



PhyloQuant approach provides insights into *Trypanosoma cruzi* evolution using a systems-wide mass spectrometry-based quantitative protein profile

Simon Ngao Mule¹, André Guilherme Costa-Martins¹, Livia Rosa-Fernandes ¹, Gilberto Santos de Oliveira ¹, Carla Monadeli F. Rodrigues¹, Daniel Quina¹, Graziella E. Rosein², Marta Maria Geraldes Teixeira¹ & Giuseppe Palmisano¹✉

The etiological agent of Chagas disease, *Trypanosoma cruzi*, is a complex of seven genetic subdivisions termed discrete typing units (DTUs), TcI-TcVI and Tcbat. The relevance of *T. cruzi* genetic diversity to the variable clinical course of the disease, virulence, pathogenicity, drug resistance, transmission cycles and ecological distribution requires understanding the parasite origin and population structure. In this study, we introduce the PhyloQuant approach to infer the evolutionary relationships between organisms based on differential mass spectrometry-based quantitative features. In particular, large scale quantitative bottom-up proteomics features (MS1, iBAQ and LFQ) were analyzed using maximum parsimony, showing a correlation between *T. cruzi* DTUs and closely related trypanosomes' protein expression and sequence-based clustering. Character mapping enabled the identification of synapomorphies, herein the proteins and their respective expression profiles that differentiate *T. cruzi* DTUs and trypanosome species. The distance matrices based on phylogenetics and PhyloQuant clustering showed statistically significant correlation highlighting the complementarity between the two strategies. Moreover, PhyloQuant allows the identification of differentially regulated and strain/DTU/species-specific proteins, and has potential application in the identification of specific biomarkers and candidate therapeutic targets.

¹Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil. ²Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, Brazil. ✉email: palmisano.gp@usp.br

T*rypanosoma cruzi* is a protistan parasite agent of Chagas disease (American trypanosomiasis), a zoonotic disease endemic in 21 Latin American countries¹ with ~25 million people at risk of infection. An estimated 6–7 million people are infected worldwide and 10,000 annual mortalities are reported^{2,3}. The hematophagous triatomine insects (Reduviidae: Triatominae) are the main vectors of *T. cruzi*. Other transmission routes include blood transfusion^{4–6}, oral transmission from food or beverage contamination^{7–10}, congenital transmission^{11,12} and organ and marrow transplant from infected donors^{13,14}. Chagas disease is currently an emergent public health concern due to the outbreaks of oral infection, especially in the Amazon region^{7–10}, and international migrations contributing to many cases of the disease in non-endemic countries including Europe, USA, Australia and Japan^{15,16}.

During its life cycle, *T. cruzi* alternates between triatomine vectors and mammalian hosts¹⁷. Epimastigotes proliferate in the insect midgut and differentiate into metacyclic trypomastigotes, the non-proliferative stage that infects mammals through skin wounds or mucosal membranes, where they invade the cells and differentiate into amastigotes. The amastigotes replicate intracellularly and differentiate into infective trypomastigotes, which are released into the bloodstream following host cell rupture¹⁸. Each developmental stage is characterized by changes in morphology, molecular and biochemical makeup^{17,19–23}.

The variable clinical outcomes of Chagas disease have been associated to *T. cruzi* heterogeneity^{24–26}. Different studies have elucidated the genetic variability in *T. cruzi* populations and shown close associations between the parasites' genetics and relevant biological, pathological and clinical characteristics^{24,25,27–29}. In 1999, a consensus nomenclature was proposed, placing *T. cruzi* strains into two major groups; *T. cruzi* I and *T. cruzi* II³⁰. In 2009, this classification was revised, distributing *T. cruzi* strains into six genetic subdivisions termed discrete typing units (DTUs)³¹. Currently, seven DTUs are recognized in *T. cruzi* (TcI–TcVI)³², along with Tcbat, which infects predominantly bats^{33–35}. The parasite's DTU has been implicated as one relevant factor influencing clinical variations of the disease^{24,27,36}, drug resistance/susceptibility^{37,38}, pathogenicity in mice³⁸, and vector competence^{18,39}. Other factors influencing disease outcome include mixed infections, infection routes, host genetics and a range of eco-geographical factors^{32,40,41}. Different molecular techniques have been used to study the genetic diversity of *T. cruzi*, such as kinetoplast DNA⁴², RAPDs^{43,44}, polymorphism of the spliced leader intergenic region^{45–48} and multilocus sequence typing^{49–52}.

The application of mass spectrometry in the typing of organisms has been previously described^{53,54}. Moreover, the comparison between proteomes of *C. elegans* and *D. melanogaster* showed that protein abundance between organisms is more comparable than mRNA transcripts and more similar than protein abundance and transcript expression within the same organism⁵⁵. Telleria et al. have described the hierarchical clustering of *T. cruzi* strains using Euclidean Distances based on protein expression levels from 2D-DIGE analysis⁵⁶. This clustering showed a high correlation between *T. cruzi* phylogenetics and levels of protein expression. In addition, *T. cruzi* strains have been assigned to their DTUs with high accuracy based on the genome-free spectral matching assay⁵⁷. In the current study, we applied a proteome-wide quantitative proteomics approach to profile and infer evolutionary relationships between *T. cruzi* DTUs (TcI–TcVI, Tcbat) and their most closely related bat trypanosomes; *Trypanosoma cruzi marinkellei*, *Trypanosoma dionisii*, and *Trypanosoma erneyi*, all classified in the subgenus *Schizotrypanum* nested into *T. cruzi* clade^{33,35,58–60}. The generalist and human-infective *T. rangeli* also belonging to the *T. cruzi* clade was included as an outgroup of *Schizotrypanum*

subgenus^{59,61}. This approach was termed “PhyloQuant”, and is herein introduced to encompass phyloproteomics approaches with other identified and unidentified biomolecular features obtained by quantitative proteomics, metabolomics, lipidomics, glycoproteomics and other omics techniques in typing and evolutionary inferences of organisms. Genetic distance matrices from PhyloQuant-based phylogenies compared to genetic distance matrices from phylogenies inferred from concatenated SSU rRNA, glycosomal glyceraldehyde 3-phosphate dehydrogenase (gGAPDH) and HSP70 gene sequences were analyzed to determine their levels of correlation. We show the strong discriminatory power of three MS-based quantitative features to accurately assign *T. cruzi* strains into their DTUs, and the suitability of this technique to reconstruct the evolutionary relationships among the DTUs of *T. cruzi*, and between this species and closely related trypanosomes. Our findings support a close congruence of the evolutionary relationships among all assayed trypanosome species and *T. cruzi* DTUs using PhyloQuant and sequence-based phylogenetic studies. Putative *T. cruzi* strain/DTU, and species-specific proteins at the exponential growth phase of the epimastigote life stage are also explored.

Results

Mass spectrometry results. The evolutionary relationships of *T. cruzi* strains representative of all DTUs and closely related trypanosomes of the subgenus *Schizotrypanum* were investigated using a comprehensive bottom-up mass spectrometry-based quantitative proteomics combined with statistical and computational analyses. A schematic flowchart of the experimental procedures applied in this study is summarized in Fig. 1. A total of 524,303 MS1 values from precursor ions with a charge state of ≥ 2 were quantified from the 17 trypanosomes in all the biological replicates evaluated. Of these, 68,828 precursor ion intensities had a minimum of three valid values in at least one group, of which 52,622 were statistically regulated using ANOVA (Supplementary Data 1). A total of 4796 proteins were identified and quantified in this study (Supplementary Data 2). Of these, 2441 and 4018 label-free quantification (LFQ) and intensity-based absolute quantification (iBAQ) values, respectively, were statistically significant (FDR < 0.05), with three valid values in at least one group (Supplementary Data 2 and 3, respectively). All the quantitative proteomic values from total and statistically significant MS1, iBAQ and LFQ features were normalized by *z* score, and maximum parsimony (MP), a phylogenetic algorithm, used to infer evolutionary relationships.

Proteome variation among trypanosomes. A multivariate data analysis based on principal component analysis (PCA) was employed to explore the variation in protein expression between trypanosomes obtained by LC-MS/MS. Variations of Log2-transformed MS1, iBAQ and LFQ were visualized by the first and second principal components (Fig. 2). PCA on MS1 intensities placed the *T. cruzi* DTUs into a non-cohesive assemblage comprising the highly phylogenetically related *T. c. marinkellei*, whereas the more distant related *T. dionisii*, *T. erneyi* and *T. rangeli* clustered separately (Fig. 2a). This clustering was based on the first and the second principal components, which represented 11% and 7% of the total variation in the dataset, respectively. Proteome variation based on iBAQ values resulted in a much more resolved clustering (Fig. 2b), showing the separation of *T. cruzi* DTUs from all allied bat trypanosomes. Clustering analysis of *T. cruzi* DTUs by PCA showed that TcI and Tcbat clustered tightly together closest to a cluster exclusive of TcIV strains based on the second principal component representing 10% of the total variation in the dataset. Strains representative of

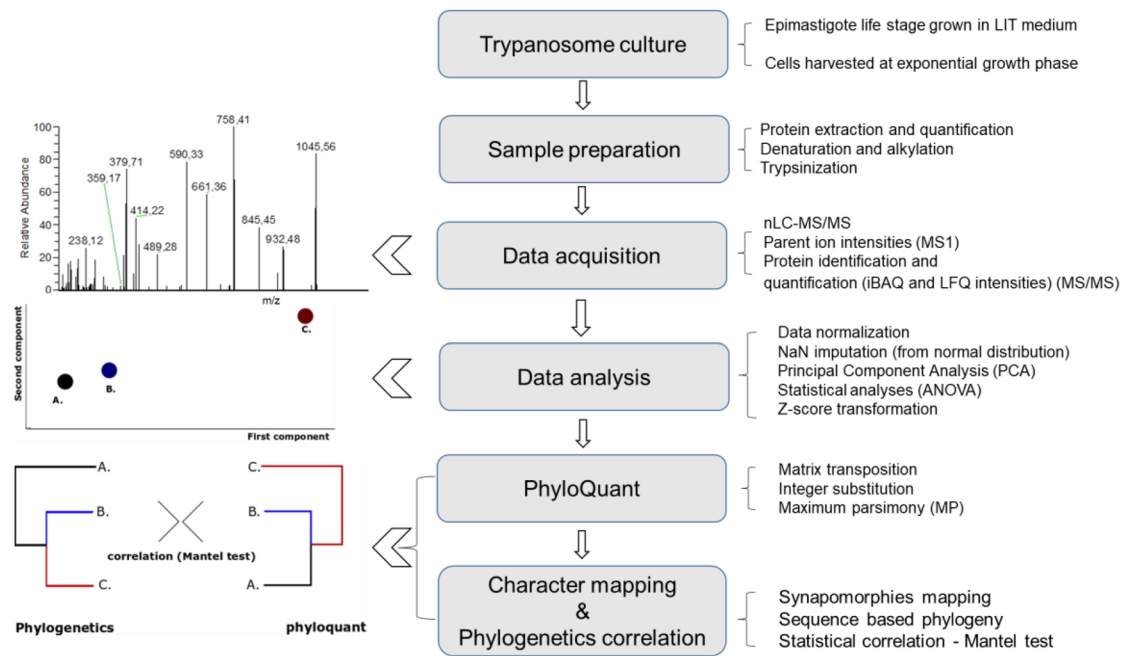


Fig. 1 Experimental workflow of the PhyloQuant approach. Trypanosomes grown in complete LIT medium were harvested at the exponential proliferative phase of the epimastigote life stage. Proteins were extracted in 8 M urea lysis buffer in the presence of protease inhibitors, digested with trypsin and the resultant peptides analyzed by nLC-MS/MS. Normalized MS1, iBAQ and LFQ intensities were used to infer evolutionary relationships using the Maximum Parsimony criterion, implemented in the TNT program. Character mapping of synapomorphies that differentiate each clade was performed in TNT program. Phylogeny based on three concatenated genes, gGAPDH, V7V8 SSU rRNA and HSP70 was chosen to perform correlation of the evolutionary relationships based on proteome-wide mass spectrometry-based features. Statistical levels of correlation between PhyloQuant and phylogenetic distance matrices were inferred using the Mantel test.

