BMC Genomics

RESEARCH

Open Access

An algorithm for annotation and classification of *T. cruzi* MASP sequences: towards a better understanding of the parasite genetic variability

Aldana Alexandra Cepeda Dean[1,2], Luisa Berná[3,4], Carlos Robello[3,5], Carlos Andrés Buscaglia[1,2*†] and Virginia Balouz[1,2*†]

Abstract

Background *Trypanosoma cruzi*, the protozoan causing Chagas disease, is responsible for a neglected tropical disease affecting millions in Latin America. Its genome contains rapidly evolving multigene families, such as mucins (TcMUC), trans-sialidases (TS), and mucin-associated surface proteins (MASP), which are essential for parasite transmission and disease mechanisms. However, methodological challenges in genome assembly and annotation have limited the characterization of these gene families, particularly MASPs.

Results We developed a bioinformatic pipeline for the automatic identification, characterization, and annotation of MASPs directly from *T. cruzi* genome assemblies. This algorithm, based on a manually curated MASP database and HMM-based identification of MASP diagnostic motifs, enables the robust classification of these molecules into canonical MASPs, MASP-related molecules (mostly pseudogenes), and chimeric sequences combining MASPs and TcMUC/TS genes. Validation against a rigorously annotated dataset demonstrated high accuracy, and allowed us to reclassify misanotated sequences and, more crucially, to accurately identify previously unrecognized canonical MASPs and MASP chimeras. This algorithm was then used to explore the MASP repertoire in the genomes of 13 parasite strains from different evolutionary lineages, revealing patterns of diversity. For instance, TcI and TcII strains exhibited higher ratios of canonical MASP/MASP-related molecules and a greater abundance of MASP chimeras, suggesting that their genomes are under strong selective pressures towards maintaining a broader panel of full-length MASP genes at the expense of pseudogenes. On the contrary, structural features of canonical MASPs, MASP-related sequences, and MASP-chimeras were largely conserved across parasite genomes.

[†]Carlos Andrés Buscaglia and Virginia Balouz contributed equally to this work and should be thus considered senior authors.

*Correspondence:
Carlos Andrés Buscaglia
cbuscaglia@iib.unsam.edu.ar
Virginia Balouz
vbalouz@iib.unsam.edu.ar

Full list of author information is available at the end of the article

Conclusions This novel pipeline automates the annotation of MASPs, a key surface protein family unique to *T. cruzi*, improving genome annotation and enabling robust comparative analyses. It provides an essential tool for exploring the evolutionary dynamics of multigene families in *T. cruzi* and could be extended to other gene families.

Keywords *Trypanosoma cruzi*, Mucin-associated surface proteins (MASP), Hidden Markov Models (HMM), Molecular signatures, Genome annotation

Background

Chagas disease is a zoonosis of great medical and socio-economic importance, caused by the protozoan *Trypanosoma cruzi* [1]. In endemic Latin American areas, infection with this parasite occurs mainly by the vectorial route, i.e., through contact of mucous membranes or wounds with the feces of infected hematophagous triatomine bugs. However, humans can also become infected with *T. cruzi* through ingestion of tainted food and fluids, contaminated blood transfusion or organ transplantation, and from mother-to-child during pregnancy/delivery. Current chemotherapy is only effective in acute cases and may display severe side effects, hence pressing the need to identify new potential diagnostic and therapeutic targets [2].

T. cruzi presents a highly structured population, with multiple strains showing differential eco-epidemiological features and extensive genetic/phenotypic diversity [3]. Certain studies have also shown a partial correlation between the parasite genotype and the clinical course of the infection [3, 4]. Biochemical and genetic typing schemes converged in the delineation of 7 evolutionary lineages or discrete typing units (DTUs), named TcI to TcVI and TcBat [5]. TcI, TcII, TcIII and TcIV have more ancient origins whereas TcV and TcVI are clusters of hybrid strains, the product of relatively recent genetic crosses between TcII and TcIII parentals. The seventh DTU named TcBat is genetically affiliated to TcI, and has been mostly found in neotropical bats [3].

The first draft of the *T. cruzi* genome (CL Brener clone, TcVI) was produced in 2005, using Sanger technology [6]. The resultant genome revealed a great complexity, with over 50% of the parasite genome lacking synteny, i.e., conservation of gene order and disposition, with those of phylogenetically related organisms such as *Trypanosoma brucei* and *Leishmania major* [6]. Due to its repetitive nature, the *T. cruzi* genome determined a highly fragmented assembly, in which chromosome number and structure could not be obtained [6]. In 2009, however, a total of 41 large contigs ('pseudo chromosomes') for each haplotype could be assembled using BAC end sequences and synteny maps with the *T. brucei* genome [7].

More recently, the application of next-generation sequencing (NGS) methods based on long reads, such as PacBio and nanopore sequencing, has significantly improved *T. cruzi* genome assembly [8–13]. In particular, these technologies allowed for the resolution of the non-syntenic fraction of the genome, which was found to segregate from the syntenic or "core" compartment based on gene composition and GC content, leading to its designation as the "disruptive" compartment [8]. The latter, turned out to be enriched in repetitive sequences, mainly represented by highly evolving gene families made up of tens or hundreds of alleles with varying degree of polymorphism [7, 14, 15]. These code for virulence factors that play pivotal roles in parasite-host interaction such as mucins (TcMUC), *trans*-sialidases (TS), Gp63, Dispersed Gene Family-1 (DGF-1), Serine-, Alanine-, Proline-rich proteins (SAP) and Retrotransposon hot spot proteins (RHS) [14, 16, 17]. Structural features of the *T. cruzi* genome were found to correlate with its genomic architecture and epigenetic modifications [18–20].

Sequencing of the *T. cruzi* genome also revealed a large and novel family of polymorphic sequences (>1,300 copies), which were termed mucin-associated surface proteins (MASP) because they were found to be preferentially linked to TcMUC genes [6]. Subsequent genetic sequencing and annotations of strains belonging to the DTUs TcI, TcII and TcVI yielded from 1,045 to 1,398 MASP sequences per genome [8, 10, 11]. As verified for TcMUC, MASP expression is largely coordinated from multiple *loci* and up-regulated on infective, trypomastigote forms [15, 21–25]. MASP deduced products are characterized by the presence of flanking regions coding for a signal peptide (SP) and a glycosylphosphatidylinositol (GPI)-anchoring signal, a diagnostic feature of *T. cruzi* surface-associated molecules [26]. Likely due to their key role in MASP surface disposition, these flanking regions are under strong selection pressure against diversification. Within the central region, however, MASP proteins display a mosaic-like structure, made up of strikingly variable and repetitive, 8 to 50 amino acid (aa)-long sequence blocks shuffled among its members [6, 15, 27]. These central regions display a biased aa composition and are predicted to undergo substantial post-translational modifications, including phosphorylation and glycosylation [28–30]. Functional studies, though restricted to a few molecules, have shown that surface-displayed MASPs may be involved in the engagement with host cell receptors, thereby contributing to parasite invasion [31, 32].

From a genetic standpoint, the MASP mosaic-like configuration suggests that new variants majorly emerge via recombination, either between alleles showing partial

sequence conservation or within a single *locus* due to the fairly common presence of repeated sequences [13, 15]. Retrotransposon-like elements frequently found in MASP *loci* vicinities may also play a role in this phenomenon, most likely by promoting the occurrence of DNA lesions [13, 33, 34]. Though not proven, the idea of recombination as the driving force in MASP evolution finds additional support in (i) the presence of multiple pseudogenes; and (ii) the identification of chimeras showing sequences from MASPs and other gene families such as TS, TcMUC or SAP [6, 15]. Further accumulation of point mutations and/or indels, particularly in the central and mature region of MASP molecules, contribute to increase their inter-allelic diversity.

Independently of the underlying mechanisms of evolution, amplification and diversification of MASP sequences, and of the multigene families in general, seems to be an adaptive trait of the *T. cruzi* genome [35, 36]. In this framework, studies aimed at assessing the differences in their dosage and extent of allele variation across strains are expected to provide relevant insights into parasite biology and pathogenesis. However, only a few of such studies, and limited to a short number of genomes, were so far undertaken. More importantly, these studies were often biased by methodological constraints, as they compared genomes assembled using different sequencing technologies and annotated using different protocols, most of which migrate errors introduced at the time of CL Brener genome annotation, thus creating a negative feedback loop [8, 9, 11–14, 35, 37]. To tackle this issue, a new Illumina read-based methodology was recently developed. This approach relies on the counting of short sequences (~30 nt-long) diagnostic of each gene family, and is thereby independent of allele-specific read mapping and of *de novo* genome assembly and annotation [38]. Implementation of this method allowed for a more accurate estimation of differences in copy number and sequence variability of MASP, TcMUC, and TS genes among *T. cruzi* strains [38]. Unfortunately, intrinsic limitations of the method, i.e. impossibility of sorting genes from pseudogenes and/or chimeras, narrowed the appreciation of the entire genetic landscape of these multigene families.

