Published in final edited form as:

Infect Genet Evol. 2018 August ; 62: 151–159. doi:10.1016/j.meegid.2018.04.025.

The role of natural selection in shaping genetic variation in a promising Chagas Disease drug target: *Trypanosoma cruzi* trans-sialidase

Joseph P. Gallant^{1,2}, Raquel Asunción Lima-Cordón¹, Silvia Justi¹, Maria Carlota Monroy³, Toni Viola¹, and Lori Stevens^{2,*}

¹Department of Biology, University of Vermont, Burlington, VT, United States

²Department of Pharmacology, University of Vermont, Burlington, VT. United States

³Biology School, Universidad de San Carlos de Guatemala, Guatemala City, Guatemala

Abstract

Rational drug design creates innovative therapeutics based on knowledge of the biological target to provide more effective and responsible therapeutics. Chagas disease, endemic throughout Latin America, is caused by *Trypanosoma cruzi*, a protozoan parasite. Current therapeutics are problematic with widespread calls for new approaches. Researchers are using rational drug design for Chagas disease and one target receiving considerable attention is the *T. cruzi* trans-sialidase protein (TcTS). In *T. cruzi*, trans-sialidase catalyzes the transfer of sialic acid from a mammalian host to coat the parasite surface membrane and avoid immuno-detection. However, the role of TcTS in pathology variance among and within genetic variants of the parasite is not well understood despite numerous studies. Previous studies reported the crystalline structure of TcTS and the TS protein structure in other trypanosomes where the enzyme is often inactive. However, no study has examined the role of natural selection in genetic variation in TcTS. To understand the role of natural selection in TcTS DNA sequence and protein variation, we examined a 471 bp portion of the TcTS gene from 48 *T. cruzi* samples isolated from insect vectors. Because there may be multiple parasite genotypes infecting one insect and there are multiple copies of TcTS per parasite genome, all 48 sequences had multiple polymorphic bases. To resolve these polymorphisms, we examined cloned sequences from two insect vectors. The data are analyzed to understand the role of natural selection in shaping genetic variation in TcTS and interpreted in light of the possible role of TcTS as a drug target. The analysis highlights negative or purifying selection on three amino acids previously shown to be important in TcTS transfer activity. One amino acid in particular, Tyr342, is a strong candidate for a drug target because it is under negative selection and amino acid substitutions inactivate TcTS transfer activity.

*Corresponding Author: lori.stevens@uvm.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Trypanosoma cruzi; Trans-sialidase; rational drug design; genetic variation; natural selection; Chagas disease; Triatoma dimidiata; Triatoma nitida

1. Introduction

Trypanosoma cruzi is a human pathogen responsible for Chagas Disease, also known as American Trypanosomiasis. The disease is endemic in Latin America and is currently affecting an estimated 8–10 million people with another 70 million people at risk [1]. Primarily transmitted by insect vectors of the Triatominae (Hemiptera: Reduviidae) sub-family, *T. cruzi* displays complex genetic diversity with resulting complexity in its pathogenic effects within and between mammalian species [2]. Various approaches have been used to characterize the *T. cruzi* diversity with the general consensus of six major subdivisions known as Discrete Typing Units (DTUs) and named TcI – TcVI. A seventh DTU (TcBat), found within TcI has been brought to recent attention, as well as a lineage found only in bats, *T. cruzi marinkelli*, that forms a sister group to TcI-TcVI [3].

This heterogeneity in *T. cruzi* has been shown to be significant in human Chagas Disease as certain progressions of the disease have been linked to particular DTUs. For example, TcI has been linked to the development of Chagasic cardiomyopathies, while other DTUs have been shown to progress to more gastrointestinal symptoms such as megaesophagus, though they may progress to cardiomyopathies as well [4].

Efforts to reduce Chagas transmission through vector control have been remarkably effective for domesticated vectors. However, in regions where sylvatic populations exist, vector control requires repeated use of insecticides around and within homes and such repeated use of insecticides can have adverse ecological and human health effects as well as lead to insecticide resistance [5]. Further, although vector control methods have interrupted transmission significantly, they fail to address the growing patient population already infected. This combination of factors means that new therapeutics are needed.

The *T. cruzi* trans-sialidase (TcTS) receptor gene has been suggested as a biological target for a rationally designed Chagas drug [6–13]. Factors contributing to its desirability as a potential drug target are the receptor is essential for parasite growth, because it is a receptor it can potentially be inhibited [6], and it has been implicated in virulence [7,8]. The TcTS receptor harvests the mammal hosts' sialic acid, a molecule critical for eukaryotic proliferation and survival that is produced by mammalian cells but not by *T. cruzi*. Among the Trypanosomatidae, which includes other human pathogens such as *Trypanosoma brucei*, the etiological agent of African sleeping sickness, only *T. cruzi* is known to be unable to produce sialic acid and thus expresses high levels of the TcTS receptor in the metacyclic trypomastigote stage (the stage infecting vertebrate blood) [7]. The TcTS receptor captures sialic acid from the blood of its mammalian host, transferring it to surface mucins coating its outer membrane. The sialidated molecule is important in evading immune detection [9]. These features of TcTS make it an excellent target for Chagas rational drug development [7,8].