TcIII clustered very near to TcII forming a cohesive cluster also comprising strains of TcV and TcVI. *T. c. marinkellei* was placed well separated from both *T. cruzi* DTUs and the other bat trypanosomes *T. dionisii* and *T. erneyi*. *T. rangeli* was placed furthest from all *Schizotrypanum* trypanosomes based on the first principal component which represented 30.9% of the total variation in the dataset.

PCA analysis based on LFQ intensities resulted in four distant clusters based on the first and second components representing 35.4% and 10.9% of the total variability in the dataset, respectively (Fig. 2c). All *T. cruzi* DTUs clustered closely together, although TcI, Tcbat and TcIV were placed at the edge of the major cluster comprising all *T. cruzi* strains, and the parental genotypes TcII and TcIII and their respective hybrid genotypes TcV and TcVI^{27–29,32} formed a very tight cluster. Similar to iBAQ based PCA, *T. c. marinkellei* was clearly separated from *T. cruzi* DTUs, *T. dionisii* and *T. erneyi*.

Protein expression-based evolutionary inference (‘PhyloQuant’). In this proof of concept study, we evaluated the potential application of quantitative proteomic-based data for DTU assignment and inference of evolutionary relationships of *T. cruzi* strains and closely phylogenetically related trypanosomes. Phylogenies were inferred by MP using numeric values from total and statistically significant unidentified precursor ion (m/z) intensities; and total and statistically significant intensities of identified MS-based proteomic features (iBAQ and LFQ intensities). Subsequently, correlation of the PhyloQuant approach was performed against phylogenies inferred from concatenated sequences of three gene loci; gGAPDH, V7V8 SSU rRNA and HSP70 (Supplementary Fig. 1). Character mapping based on statistically significant LFQ-based PhyloQuant were mapped, showing the proteins and their expression profiles which support

the different monophyletic clades (Supplementary Data 4). PhyloQuant clustering inferred from statistically significant MS1 m/z intensities showed the separation of all *T. cruzi* DTUs including Tcbat from the closely phylogenetically related trypanosomes strongly linked to bats; *T. c. marinkellei*, and the sister clade comprising *T. erneyi* and *T. dionisii* (Fig. 3a). *T. c. marinkellei* formed the basal clade to *T. cruzi* DTUs, supported by a high bootstrap value. *T. cruzi* strains were grouped in two major clusters; one composed of TcI-Tcbat/IV/III strains, and the second composed of TcII/III/V/VI. Strains representative of TcI, TcIV and TcVI clustered to form monophyletic clades. TcII strains and one strain of TcIII clustered with those of TcV and TcVI, consistent with TcII and TcIII hybridization events given origin to TcV and TcVI. In addition, corroborating relevant intra-DTU TcIII genetic diversity, the TcIII strain M-6241 clustered with those of TcIV. Tcbat was placed sister to TcI forming a clustering congruent with evolutionary relationships inferred by various genetic markers including cytb, gGAPDH, SSU rRNA, H2B and many other gene sequences^{33–35}. PhyloQuant clustering based on total MS1 intensities strongly supported the separation of *T. cruzi* DTUs from their allied trypanosome species, including *T. c. marinkellei*, but showed less inter DTU discriminatory power (Supplementary Fig. 2a). Although only TcIV and TcVI strains formed DTU-specific clades, the analysis showed the expected close relationships among TcII, TcIII, TcV and TcVI.

PhyloQuant clustering based on the identified MS features, i.e., statistically significant iBAQ and LFQ intensities (Fig. 3b, c), and total iBAQ and LFQ intensities (Supplementary Fig. 2b, c), showed very well resolved clustering of the *T. cruzi* strains and closely related trypanosomes with higher bootstrap support values. PhyloQuant based on these features supported with high bootstrap values the separation of all *T. cruzi* DTUs from *T. c. marinkellei*, *T. dionisii* and *T. erneyi* and the outgroup *T. rangeli*. Notable, these quantitative features showed a high inter DTU

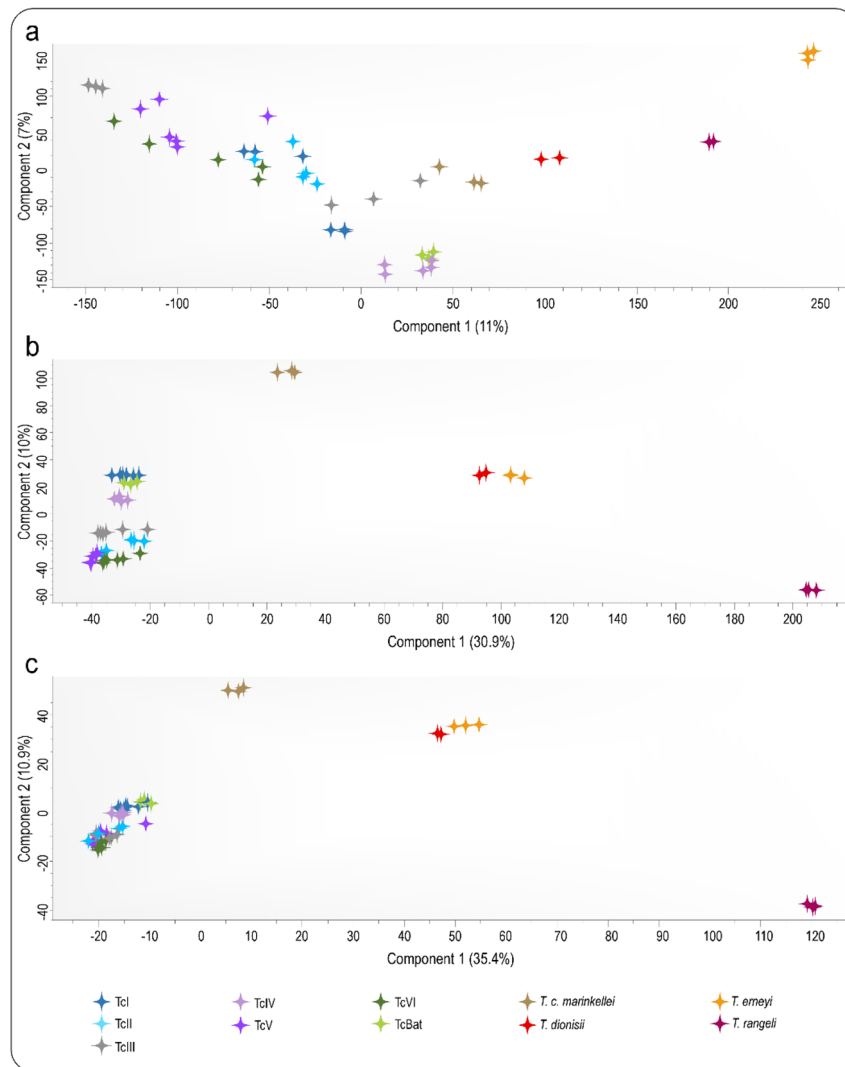


Fig. 2 Principal component analysis of *T. cruzi* DTUs and bat allied trypanosomes; *T. c. marinkellei*, *T. dionisii*, *T. erneyi*, and the generalist *T. rangeli*. The first and second principal components based on normalized **a** MS1 values, **b** iBAQ and **c** LFQ proteomic features representing the total proteome variations in the three datasets are presented.

discriminatory power, enabling the assignment of *T. cruzi* strains to their DTU clades, with the exception of TcII strains, which remained intermingled with TcV or TcVI. Three main clusters were formed with *T. cruzi* strains of the seven DTUs; TcI-TcBat were placed as sister clades, and formed the basal clade to *T. cruzi* DTUs. TcIV strains clustered separated between the strongly supported clades TcI-TcBat and TcIII. The parental TcIII was basal to a major clade harboring monophyletic TcV and TcVI strains, and the paraphyletic TcII. Corroborating different strains given origin hybrid DTUs, *T. cruzi* Esmeraldo was placed closer to TcV, while *T. cruzi* Y showed a close relationship with TcVI. Phylogenetic character mapping of the proteins supporting each clade (synapomorphies), and their respective expression profiles, were determined using the TNT program, and are presented in Supplementary Data 4.

PhyloQuant shows high correlation with phylogenetics. To infer evolutionary correlation between PhyloQuant and phylogenetics, the replicate z score values of each strain were concatenated and used to infer PhyloQuant-based trees based on statistically significant and total MS1, iBAQ and LFQ intensities.

Subsequently, distance matrices from PhyloQuant and molecular-based evolutionary trees inferred by MP based on concatenated gGAPDH, SSU rRNA and HSP70 gene sequences (Supplementary Fig. 1) were analyzed for levels of congruency. Phylogenies inferred by the concatenated DNA sequences have previously been employed to infer phylogenies aimed at resolving relationships between trypanosome species of the subgenus *Schizotrypanum* lineage and the Tra [Tve-Tco] lineage⁶¹ comprising *T. rangeli*, *T. conorhini*, *T. vespertilionis* and other Old World trypanosomes.

