in methods of parasite control.

In summary, we show that natural selection has played a larger role in the evolution of *T. cruzi* proteins than in the evolution of *Leishmania* spp. proteins. We propose that this difference is the result of the greater versatility of *T. cruzi* in terms of mammal species it can infect, its cell tropism and its intracellular invasion mechanisms. We provide a list of proteins that have evolved under positive selection that could be evaluated in methods of parasite control in both human diseases. Identification of a protein that was recently proposed as a Leishmaniasis vaccine using our evolutionary analyses confirms the importance that comparative evolutionary studies can have on the exploration of vaccine candidates.

## Author contributions

Conceived and designed the experiments: CAF-L CAM. Performed the experiments: CAF-L. Analyzed the data: CAF-L CAM. Contributed reagents/materials/analysis tools: CAM. Wrote the paper: CAF-L CAM.

## Acknowledgments

We thank Maximilian J. Telford for help on the translation/alignment analyses, and Kawther Abdilleh, Kevin Nyberg, Guilherme Targino Valente and two anonymous reviewers for constructive suggestions. Work partially supported by University of Maryland startup funds and by NSF award MCB-1026200 to CAM.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2015.04.008.

## References

Abascal, F., Zardoya, R., Telford, M.J., 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38, W7–13.
Akopyants, N.S., Kimblin, N., Secundino, N., Patrick, R., Peters, N., et al., 2009. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. Science 324, 265–268.
Alexander, J., Bryson, K., 2005. T helper (h)1/Th2 and *Leishmania*: paradox rather than paradigm. Immunol. Lett. 99, 17–23.
Alvar, J., Velez, I.D., Bern, C., Herrero, M., Desjeux, P., et al., 2012. Leishmaniasis worldwide and global estimates of its incidence. PLoS One 7, e35671.
Anisimova, M., Bielawski, J.P., Yang, Z., 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18, 1585–1592.
Barraclough, T.G., Fontaneto, D., Ricci, C., Herniou, E.A., 2007. Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. Mol. Biol. Evol. 24, 1952–1962.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.
Bethony, J.M., Cole, R.N., Guo, X., Kamhawi, S., Lightowlers, M.W., et al., 2011. Vaccines to combat the neglected tropical diseases. Immunol. Rev. 239, 237–270.
Brisse, S., Henriksson, J., Barnabe, C., Douzery, E.J., Berkvens, D., et al., 2003. Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. Infect. Genet. Evol. 2, 173–183.
Burton, D.R., Poignard, P., Stanfield, R.L., Wilson, I.A., 2012. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. Science 337, 183–186.
Buscaglia, C.A., Campo, V.A., Frasch, A.C., Di Noia, J.M., 2006. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. Nat. Rev. Microbiol. 4, 229–236.
Castillo-Davis, C.I., Hartl, D.L., 2003. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19, 891–892.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
Charlesworth, D., Wright, S.I., 2001. Breeding systems and genome evolution. Curr. Opin. Genet. Dev. 11, 685–690.
Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2, e383.
Daifalla, N.S., Bayih, A.G., Gedamu, L., 2012. *Leishmania donovani* recombinant iron superoxide dismutase B1 protein in the presence of TLR-based adjuvants induces partial protection of BALB/c mice against *Leishmania major* infection. Exp. Parasitol.
De Pablos, L.M., Osuna, A., 2012. Multigene families in *Trypanosoma cruzi* and their role in infectivity. Infect. Immun. 80, 2258–2264.
Desjeux, P., 2001. The increase in risk factors for leishmaniasis worldwide. Trans. R. Soc. Trop. Med. Hyg. 95, 239–243.
Dos Reis, M., Yang, Z., 2013. Why do more divergent sequences produce smaller nonsynonymous/synonymous rate ratios in pairwise sequence comparisons? Genetics 195, 195–204.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., et al., 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309, 409–415.
Fay, J.C., 2011. Weighing the evidence for adaptation at the molecular level. Trends Genet. 27, 343–349.
Fernandes, M.C., Andrews, N.W., 2012. Host cell invasion by *Trypanosoma cruzi*: a unique strategy that promotes persistence. FEMS Microbiol. Rev. 36, 734–747.
Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., et al., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247–251.
Flores-Lopez, C.A., Machado, C.A., 2011. Analyses of 32 loci clarify phylogenetic relationships among *Trypanosoma cruzi* lineages and support a single hybridization prior to human contact. PLoS Negl. Trop. Dis. 5, e1272.
Frank, S.A., 2002. Immunology and evolution of infectious disease. Princeton University Press, Princeton.
Franzen, O., Ochaya, S., Sherwood, E., Lewis, M.D., Llewellyn, M.S., et al., 2011. Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio X10/1 and comparison with *T. cruzi* VI CL Brener. PLoS Negl. Trop. Dis. 5, e984.
Gaunt, M.W., Yeo, M., Frame, I.A., Stothard, J.R., Carrasco, H.J., et al., 2003. Mechanism of genetic exchange in American trypanosomes. Nature 421, 936–939.
George, R.D., McVicker, G., Diederich, R., Ng, S.B., MacKenzie, A.P., et al., 2011. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. Genome Res. 21, 1686–1694.
Glemin, S., 2007. Mating systems and the efficacy of selection at the molecular level. Genetics 177, 905–916.
Gu, M., Liu, W., Xu, L., Cao, Y., Yao, C., et al., 2011. Positive selection in the hemagglutinin–neuraminidase gene of Newcastle disease virus and its effect on vaccine efficacy. Virol. J. 8, 150.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