The TcTS genes are part of a large superfamily of over 1400 genes containing the conserved VTVxNVxLYNR motif [14]. These genes form eight groups, of which only group 1 codes for active trans-sialidases [8,15] with estimates of 1–32 enzymatically active TcTS gene copies per haploid genome [10]. Of the remaining genes about 700 are functional but produce inactive TS, and approximately 700 are pseudogenes [7]. This high diversity possibly accounts for the relative lack of success in previous studies utilizing high-throughput screens looking for inhibitors of the TcTS receptor [7].

The TcTS protein has an N-terminal catalytic domain (amino acids 1-371) that ends in a six-bladed beta propeller, an alpha helix (amino acids 372-394) connects the catalytic domain to the C-terminal lectin-like domain (amino acids 395-632). Another alpha helix near the C-terminus (amino acids 614-626) is sometimes followed by a variable repetitive (100–500 amino acids) hydrophilic, 12 residue shed acute phase antigen (SAPA) motif [9].

Within the N-terminal catalytic domain, group I active TcTS differ from inactive forms by a single amino acid replacement, Tyr342His. The transfer of sialic acid from host to parasite glycoconjugates catalyzed by TcTS involves an additional seven amino acids known to bind to sialic acid. During the transfer reaction, Tyr342 is crucial in forming the covalent intermediate resulting from cleaving the sialic acid from the host glycoconjugate before it is transferred to a parasite glycoconjugate [11].

Identifying and further exploring the genetic diversity of TcTS is important towards drug development since the structure and function of proteins has been shown to be sensitive to amino acid changes anywhere in a peptide sequence, not just active domains (e.g., [14]). Therefore, genetic variation in TcTS may affect susceptibility to inhibitors, and may need to be considered when evaluating TcTS for rational drug design.

Rational drug design incorporates biological information and aims to create more effective and longer lasting therapies. Since drugs act as selective agents, it is important to understand the evolutionary processes acting on a drug target [17]. There are several statistical tests to determine if DNA sequence variation results from neutral processes, or negative or positive selection [17, 18]. Negative, or purifying selection, is the removal of disadvantageous mutations. Positive selection, or adaptive evolution, includes the increase in frequency of advantageous mutations. Finally balancing selection (also called heterozygote advantage or overdominance) and negative frequency-dependent selection (i.e., rare alleles are favored) are types of selection thought to be important in host parasite interactions. These types of selection have different implications for a drug target. It has been suggested that positive selection is not beneficial for a drug target while negative selection does add to a targets justification [19].

We assessed DNA sequence and amino acid variability in a 471 bp DNA portion of the TcTS gene of *T. cruzi*, encompassing 157 amino acids (sites 275–431) in the TcTS protein sequence. This region includes five amino acid residues in the catalytic domain important in the function and structure of the protein: Pro283, Trp312, Arg314, Tyr342 and Glu357 [9] including the critical amino acid Tyr342His that distinguishes the active and inactive forms, but not including the conserved VTVxNVxLYNR motif [14].

The samples were isolated from the abdomens of *Triatoma dimidiata*, the major Chagas disease vector in Central America and Mexico, and *Triatoma nitida*, a species with a more restricted geographic range that is sometimes found in sympatry with *T. dimidiata*. Our *T. dimidiata* specimens were from Guatemala and El Salvador, and the *T. nitida* from Guatemala. The sampling represents a region that is understudied with respect to the genetic diversity of *T. cruzi*, over 90% of DTU identifications are based on samples from South America [2]. In addition to our new data, we expand geographic coverage by also examining TcTS genetic variation within a single genome using publicly available data. Our analysis utilized various methods to evaluate mixed signatures of selection.

2. Methods

2.1 TcTS DNA sequence sampling and TcTS sequencing

The overall experimental design is shown in Fig. 1. We examined parasites from two species of insect vectors, *Triatoma dimidiata* and *Triatoma nitida* (S1). The *T. dimidiata* were collected from three departments (20 from Huehuetenango, one from Quiché and two from Jutiapa) in Guatemala and four in El Salvador (two from Morazan, 10 from Santa Ana, two from Ahuachapán, and four from Sonsonate). The *T. nitida* were from two departments in Guatemala (one from Huehuetenango and six from Chiquimula). To expand the genetic and geographic range examined we also examined TcTS genes from the nine *Trypanosoma cruzi* genomes in publicly available databases.

The *T. cruzi* DNA was extracted from the last three segments of each insect abdomen using previously described methods [20]. The PCR amplification used primers (TS 31: 5'-TCACGCAGCGGTACGCATCCT-3', TS 51: 5'-GGAGGCTGTCGGCACGCTCTC-3'), reported to be specific to group I TS genes that make active trans-sialidases; however it not know if all active trans-sialidases are amplified [10]. The PCR conditions were as previously reported [10] and PCR results were confirmed via agarose gel electrophoresis. Samples showing the appropriate size PCR product were Sanger sequenced by a commercial facility (Genewiz, South Plainfield, NJ).