Levels of correlation were statistically evaluated using the nonparametric Mantel test using analyses of phylogenetics, implemented in R statistical software. Mantel test gave a statistically significant correlation (p value of 0.0009), showing close correlation between three evolutionary trees based on MS1, iBAQ and LFQ and phylogenetics. Levels of correlation were determined based on 1000 permutations. The statistically significant correlation between phylogenies derived from protein expression levels and molecular-based phylogenies illustrate the high congruence between *T. cruzi* protein expression and intraspecific genetic subdivisions, as earlier illustrated by Telleria et al.⁵⁶

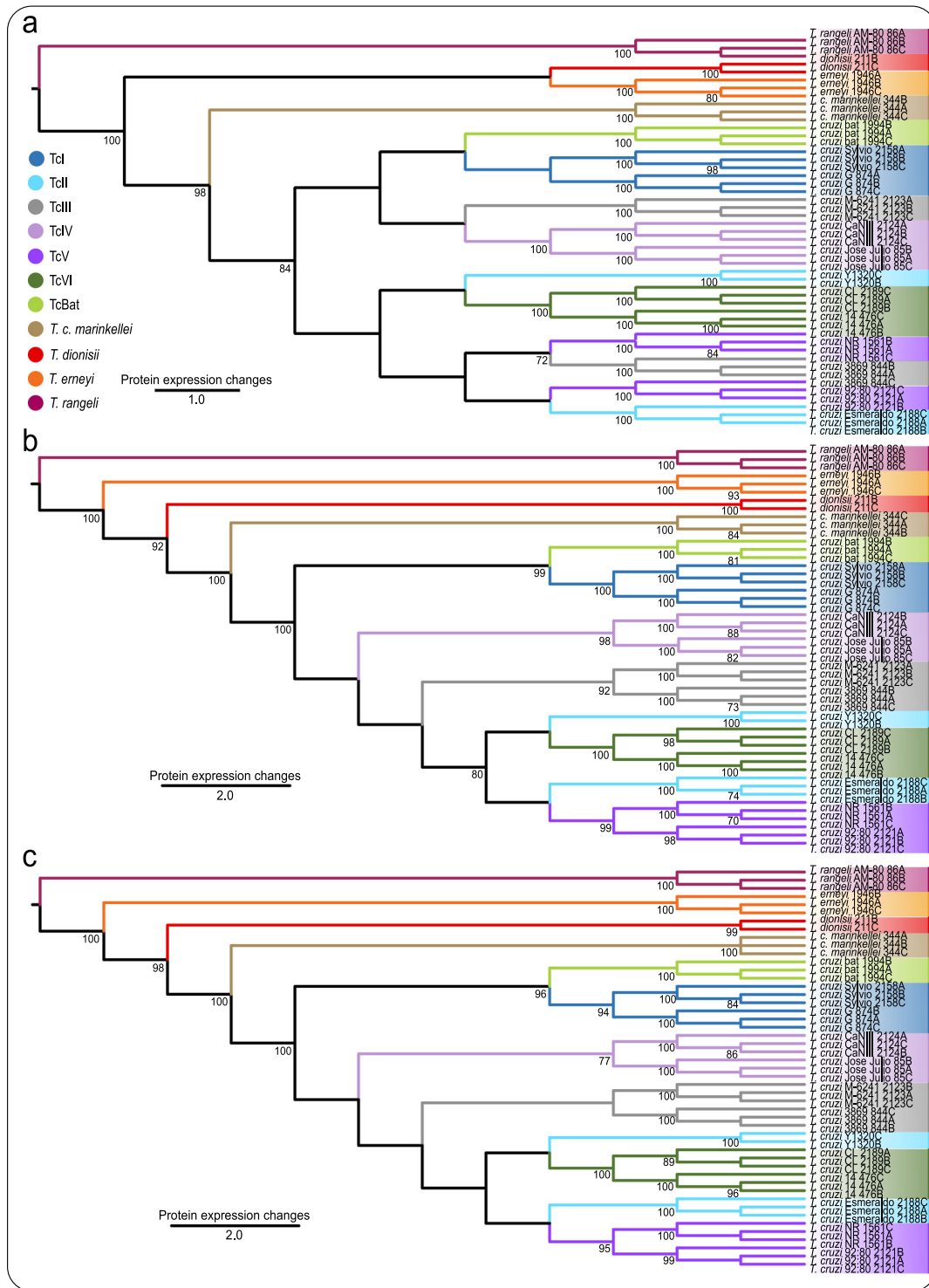


Fig. 3 PhyloQuant analysis of trypanosomes of the *Schizotrypanum* subgenus based on statistically regulated MS features. **a** MS1, **b** iBAQ and **c** LFQ values inferred by the parsimony evolutionary algorithm. Numbers on branches represent bootstrap support above 70% estimated with 1000 replicates.

DTU and species-specific proteins. Species, DTU and strain exclusive proteins were identified in this study. It should be noted that the identification of exclusive proteins does not imply the absence of the protein in other groups, but can be related to the abundance below the limit of analytical detection, the presence of modified peptides and/or unique peptide sequences. The number of strain/species and DTU-specific proteins of the human-infective strains is summarized in Fig. 4a, and the proteins and the number of unique peptides for each protein is presented in

Supplementary Data 5. Among the human-infective *T. cruzi* strains, *T. cruzi* 3869, a TcIII strain, had the highest numbers of exclusively identified proteins, of which nine were assigned to single proteins. Of these, four proteins were assigned to the multigene family retrotransposon hot spot (RHS) protein (Q4D2F5, Q4DM62, V5AUD1 and V5BBB7). *Tcbat*TCC1994, a representative strain of the bat restricted *T. cruzi* DTU, had 17 exclusive proteins, of which seven were assigned to single proteins including chaperonin HSP60/CNP60, putative (Fragment)

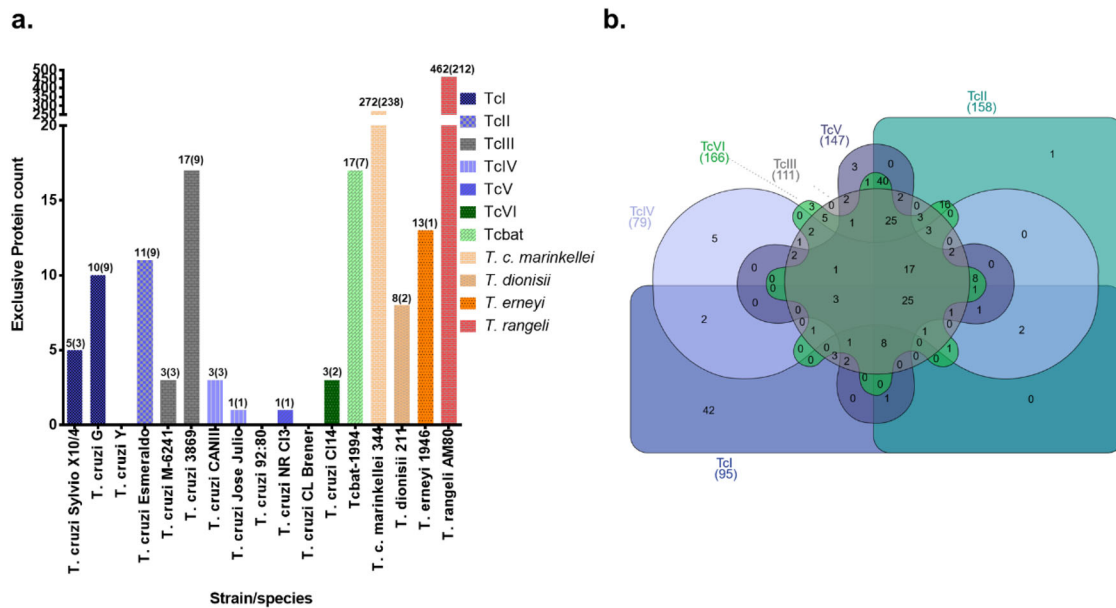


Fig. 4 Analysis of strain/DTU and species-specific proteins identified and quantified by nLC-MS/MS analysis of epimastigote stage parasites at the exponential growth phase. **a** The number of proteins identified exclusively for each *T. cruzi* strain and closely related trypanosome species, (proteins identified exclusively and not in protein groups are highlighted in parenthesis). **b** Venn diagram of DTU-specific proteins identified and quantified in this study.

(K2MB79) and a putative vesicle-associated membrane protein (K2MYF8). In addition, three trans-sialidase protein groups were identified exclusively to Tcbat. *T. c. marinkellei* and *T. rangeli* had the highest numbers of exclusive proteins, with 272 and 462, respectively. On the other hand, a total of 8 and 13 proteins were identified and quantified exclusive for *T. dionisii* and *T. erneyi*, respectively (Supplementary Data 5). The analysis of exclusive human-infective DTU-specific proteins revealed varying numbers of proteins exclusively identified as DTU-specific proteins/protein groups (Fig. 4b). TcI had the highest number (42) of exclusively identified DTU-specific proteins (Supplementary Data 5). One protein, FAD-binding PCMH-type domain-containing protein (Q4DPS1), was identified exclusive to TcII with three unique peptide counts, while no TcIII exclusive protein(s) was identified based on a cutoff of five valid values per DTU. TcIV had five exclusive proteins, including two RHS proteins (Q4DII2/Q4DMF9 and V5B362) and a putative Thiol-dependent reductase 1 (K2N3V3) identified with 17 unique peptides. All three proteins identified exclusive to TcV were putative RHS proteins (Q4DBD1, Q4CKU7 and Q4CU66), while three proteins were identified exclusive to TcVI; RHS protein group (Q4CQ60/V5B746/V5B8V7/V5B362), a putative Hsc70-interacting protein (Hip) (Q4DAT6) and an uncharacterized protein (Q4E501/A0A061IZW3).