Henry, L., Schwander, T., Crespi, B.J., 2012. Deleterious mutation accumulation in asexual Timema stick insects. Mol. Biol. Evol. 29, 401–408.
Hotez, P.J., Molyneux, D.H., Fenwick, A., Kumaresan, J., Sachs, S.E., et al., 2007. Control of neglected tropical diseases. New Engl. J. Med. 357, 1018–1027.
Inbar, E., Akopyants, N.S., Charmoy, M., Romano, A., Lawyer, P., et al., 2013. The mating competence of geographically diverse *Leishmania major* strains in their natural and unnatural sand fly vectors. PLoS Genet. 9, e1003672.
Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., et al., 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science 309, 436–442.
Liu, D., Uzonna, J.E., 2012. The early interaction of *Leishmania* with macrophages and dendritic cells and its influence on the host immune response. Front. Cell. Infect. Microbiol. 2, 83.
Lukes, J., Mauricio, I.L., Schonian, G., Dujardin, J.C., Soteriadou, K., et al., 2007. Evolutionary and geographical history of the *Leishmania donovani* complex with a revision of current taxonomy. Proc. Natl. Acad. Sci. U.S.A. 104, 9375–9380.
Machado, C.A., Ayala, F.J., 2001. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. Proc. Natl. Acad. Sci. U.S.A. 98, 7396–7401.
Machado, C.A., Ayala, F.J., 2002. Sequence variation in the dihydrofolate reductase-thymidylate synthase (DHFR-TS) and trypanothione reductase (TR) genes of *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 121, 33–47.
Manso-Alves, M.J., Arruda Mortara, R., 2009. A century of research: what have we learned about the interactions of *Trypanosoma cruzi* with host cells? Mem. Inst. Osw. Cruzi 104, 76–88.
McCann, H.C., Nahal, H., Thakur, S., Guttman, D.S., 2012. Identification of innate immunity elicitors using molecular signatures of natural selection. Proc. Natl. Acad. Sci. U.S.A. 109, 4215–4220.
Moradin, N., Descoteaux, A., 2012. *Leishmania* promastigotes: building a safe niche within macrophages. Front. Cell. Infect. Microbiol. 2, 121.
Mottram, J.C., Coombs, G.H., Alexander, J., 2004. Cysteine peptidases as virulence factors of *Leishmania*. Curr. Opin. Microbiol. 7, 375–381.
Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., et al., 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3, e170.
Noireau, F., Diosque, P., Jansen, A.M., 2009. *Trypanosoma cruzi*: adaptation to its vectors and its hosts. Vet. Res. 40, 26.
Noyes, H.A., Chance, M.L., Croan, D.G., Ellis, J.T., 1998. *Leishmania* (sauroleishmania): a comment on classification. Parasitol. Today 14, 167.
Oleksyk, T.K., Smith, M.W., O'Brien, S.J., 2010. Genome-wide scans for footprints of natural selection. Philos. Trans. R. Soc. Lond. Ser. B, Biol. Sci. 365, 185–205.
Ollivier, M., Gabaldon, T., Poulain, J., Gavory, F., Leterme, N., et al., 2012. Comparison of gene repertoires and patterns of evolutionary rates in eight aphid species that differ by reproductive mode. Genome Biol. Evol. 4, 155–167.
Osier, F.H., Polley, S.D., Mwangi, T., Lowe, B., Conway, D.J., et al., 2007. Naturally acquired antibodies to polymorphic and conserved epitopes of *Plasmodium falciparum* merozoite surface protein 3. Parasite Immunol. 29, 387–394.
Paland, S., Lynch, M., 2006. Transitions to asexuality result in excess amino acid substitutions. Science 311, 990–992.
Peacock, C.S., Seeger, K., Harris, D., Murphy, L., Ruiz, J.C., et al., 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat. Genet. 39, 839–847.
Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M., Nielsen, R., 2007. Genes under positive selection in Escherichia coli. Genome Res. 17, 1336–1343.
Posada, D., 2008. JModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.
Privman, E., Penn, O., Pupko, T., 2011. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol. Biol. Evol. 29, 1–5.