The PCR reaction amplified 540 bp of the TcTS protein chosen to include five amino acids important to the function and structure of the protein in the catalytic domain and, for comparison, part of the non-catalytic region (Table 1). The 48 sequences from the PCR products were trimmed to 471 bp removing the primer regions, edited using Sequencher (V5) and are hereafter referred to as PCRp sequences (Fig. 1). Polymorphic peaks in the PCRp sequence from a single insect sample represented either infection of that insect with multiple *T. cruzi* genotypes or from multiple copies of TcTS within a single *T. cruzi* genome. To resolve the polymorphism and identify unique haplotypes, we cloned the PCRp from two insect vectors so that a single copy of *T. cruzi* DNA would be expressed in the clonal samples after transformation (Fig. 1, Table 1). Cloning used the pGEM-T Easy cloning kit (Promega) following the manufacturer's instructions for both ligation and transformation, specifically using the pGEM-T Easy Vector, T4 DNA Ligase, and JM109 High Efficiency Competent Cells (Promega). In order to confirm the success of the transformation both ampicillin selection and a blue-white screen were used. Harvested white

colonies were boiled to lyse the cells and extract DNA which was then PCR- amplified and sequenced as described above.

2.2 TcTS phylogenetic reconstruction

Our cloned and PCRp sequences were combined with GenBank TcTS sequences and other *Trypanosoma* species trans-sialidase sequences (S1) for a total of 92 sequences; 58 new from this study and 34 from Genbank. The 58 sequences from this study include 48 PCRp sequences and 10 cloned sequences (four from one and six from a second of the 48 PCRp sequences) (S1).

Phylogenetic analysis was used to confirm the TcTS group of the 58 sequences new from this study by comparison with reference sequences [14]. The nucleotide sequences were translated into amino acid using MEGA (V7)[21]. Alignment of amino acid sequences was performed using PSI/TM-Coffee algorithm from T-Coffee [22–24].

The amino acid alignment was analyzed in ProtTest (V2.4) to identify the best fit evolutionary model for the data. The phylogenetic tree was reconstructed using RaxML(V8.0.0) [25] on CIPRES Science Gateway V3.3 [26] using the evolutionary model GTR + G + I with 100 bootstrap pseudo replicates. The optimal tree output was then drawn in FigTree (V1.4.3).

2.3 TcTS sequences from published genomes

We extracted multiple copies of TcTS from each of the nine published genomes for *T. cruzi* to investigate copy number per genome and compare selection on these sequences with our results. The 471 bp sequence from one clone sample (A10055_RE2, GenBank Accession MG197269) was searched against each of the genomes using the BLAST algorithm (citation). The BLAST results were filtered to remove pseudogenes (i.e., sequences with stop codons) and paralogs (i.e., inactive forms Tyr342His). Selection analysis was done for each genome with 10 or more copies of the TcTS gene. The genomic data include four TcI genomes, one TcII and TcIII and three TcVI, expanding the geographic and genetic scope of our analysis.

2.4 Tests for natural selection

We tested for natural selection for the 48 PCRp sequences, 10 cloned sequences and data from the nine published *T. cruzi* genomes (S3). The sequence data were analyzed for natural selection at both the level of the gene and the level of individual amino acids. First, the McDonald Kreitman (MK) test, which considers an entire gene or region of DNA, was used to test for selection within homogeneous regions (catalytic vs non-catalytic) [27,28]. Second, we used Selecton, an amino acid site-specific test [29,30]. The 10 cloned sequences were examined with both the MK test and Selecton. The 48 PCRp sequences and 87 sequences from the genomes were only examined with Selecton.

We used an online version of the MK test [27,28] with *Trypanosoma brucei* group 1 TS (GenBank Accession AF310232) as the outgroup [14]. The MK test assumes homogeneity across the gene, but because different regions of the TS gene may evolve differently, we used

the multi- locus MK test to test for heterogeneity of part of the region with the N terminal catalytic domain (amino acids 275 to 371) compared to the region with the alpha helix (372–394) and part of the C-terminal lectin- like domain region (amino acids 395 to 431). Then we used the standard MK test to test for selection within homogeneous regions.

The MK test compares the within (polymorphic) *T. cruzi* non-synonymous to synonymous (neutral) changes (Pn/Ps [27]) to the between species (divergent) *T. cruzi* - *T. brucei* non-synonymous to synonymous changes (Dn/Ds) using a Chi-square test. The Neutrality Index (NI) indicates the strength of departure from the neutral model [23]: 1 indicates the data are consistent with a neutral model of evolution, > 1 indicates either Dn is lower than expected due to purifying selection against harmful mutations or Pn is higher than expected due to balancing selection, and < 1 indicates Dn is high due to an excess of fixation of non-neutral replacements from adaptive evolution or Ps is low [31].