In this work, we introduce an automatic algorithm designed for the identification, classification and annotation of MASPs directly from *T. cruzi* genome datasets. This algorithm leverages a combination of bioinformatics and molecular parasitology strategies, and may be easily extended to the study of other *T. cruzi* multigene families. This gene annotation-independent approach lays the foundation for robust comparative and evolutionary genomics studies in this relevant pathogen.

Methods

Database compilation, curation, and protein analysis

MultiFASTA files containing annotated proteins from *T. cruzi* Brazil A4 and TCC strains, available in the public database TriTrypDB [39], were compiled. Python scripts were used to filter sequences based on the information provided in their headers, selecting those that contained the acronym 'MASP'. Redundant sequences were identified by GenomeTools using the sequniq command [40]. Protein alignments were generated in ClustalW and visualized in Jalview [41, 42]. Divergent MASP sequences, characterized as those lacking a defined terminal end—either due to being chimeric, truncated, or associated with potential pseudogenes—were manually identified in the alignments and either discarded or edited prior to inclusion in the final database. Additionally, sequences containing the MASP GPI signal but extending downstream due to stop codon shifts were included. To validate these cases, raw genomic data were mapped to the corresponding contigs using Artemis v.17.0.1 [43] confirming frameshifts and the absence of stop codons within the reading frame of the annotated MASP sequence. The presence of functional SP and GPI signals was assessed using the SignalP 6.0, predGPI and NetGPI 1.1 servers [44, 45]. The aa composition of sequences was calculated using a custom Python script. To validate the novel MASP sequences, multiple sequence alignments were conducted using MAFFT v.7 [46] (web interface) with the BLOSUM45 scoring matrix and leaving all the remaining parameters as default values. The resulting alignment was encoded using the same scoring matrix, processed through custom Python scripts. Principal component analysis (PCA) was then performed on this encoded data using the PCA tool available in the scikit-learn library [47]. The MASP database used for comparative purposes included the 1,249 curated sequences from the TCC and Brazil A4 strains. Putative parental sequences for MASP-chimeras were identified through BLASTn (default parameters) analysis against the available *T. cruzi* database, selecting the best hits based on the lowest *E*values. Chimerization events were further analyzed using RDP4 v.4.101 [48], with default parameters, by gene-wide pairwise comparison between MASP-chimeras and putative parental sequences. For the generation of TcMUC and TS databases, the same curation pipeline described for MASP was employed. In these cases, annotated MultiFASTA files from TCC, Brazil A4 and CL Brener strains were compiled and sequences containing the acronyms 'trans-sialidase' and 'TcMUC' on their headers were selected.

Identification of molecular signatures and generation of probability matrices

N- and C-terminal ends (30 and 40 aa, respectively) from MASPs, TS or TcMUC sequences included in curated databases were extracted separately, filtered by redundancy using GenomeTools and aligned with ClustalW. For MASPs, repeated sequences from the conserved N- and C-terminal ends (23 and 35 aa, respectively) were filtered out by redundancy, re-aligned and the aa variability assessed using probability matrices generated with PyHMMER 0.10.14 [49]. For TcMUC and TS, Regular Expressions (RegEx) were generated upon tables with aa per position using a custom Python script. Sequence logos were created using WebLogo3 [50].

MASP classification algorithm assembly and annotation pipeline

The algorithm was developed using Python 3.11.5, incorporating the Pandas 1.4.2 and PyHMMER 0.10.14 libraries. The pipeline consists of several modules, each designed for specific tasks in the assembly and annotation process. The annotation module takes any protein multiFASTA file and processes it to search for MASP molecular signatures as described below. Since genomic datasets were the main target of our analysis, a pre-processing pipeline is previously executed to find, extract, translate, and compile ORFs > 120 bp from genomic assemblies into a multiFASTA file. This task is performed using the GetORF tool from EMBOSS [51], with the following settings: -find 1; -minsize 120. The multiFASTA file containing the translated ORFs is then used as input for our annotation program. The algorithm first takes each ORF and scrutinizes it for internal Met residues that could define shorter polypeptides. The ORF predicted by GetORF and the set of ORFs derived from it (and being > 40 aa), are compiled in a multiFASTA file as members of the same 'holo-ORF'. It must be noted that all ORFs belonging to the same holo-ORF are encoded in the same frame and share the same STOP codon. Information for each holo-ORF (sequences, genomic coordinates) is also stored. In the second step, each ORF is scanned for MASP diagnostic motifs within its N- and C-terminal regions (30 and 40 aa, respectively) using probability matrices generated with PyHMMER. Based on this analysis, ORFs are classified as 'MASP' (if motifs are present at both termini), 'MASP-related' (if a motif is found at either terminus), or 'non-MASP' (if motifs are absent at both termini). MASP-related molecules are re-scanned on their terminal ends, now looking for TcMUC or TS specific signatures. In case of positive results, they are classified as 'MASP-chimera'; otherwise, they remain classified as 'MASP-related'. To prevent overestimation, we implemented a hierarchical ranking system to annotate one ORF per holo-ORF: (i) MASP; (ii) MASP-chimera; (iii) MASP-related and (iv) Non-MASP. The ORF with the highest-ranked classification is annotated at the corresponding genomic position. In cases where the highest-ranked classification is shared by two or more members of the same holo-ORF, the longest sequence is selected for annotation. To further prevent overestimation, if two MASP-related sequences are located on the same DNA strand, within less than 1,800 bp of each other -which approximates the maximal length observed for annotated MASP genes [8, 11]- and one sequence contains only the N-terminal motif while the other contains only the C-terminal motif, they are considered parts of the same sequence. These sequences are concatenated, generating a single sequence that possesses both MASP termini and classified as "MASP-related". The algorithm produces several outputs:

1. Comprehensive CSV Table: Contains detailed information for each sequence, including identifiers, genomic coordinates, strand orientation, final classification, N- and C-terminal types, *E* values, and bit scores for each MASP match.
2. multi-FASTA and GFF Files: Lists sequences according to their MASP classification—MASP, MASP-related, and MASP-chimeras.
3. multi-FASTA File of Holo-ORFs: Includes all holo-ORFs generated by the algorithm.
4. README File: Summarizes user-provided information and presents the final count of classified MASP types.

*E*value and bit score cutoff settings

We separated the annotated proteins from the *T. cruzi* Dm28c strain [40] into two categories "MASP" and "non-MASP" according to their existing annotations. The complete proteome was combined with 90,000 random peptides (70 aa-long each) in a multiFASTA file. These peptides were generated using a custom Python script and labeled as 'non-MASP' and created randomly without any specific selection criteria. The generated peptides and the Dm28c proteins were analyzed exclusively using the HMMER module, omitting the internal methionine (Met) scanning step described earlier. Default parameters were applied (*E* value ≤ 10; bit score > 0). The HMMER search was restricted to the N-terminal (first 30 aa) and C-terminal (last 40 aa) regions of each input sequence. For each match, the *E* value and bit score were recorded and plotted using GraphPad Prism v8.0.2. Based on these plots, optimal cutoff values were established manually: for the N-terminal region: *E* value $< 10^{-5}$ and bit score > 25, and for the C-terminal region: *E* value $< 10^{-7}$ and bit score > 30.