Discussion

In this study, an approach to typing and evolutionary inference between organisms based on systems-wide protein expression profiles is described, and was applied to trypanosomes of the subgenus *Schizotrypanum*, a monophyletic assemblage nested in the clade *T. cruzi* comprising mostly trypanosomes of bats, and two species infective to mammals in general, including humans: the pathogenic *T. cruzi* and the non-pathogenic *T. rangeli*^{34,35,59–62}. This approach is based on different biomolecular features obtained by bottom-up large scale mass spectrometry-based quantitative proteomics analysis (MS1, iBAQ and LFQ) followed by MP, an evolutionary criterion to infer evolutionary distances between

strains/species. This analysis requires a nomenclature that describes evolutionary inferences of species or strains using quantitative mass spectrometry data. We propose the term “PhyloQuant” to indicate methodologies which use biomolecular features obtained by mass spectrometry-based methods to infer evolutionary relationships between organisms. Historically, evolutionary relationships are deduced from morphology, or DNA/RNA and predicted protein sequence data based respectively on phenotype, or nucleotide and amino acid sequences from different organisms for their comparisons. Evolutionary trees build using phylogenetic approaches use MP⁶³, maximum likelihood (ML)⁶⁴ and Bayesian inference (BI) methods⁶⁵. PhyloQuant approach uses MP to infer evolutionary relationships between organisms. MP uses the lowest number of evolving steps as a criterion to drive the search for the evolutionary trees, and does not imply the use of an evolutionary model for clustering, which are required for ML and BI inferences. In addition, MP can track changes of a character (synapomorphies), herein differences in expression of different proteins that support each clade. The proteins that support the different trypanosome clades and their differential expression profiles based on LFQ intensities, are shown in Supplementary Data 4. Phyloproteomics compare biological clustering using phylogenetic algorithms based on mass spectrometry protein/peptide peaks appearance or disappearance. Abu-Asad et al. used the term phyloproteomics to classify serum cancer and non-cancerous samples using parsing and a phylogenetic algorithm⁶⁶. In particular, SELDI-TOF MS data were obtained from serum collected from cancer and non-cancer patients and the m/z values were sorted for polarity in order to identify derived vs ancestral peaks using MIX, a phylogenetic algorithm. Another study investigated *C. jejuni* to develop a phyloproteomic typing scheme that combines the analysis of variable masses observed during intact cell MALDI-TOF MS with ribosomal MLST and whole genome MLST database-deduced isoform lists⁶⁷. Another study used the term phyloproteomics to infer evolutionary relationships of bacteria based on mass spectrometry-based identification and protein sequence information⁶⁸. The term PhyloQuant proposed in the present study unifies any method that uses quantitative omics data derived from

mass spectrometry to infer evolutionary relationships. It is important to note that the correspondence between any PhyloQuant and phylogenetic evolutionary relationships should be evaluated in each organism under study, and differences or similarities arising from this comparison holds important evolutionary and/or biological relevance.

PCA based on quantitative MS1 data clustered *T. c. marinkellei* closer to the major clade comprising all *T. cruzi* strains, while PCA based on iBAQ and LFQ intensities placed *T. c. marinkellei* between the clearly separated *T. cruzi* DTUs and the rest of the closely related trypanosome species (Fig. 2). Full genome analyses of generalist *T. cruzi* and *T. c. marinkellei* of bats revealed that they share majority of the core genes, and that *T. c. marinkellei* is also able to invade non-bat cells^{60,69}, including human cells, in vitro. However, although infection with *T. c. marinkellei* appears to be highly restrict to bat, and are unable of infecting mice⁷⁰, differences among the two genomes were mainly a reduced number of repetitive sequences in *T. c. marinkellei* genome compared to *T. cruzi*, with overall small nucleotide polymorphisms, and the identification of several putative exclusive gene sequences including 11 TcMU-CII and an acetyltransferase gene in *T. c. marinkellei* absent in *T. cruzi* Sylvio X10 and CL Brener genomes⁶⁹. *T. c. marinkellei* has been placed in many phylogenetic studies as an outgroup to *T. cruzi* DTUs^{34,59–61,69–71}. Phylogenies inferred from multiple genes on bat trypanosomes clustered *T. cruzi*, *T. c. marinkellei* and the cosmopolitan *T. dionisii* in the *Schizotrypanum* subgenus^{34,59–61,70}, which also included *T. erneyi*, a bat restricted trypanosome placed closest to *T. c. marinkellei*⁶⁰. *T. rangeli*, included as the outgroup for *Schizotrypanum* species, has a wide range of mammalian hosts, which are shared with *T. cruzi* including man, but are not pathogenic to mammalian hosts. Mixed infections of *T. rangeli* with *T. cruzi* have been reported in humans and other mammals, as well as in their shared triatomine vectors^{61,72}. *T. rangeli* belongs to the Tra [Tve-Tco] lineage, which is constituted by trypanosomes unable to develop intracellularly in vitro of unknown life cycle in the mammalian hosts, and is more closely related to Old World *T. vespertilionis* of bats and *T. conorhini* of rats (infective to non-human primates) than to the *Schizotrypanum* species^{61,62}. Based on our study, PCA analysis based on quantitative protein expression datasets from MS1, iBAQ and LFQ quantitative features showed congruence to well-established evolutionary distances separating *T. cruzi* from *T. c. marinkellei*, *T. dionisii*, *T. erneyi* and *T. rangeli*^{60–62}.

Evolutionary relationships based on all three quantitative mass spectrometry features showed that *T. erneyi*, *T. dionisii* and *T. c. marinkellei* diverged earlier compared to *T. cruzi* DTUs^{60–62,69,70}. *T. c. marinkellei* formed the basal clade to *T. cruzi* DTUs, supported by high bootstrap support values in all analyses (Fig. 3 and Supplementary Fig. 2). High levels of correlation between PhyloQuant and phylogenetic distance matrices were evidenced by Mantel test, and are in agreement with findings by Telleria et al.⁵⁶ who demonstrated the high correlation between gene expression, as shown by proteomic analysis using 2D-DIGE, and phylogenetic diversity assayed by multilocus enzyme electrophoresis. Our study demonstrates the potential application of quantitative systems-wide proteomics as a complementary method for taxonomic assignment of *T. cruzi* DTUs and closely related trypanosomes, and provides a valuable link between phylogenetic information and large-scale protein expression in the highly diverse *T. cruzi* strains.

The clustering of all trypanosome replicates in the same clades highlighted the reproducibility of our procedure. Compared to MS1 intensities, the evolutionary relationships of *T. cruzi* strains and closely related trypanosome species based on iBAQ and LFQ

values showed the highest correlation to evolutionary relationships of *T. cruzi* inferred from phylogenies based on multilocus sequence typing, the golden standard proposed for the genotyping of *T. cruzi* in population studies³². Using 32 different gene loci, both separately and in concatenation, Flores-Lopez and Machado⁵¹ reported the paraphyletic nature of *T. cruzi* II strains (now TcII-TcVI). Based on MS1, iBAQ and LFQ intensities, our analyses are in agreement with the clustering of *T. cruzi* TcII-TcVI. Two major clades of *T. cruzi* were described by Flores-Lopez and Machado⁵¹; one with TcI, TcIV, TcIII, and one haplotype from each TcV and TcVI; and the second clade harboring TcII, and the other haplotypes from TcV and TcVI. This branching pattern was corroborated by multilocus analyses including gene sequences from the genome of Tcbat, which was strongly supported as the DTU more closely related to TcI³⁵. In our analyses, statistically significant MS1 based clustering showed a topology with two major groups; one composed of TcI-Tcbat, TcIV, and one representative strain of TcIII, and the second clade composed of TcII, TcV and TcVI, and the second representative strain of TcIII. The close clustering of TcIII and TcI has also been shown based on nuclear, mitochondrial and satellite DNA sequence^{35,50,51,73}.

Two hypotheses have been proposed to explain the evolutionary relationships of *T. cruzi* strains and the origin of the hybrid lineages, TcV and TcVI^{27,28,32,74,75}. Based on PhyloQuant analysis (Fig. 3 and Supplementary Fig. 2, respectively), the close relationship between the parental TcII (*T. cruzi* Esmeraldo and *T. cruzi* Y) and the hybrid TcV and TcVI, was corroborated, in congruence to the well-accepted evolutionary history of *T. cruzi* DTUs. Based on iBAQ and LFQ intensities, TcIII strains formed a clade basal to the clade comprised of parental (TcII) and hybrid (TcV and TcVI) genotypes. Based on MS1 clustering, however, TcIII strains cluster with TcIV and TcV (Fig. 3a). Alternative topologies for TcIII strains have been reported in phylogenetic analysis using different gene markers^{35,50–52}. The close relationship of TcI, TcIII and TcIV has been proposed in the ‘Two hybridization’ model in which initially TcI and TcII hybridized and gave origin to TcIII and TcIV, and a putative hybrid TcIII hybridizes with TcII, with secondary events of hybridization giving origin to TcV and TcVI strains^{32,73–75}. Contributing to clarifying the evolutionary history of DTUs, TcIII was not confirmed as the result of a hybridization event between TcI and TcII based on multiple gene sequences^{50–52,61,71} and analysis of satellite DNA⁷³. The placement of DTU TcIV strains in a single clade is supported by phylogenies based on different genetic markers^{35,50,51,71}, suggesting a possible independent origin of this DTU⁷³.

Phyloquant offers the opportunity to include in a phylogenetic clustering different protein features. This implies that genes exposed or not to selective pressure are jointly considered. Since differentially expressed proteins can be associated to genes exposed to selective pressure, and consequently poor selective markers, using total quantitative proteomic features should leverage this issue. Based on total and statistically regulated iBAQ and LFQ quantitative features, highly similar tree topologies were obtained (Fig. 3 and Supplementary Fig. 2). However, based on unidentified features, statistically significant MS1 features had better resolved trees. Indeed, this method offers the possibility to identify strain/species-specific quantitative markers for discrimination. In our analyses, a total of 4796 proteins were identified and quantified, of which 2733 proteins were identified with three valid values in at least one condition, and used for PhyloQuant analysis based on LFQ intensities. Of these, 2441 proteins were statistically regulated while 292 were not differentially regulated (~10.68%), meaning that the differentially expressed protein pool contributes to the majority of proteins resolving the

differences between the strains compared to the non-regulated proteins. Imputation of non-assigned numbers (NaN) was performed prior to statistical and PhyloQuant analysis. A mass spectrometry data-dependent acquisition method was used, in which precursor ions are selected in a stochastic manner creating missing values between replicates and samples. A way to overcome this limitation would be the use of data independent acquisition where all precursor ions within a certain m/z range will be selected and fragmented having a deeper coverage with less missing values.