In addition, selection at each individual amino acid site was examined and correlated with the protein structure using the Selecton 2007 online server (<http://selecton.tau.ac.il/>) [29,30]. The test uses Bayesian inference to calculate the dN/dS ratio and used a likelihood ratio test to compare a model with positive selection (their M8) with a null model (M8a) that assumes only purifying selection and neutral changes. The selection values for each amino acid ranged from 1 (strong negative selection) to 7 (strong positive selection) and were projected on the 3D structure of the protein using the tool First Glance in Jmol (FGiJ, <http://firstglance.jmol.org>, PDB 1S0I) implemented in Selecton [29,30].

In interpreting the results, for the MK test, $NI < 1$ (“positive” selection) refers to adaptive evolution through directional selection; whereas $NI > 1$ (“negative” selection) includes both purifying and balancing selection. For the Selecton test, negative selection refers to purifying selection, while positive selection refers to adaptive evolution through directional and balancing selection. Finally, to compare the variation among the cloned sequences with the variation among the PCRp sequences, we calculated the conservation for each amino acid, on a scale of 0 to 1, where no variation = 1. Using the most common amino acid at each site as the reference, each amino acid was assigned a value of 1 if it was identical to the reference, 0.5 if the variation did not change the amino acid class (e.g., polar to polar or negative charge to negative charge), and 0 if the variation caused a substitution of a different amino acid class (e.g., polar to negative charge). The sum for each site was divided by the number of total sequences (10 for clones and 48 for PCRp) creating a conservation score between 0 and 1 for each site.

3. Results

Briefly we found consistent purifying selection for 3 of the amino acids. The 58 new TcTS sequences from this study belong to group 1 of the trans-sialidase superfamily, the only one of the eight groups in the superfamily with active TS and all 58 of our sequences are TcI. Examination of the cloned TcTS sequences with the MK test indicated significant balancing selection in the catalytic-domain of the TcTS gene. The variation was not significantly different from a neutral model of evolution for the part of the C-terminal lectin-like domain region examined. In addition, for both the cloned, PCRp and genome sequences, the test for

selection at each individual amino acid site found evidence of positive and negative selection within the regions examined. Details of these results are provided below.

3.1 Confirmation of TcTS sequences as TS group 1

Phylogenetic analysis indicated that the sequences are from TcTS group 1 genes that code for active TS (Fig. 2) with strong statistical support (100% bootstrap support). As expected, all 58 sequences from this study had Tyr342, the critical amino acid that distinguishes the active and inactive forms of TS.

3.2 Tests for natural selection with the PCRp and cloned sequences

The multi-locus MK test indicated the two regions of the TcTS gene were heterogeneous (Table 2, $\omega_{MK} = 2.077$, $\chi^2 = 6.205$, $p < 0.02$). We therefore analyzed each region separately with the standard MK test.

We detected statistically significant negative and/or balancing selection in the catalytic domain (Table 2, Neutrality index $NI = 4.940$, $\chi^2 = 14.348$, $p < 0.001$). Examination of the contingency table (Table 2) shows a higher number of non-synonymous polymorphisms than synonymous polymorphisms for this region ($P_n = 15 > P_s = 8$) suggesting balancing selection.

For the non-catalytic region that included the alpha helix and part of the C-terminal lectin-like domain, the pattern of variation is consistent with the neutral model of evolution (Table 2, Neutrality index $NI = 1.091$, $\chi^2 = 0.047$, $p > 0.05$).

The test using Selecton that examined selection at each individual amino acid, showed evidence of both positive and negative selection over the region examined (Fig. 3). Among the 157 amino acids studied, Selecton identified 48 sites under positive selection, including both directional and balancing selection, 86 sites under purifying selection and 23 neutral sites. Overall the model including positive selection is a better fit to the data than the null model that includes only purifying selection and neutral synonymous changes (log likelihood -1227.73 , delta-log-likelihood -3.35 , $p\text{-value} < 0.02$).

The conservation metric (Fig. 3) provides an estimate of variation for a particular amino acid. Over the region examined, the conservation metric ranges from 0.5 to 1, with an average value of 0.962. The catalytic domain showed higher average conservation (0.976) than the non-catalytic region (0.917).

Selecton values span the possible range 1 to 7 (1 indicates strongest positive selection, 4 is neutral, 7 is strongest purifying selection). The average selection value over the region examined was 4.637, the catalytic domain was toward more purifying selection (average of 4.969) while the non-catalytic region indicated more positive selection (average of 4.095).

With respect to the amino acids essential to catalytic activity, sites 283, 342 and 357 showed negative selection, having both high conservation values and Selecton values (Table 1). Site 312 was neutral. Site 314 gave conflicting results, although this site was not identified as polymorphic by the PCRp sequence, one of the four clones was variable. Possible

explanations include the polymorphism is an artifact of polymerase error during PCR amplification or that the peak representing the second allele was below our threshold of detection. It should be noted that this is the only case in the study where we found variation among the clones that inconsistent with the PCRp sequence and suggested possible positive or balancing selection in the clone data set and purifying selection among the PCRp samples.