Complete genomes

Genomes of *T. cruzi* isolates Brazil A4 (TcI), Dm25 (TcI), Dm28c (TcI), Bug2148 (TcI), Berenice (TcII), YC6 (TcII), TCC (TcVI), Tula cl4 (TcVI) and of *T. brucei* Lister 437 strain were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/genome) and TritrypDB (http://tritrypdb.org/tritrypdb). For the evaluation of the algorithm, the Dm28c -version 66- annotated protein and GFF files were used. Though initially reported as TcV [52], multiple evidences indicate that Bug2148 belongs to TcI [53–56]. We also included in the analysis the genomes of *T. cruzi* RA (TcVI), which was recently obtained at IIBio-UNSAM/CONICET using PacBio RSII technology (Balouz et al., unpublished data), and the genomes of *T. cruzi* Merjo (TcIII), MT3663 (TcIII), Jose Julio (TcIV) and BolFc10A (TcV), which were sequenced at the Instituto Pasteur using PacBio and Nanopore technologies (Greif et al., unpublished data). A list of manually identified Dm28c MASP pseudogenes was kindly provided by Florencia Diaz-Viraqué (Institut Pasteur de Montevideo, Uruguay).

Results

An algorithm for MASP identification, classification and annotation

In order to identify MASP diagnostic molecular signatures, we started by generating a MASP database from the *T. cruzi* Brazil A4 and TCC isolates. These strains represent different parasite DTUs (TcI and TcVI, respectively), with high-quality genomes [8, 11]. Due to its hybrid nature (bearing TcII-like and TcIII-like haplotypes), the inclusion of TCC was expected to contribute further to the diversity of the database. Indeed, from the total of 1,423 MASPs initially retrieved, 941 were from TCC and 482 from Brazil A4 (Table 1 and Additional Table 1).

After redundancy filtering, unique sequences were aligned and those showing evident structural divergence were pinpointed. Some of them presented deletions or truncations, which preferentially affected their C-terminal region. They were more frequent in Brazil A4 ($n=90$, 17.6%) than in TCC ($n=15$, 1.6%), and all of them were discarded (Table 1). Other divergent sequences ($n=42$, all from TCC) displayed the insertion of a peptide of variable length upstream of the conserved sequence block at the N-terminal end of most MASPs (Additional Fig. 1). These were manually edited, i.e. the N-terminal end was removed and an internal Met residue that coincides with the one determined as the initial for the bulk of MASPs was assigned as the translation initiation site. Following the N-terminal editing, neo-sequences were predicted to have gained a functional SP, hence supporting the validity of this procedure (Additional Fig. 1).

Although in smaller numbers, we also identified MASP sequences ($n=5$, all from Brazil A4) that exhibited a slightly divergent C-terminus. A closer inspection of their DNA sequences revealed that these MASPs turned out not to be actual sequences but artifacts generated during genome annotation [11], and were therefore excluded from our database (Table 1). Finally, and considering the proposal of chimeric MASPs [6, 15], the N- and C-terminal ends of sequences showing polymorphisms at these otherwise highly conserved blocks were subjected to BLAST analysis against the parasite protein database. This exercise revealed 30 MASPs displaying > 90% identity with members of TcMUC or TS at either terminal end, which were also removed (Table 1 and Additional Fig. 1).

Our final set of MASPs included 1,249 sequences, 877 from TCC and 372 from Brazil A4 (Table 1 and Additional Table 1). These molecules bore as low as ~20% sequence identity between them, though they were unified by certain structural features. The length range and aa composition, for instance, were very similar between strains, and closely matched those calculated for an independently annotated MASP dataset (Additional Fig. 1) [11]. In addition, all of them showed the typical modular design of *T. cruzi* MASPs, with highly conserved N- and C-ends and a strikingly variable central region (Additional Fig. 2). Considering this general structure, generationof molecular signatures was focused on the terminal regions. To that end, sequences from the N- and C-end of each protein were extracted and compiled in separate lists, both of which were then filtered for redundancy and

Table 1 Generation of MASP database

Step	Procedure	TCC	Brazil A4	Total
Collection of sequences	Extraction of 'MASP' annotated sequences	941	482	1,423
Curation of divergent sequences	Redundancy	30*	9*	39*
	Truncations/deletions	15*	85*	100*
	N-terminal extensions	42**	0	42**
	Terminal frameshifts	0	5*	5*
	Chimeras	19*	11*	30*
Final database		877	372	1,249

* These sequences were not included in the final MASP database

** These sequences were manually edited and included in the final MASP database

aligned to generate probability matrices using HMMER (Additional Fig. 2).

To enable the algorithm to not only identify MASPs but also detect MASP chimeras, we also looked for molecular signatures in the flanking regions of TcMUC and TS. Briefly, we generated curated databases for TcMUC and TS molecules following basically the same protocol described above (Additional Files 1 and 2). These datasets were independently aligned and sequence logos were derived from terminal ends (Additional Figs. 3 and 4). At variance with what has been observed in MASPs, TcMUC and TS molecules presented more variability in their flanking regions, which allowed for the formation of several homology groups. Maximal heterogeneity was found in the C-terminal region of TS, upon which 9 clusters could be defined (Additional Fig. 4). Based on these sequence clusters, TcMUC and TS molecular signatures were generated using RegEx. Of note, and despite some extent of overlapping, our grouping of TcMUC and TS based on their flanking regions did not exactly match previous attempts of clusterization of these families based on whole-sequence alignments [57–59]. A diagram showing the collection and curation of MASP, TcMUC and TS sequences, and the generation of molecular signatures upon them is presented in Fig. 1A.

When implemented on protein datasets, the algorithm directly takes input sequences and scans them for MASP signatures. Considering MASP overall structure (Additional Fig. 2), the search space was restricted to 30 and 40 aa from the N- and C-terminal regions, respectively. Proteins showing hits at both ends are classified as 'MASP', whereas those yielding double negative results are classified as 'non-MASP'. In a third possible scenario, proteins may be recognized as MASP solely by one end. Such molecules are annotated as 'MASP-related' and re-scanned on their terminal ends, now looking for TcMUC or TS specific signatures. In the case of positive results, they are classified as 'MASP-chimera'; otherwise, they remain classified as 'MASP-related' (Fig. 1B).

We also developed a module for MASP identification and classification directly from genome assemblies. On a first step, ORFs > 120 bp are identified throughout the dataset, translated and compiled using the GetORF tool from EMBOSS. The resulting file is used as input of the algorithm, and each translated ORF is next scrutinized for Met residues that could mark the initial position of internal, shorter polypeptides. The original ORF predicted by GetORF and the set of ORFs > 40 aa derived from it, are compiled as members of the same 'holo-ORF'. On a second step, each ORF is scanned for MASP diagnostic motifs in its flanking regions (and eventually for TcMUC/TS signatures) and classified as described above. Following the assessment of all ORFs, classifications obtained by the members of the same holo-ORF are ranked using the following hierarchical order: (1) MASP, (2) MASP-chimera, (3) MASP-related, (4) Non-MASP. The ORF displaying the best ranked classification is annotated in the corresponding genomic position. In cases where the best ranked classification is shared by two or more members of the holo-ORF, the longest of them becomes annotated (Fig. 1B).

Calibration of the algorithm parameters

The resolution power of our algorithm relies on the positive/negative recognition of signatures within input sequences, which in turn depends on the established homology cutoffs. Therefore, we first calibrated HMMER parameters (*E* value and bit score) by assessing the accuracy of predictions on the annotated proteome of the *T. cruzi* Dm28c strain [8]. For simplicity, we re-categorized this well-defined dataset, made up of 15,319 proteins, into two groups of molecules: MASPs ($n = 736$) and non-MASPs ($n = 14{,}583$), the latter comprising all annotated proteins with no reference to 'MASP' in the header and/or description. To simulate more stringent conditions, such as those encountered when evaluating genomic datasets, we supplemented the non-MASP group with 70 aa-long random peptides ($n = 90{,}000$). The flanking regions of each sequence from either group were evaluated using HMMER default parameters (*E* value cut-off ≤ 10; bit score > 0), and the *E* value and bit score of the actual matches were informed.

For the N-terminus, 673 out of 736 MASPs (91.4%) yielded positive results, with mean *E* values of 6.3×10^{-14} and mean bit scores of 53.89 (Fig. 2). Manual inspection of the 63 apparent 'false negative' results, i.e. sequences annotated as MASPs but not identified by our algorithm through N-terminal scanning, revealed that they correspond either to chimeras ($n = 12$), all of them displaying a typical TcMUC N-terminal region and a MASP C-terminus, or to MASP sequences likely translated from a premature START codon ($n = 51$), similar to those found during database curation (Table 1 and Additional Fig. 1). As for the 104,583 non-MASP sequences (14,583 proteins from Dm28c + 90,000 random peptides), only 1 of them yielded a positive match (Fig. 2). This corresponded to a chimera displaying MASP N-terminal region and TS C-terminal domain, that has been annotated as TS (Fig. 2, inset 1).