Epimastigote stage trypanosomes in the *Schizotrypanum* clade are confined to the digestive tract of triatomine and cimicid bugs, where they replicate in the midgut, and differentiate into metacyclic trypomastigotes in the hindgut from where they are passed through feces contamination to mammalian hosts during a blood meal^{33,60,70}. On the contrary, *T. rangeli* is transmitted via salivarian route. Epimastigotes pass from the triatomine gut to the haemocoel, and then invade the salivary glands from where metacyclic trypomastigotes are transferred to mammalian hosts by bite during a blood meal^{61,72}. The identification of differentially regulated proteins between *T. cruzi*, other species of *Schizotrypanum* and *T. rangeli* in the invertebrate hosts, which are all obligate hematophagous hemipterans, could help elucidate differences in adaptations of the parasites to survive, proliferate, and differentiate exclusively in the gut (*Schizotrypanum*), or invade the hemolymph and then reach the salivary glands (*T. rangeli*). Different strains of *T. cruzi* develop in a wide range of species of several triatomine genera. However, Peterson et al.⁷⁶ showed significant variations in *Rhodnius prolixus* development following infections with different TcI strains. Contrasting with the well-known reference strains of DTUs TcI-TcVI, which are all infective to a range of experimentally infected triatomine species despite differences in both infection and metacyclogenesis levels, Tcbat was incapable of infecting triatomines from laboratory colonies, and its vectors remain to be identified^{32,33,35}. Also controversial is the exclusive transmission of the widespread *T. c. marinkellei* by rare triatomines of the genus *Cavernicola*^{69,70}. The cosmopolitan *T. dionisii* develops similarly to *T. cruzi*, but, apparently exclusively in bat-associated cimicid bugs, which most likely are also vectors of *T. erneyi*, a species so far restricted to bats from Africa, where triatomines are absent^{59–62,70}. The vectors of *T. rangeli* are triatomines of the genus *Rhodnius*, and biogeographical studies have supported strong association between complexes of *Rhodnius* species and the ability of salivary glands invasion by distinct phylogenetic lineages of *T. rangeli*^{61,72}. A proteome-wide characterization of membrane proteins (surfaceome) revealed *T. rangeli* and *T. cruzi*-specific membrane proteins, including Gim5a and Lanosterol 14 α -demethylase, exclusively in the epimastigote stage, and gp63 and Group II trans-sialidases in both epimastigote and trypomastigote life stages⁷⁷. In our analysis, proteins that support separation between *T. rangeli* and the *Schizotrypanum* subgenus were mapped (Supplementary Data 4) based on the statistically significant LFQ intensities. A total of 456 protein expression profiles supported the large divergence of *T. rangeli* from species of the *Schizotrypanum* clade. Differential protein expression between *T. cruzi* and its closest relative trypanosomes from bats (stercorarian) and *T. rangeli* (salivarian) forms a platform to gain new insights into parasite-vector selection, development in the gut, haemocoel and salivary glands, and the regulated processes involved in these interactions.

Proteins exclusively identified in TcI included glucose-regulated protein 78 (V5B942) (Supplementary Data 5). The grp-78 is a member of the HSP70 family in the endoplasmic reticulum, and has been shown to be highly immunogenic against sera from *T. cruzi* infected mice⁷⁸. Cyclophilin (V5B9Y6), a peptidyl-prolyl cis/

trans isomerase enzyme with diverse biological functions, was exclusively identified in *T. cruzi* strain MT3869 (TcIII) as a protein family (V5B9Y6/Q4DJ19/Q4DM35), sharing high sequence similarity. In addition, the expression profile of cyclophilin (V5D7H4, Q4E4G0) was also identified among the synapomorphies which support TcIII clade, and Tcbat (Supplementary Data 4). Cyclophilin is involved in protein folding, stress response, immune modulation, regulation of parasite cell death, and signal transduction. In *T. cruzi*, the cyclophilin gene family is comprised of 15 paralogous genes⁷⁹. The upregulation of a *T. cruzi* cyclophilin, TcCyP19, was demonstrated in benzimidazole resistant populations^{80,81}. DTU-specific RHS proteins for TcI (V5BCL4), TcIV (Q4DII2/Q4DMF9 and V5B362) and TcVI (Q4CQ60/V5B746/V5B8V7) were also identified, in addition to several strain-specific RHS (Supplementary Data 5). Furthermore, *T. cruzi* strain-specific and trypanosome species-specific multigene families including trans-sialidases were also identified. The strain-specific trans-sialidase identified included (K2MR99) exclusive to *T. cruzi* M-6241 (TcIII), and a trans-sialidase protein group (Q4D8B3/V5B8Y6/K2MQD1/K2LZT0) exclusive to *T. cruzi* cl14 strain (TcVI). Three Tcbat (Q4CZE6, Q4D095 and Q4E2A1), one *T. dionisii* (A0A422MS29) and four *T. rangeli* (A0A3R7LX62, A0A3R7M8Q8, A0A422MQ19 and A0A422MT64) exclusive trans-sialidases were also identified. From the total proteins identified and quantified in this study (Supplementary Data 2), multigene families included 122 RHS proteins, 37 trans-sialidases and 20 dispersed gene family protein 1. Furthermore, several multigene family protein expression profiles were identified as synapomorphies that differentiate *T. cruzi* DTUs and related trypanosome species (Supplementary Data 4). The identification of multigene families among the most represented proteins and among the mapped synapomorphies indicate their diversity in expression according to hosts, life stages and developmental niches. This diversity might represent a unique signature for each strain/DTU interacting with its vertebrate and invertebrate hosts in a differential manner. The multigene families constitute 50% of the whole genomes of *T. cruzi* strains, and their content has been shown to be different in phylogenetically distant strains⁸². The differential expression profiles of multigene families in different DTUs and closely phylogenetically related trypanosomes might regulate the infectivity, host restriction, host immune response, intracellular proliferation and transmission⁸².

Three putative trans-sialidases (Q4CZE6, Q4D095, Q4E2A1), proteins with exo- α -sialidase activity involved in cell-invasion and intracellular survival of *T. cruzi* with important roles in the pathogenesis of Chagas disease were identified exclusively in Tcbat. A study by Maeda et al.⁸³ reported the surface molecules of Tcbat shared with TcI and corroborated the infectivity of metacyclic trypomastigotes to human HeLa cells and mice despite lower parasitemia compared to virulent CL and even to the low parasitemic *T. cruzi* G of TcI^{33,84}. Gp82, involved in host cell invasion by the induction of lysosomal exocytosis, was shown to be expressed and highly conserved in Tcbat, CL and G strains⁸³. Mixed infection of Tcbat and TcI has also been reported by DNA detection in a child from northwestern Colombia⁸⁴, and Tcbat DNA has also been identified in mummies from Atacama desert by molecular paleoepidemiological tools⁸⁵. The identification of putative Tcbat proteins or differential protein expression profiles can help to understand the biology and host-parasite interaction of this atypical DTU of *T. cruzi*.

Superoxide dismutase was among the proteins identified as exclusive to *T. c. marinkellei* (K2MYD1) with two unique peptides from seven identified peptides, and *T. rangeli* (A0A3S5IRU3) with two unique peptides from a total of six peptides. Based on a 2DIGE coupled to MS/MS study by Telleria et al.⁵⁶, a superoxide dismutase (Q4DI29) was identified

Table 1 *T. cruzi* strains and closely related trypanosome species analyzed in this study.

TCC	DTU	Trypanosome strain/species	Host	Locality state/country
2158	TcI	<i>T. cruzi</i> Sylvio X10/4	<i>Homo sapiens</i>	Pará, Brazil
874		<i>T. cruzi</i> G. mucurae	<i>Didelphis marsupialis</i>	Amazonas, Brazil
1320	TcII	<i>T. cruzi</i> Y	<i>Homo sapiens</i>	Rio Grande do Sul, Brazil
2188		<i>T. cruzi</i> Esmeraldo	<i>Homo sapiens</i>	Bahia, Brazil
2123	TcIII	<i>T. cruzi</i> M-6241 cl6	<i>Homo sapiens</i>	Pará, Brazil
844		<i>T. cruzi</i> 3869	<i>Homo sapiens</i>	Amazonas, Brazil
2124	TcIV	<i>T. cruzi</i> CANIII cl1	<i>Homo sapiens</i>	Para, Brazil
85		<i>T. cruzi</i> Jose Julio	<i>Homo sapiens</i>	Amazonas, Brazil
1561	TcV	<i>T. cruzi</i> NR cl3	<i>Homo sapiens</i>	Santa Cruz, Bolivia
2121		<i>T. cruzi</i> 92:80 cl2	<i>Homo sapiens</i>	Santa Cruz, Bolivia
476	TcVI	<i>T. cruzi</i> cl14	<i>Triatoma infestans</i>	Rio Grande do Sul, Brazil
2189		<i>T. cruzi</i> CL Brener cl1	<i>Triatoma infestans</i>	Rio Grande do Sul, Brazil
1994	Tcbat	<i>T. cruzi</i> bat cl1.1	<i>Myotis levis</i>	São Paulo, Brazil
344		<i>T. c. marinkellei</i>	<i>Carollia persipicillata</i>	Rondônia, Brazil
211		<i>T. dionisii</i>	<i>Epitesicus brasiliensis</i>	São Paulo, Brazil
1946		<i>T. erneyi</i>	<i>Mops condylurus</i>	Sofala, Mozambique
86		<i>T. rangeli</i> AM80	<i>Homo sapiens</i>	Amazonas, Brazil

TCC Trypanosomatid Culture Collection, Department of Parasitology, University of São Paulo.

exclusively in *T. c. marinkellei*. Superoxide dismutase is a metabolic enzyme that protects the parasites from damage by oxidative stress, and is among proteins shown to have chemotherapeutic promise⁸⁶. Structural proteins including paraflagellar rod proteins were also exclusively identified in *T. c. marinkellei* (K2LTU6) and *T. rangeli*. A Tubulin alpha chain (R9TNG7/A0A061J3L8/K2N695/V5AC55) and two paraflagellar rod 2C (A0A3R7KDB9 and A0A3S5IRI8) proteins were also identified exclusively in *T. rangeli*. Based on 2D-DIGE coupled to mass spectrometry, α -tubulins were identified as exclusive proteins for TcII and TcIV, while β -tubulins were identified exclusive to TcI and TcIII⁵⁶.

In conclusion, the PhyloQuant approach to profile and infer evolutionary relationships of organisms based on large scale quantitative mass spectrometry intensities was described, and validated by the high resolution of the DTUs harboring the highly heterogenetic *T. cruzi* strains, and their closest related trypanosome species from bats. This approach showed high-congruence to well-established phylogenetic clustering. Importantly, the phylogenetic character mapping of the evolutionary trees generated by the PhyloQuant approach allowed the identification of proteins and their expression profiles which support the different *T. cruzi* DTUs and the distinction of all closely related trypanosome species.

Methods

Trypanosome culture. Two representative strains for the six human-infective *T. cruzi* DTUs (TcI–TcVI), one from Tcbat (TcVII), *T. cruzi marinkellei*, *Trypanosoma dionisii*, *Trypanosoma erneyi* and *Trypanosoma rangeli* (Table 1) were comparatively analyzed in this study. These parasites have been previously characterized using genetic markers^{35,61}. The epimastigotes were grown in liver infusion tryptose medium supplemented with 10% fetal calf serum at 28 °C. The cells were harvested at the exponential growth phase by centrifugation at 3000 × g for 10 min and washed twice with ice cold PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, 1.8 mM KH₂PO₄, pH 7.4). Three biological replicates of epimastigote cultures were acquired for each trypanosome to guarantee the reproducibility of the results.