Purkait, B., Kumar, A., Nandi, N., Sardar, A.H., Das, S., et al., 2012. Mechanism of amphotericin B resistance in clinical isolates of *Leishmania donovani*. Antimicrob. Agents Chemother. 56, 1031–1041.
Rassi Jr., A., Rassi, A., Marin-Neto, J.A., 2010. Chagas disease. Lancet 375, 1388–1402.
Ready, P.D., 2013. Biology of phlebotomine sand flies as vectors of disease agents. Annu. Rev. Entomol. 58 (58), 227–250.
Rocha, E.P., Smith, J.M., Hurst, L.D., Holden, M.T., Cooper, J.E., et al., 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 239, 226–235.
Rogers, M.B., Downing, T., Smith, B.A., Imamura, H., Sanders, M., et al., 2014. Genomic confirmation of hybridisation and recent inbreeding in a vector-isolated *Leishmania* population. PLoS Genet. 10, e1004092.
Rougeron, V., De Meeus, T., Hide, M., Waleckx, E., Bermudez, H., et al., 2009. Extreme inbreeding in *Leishmania braziliensis*. Proc. Natl. Acad. Sci. U.S.A. 106, 10224–10229.
Schmidt, G.D., Roberts, L. S., 2005. Foundations of Parasitology.
Sibley, L.D., 2011. Invasion and intracellular survival by protozoan parasites. Immunol. Rev. 240, 72–91.
Simpson, A.G., Inagaki, Y., Roger, A.J., 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol. Biol. Evol. 23, 615–625.
Soyer, Y., Orsi, R.H., Rodriguez-Rivera, L.D., Sun, Q., Wiedmann, M., 2009. Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected Salmonella serotypes. BMC Evol. Biol. 9, 264.
Suzuki, Y., 2004. Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. Gene 328, 127–133.
Swanson, W.J., Nielsen, R., Yang, Q., 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol. Biol. Evol. 20, 18–20.
Tibayrenc, M., Ayala, F.J., 1988. Isozyme variability in *Trypanosoma cruzi*, the agent of Chagas' disease: genetical, taxonomical, and epidemiological significance. Evolution 42, 277–292.
Tibayrenc, M., Ayala, F.J., 2002. The clonal theory of parasitic protozoa: 12 years on. Trends Parasitol. 18, 405–410.
Tibayrenc, M., Ayala, F.J., 2013. How clonal are *Trypanosoma* and *Leishmania*? Trends Parasitol. 29, 264–269.
Tibayrenc, M., Ward, P., Moya, A., Ayala, F.J., 1986. Natural populations of *Trypanosoma cruzi*, the agent of Chagas disease, have a complex multiclonal structure. Proc. Natl. Acad. Sci. U.S.A. 83, 115–119.
Weedall, G.D., Conway, D.J., 2010. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. Trends Parasitol. 26, 363–369.
WHO, 2002. Control of Chagas disease. World Health Organ. Tech. Rep. Ser. 905 (i–vi), 1–109, back cover.
Wolf, J.B., Kunstner, A., Nam, K., Jakobsson, M., Ellegren, H., 2009. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome Biol. Evol. 1, 308–319.
Xu, Z., Chen, H., Zhou, R., 2011. Genome-wide evidence for positive selection and recombination in *Actinobacillus pleuropneumoniae*. BMC Evol. Biol. 11, 203.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yang, Z., Swanson, W.J., 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. 19, 49–57.
Yeo, M., Acosta, N., Llewellyn, M., Sanchez, H., Adamson, S., et al., 2005. Origins of Chagas disease: Didelphis species are natural hosts of *Trypanosoma cruzi* I and armadillos hosts of *Trypanosoma cruzi* II, including hybrids. Int. J. Parasitol. 35, 225–233.
Zhai, W., Nielsen, R., Goldman, N., Yang, Z., 2012. Looking for Darwin in genomic sequences – validity and success of statistical methods. Mol. Biol. Evol. 29, 2889–2893.
Zhang, Y., Zhang, H., Zhou, T., Zhong, Y., Jin, Q., 2011. Genes under positive selection in *Mycobacterium tuberculosis*. Comput. Biol. Chem. 35, 319–322.
Zingales, B., Andrade, S.G., Briones, M.R., Campbell, D.A., Chiari, E., et al., 2009. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem. Inst. Oswaldo Cruz 104, 1051–1054.
Zingales, B., Miles, M.A., Campbell, D.A., Tibayrenc, M., Macedo, A.M., et al., 2012. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. Infect. Genet. Evol. 12, 240–253.