A larger fraction of MASPs (725/736, 99%) were recognized by the C-terminus (mean values of 2.65×10^{-19} and 71.21 for *E* value and bit score, respectively; Fig. 2). Of the 11 apparent 'false negatives', 4 turned out to be MASP-chimeras displaying TS C-terminal sequences, 3 corresponded to MASP sequences truncated at their C-terminal region and the remaining 4 to MASP sequences displaying an insertion downstream of the GPI-anchor motif. In addition, 1 of the 725 positive hits

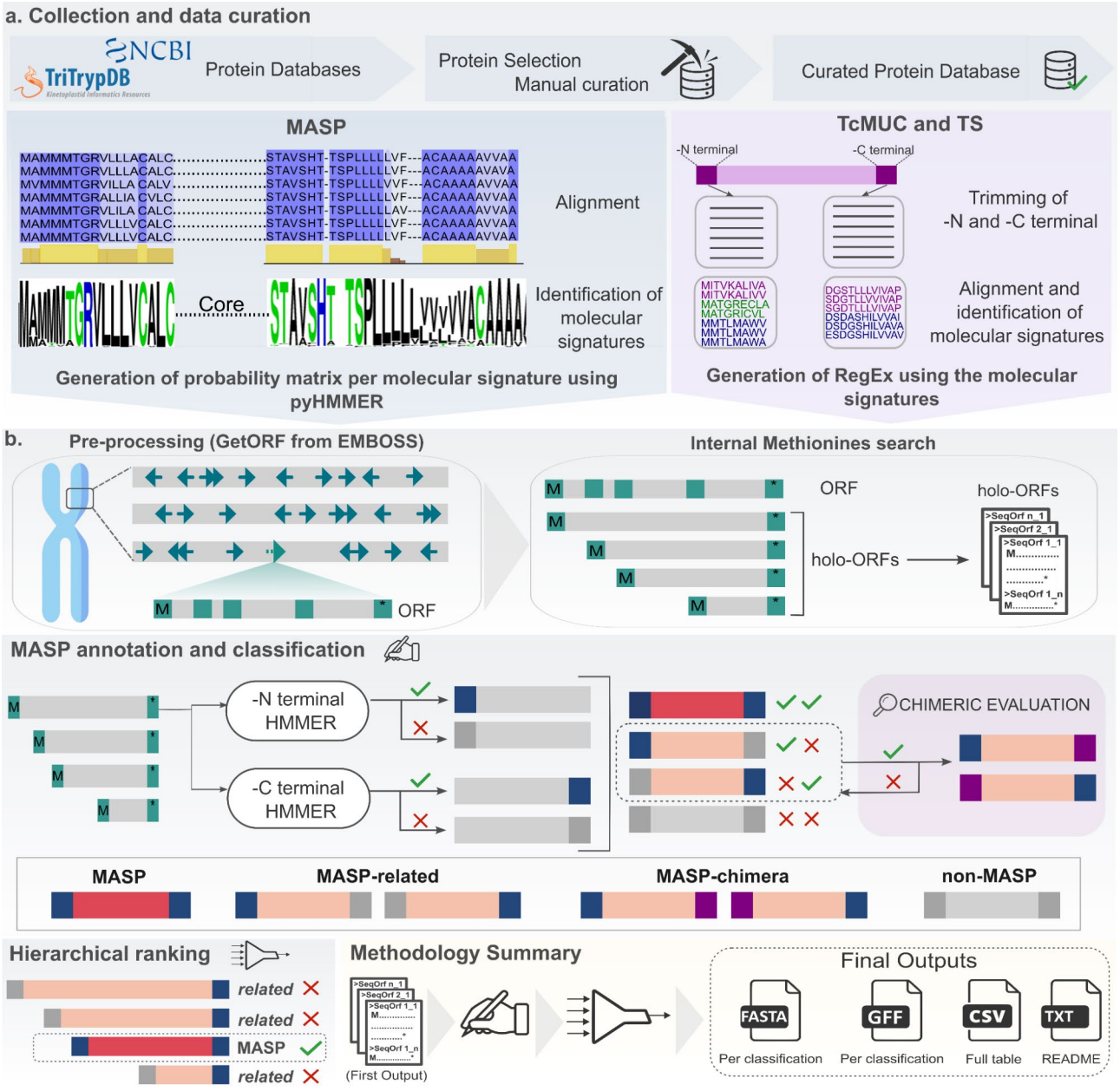

Fig. 1 Protocol for MASP identification, classification and annotation. (**A**) Diagram of the generation of sequence databases and derived molecular signatures. (**B**) Schematic representation of the annotation pipeline and functioning of the algorithm. Terminal regions of the proteins are shown in blue (MASP signature), purple (TS/TcMUC signature) or gray (no match), and central regions in red (MASP classification), pink (MASP-related or MASP-chimera classification) or gray (non-MASP classification)

showed very poor scores (Fig. 2). A closer inspection of this sequence revealed a MASP molecule with a divergent C-terminal region, which is nevertheless predicted to encompass a functional GPI-signal (Fig. 2, inset 2). Further analysis of genomic sequences allowed us to identify part of the canonical MASP C-terminal signature and the STOP codon in a different reading frame (Fig. 2, inset 2). This finding suggests the occurrence of a 'terminal' frameshift, caused by a sequencing error and/or a single nt indel, underlying the variability of this allele. A similar phenomenon could be invoked to explain the above-mentioned 4 MASP sequences displaying an insertion downstream of the GPI-anchor motif, and also the 5 discordant Brazil A4 MASPs identified during database curation (Table 1). Out of the 104,583 non-MASP proteins, 13 (0.01%) were recognized by the MASP C-terminal signature, though with very poor homology scores (*E* values > 1 and bit scores < 10, Fig. 2). These 'false positives' corresponded to TS, Gp63 or TASV [60] sequences