Protein extraction and digestion. The cell pellets were resuspended in protein extraction buffer containing 8 M urea and 1 × protease inhibitor cocktail. The cells were lysed by three cycles of probe tip sonication on ice for 15 s with 15 s cooling intervals. The lysates were briefly vortexed and spun down, and the extracted proteins quantified by Qubit fluorometric detection method. Ammonium bicarbonate was added to a final concentration of 50 mM, followed by reduction using dithiothreitol (DTT) at a final concentration of 10 mM. The reactions were incubated for 45 min at 30 °C, and alkylated by the addition of iodoacetamide to a final concentration of 40 mM and incubation for 30 min in the dark at room

temperature. DTT at a final concentration of 5 mM was added to quench the reaction, and the urea diluted ten times by the addition of NH₄HCO₃. Porcine trypsin was used to digestion of the proteins at a ratio of 1:50 (μ g trypsin/ μ g protein), and the mixture incubated overnight at 30 °C. The resultant peptide mixtures were centrifuged at 10,000 × g for 10 min, and 20 μ g for each sample desalted separately with an in-house C18 micro-column of hydrophilic–lipophilic-balanced solid phase extraction. The peptides were eluted with 100 μ L of 50% (v/v) acetonitrile and 0.1% (v/v) trifluoroacetic acid (TFA).

Nano LC-MS/MS analysis. LC-MS/MS analysis was performed on an EASY-Spray PepMap[®] 50 cm × 75 μ m C18 column using an Easy nLC1000 nanoflow system coupled to Orbitrap Fusion Lumos mass spectrometer (Thermo Fischer Scientific, Waltham, MA, USA). The HPLC gradient was 5–25% solvent B (A = 0.1% formic acid; B = 100% ACN, 0.1% formic acid) in 74 min at a flow of 300 nL/min. The most intense precursors selected from the FT MS1 full scan (resolving power 120,000 at m/z 200) were quadrupole-isolated and fragmented by higher-energy collision dissociation (HCD) and detected in the dual-pressure linear ion trap with 30 as normalized collision energy. The MS1 scan range was between 375 and 1600 m/z, the AGC target was set to 5 × 10⁵, the MS2 ion count target was set to 5 × 10⁴, and the max injection time was 50 and 54 ms for MS1 and MS2, respectively. The dynamic exclusion duration was set to 45 s with a 10-ppm mass tolerance around the selected precursor and its isotopes. The maximum total cycle time was confined to 3 s²³. All raw data have been submitted to PRIDE archive⁸⁷, project accession: PXD017228.

Peptide and protein identification and quantification. LC-MS/MS raw files were analyzed using MaxQuant v1.5.2.8 for identification and LFQ of proteins and peptides. Using the following parameters, MS/MS spectra were searched against the combined reference Uniprot proteome databases of *T. cruzi* CL Brener, *T. cruzi* Dm 28c, *T. cruzi marinkellei* B7 and *T. rangeli* (strains SC58 and AM80) (Released, March, 2020; 50,075 entries) and common contaminants protein database with a mass tolerance level of 4.5 ppm for MS and 0.5 Da for MS/MS. Enzyme specificity was set to trypsin with a maximum of two missed cleavages. Carbamidomethylation of cysteine (57.0215 Da) was set as a fixed modification, and oxidation of methionine (15.9949 Da), deamidation NQ (+0.9840 Da) and protein N-terminal acetylation (42.0105 Da) were selected as variable modifications. The minimum peptide length was set to seven amino acid residues. The ‘match between runs’ feature in MaxQuant which enables peptide identifications between samples based on their accurate mass and retention time was applied with a match time window of 0.7 min and an alignment time window of 20 min. All identifications were filtered in order to achieve a protein false discovery rate (FDR) <1%^{23,57}. Proteins identified in the reverse database, contaminants and proteins identified only by site were excluded prior to performing statistical analysis and evolutionary inferences.

Evolutionary resolution of *Schizotrypanum* lineage. Protein expression profiles evidenced by bottom-up large scale mass spectrometry-based quantitative proteomic features were analyzed to discriminate *T. cruzi* DTUs and closely related trypanosomes, and to infer the evolutionary relationships of the highly phylogenetically related trypanosomes in the subgenus *Schizotrypanum*. *T. rangeli* was included as the outgroup.

The *m/z* values of the parent ions (MS1) are unidentified features which are independent of database search, peptide identification and genome information. The MS1 intensities in the matched features file were filtered to include all peptide intensities with a charge state of ≥ 2 . The cutoff of charge state 1 served to differentiate the majority of peptide ions from non-peptide ions which are single-charged. The filtered MS1 intensity values were normalized based on the total MS1 intensities for each replicate and processed further in the Perseus computational platform version 1.6.10.43⁸⁸. The *m/z* peptide ion intensities were transformed to \log_2 values, and filtered based on three valid values in at least one replicate. The NaN were imputed from normal distribution using a down-shift of 1.8 and distribution width of 0.3. To visualize the variability of the MS1 intensities, PCA was performed on the \log_2 -transformed MS1 intensities. Statistically significant MS1 intensities were determined by analysis of variance (ANOVA) with Benjamini–Hochberg-based FDR correction at an FDR <0.05. To infer evolutionary relationships between the trypanosome species based on MS1 intensities, total and statistically regulated MS1 intensities were normalized by *z* score transformation (standard deviations from means) in Perseus. To build character-based matrix for MP method, the proteomic matrix was transposed to present operational taxonomic units (OTUs) in rows and the intensities in columns. Subsequently, *z* scores were transformed into character states by rounding decimal fractions to the nearest integers, and the negative integers were substituted by corresponding letters. Finally, Tree Analyses using New Technology (TNT)⁸⁹ version 1.5⁹⁰ was employed to infer evolutionary relationships using MP. Branch statistical supports were obtained as implemented in TNT using 1000 replicates. The second quantitative measure used in this study was the iBAQ. This is a measure of absolute protein amounts calculated as the sum of all peptide peak intensities divided by the number of theoretically observable tryptic peptides⁹¹. The iBAQ values generated by MaxQuant software were normalized based on the total intensities for each replicate, loaded onto Perseus, and PCA used to visualize the variability of the proteins based on iBAQ intensities. Statistically significant iBAQ intensities were determined by ANOVA with Benjamini–Hochberg-based FDR correction at an FDR <0.05. To infer evolutionary relationships based on total and statistically significant iBAQ values, MP as implemented in TNT was used as previously described. Evolutionary trees were constructed based on total and statistically significant iBAQ values.

Thirdly, proteome-wide quantitative proteomics based on LFQ of fragmented ions (MS2) were evaluated to determine the discriminatory power of the total and significantly expressed proteins between the strains and species, and to reconstruct the evolutionary relationships of the trypanosomes. The LFQ values were analyzed by Perseus as previously described. To determine proteins with significant changes in abundances, ANOVA was applied with Benjamini–Hochberg-based FDR correction at an FDR <0.05. The total and significantly expressed protein LFQs were normalized using the *z* score and the evolutionary relationships inferred using parsimony as previously described. Phylogenetic character mapping⁹² of the proteins supporting different clades, and their expression profiles, were inferred using the TNT program.

Phylogenetic clustering. Three genes; gGAPDH, V7V8 variable region of the SSU (small subunit) rRNA and heat shock protein-70 (HSP70) were selected and concatenated to construct a phylogenetic reference tree. These gene sequences have been routinely used to discriminate and infer phylogenetic relationships among trypanosomatids in general, resulting in the well-resolved phylogenetic inferences of *Schizotrypanum* trypanosomes and *T. rangeli*⁶¹. Orthologs of the three genes were obtained from the available genomes (complete and drafts) using a standalone BLASTn (v2.2.31+). The genomes from *T. cruzi* (Sylvio X10), *T. cruzi* (CL Brener) and *T. cruzi* (Esmeraldo) are available in TriTrypDB⁹³. *T. cruzi* (Y) genome is available in NCBI-GenBank⁹⁴. The genome drafts from *T. dionisii* (TCC211), *T. erneyi* (TCC1946), *T. rangeli* AM80 (TCC86), *T. cruzi marinkellei* (TCC344), *T. cruzi G* (TCC874) and *T. cruzi M-6241* (TCC2123) have been generated in our laboratories within the ATOL (Assembly of Life, NSF-USA) and TCC-USP (Brazil) projects and were obtained as previously described^{62,71}. Genes for *T. cruzi* MT3869 (TCC844), *T. cruzi* CANIII (TCC2124), *T. cruzi* Jose Julio (TCC85), *T. cruzi* 92:80 (TCC2121) were retrieved from NCBI repository using BLASTn.

SSU rRNA gene from *T. cruzi* CL Brener (TCC2189), gGAPDH genes from *T. cruzi* NR Cl1 (TCC1561) and *T. cruzi* Cl14 (TCC476), and HSP70 genes from *T. cruzi* 92:80 (TCC2121), *T. cruzi* NR Cl1 (TCC1561) and *T. cruzi* Cl14 (TCC476) were amplified by PCR as previously described^{61,95,96}, and subsequently sequenced. All gene sequences were aligned using Clustal Omega within the Seaview v.4.7 graphical tool software⁹⁷. Manual refining of the alignment was performed prior to inferring phylogenies by the parsimony method using TNT. Branch statistical supports were obtained using bootstrap with 1000 replicates.

Phylogenetic and PhyloQuant correlation. Evolutionary trees obtained using PhyloQuant and phylogenetics were converted into distance matrices using Patristic Distance Matrix and their correlation evaluated using the Mantel test using the Analysis of Phylogenetics and Evolution package of the R statistical language environment for statistical computing^{98,99}.

Putative strain/DTU/species-specific protein identification. Putative strain/DTU/species-specific proteins exclusively expressed in epimastigotes harvested in the exponential growth phase of the trypanosomes were determined using the average groups and numeric Venn diagram features implemented in Perseus computational platform. A cutoff of two valid values in at least one group for strain/species were considered for this analysis. Strain/species-specific proteins were represented on a bar graph. Proteins identified and quantified exclusively in the six human-infective DTUs (and not in other trypanosomes) were considered for subsequent analysis to determine specific DTU exclusive proteins. A cutoff of five valid values in at least 1 DTU was considered and mapped on a Venn diagram using InteractiVenn¹⁰⁰.

Statistics and reproducibility. The trypanosome strains and statistical methods applied have been described in the relevant “Methods” section. The sample size was 17 (Table 1), with each sample analyzed in biological triplicates. One replicate each for *T. cruzi* Y strain (TCC1320) and *T. dionisii* (TCC211) were excluded from the final analysis. All PhyloQuant clustering results and statistical correlation can be reproduced according to the parameters described in the “Methods” section.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All mass spectrometry raw data have been submitted to PRIDE archive, project accession: PXD017228. The sequenced genes have been deposited in GenBank under the accession numbers MW345242, MW345243, MW345244, MW345245, MW345246 and MW325707.