Comparison of the Selecton values of the 10 clones with the 48 PCRp sequences shows the PCRp sequences show fewer sites of positive selection (Fig. 3, S2). The Selecton results of the PCRp sequences identified only 49 positively selected sites, 26 neutral sites, and 91 negatively selected sites (Fig. 3). The average Selecton value for the entire region examined was 4.650, as well using the same division in the sequence as above, the catalytic region was more indicating of purifying selection (average of 4.941) than the non-catalytic region (average of 4.174). Similarly, the average conservation score for the entire region analyzed was 0.956. While the catalytic domain region again showed higher conservation (0.977) than the non-catalytic region (0.921).

3.3 Analysis of multiple copies of TcTS within a single genome

The genomic analysis extended our results to include TcII, TcIII and TcVI and widened the geographic range. In contrast to the 1–32 copies of TcTS per genome previously reported (Burgos et al., 2013), we identified 1–23 copies of the 498 bp fragment of TcTS examined in this study (S3). Overall we found about half the number of TcTS gene copies (19.375 vs 9.667) the biggest difference being fewer copies per TcI genome. For the one strain examined previously that also had a sequenced genome (TcVI, CL Brenner, AAHK01) we found about half as many copies of TcTS group I genes (29 vs 15).

Sequencing technologies varied among the genomes including from long reads with high confidence but few copies, short reads with relatively high confidence and many copies, to long reads with low confidence and multiple copies or a combination of these approaches. Algorithms for assembling and correcting errors may reduce differences among gene copies to some extent; however, the genomes with short reads only (e.g., less than the length of the ~2000 bp TcTS gene) tend to report fewer genes.

Analysis of the TcI, TcIII and two TcVI genomes with 10 or more copies of the TcTS gene (the recommended minimum sample size) indicated that for each genome at least one haplotype occurred more than once (Fig. 1); however most copies represented a unique haplotype. In addition, some haplotypes occurred in more than one genome.

The Selecton values from the four genomes that had more than 10 copies of TcTS group 1 were similar to those found in our Clone sequences (Table 1, Fig. 4). All of the sites were invariant and four of the five amino acids essential to catalytic activity had high Selecton values indicating purifying selection (sites 283, 314, 342 and 357) with Tyr342 consistently having the maximum value (Table 1). The only disagreement was site 314 which showed positive selection in the clone samples, but consistently purifying selection within and among genomes. In contrast site 312 was consistently neutral.

4. Discussion

In this study, our analysis detected natural selection on the TcTS gene and examined its genetic variation. The region analyzed starts in the N terminal catalytic domain (amino acids 275 to 371), extends into the alpha helix (372–394) and ends in the C-terminal lectin-like domain (amino acids 373 to 431). This region includes five amino acids important in the binding of sialic acid, Pro283, Trp312, Arg314, Tyr342, and Glu357 (Table 1). Here we discuss Tyr342 as a site to further study for potential drug design due to the functional significance and consistently strong selection signature (Table 1). In addition, our results and analysis of nine *T. cruzi* genomes publicly available showed that the sequences were highly conserved at the DNA and amino acid level at most of the sites essential to catalytic activity, however site Trp312 was neutral and one site Arg314 showed weak evidence of diversifying or balancing selection in one of the seven data sets (Fig. 4).

Although our sampling from Central America found only TcI and represents a region with particularly low diversity of *T. cruzi*, by analyzing publicly available sequenced genomes we expanded our study to include TcII, TcIII and TcVI covering the geographic range of vector transmitted Chagas disease in Latin America.

TcTS group 1 is the only group within the trans-sialidase protein superfamily that produces active trans-sialidase [15] and *T. cruzi* is the only trypanosome reported to have transfer activity. Our phylogenetic analysis confirmed that our PCRp and cloned sequences represent TcTS protein group 1. The sequence alignments for the PCRp, clone samples and genomic data showed no DNA variation for four of the five amino acids essential to TS activity (Table 1). In a single TcI, TcIII or TcVI genome, the copy number of TcTS ranges from 9 – 32 (S3); however, TcII had only three copies and also some strains of TcI and TcVI had few copies (1–4). Our finding of little to no variation for the amino acids important for TS activity suggest the apparent redundancy of multiple copies of the TcTS gene does not seem to be a source of variation for the parasite.

The TcTS region analyzed included a catalytic and non-catalytic domain, where signatures of natural selection and genetic variation were evaluated. With respect to the catalytic domain, signatures of balancing and negative or purifying selection were supported by the MK test. The detailed analysis of individual amino acids with Selecton confirmed the “negative” selection detected by the MK test to be a combination of negative and balancing selection. The 3D structure resolved for the TcTS protein (Fig. 5) reveals that the region of the catalytic domain sequenced in this study contains 64 of the overall 91 negatively selected sites, indicating that there is strong constraint on the TcTS structure. It is important to mention that three out of the five amino acids involved in the catalytic activity showed signs of negative selection (e.g., sites 283, 342 and 357) consistent with the explanation of lack of variation because most mutations are deleterious and recessive and are removed by selection or linger in the population at low frequency [31]. Site 312 has a neutral Selecton value although it was not variable and likely neutral because of its location next to a site under positive selection.