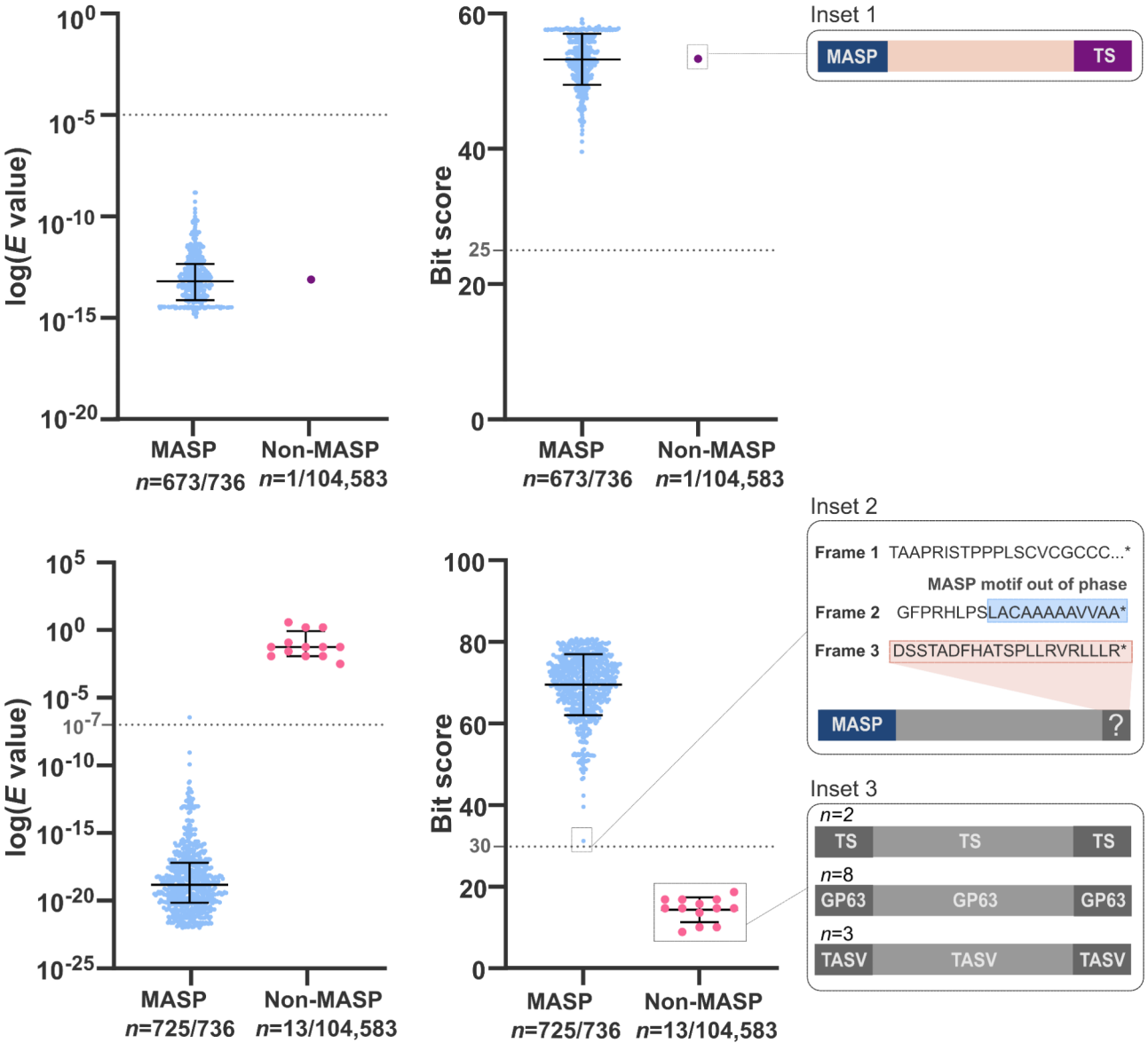

Fig. 2 Setting the parameters of the algorithm. Dm28c proteins annotated as MASPs ($n=736$) or non-MASPs ($n=14{,}583$) were evaluated by the HMMER module, and for those displaying a MASP signature at the N-terminal (upper panels) or C-terminal region (bottom panels), the *E* value and bit score of the corresponding match are shown (light blue dots). The mean value ± SD for each population is also shown. The non-MASP group was added to 90,000 random proteins, thus totaling 104,583 sequences. Pink dots represent manually inspected sequences without MASP signatures, while violet dots indicate chimeras. The cutoff values are indicated with dotted lines. Insets 1 to 3 provide structural details of particular cases (see text). Terminal regions of the proteins schematized in the insets are colorized as in Fig. 1B

displaying a GPI-anchoring motif with a rather similar structure to that of MASPs (Fig. 2, inset 3).

Based on these results, we established the following cutoff values for our HMMER searches: N-terminal region, *E* value $< 10^{-5}$ and bit score > 25; C-terminal region, *E* value $< 10^{-7}$ and bit score > 30 (Fig. 2). Disregarding erroneous annotation cases already discussed, these cutoffs allowed the algorithm to perform with maximal accuracy (100% sensitivity and 100% specificity) on the Dm28c annotated proteome while, at the same time, they are not so stringent as to preclude the capture of MASP diversity.

Evaluation of the algorithm

We next assessed the overall functioning and output of the algorithm by carrying out *de novo* MASP identification and classification upon the genome of the Dm28c strain. As was the case for TCC and Brazil A4 strains, this is a high-quality genome, obtained using NGS technologies [8, 11]. Importantly, this genome has been solved using PacBio, based on long-reads technology, which

results in less collapsing of repetitive regions and multigene families (such as MASPs) during the assembly.

A total of 303,265 ORFs were *de novo* identified by GetORF on the Dm28c genome. Each ORF was further expanded to a holo-ORF through internal Met scanning, thereby broadening the repertoire of sequences to be evaluated, and each holo-ORF was finally classified as MASP, MASP-related, MASP-chimera or non-MASP as described above (Fig. 1B). This classification was then contrasted with the complete annotated dataset from Berná et al. [8], which consisted of 736 MASPs and 18,210 non-MASPs. The latter included 14,583 non-MASP proteins, 1,629 pseudogenes and 1,998 sequences displaying other types of annotations such as tRNAs, ncRNAs, rRNAS, transposons, etc. For a better comparison, a set of manually identified MASP pseudogenes ($n=249$) was included in the analysis as a separate category.

From the 790 MASPs identified, 713 have been already annotated as MASPs (and at the same genomic coordinates), whereas 77 were 'novel', rescued either from the non-MASP pool ($n=4$), the non-annotated fraction of the genome ($n=1$) or, mainly ($n=72$), from the set of MASP pseudogenes (Fig. 3A). Further characterization of the 77 'novel' MASPs showed that they bear the overall structure (length, aa composition) of canonical MASPs (Fig. 3B). A principal component analysis (PCA) also revealed that the novel MASP sequences cluster with sequences from our curated MASP dataset, suggesting the presence of related molecules in the TCC and/or Brazil A4 strains (Fig. 3B). However, PCA results should be taken cautiously due to the limitation of the represented variances. Further details of these novel Dm28c MASPs are provided in Additional Table 2.

Our algorithm also allowed for the identification of 317 MASP-related molecules in the Dm28c genome (Fig. 3A), which presented a different overall structure than canonical MASPs (Fig. 3C, left panel). Most of them were found in the non-annotated fraction of the genome ($n=141$) or were listed as MASP pseudogenes ($n=130$), with only a minor fraction ($n=23$) coming from the non-MASPs pool. In addition, 23 MASP-related sequences identified by our algorithm have been annotated as MASP by Berná et al. (Fig. 3A), hence warranting further inspection. A closer look at these sequences supported our 'MASP-related' classification: 8 of them presented C-terminal truncations (and thereby lacked a functional GPI anchor) and the remaining 15 corresponded to MASP-chimeras (Fig. 3C).

As for the non-MASPs, we found 43 minor 'discrepancies', i.e. sequences classified as non-MASP by our algorithm that were included in the list of MASP pseudogenes (Fig. 3A). Though not further analyzed, these may correspond to heavily corrupted MASP sequences, displaying remnants of the central and mature region but lacking recognizable terminal signatures.

A total of 27 MASP-chimeras were identified by our algorithm, either among MASP genes ($n=15$, see above), MASP pseudogenes ($n=10$) or the non-MASP pool ($n=2$). One of the latter corresponded to the MASP-chimera displaying a typical TS C-terminal domain and annotated as TS previously described (Fig. 2, inset 1). Notably, the configurations of Dm28c MASP-chimeras were not random: TcMUC signatures ($n=15$) were invariably detected at the N-terminus and always involved the most abundant type of N-terminal TcMUC motif (#2, Additional Fig. 3). TS signatures ($n=12$), on the other hand, were always found at the C-terminus of MASP-chimeras, and corresponded to TS C-terminal motifs #8 or #9 (Additional Fig. 4), which largely corresponded to previously described TS groups IV and V [58].

To get further insights into MASP-chimeras, we performed genome-wide BLAST searches using these sequences as bait. For two of them (one TcMUC-MASP and one MASP-TS), a detailed similarity analysis with putative 'parental' MASP and TS/TcMUC sequences retrieved in our screenings was undertaken. Both MASP-chimeras display a mosaic structure, with TS/TcMUC sequences extending well into the central region of the molecule (Fig. 3D).

Assessing MASP diversity in *T. cruzi*

We finally used our algorithm to explore the global diversity of the MASP family in *T. cruzi*. NGS genomes sequenced using long-read-based methodologies from 13 parasite isolates from six major DTUs were analyzed using the pipeline described in Fig. 1B. As a control, we also scanned the *T. brucei* Lister 437 genome. Though devoid of MASPs, the *T. brucei* genome is rich in SP- and/or GPI-containing molecules that may display structural resemblance to MASP terminal signatures [61]. For each genome, ORFs were *de novo* predicted, further expanded to holo-ORFs through internal Met scanning and classified.