Received: 27 January 2020; Accepted: 24 January 2021;

Published online: 11 March 2021

References

- Pérez-Molina, J. A. & Molina, I. Chagas disease. *Lancet* **391**, 82–94 (2018).
- WHO. *Chagas disease (also known as American trypanosomiasis)* (WHO, 2020) <https://www.who.int/news-room/fact-sheets/detail/chagas-disease-american-trypanosomiasis>.
- WHO. *Chagas disease (American trypanosomiasis)* (WHO, 2020). <https://www.who.int/news-room/q-a-detail/chagas-disease>.
- Pedreira De Freitas, J. L. et al. First tests on the accidental transmission of Chagas disease to man by blood transfusion. *Rev. Paul. Med.* **40**, 36–40 (1952).
- Schmunis, G. A. Chagas’ disease and blood transfusion. *Prog. Clin. Biol. Res.* **182**, 127–145 (1985).
- Schmunis, G. A. Prevention of transfusional *Trypanosoma cruzi* infection in Latin America. *Mem. Inst. Oswaldo Cruz* **94**, 93–101 (1999).
- Nobrega, A. A. et al. Oral transmission of Chagas disease by consumption of acai palm fruit, Brazil. *Emerg. Infect. Dis.* **15**, 653–655 (2009).
- Shikanai-Yasuda, M. A. & Carvalho, N. B. Oral transmission of Chagas disease. *Clin. Infect. Dis.* **54**, 845–852 (2012).
- Coura, J. R. The main sceneries of Chagas disease transmission. The vectors, blood and oral transmissions—a comprehensive review. *Mem. Inst. Oswaldo Cruz* **110**, 277–282 (2015).
- Santos, V. et al. Acute chagas disease in the state of Pará, Amazon region: is it increasing? *Mem. Inst. Oswaldo Cruz* **113**, e170298 (2018).
- Bern, C., Martin, D. L. & Gilman, R. H. Acute and congenital Chagas disease. *Adv. Parasitol.* **75**, 19–47 (2011).
- Martins-Melo, F. R., Lima Mda, S., Ramos, A. N. Jr., Alencar, C. H. & Heukelbach, J. Prevalence of Chagas disease in pregnant women and congenital transmission of *Trypanosoma cruzi* in Brazil: a systematic review and meta-analysis. *Trop. Med. Int. Health* **19**, 943–957 (2014).
- Gomez, P. C., Mantilla, H. J. & Rodriguez-Morales, A. J. Fatal chagas disease among solid-organ transplant recipients in Colombia. *Open Forum Infect. Dis.* **1**, ofu032 (2014).
- Anghoben, A. et al. Reactivation of Chagas disease after a bone marrow transplant in Italy: first case report. *Blood Transfus.* **10**, 542–544 (2012).
- Antinori, S. et al. Chagas disease in Europe: a review for the internist in the globalized world. *Eur. J. Intern. Med.* **43**, 6–15 (2017).
- Bern, C., Messenger, L. A., Whitman, J. D. & Maguire, J. H. Chagas disease in the United States: a public health approach. *Clin. Microbiol. Rev.* **33**, <https://doi.org/10.1128/cmr.00023-19> (2019).
- Tyler, K. M. & Engman, D. M. The life cycle of *Trypanosoma cruzi* revisited. *Int. J. Parasitol.* **31**, 472–481 (2001).
- Brener, Z. Biology of *Trypanosoma cruzi*. *Annu. Rev. Microbiol.* **27**, 347–382 (1973).

19. Contreras, V. T. et al. Early and late molecular and morphologic changes that occur during the in vitro transformation of *Trypanosoma cruzi* metacyclic trypomastigotes to amastigotes. *Biol. Res.* **35**, 47–58 (2002).
20. Rodrigues, J. C., Godinho, J. L. & de Souza, W. Biology of human pathogenic trypanosomatids: epidemiology, lifecycle and ultrastructure. *Sub-Cell. Biochem.* **74**, 1–42 (2014).
21. Minning, T. A., Weatherly, D. B., Atwood, J. 3rd, Orlando, R. & Tarleton, R. L. The steady- state transcriptome of the four major life-cycle stages of *Trypanosoma cruzi*. *BMC Genom.* **10**, 370 (2009).
22. Li, Y. et al. Transcriptome remodeling in *Trypanosoma cruzi* and human cells during Intracellular Infection. *PLoS Pathog.* **12**, e1005511 (2016).
23. Avila, C. C. et al. Proteome-wide analysis of *Trypanosoma cruzi* exponential and stationary growth phases reveals a subcellular compartment-specific regulation. *Genes* **9**, <https://doi.org/10.3390/genes9080413> (2018).
24. Dvorak, J. A. The natural heterogeneity of *Trypanosoma cruzi*: biological and medical implications. *J. Cell. Biochem.* **24**, 357–371 (1984).
25. Messenger, L. A., Miles, M. A. & Bern, C. Between a bug and a hard place: *Trypanosoma cruzi* genetic diversity and the clinical outcomes of Chagas disease. *Expert Rev. Anti Infect. Ther.* **13**, 995–1029 (2015).
26. Messenger, L. A. & Miles, M. A. Evidence and importance of genetic exchange among field populations of *Trypanosoma cruzi*. *Acta Tropica* **151**, 150–155 (2015).
27. Berry, A. S. F. et al. Sexual reproduction in a natural *Trypanosoma cruzi* population. *PLoS Negl. Trop. Dis.* **13**, e0007392–e0007392 (2019).
28. Schwabl, P. et al. Meiotic sex in Chagas disease parasite *Trypanosoma cruzi*. *Nat. Commun.* **10**, 3972–3972 (2019).
29. Monje-Rumi, M. M. et al. Potential association of *Trypanosoma cruzi* DTUs TcV and TcVI with the digestive form of Chagas disease. *Infect. Genet. Evol.* **84**, 104329 (2020).
30. Luquetti, A. et al. Recommendations from satellite meeting. *Mem. Inst. Oswaldo Cruz.* (Suppl 1):1–6 (1999).
31. Zingales, B. et al. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. *Mem. Inst. Oswaldo Cruz* **104**, 1051–1054 (2009).
32. Zingales, B. et al. The revised *Trypanosoma cruzi* subspecific nomenclature: Rationale, epidemiological relevance and research applications. *Infect. Genet. Evol.* **12**, 240–253 (2012).
33. Marcili, A. et al. A new genotype of *Trypanosoma cruzi* associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA. *Parasitology* **136**, 641–655 (2009).
34. Pinto, C. M., Kalko, E. K. V., Cottontail, I., Wellinghausen, N. & Cottontail, V. M. TcBat a bat-exclusive lineage of *Trypanosoma cruzi* in the Panama Canal Zone, with comments on its classification and the use of the 18S rRNA gene for lineage identification. *Infect. Genet. Evol.* **12**, 1328–1332 (2012).
35. Lima, L. et al. Genetic diversity of *Trypanosoma cruzi* in bats, and multilocus phylogenetic and phylogeographical analyses supporting Tcbat as an independent DTU (discrete typing unit). *Acta Tropica* **151**, 166–177 (2015).
36. Lewis, M. D., Francisco, A. F., Taylor, M. C., Jayawardhana, S. & Kelly, J. M. Host and parasite genetics shape a link between *Trypanosoma cruzi* infection dynamics and chronic cardiomyopathy. *Cell. Microbiol.* **18**, 1429–1443 (2016).
37. Toledo, M. J. et al. Chemotherapy with benznidazole and itraconazole for mice infected with different *Trypanosoma cruzi* clonal genotypes. *Antimicrobial Agents Chemother.* **47**, 223–230 (2003).
38. Monteiro, W. M. et al. *Trypanosoma cruzi* I and IV stocks from Brazilian Amazon are divergent in terms of biological and medical properties in mice. *PLoS Negl. Trop. Dis.* **7**, e2069 (2013).
39. Perlowagora-Szumlewicz, A., Muller, C. A. & Moreira, C. J. Studies in search of a suitable experimental insect model for xenodiagnosis of hosts with Chagas' disease. 4–The reflection of parasite stock in the responsiveness of different vector species to chronic infection with different *Trypanosoma cruzi* stocks. *Rev. Saude Publica* **24**, 165–177 (1990).
40. Llewellyn, M. S. et al. Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. *PLoS Pathog.* **5**, e1000410–e1000410 (2009).
41. Barros, J. H. S. et al. Identification of novel mammalian hosts and Brazilian biome geographic distribution of *Trypanosoma cruzi* TcIII and TcIV. *Acta Tropica* **172**, 173–179 (2017).
42. Telleria, J. et al. *Trypanosoma cruzi*: sequence analysis of the variable region of kinetoplast minicircles. *Exp. Parasitol.* **114**, 279–288 (2006).
43. Steindel, M., Neto, E. D., de Menezes, C. L. P., Romanha, A. J. & Simpson, A. J. G. Random amplified polymorphic DNA analysis of *Trypanosoma cruzi* strains. *Mol. Biochem. Parasitol.* **60**, 71–79 (1993).
44. Souto, R. P., Fernandes, O., Macedo, A. M., Campbell, D. A. & Zingales, B. DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. *Mol. Biochem. Parasitol.* **83**, 141–152 (1996).
45. Perez, E. et al. Predominance of hybrid discrete typing units of *Trypanosoma cruzi* in domestic *Triatoma infestans* from the Bolivian Gran Chaco region. *Infect. Genet. Evol.* **13**, 116–123 (2013).
46. Cura, C. I. et al. *Trypanosoma cruzi* I genotypes in different geographical regions and transmission cycles based on a microsatellite motif of the intergenic spacer of spliced-leader genes. *Int. J. Parasitol.* **40**, 1599–1607 (2010).
47. Tomasini, N. et al. Interest and limitations of Spliced Leader Intergenic Region sequences for analyzing *Trypanosoma cruzi* I phylogenetic diversity in the Argentinean Chaco. *Infect. Genet. Evol.* **11**, 300–307 (2011).
48. Mejia-Jaramillo, A. M. et al. Geographical clustering of *Trypanosoma cruzi* I groups from Colombia revealed by low-stringency single specific primer-PCR of the intergenic regions of spliced-leader genes. *Parasitol. Res.* **104**, 399–410 (2009).
49. Messenger, L. A., Yeo, M., Lewis, M. D., Llewellyn, M. S. & Miles, M. A. Molecular genotyping of *Trypanosoma cruzi* for lineage assignment and population genetics. *Methods Mol. Biol.* **1201**, 297–337 (2015).
50. Yeo, M. et al. Multilocus sequence typing (MLST) for lineage assignment and high resolution diversity studies in *Trypanosoma cruzi*. *PLoS Negl. Trop. Dis.* **5**, e1049 (2011).
51. Flores-Lopez, C. A. & Machado, C. A. 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 (2011).
52. Diosque, P. et al. Optimized multilocus sequence typing (MLST) scheme for *Trypanosoma cruzi*. *PLoS Negl. Trop. Dis.* **8**, e3117 (2014).
53. Sauer, S. & Kliem, M. Mass spectrometry tools for the classification and identification of bacteria. *Nat. Rev. Microbiol.* **8**, 74–82 (2010).
54. Önder, Ö., Shao, W., Kemps, B. D., Lam, H. & Brisson, D. Identifying sources of tick blood meals using unidentified tandem mass spectral libraries. *Nat. Commun.* **4**, 1746 (2013).
55. Schimpf, S. P. et al. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* **7**, e48 (2009).
56. Telleria, J. et al. Phylogenetic character mapping of proteomic diversity shows high correlation with subspecific phylogenetic diversity in *Trypanosoma cruzi*. *Proc. Natl Acad. Sci. USA.* **107**, 20411–20416 (2010).
57. de Oliveira, G. S. et al. Development of a *Trypanosoma cruzi* strain typing assay using MS2 peptide spectral libraries (Tc-STAMS2). **12**, e0006351, <https://doi.org/10.1371/journal.pntd.0006351> (2018).
58. Baker, J. R., Miles, M. A., Godfrey, D. G. & Barrett, T. V. Biochemical characterization of some species of *Trypanosoma* (Schizotrypanum) from bats (Microchiroptera). *Am. J. Trop. Med. Hyg.* **27**, 483–491 (1978).
59. Hamilton, P. B., Teixeira, M. M. & Stevens, J. R. The evolution of *Trypanosoma cruzi*: the 'bat seeding' hypothesis. *Trends Parasitol.* **28**, 136–141 (2012).
60. Lima, L. et al. Evolutionary insights from bat trypanosomes: morphological, developmental and phylogenetic evidence of a new species, *Trypanosoma* (Schizotrypanum) *erneyi* sp. nov., in African bats closely related to *Trypanosoma* (Schizotrypanum) *cruzi* and allied species. *Protist* **163**, 856–872 (2012).
61. Espinosa-Alvarez, O. et al. *Trypanosoma rangeli* is phylogenetically closer to Old World trypanosomes than to *Trypanosoma cruzi*. *Int. J. Parasitol.* **48**, 569–584 (2018).
62. Bradwell, K. R. et al. Genomic comparison of *Trypanosoma conorhini* and *Trypanosoma rangeli* to *Trypanosoma cruzi* strains of high and low virulence. *BMC Genom.* **19**, 770 (2018).
63. Fitch, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Biol.* **20**, 406–416 (1971).
64. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
65. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
66. Abu-Asab, M., Chaouchi, M. & Amri, H. Phyloproteomics: what phylogenetic analysis reveals about serum proteomics. *J. Proteome Res.* **5**, 2236–2240 (2006).
67. Zautner, A. E., Masanta, W. O., Weig, M., Gross, U. & Bader, O. Mass Spectrometry-based PhyloProteomics (MSPP): a novel microbial typing Method. *Sci. Rep.* **5**, 13431 (2015).
68. Deshpande, S. V. et al. ABOid: a software for automated identification and phyloproteomics classification of tandem mass spectrometric data. *J. Chromatogr. Separat. Techniq.* **55:001**, <https://doi.org/10.4172/2157-7064> (2011).
69. Franzen, O. et al. Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi* marinkellei. *BMC Genom.* **13**, 531 (2012).
70. Cavazzana, M. Jr. et al. Phylogeographical, ecological and biological patterns shown by nuclear (ssrRNA and gGAPDH) and mitochondrial (Cyt b) genes of trypanosomes of the subgenus *Schizotrypanum* parasitic in Brazilian bats. *Int. J. Parasitol.* **40**, 345–355 (2010).