As mentioned before, we detected strong negative selection for the amino acid that distinguishes active and inactive TcTS, Tyr342. Overall these results inform rational drug design because regions under negative selection are considered strong targets [19].

Although most amino acids involved in the catalytic activity are under negative selection, one site (Arg314) indicated balancing selection. For the case of the variation at Arg314, there is variation within the clonal sequences for one sample. However, all other clonal samples, the 48 PCRp sequences and the 87 sequences from the genomic data are conserved for this site with no DNA variation. This disparity in clonal variation arising where no PCRp variation was seen is either from polymerase error or low peak height of the alternative nucleotide in the sequencing electropherogram (Fig. 1). In terms of rational drug design, regions of positive or balancing selection indicating adaptive evolution are not desirable drug targets [19]. This is because these regions show variability indicating the potential of the parasite to develop drug resistance [19]. Although we found one variable nucleotide from 145 sequences, based on these results, drug targets should avoid sites of possible positive or balancing selection such as Arg314, an amino acid important in the carboxylation step of TcTS activity.

Drug design should work to develop inhibitors acting specifically at sites such as Tyr342 that are crucial to the protein function and under negative (purifying) selection. Specificity ensures an inhibitor is acting on the desired amino acid residue and can be measured experimentally. A promising inhibitor believed to be specific for Tyr342 could be tested by mutating or removing Tyr342 (e.g., through site-directed mutagenesis) and determining if the inhibitor exhibits lower affinity. For example, a recent study using drug repositioning to identify possible trypanocidal agents acting on TcTS [13] could also test the specificity of promising compounds with respect to identified critical and negatively selected amino acids. Trypanocidal agents specific to sites under negative selection would be preferential for long term drug efficacy compared to agents specific to sites under positive or balancing selection.

In contrast to the catalytic domain, we were not able to reject the model of neutrality, proposed based on the MK test, for the non-catalytic domain. That is either because the non-catalytic domain follows the expectations of a neutral model of evolution, or there are signatures of both positive and negative selection in about equal numbers. The Selecton analysis supports the latter idea, in that there were 27 negatively selected sites, 12 neutral sites, and 24 positive sites, compared to the catalytic region with respectively 64, 14, and 25.

In summary, this study analyzing natural selection and genetic variation in the TcTS gene identified regions of purifying (negative), balancing and positive selection in the TcTS protein. Rational drug design should consider this variation to increase the likelihood of developing effective drugs with lower chances for the evolution of drug resistance [19].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NSF grant BCS-1216193 as part of the joint NSF-NIH-USDA Ecology and Evolution of Infectious Diseases program and by NIH grant R03AI26268/1-2. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding organizations. Additional funding received via University of Vermont APLE Grant.