E values and bit scores obtained for every match found in each set of classifications (MASP, MASP-related and non-MASP molecules) is provided in Fig. 4. Notwithstanding subtle differences, all these profiles (value ranges, mean/median values, dispersion of the data) were highly conserved across *T. cruzi* strains. This trend applied to TcI and TcVI isolates, i.e. strains bearing genotypes 'related' to the genomes used to assemble our MASP database, as well as to those bearing more distant genotypes such as Berenice (TcII), YC6 (TcII), Merjo (TcIII), MT3663 (TcIII), Jose Julio (TcIV) and BolFc10A (TcV). Though preliminary, these findings suggest that most of the variability of MASP terminal regions was contemplated in our curated repository, and hence in the

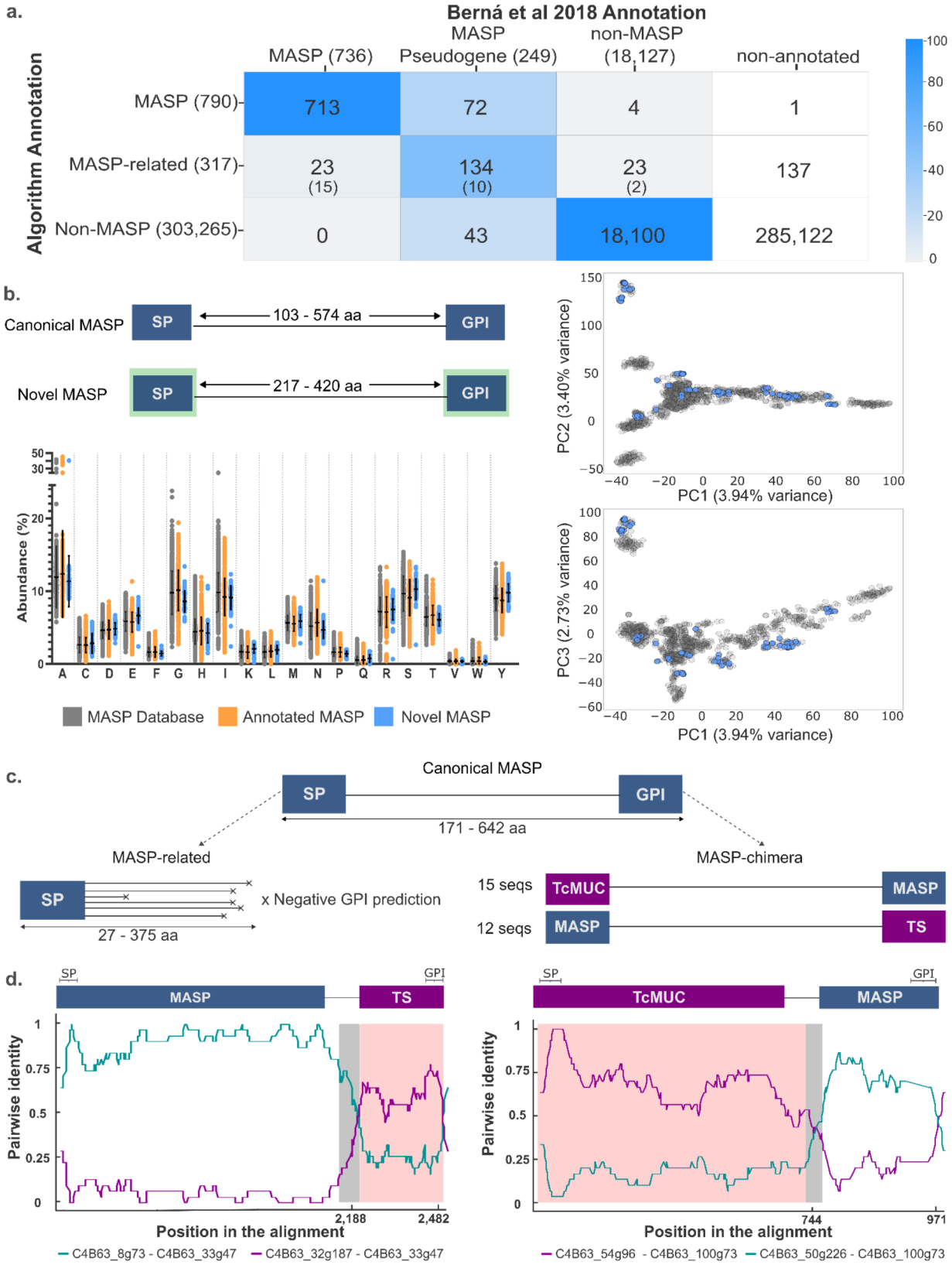

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Evaluation of the algorithm. (**A**) Total counting of MASP classifications on the genome of the Dm28c strain as assessed by our algorithm and by Berná et al. (2018) [8]. The percentage of agreement between classifications is indicated with a color scale. MASP-chimeras are indicated between parentheses, as part of the algorithm-predicted MASP-related group. (**B**) Novel MASPs identified in the Dm28c genome ($n=77$) display quite similar structure (functionally predicted SP and GPI are highlighted in green) (i) and aa composition (ii) than Dm28c MASPs previously annotated by Berna et al. ($n=713$), and to those comprised on our curated MASP database ($n=1,249$). The abundance of each aa was determined by summing its total occurrences across the MASP sequence and dividing the result by the sequence length and expressed as percentage. iii) PCA analyses showing the relationship between MASP proteins from our curated database (gray dots) and 'novel' MASP proteins found in Dm28c (blue dots). (**C**) Schematic illustration of the structure and length-range of canonical MASP, MASP-related molecules and MASP-chimeras found in Dm28c. SP, Signal peptide; GPI, GPI-anchoring signal. (**D**) Sliding-window Simplot graphs showing changing patterns of sequence similarity between MASP-chimeras (C4B63_33g47, left, C4B63_100g73, right) and putative 'parental' MASP, TS and TcMUC genes. Simplots were generated using sequence alignment of the three indicated genes with a window size of 200 nt and a step size of 20 nt. Diagrammatic representations of the ensuing MASP-chimeras are shown above each panel. Breakpoint confidence intervals (99%) are indicated in gray, and the track of the 'non-MASP' part of the sequence is highlighted in pink

HMMER matrices derived thereupon. More interestingly, our results also revealed close similarity between the data profiles recorded for MASP and MASP-related molecules, independently of the strain and/or the signature analyzed in Fig. 4). These findings suggest that the secretory signals for both kinds of molecules are under similar selection pressures against diversification.

Though more heterogeneous, the profiles recorded for non-MASPs were also similar among strains, and also similar to *T. brucei* which, as expected, yielded only non-MASP classifications (Fig. 4). These non-MASP matches were notably reduced in numbers and, most importantly, they presented much worse homology scores as compared to those of MASP and MASP-related molecules (Fig. 4). As shown, the vast majority of *E* values and bit