71. Caballero, Z. C. et al. Phylogenetic and syntenic data support a single horizontal transference to a *Trypanosoma* ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences. *Parasites Vectors* **8**, 222 (2015).
72. Da Silva, F. M. et al. Phylogeny, taxonomy and grouping of *Trypanosoma rangeli* isolates from man, triatomines and sylvatic mammals from widespread geographical origin based on SSU and ITS ribosomal sequences. *Parasitology* **129**, 549–561 (2004).
73. Ramirez, J. C., Torres, C., Curto, M. L. A. & Schijman, A. G. New insights into *Trypanosoma cruzi* evolution, genotyping and molecular diagnostics from satellite DNA sequence analysis. *PLoS Negl. Trop. Dis.* **11**, e0006139 (2017).
74. de Freitas, J. M. et al. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. *PLoS Pathog.* **2**, e24 (2006).
75. Westenberger, S. J., Barnabe, C., Campbell, D. A. & Sturm, N. R. Two hybridization events define the population structure of *Trypanosoma cruzi*. *Genetics* **171**, 527–543 (2005).
76. Peterson, J. K., Graham, A. L., Dobson, A. P. & Chavez, O. T. Rhodnius prolixus life history outcomes differ when infected with different *Trypanosoma cruzi* I strains. *Am. J. Trop. Med. Hyg.* **93**, 564–572 (2015).
77. Wagner, G. et al. The *Trypanosoma rangeli* trypomastigote surfaceome reveals novel proteins and targets for specific diagnosis. *J. Proteom.* **82**, 52–63 (2013).
78. Tibbetts, R. S. et al. Molecular cloning and characterization of the 78-kilodalton glucose- regulated protein of *Trypanosoma cruzi*. *Infect. Immun.* **62**, 2499–2507 (1994).
79. Perrone, A. E., Milduberg, N., Fuchs, A. G., Bustos, P. L. & Bua, J. A functional analysis of the cyclophilin repertoire in the protozoan parasite *Trypanosoma Cruzi*. *Biomolecules* **8**, <https://doi.org/10.3390/biom8040132> (2018).
80. Andrade, H. M. et al. Proteomic analysis of *Trypanosoma cruzi* resistance to Benzimidazole. *J. Proteome Res.* **7**, 2357–2367 (2008).
81. Rego, J. V. et al. Molecular characterization of Cyclophilin (TcCyP19) in *Trypanosoma cruzi* populations susceptible and resistant to benzimidazole. *Exp. Parasitol.* **148**, 73–80 (2015).
82. De Pablos, L. M. & Osuna, A. Multigene families in *Trypanosoma cruzi* and their role in infectivity. *Infect. Immun.* **80**, 2258–2264 (2012).
83. Maeda, F. Y., Alves, R. M., Cortez, C., Lima, F. M. & Yoshida, N. Characterization of the infective properties of a new genetic group of *Trypanosoma cruzi* associated with bats. *Acta Tropica* **120**, 231–237 (2011).
84. Ramirez, J. D. et al. First report of human *Trypanosoma cruzi* infection attributed to TcBat genotype. *Zoonoses Public Health* **61**, 477–479 (2014).
85. Guhl, F., Auderheide, A. & Ramirez, J. D. From ancient to contemporary molecular eco- epidemiology of Chagas disease in the Americas. *Int. J. Parasitol.* **44**, 605–612 (2014).
86. Olmo, F. et al. An in vitro iron superoxide dismutase inhibitor decreases the parasitemia levels of *Trypanosoma cruzi* in BALB/c mouse model during acute phase. *Int. J. Parasitol. Drugs Drug Resist* **5**, 110–116 (2015).
87. Vizcaino, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
88. Tyanova, S., Temu, T. & Sinitcyn, P. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
89. Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).
90. Pignone, M., Greth, K. M., Cooper, J., Emerson, D. & Tang, J. Identification of mycobacteria by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry. *J. Clin. Microbiol.* **44**, 1963–1970 (2006).
91. Schwanhausser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
92. Avise, J. C. *Molecular markers, natural history and evolution*. 2nd edn, (Chapman & Hall, 2011).
93. Aslett, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* **38**, D457–D462 (2010).
94. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **45**, D12–D17 (2017).
95. Noyes, H. A., Stevens, J. R., Teixeira, M., Phelan, J. & Holz, P. A nested PCR for the ssrRNA gene detects *Trypanosoma binneyi* in the platypus and *Trypanosoma* sp. in wombats and kangaroos in Australia. *Int. J. Parasitol.* **29**, 331–339 (1999).
96. Borghesan, T. C. et al. Molecular phylogenetic redefinition of Herpetomonas (Kinetoplastea, Trypanosomatidae), a genus of insect parasites associated with flies. *Protist* **164**, 129–152 (2013).
97. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer Appl. Biosci.* **12**, 543–548 (1996).
98. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
99. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
100. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinforma.* **16**, 169 (2015).

Acknowledgements

We acknowledge the use of the Willi Hennig Society edition of TNT. We are thankful to Prof. Erney Camargo for his constant support discussing the project and results, and Marta Campaner for trypanosome cultures. The National Science Foundation supported genome sequencing of some trypanosomes included in the present study (PI Gregory A. Buck, Virginia Commonwealth University, NSF DEB-0830056 Assembling the Tree of Life program: Phylum Euglenozoa). The work was supported by grants and fellowships from FAPESP (2014/06863-3, 2018/18257-1, 2018/15549-1, 2014/25494-9 to GP, 2018/13283-4 to GSDO and SNM (2017/04032-5)), CNPq (DOQ and “Productivity fellowship” to GP), CNPq (Universal Program and “Productivity fellowship” to MGMT) and postdoctoral fellowships from CAPES (AGCM and LRF).

Author contributions

Conceptualization and experimental design: G.P. and S.N.M.; Methodology: S.N.M., A.G.C.-M., L.R.-F., G.S.d.O., C.M.F.R., D.Q., G.E.R., M.M.G.T. and G.P.; Data curation: S.N.M., A.G.C.-M., L.R.-F., G.S.d.O. and G.P.; Formal analysis: S.N.M., A.G.C.-M. and G.P.; Visualization: S.N.M., L.R.-F. and C.M.F.R.; Funding acquisition: G.P.; Project administration: G.P.; Resources: G.E.R., G.P. and M.M.G.T.; Software: S.N.M. and A.G.C.-M.; Supervision: G.P.; Writing—original draft: S.N.M., A.G.C.-M., M.M.G.T. and G.P.; Writing—review and editing: S.N.M., A.G.C.-M., L.R.-F., C.M.F.R., G.E.R., M.M.G.T. and G.P. All authors edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-01762-6>.

Correspondence and requests for materials should be addressed to G.P.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021