References

1. WHO. Chagas disease in Latin America: an epidemiological update based on 2010 estimates. 2015.
2. Brenière SF, Waleckx E, Barnabé C. Over six thousand *Trypanosoma cruzi* strains classified into Discrete Typing Units (DTUs): Attempt at an inventory. *PLoS neglected tropical diseases*. 2016; 10(8):e0004792. [PubMed: 27571035]
3. Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MMG, et al. The revised *Trypanosoma cruzi* subspecific nomenclature: Rationale, epidemiological relevance and research applications. *Infection, Genetics and Evolution*. 2012; 12(2):240–53.<http://dx.doi.org/10.1016/j.meegid.2011.12.009>
4. Zingales B, Miles MA, Moraes CB, Luquetti A, Guhl F, Schijman AG, et al. Drug discovery for Chagas disease should consider *Trypanosoma cruzi* strain diversity. *Memórias do Instituto Oswaldo Cruz*. 2014; 109(6):828–33. DOI: 10.1590/0074-0276140156 [PubMed: 25317712]
5. Moncayo A. Chagas disease: current epidemiological trends after the interruption of vectorial and transfusional transmission in the Southern Cone countries. *Memórias do Instituto Oswaldo Cruz*. 2003; 98(5):577–91. [PubMed: 12973523]
6. Buscaglia CA, Campo VA, Frasch ACC, Di Noia JM. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. *Nat Rev Micro*. 2006; 4(3):229–36.
7. Neres J, Bryce RA, Douglas KT. Rational drug design in parasitology: trans-sialidase as a case study for Chagas disease. *Drug Discovery Today*. 2008; 13(3–4):110–7.<http://dx.doi.org/10.1016/j.drudis.2007.12.004> [PubMed: 18275908]
8. Escalante AA, Cornejo OE, Rojas A, Udhayakumar V, Lal AA. Assessing the effect of natural selection in malaria parasites. *Trends in parasitology*. 2004; 20(8):388–95.1. [PubMed: 15246323]
9. Freire-de-Lima L, Fonseca L, Oeltmann T, Mendonça-Previato L, Previato J. The trans-sialidase, the major *Trypanosoma cruzi* virulence factor: three decades of studies. *Glycobiology*. 2002; 10(4):757–68.[http://dx.doi.org/10.1016/S1097-2765\(02\)00680-9](http://dx.doi.org/10.1016/S1097-2765(02)00680-9) [PubMed: 12419220]
10. Burgos JM, Risso MG, Brenière SF, Barnabé C, Campetella O, Leguizamón MS. Differential Distribution of Genes Encoding the Virulence Factor Trans-Sialidase along *Trypanosoma cruzi* Discrete Typing Units. *PLOS ONE*. 2013; 8(3):e58967.doi: 10.1371/journal.pone.0058967 [PubMed: 23536842]
11. Watts AG, Damager I, Amaya ML, Buschiazzi A, Alzari P, Frasch AC, et al. *Trypanosoma cruzi* Trans-sialidase Operates through a Covalent Sialyl–Enzyme Intermediate: Tyrosine Is the Catalytic Nucleophile. *Journal of the American Chemical Society*. 2003; 125(25):7532–3. [PubMed: 12812490]
12. De Celis SSCR. Surface Topology Evolution of *Trypanosoma* Trans-Sialidase. In: Santos ALS, Branquinho MH, d'Avila-Levy CM, Kneipp LF, Sodré CL, editors *Proteins and Proteomics of Leishmania and Trypanosoma*. Dordrecht: Springer Netherlands; 2014. 203–16.
13. Lara-Ramirez EE, López-Cedillo JC, Noguera-Torres B, Kashif M, Garcia-Perez C, Bocanegra-Garcia V, et al. An in vitro and in vivo evaluation of new potential trans-sialidase inhibitors of *Trypanosoma cruzi* predicted by a computational drug repositioning method. *European Journal of Medicinal Chemistry*. 2017; 132:249–61.<http://dx.doi.org/10.1016/j.ejmech.2017.03.063> [PubMed: 28364659]
14. Chiurillo MA, Cortez DR, Lima FM, Cortez C, Ramírez JL, Martins AG, et al. The diversity and expansion of the trans-sialidase gene family is a common feature in *Trypanosoma cruzi* clade members. *Infection, Genetics and Evolution*. 2016; 37:266–74.<http://dx.doi.org/10.1016/j.meegid.2015.11.024>

15. Freitas LM, Dos Santos SL, Rodrigues-Luiz GF, Mendes TAO, Rodrigues TS, Gazzinelli RT, et al. Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of *Trypanosoma cruzi* reveal an undetected level of complexity. *PLoS One*. 2011; 6(10):e25914. [PubMed: 22039427]
16. Dong Y, Somero GN. Temperature adaptation of cytosolic malate dehydrogenases of limpets (genus *Lottia*): differences in stability and function due to minor changes in sequence correlate with biogeographic and vertical distributions. *Journal of Experimental Biology*. 2008; 212(2):169.
17. Suzuki Y. Natural Selection on the Influenza Virus Genome. *Molecular Biology and Evolution*. 2006; 23(10):1902–11. DOI: 10.1093/molbev/msl050 [PubMed: 16818477]
18. Suzuki Y. Detection of positive selection eliminating effects of structural constraints in hemagglutinin of H3N2 human influenza A virus. *Infection, Genetics and Evolution*. 2013; 16:93–8. <https://doi.org/10.1016/j.meegid.2013.01.017>
19. Searls DB. Pharmacophylogenomics: genes, evolution and drug targets. *Nature Reviews Drug Discovery*. 2003; 2(8):613–23. [PubMed: 12904811]
20. Pizarro JC, Stevens L. A new method for forensic DNA analysis of the blood meal in Chagas disease vectors demonstrated using *Triatoma infestans* from Chuquisaca, Bolivia. *PLoS One*. 2008; 3(10):e3585. [PubMed: 18974787]
21. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution*. 2016 Mar 22; 33(7):1870–4. [PubMed: 27004904]
22. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*. 2000; 302(1):205–17. [PubMed: 10964570]
23. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic acids research*. 2015; 43(W1):W580–W4. [PubMed: 25845596]
24. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis tool web services from the EMBL-EBI. *Nucleic acids research*. 2013; 41(W1):W597–W600. [PubMed: 23671338]
25. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014 Jan 21; 30(9):1312–3. [PubMed: 24451623]
26. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010. *GCE 2010*. 2010
27. Egea R, Casillas S, Barbadilla A. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic acids research*. 2008; 36(suppl_2):W157–W62. [PubMed: 18515345]
28. McDonald JH, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*. 1991; 351(6328):652. [PubMed: 1904993]
29. Doron-Faigenboim A, Stern A, Mayrose I, Bacharach E, Pupko T. Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*. 2005; 21(9):2101–3. [PubMed: 15647294]
30. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic acids research*. 2007; 35(suppl_2):W506–W11. [PubMed: 17586822]
31. Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Molecular biology and evolution*. 1996; 13(6):735–48. [PubMed: 8754210]