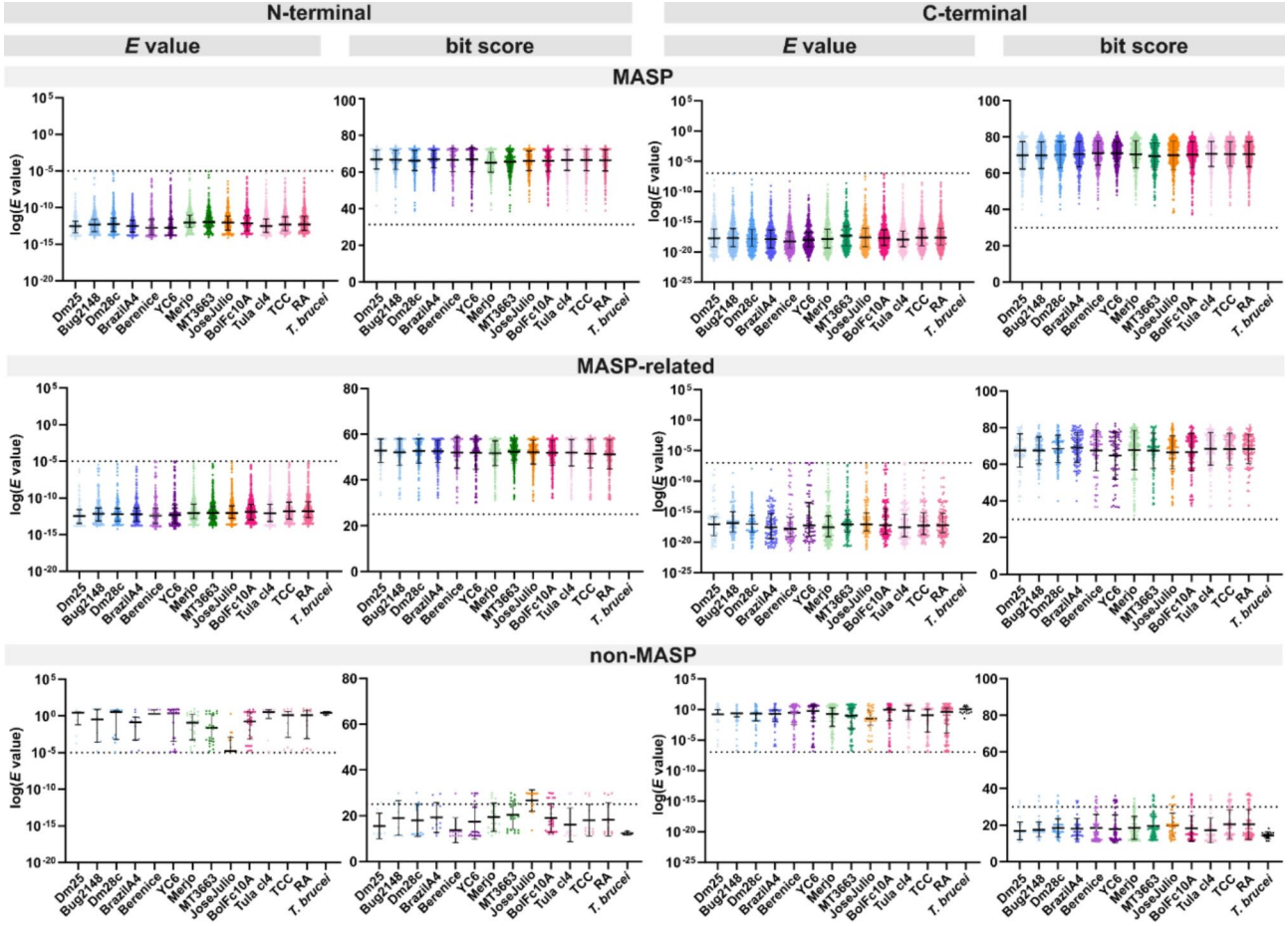

Fig. 4 Diversity of MASP signatures found in *T. cruzi*. ORFs were *de novo* predicted upon the genome of the indicated *T. cruzi* strains, evaluated by our algorithm and annotated accordingly. Individual *E* values and bit scores of MASP signatures at the N- or C-terminal region of each classification group (MASPs, MASP-related or non-MASPs) is shown. For each data population, the median ± interquartile range for the *E* values and the mean ± SD bit scores are also shown. The cutoffs established during the calibration of the algorithm are indicated with dotted lines

scores recorded for non-MASP matches from every population fell well below the cutoffs established during the calibration of the algorithm (indicated with dotted lines). Moreover, the few that displayed 'positive' bit scores were still classified as non-MASP due to 'negative' *E* values, thus reinforcing the importance of in parallel evaluation of both homology parameters during HMMER searches (Fig. 4).

The total counting of MASP classifications obtained from the genomes is shown in Fig. 5A; Table 2. Even though quantitative/qualitative differences on the MASP repertoire among strains may be due in part to differences in the accuracy of genome sequencing and/or assembly, certain strain- or DTU-specific patterns may be appreciated. For instance, and as previously suggested [8, 9, 11, 37], hybrid DTUs (TcV and TcVI) displayed a higher MASP dosage (including MASPs, MASP-related and MASP-chimeras) as compared to TcI, TcII, TcIII and TcIV lineages. These dosages ranged from 1,307 to 1,511 sequences in TcV/TcVI to 952-1,469 sequences in ancestral DTUs (Table 2). However, it should be noted that this trend was reversed after normalization by genome size. As shown in Table 2, hybrid strains, except for Tula cl4 (26.97 sequences/Mb), exhibited lower MASP densities than ancestral TcI-TcIV strains (16.55–19.88 vs. 20.78–23.63 sequences/Mb). Of note, ancestral strains Dm25 (TcI, 17.47 sequences/Mb) and MT3663 (TcIII, 17.22 sequences/Mb) displayed MASP densities well within the range observed for hybrid strains (Table 2).

With the aim of conducting an exploratory qualitative study, we analyzed the MASP/MASP-related ratios across DTUs. TcI (except for Brazil A4) and TcII strains displayed the highest MASP/MASP-related ratios (1.79–2.72) (Table 2). TcIII and TcIV strains, on the other hand, presented much lower MASP/MASP-related ratios (ranging from 0.58 to 0.67), whereas hybrid TcV and TcVI

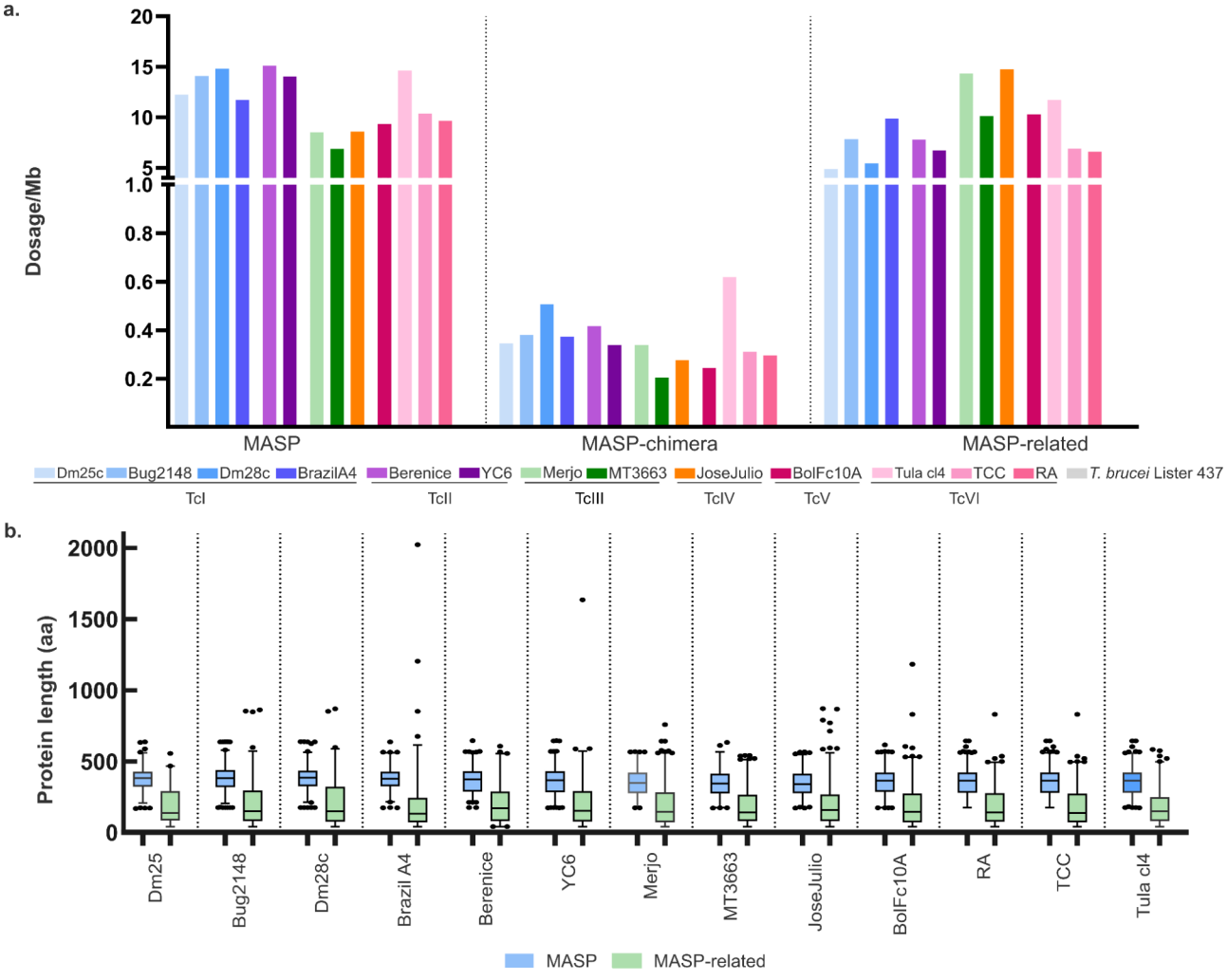

Fig. 5 MASP diversity in ***T. cruzi***. (**A**) Dosage (relativised to genome size, in Mb) of MASP, MASP-chimeras and MASP-related molecules, as determined by our algorithm on the genomes of the indicated *T. cruzi* isolates and the *T. brucei* Lister 437 genome. (**B**) Box-plots showing the length-range of MASP and MASP-related molecules found in *T. cruzi* genomes. The median with 1–99 percentile for each population is indicated. Outliers are indicated with dots