Author Summary

Chagas disease is caused by the protozoan parasite *Trypanosoma cruzi* and transmitted to humans and other mammals primarily by Triatomine insects. Being endemic in many South and Central American countries and affecting millions of people the need for new more effective and safe therapies is evident. Here, we examine genetic variation and natural selection on DNA (471 bp) and amino acid (157 aa) sequence data of the *T. cruzi* trans-sialdiase (TcTS) protein, often suggested as a candidate for rational drug design. In our surveyed region of the protein there were five amino acid residues that have been shown to be integral to the function of TcTS. We found that three were under strong negative selection making them ideal candidates for drug design; however, one was under balancing selection and should be avoided as a drug target. Our study provides new information into identifying potential targets for a new Chagas drug.

Highlights

- *Trypanosoma cruzi* trans-sialidase has been suggested as a target for rational drug design.
- We found genetic variation in a 498 nucleotide portion of this gene within and between genomes.
- Selection signatures in this portion vary between the catalytic and non-catalytic regions.
- There is strong purifying selection at three of the five amino acids involved in catalysis.
- Drug design is discussed with respect to variation and selection.

Fig. 1.

Experimental design. (A) DNA was extracted from the abdomens of infected insect vectors. (B) There could be one (top panel) or multiple (lower panel) parasite genotypes per insect indicated by numbers 1 and 2. (C) Within each parasite nucleus there can be 1–>20 copies of the TcTS group 1 gene indicated by letters a-i. Based on the analysis of published genomes, most of these are unique haplotypes, but some haplotypes occur multiple times in a genome or in multiple parasite genotypes. (D) PCRp sequence with heterozygous peaks. (E) Sequences from clones with heterozygous peaks resolved.

* All samples unique to this study resolve in TS Group 1. See Table 1 for list of samples.

Fig. 2.
Best Maximum Likelihood tree reconstruction for the trans-sialidase (TS) protein family constructed from *T. cruzi* TS samples and GenBank reference samples for each representative group for the TS protein family. Bootstrap values (0–100) are indicated at branch nodes.

Fig. 3.

Summary of results showing DNA and amino acid sequence variation, and results of the analysis of Selection test for natural selection based on analysis of the 10 clone sequences. Rows (identified with ALL CAPITAL LETTERS) indicate: SELECTION, the Selection scores represented by color (see key on Figure); POSITION, the amino acid position in the TcTS protein (our sequences cover amino acids 275-431); AA, the amino acid for each of the 10 clones; and DNA, the corresponding nucleotides for each amino acid. The blue line shows the Selection score for each amino acid with the red horizontal line indicating neutral evolution or no selection. The blue histograms show the conservation of each amino acid site (1 = no variation, 0 = all sites unique).

Fig. 4.

Comparison of selection values and type of selection detected for analysis of clonal sequences, PCRp sequences, and multiple genomes.

Fig. 5. Three-dimension projections of *Trypanosoma cruzi* trans-sialidase (TcTS) molecule with amino acids in the sequenced region shaded to indicate Selecton scores.

Table 1

Genetic variation and selection for the TcTS amino acids essential to catalytic activity in the region examined in this study.

Data set or Genome	Pro 283	Trp 312	Arg 314	Tyr 342	Glu 357
original_10_clones	Transglycosylation/Hydroxylation	Aromatic Sandwich/Hydrogen Bonding	Carboxylate Fixation	Enzymatic Nucleophile	Catalysis
variation	Purifying	Neutral	Positive	Purifying	Purifying
selection value	None	None	Non-synonymous	None	None
Original_48_PCRp	6	4	2	7	7
variation	Purifying	Neutral	Purifying	Purifying	Purifying
selection value	None	None	None	None	None
AAHK01019540.1	6	4	6	7	6
variation	Purifying	Neutral	Purifying	Purifying	Purifying
selection value	None	None	None	None	None
MBSY01000090.1	6	4	5	7	7
variation	Purifying	Neutral	Purifying	Purifying	Purifying
selection value	None	None	None	None	None
NMZN01000002.1	6	4	6	7	6
variation	Purifying	Neutral	Purifying	Purifying	Purifying
selection value	None	None	None	None	None
OGCJ01000744.1	6	4	6	7	7
variation	Purifying	Neutral	Purifying	Purifying	Purifying
selection value	None	None	None	None	None
	6	4	5	7	7

Table 2

Results of multi- locus McDonald Kreitman test for the N-terminal catalytic domain (catalytic region) and the alpha helix domain and C-terminal lectin like domain (non-catalytic region)

Region Analyzed in MKT	Polymorphism		Divergence	Total	Neutrality Index	Chi Squared	P-value
	Neutral	Non-Neutral					
Catalytic Domain	8	15	209.8	217.8	4.94	14.348	<0.000
			79.62	94.62			
	Total	23	289.42	312.42			
Non-Catalytic Domain	13		32.76	45.76	1.09	0.047	0.828
	Non-Neutral	26	60.04	86.04			
	Total	39	92.8	131.8			