Table 2 MASP repertoire in *T. cruzi* strains

DTU	Isolate	Genome size (bp)	MASP*	MASP-chimera*	MASP-related*	Non-MASP*	MASP/MASP-related ratio
TcI	Dm25	84,079,963	1,029 / 12.24	29 / 0.34	411 / 4.89	469,755 / 5,587.00	2.50
	Bug2148	55,157,397	777 / 14.09	21 / 0.38	432 / 7.83	306,597 / 5,558.58	1.80
	Dm28c	53,271,887	790 / 14.83	27 / 0.51	290 / 5.44	303,265 / 5,692.78	2.72
	Brazil A4	45,556,784	534 / 11.72	17 / 0.37	428 / 9.39	256,507 / 5,630.49	1.25
TcII	Berenice	40,801,262	617 / 15.12	17 / 0.42	318 / 7.79	225,861 / 5,535.64	1.94
	YC6	47,218,089	663 / 14.04	16 / 0.34	317 / 6.71	268,520 / 5,686.80	2.09
TcIII	Merjo	55,977,494	477 / 8.52	19 / 0.34	802 / 14.33	314,118 / 5,611.51	0.59
	MT3663	63,532,971	437 / 6.88	13 / 0.20	644 / 10.14	358,474 / 5,642.33	0.68
TcIV	Jose Julio	58,064,173	499 / 8.59	16 / 0.28	857 / 14.76	325,978 / 5,614.10	0.58
TcV	BolFc10A	73,971,694	692 / 9.35	18 / 0.24	761 / 10.29	413,732 / 5,593.11	0.91
TcVI	Tula cl4	48,462,332	709 / 14.63	30 / 0.62	568 / 11.72	270,698 / 5,585.74	1.25
	TCC	87,060,361	904 / 10.38	27 / 0.31	600 / 6.89	473,826 / 5,442.50	1.50
	RA	91,340,476	882 / 9.66	27 / 0.30	602 / 6.59	468,041 / 5,124.14	1.47
T.brucei	Lister 437	50,081,021	0 / 0	0 / 0	0 / 0	201,109 / 4,015.67	N.A.

* Values are expressed as dosages (above) and densities (in copies/Mb, below)

strains displayed an intermediate scenario, with MASP/MASP-related ratios ranging from 0.90 to 1.51 (Table 2). These findings suggest that the genomes of hybrid strains compile the MASP allele repertoire from both parental strains and, more importantly, that TcI and TcII lineages are under strong selective pressures towards maintaining a broader panel of full-length MASP genes at expenseof MASP-related molecules (putatively pseudogenes). A preliminary characterization of the complete set of MASP and MASP-related molecules identified by our algorithm revealed that they exhibit quite similar structural features across strains (Fig. 5B).

MASP-chimeras were detected in every analyzed strain (Fig. 5A). In quantitative terms, TcI and TcII isolates showed the highest (0.34–0.51 copies/Mb) and TcIII and TcIV had the lowest MASP-chimera densities (0.20–0.34 copies/Mb). As verified for the MASP/MASP-related ratios, hybrid strains showed intermediate values of MASP-chimeras (0.24–0.31 copies/Mb) compared to TcII and TcIII (Table 2). Of note, Tula cl4 (TcVI), which showed an extremely high MASP dosage, also presented the highest density of MASP-chimeras (0.62 copies/Mb, Table 2). In qualitative terms, preliminary analyses indicate that most (but not all) of the identified MASP-chimeras displayed the same structural features as those observed during database curation and algorithm evaluation, i.e. TcMUC-MASP and MASP-TS arrangements (Cepeda Dean et al., unpublished data). The comprehensive evolutionary, structural, and functional characterization of the *T. cruzi* MASP repertoire across strains, including MASP chimeras, is currently underway.

Discussion

The generation and upholding of MASP genetic variability seems to be under strong selective pressure. Though not experimentally proven, the display of variable MASP molecules on the *T. cruzi* surface coat may contribute in the undermining of the mammalian immune system and/or in the exploration of a broad range of replication niches [16, 22, 32, 62]. Indeed, functional studies carried out on selected MASP molecules have shown that they are involved in the engagement with host cell receptors [31, 32]. In this framework, a deeper understanding of the MASP repertoire, and of its diversity across parasite strains, is expected to provide valuable insights into *T. cruzi* biology and pathogenesis.

As a first step towards this goal, we herein undertook the assembly of a MASP repository based on the annotated proteomes of strains belonging to extant parasite lineages (TCC and Brazil A4). The intrinsic diversity of the MASP family and the lack of functional information led us to adopt structural criteria to guide sequence curation: (i) positive prediction of functional surface localization or secretion signals and (ii) length, aa composition, and overall homology with currently known MASPs. Within the set of collected MASPs we identified several molecules exhibiting structural divergence, i.e. truncations, insertions or frameshifts. The levels of representation of each type of event were found to be strain-specific and appear to be the result of annotation issues rather than genomic evolution. For instance, a clear bias in the frequency of artifactual sequences displaying atypical C-terminal sequences was identified in the proteome of Brazil A4, in which annotation was based on DNA-level identity searches [11]. On the other hand, an enrichment in MASP proteins lacking a functional SP was revealed in the TCC proteome. These sequences were most likely translated from a premature START codon, as they displayed a peptide of variable length upstream of the conserved sequence block at the N-terminal end of most MASPs. In addition, and as previously reported [6, 14, 15], both datasets contained putative chimeric MASPs.

These molecules showed TcMUC or TS fingerprints at their N- and C-terminal regions, respectively. No putative chimeric MASP with SAP or any other *T. cruzi* gene family was observed.

As a result of this comprehensive curation protocol, which incorporated bioinformatics tools and manual corrections, we obtained a high-quality MASP protein database that accurately represents the global diversity of this family. This repository contains only 77.18% of the annotated sequences in Brazil A4 and 84.82% of those in TCC, further stressing the impact of errors introduced during CL Brener genome annotation and the necessity for tools that facilitate the accurate identification and classification of MASPs.

Sequences from MASP conserved terminal blocks were used to generate HMMER probability matrices. This method detects homology by comparing a sequence with an ad hoc constructed Hidden Markov model, thus allowing a certain degree of flexibility and enabling positive recognition of 'novel' MASP signatures, not strictly represented in the database [49]. In addition, HMMER provides two independent scoring values (*E* value and bit score) that can be fine-tuned to suit the specific goals of the study. The relevance of using both parameters becomes evident when analyzing complex datasets such as parasite genomes, in which up to 5×10^5 holo-ORFs (representing $> 1 \times 10^6$ actual sequences) need to be evaluated. As shown in Fig. 4, certain terminal sequences coming from non-MASP molecules and exhibiting a bit score that met the established cutoff criteria were nonetheless assessed as false positives due to their lower-than-cutoff *E* values.

A major advantage of our algorithm is that every input sequence is scanned at both terminal ends for the presence of MASP diagnostic motifs. This strategy enables the robust identification of canonical MASPs, but it also allows for the classification of novel categories, such as MASP-related molecules and MASP-chimeras. In addition to contributing to a better understanding of MASP diversity, the distinction of these novel classifications (particularly MASP-chimeras) will pave the way for functional studies on these potentially relevant molecules. Indeed, our study shows that *bona fide* MASP-chimeras with similar structural features can be found in every analyzed strain, suggesting that they emerged early during MASP evolution and were conserved across the parasite lineages. Moreover, our molecular characterizations strongly suggest that MASP-chimeras are genuine chimeric genes that emerged by recombination-mediated event(s) between members of distinct multi-gene families rather than by accumulation of mutations leading to TS/TcMUC sequence convergence on their terminal ends. Preliminary analysis of the 5' and 3' UTR regions of MASP-chimeras further support this hypothesis (Cepeda Dean et al, unpublished data).

As for the MASP-related molecules, and despite the variety of sequences that can be found within this classification, it is worth noting that they mostly correspond to pseudogenes, a major signature of multigene families in trypanosomatids [33]. Interestingly, we found a close similarity between MASP signatures present in MASP and MASP-related molecules (Fig. 4), thereby suggesting either a recent origin for MASP-related molecules or, more likely, that their secretory signals are under strong selection pressure against diversification. These findings support the idea that MASP-related sequences are not merely by-products of MASP evolution but rather an additional reservoir of variability for the generation of novel and functional variants.

In order to assess the overall functioning and output of our algorithm, we applied our pipeline to the Dm28c strain genome and compared the output with the previously annotated proteome [8]. Both methods exhibited a high overall concordance. However, our results underscored the relevance of incorporating new classifications into our algorithm, which enabled us to reclassify misanotated sequences, e.g. MASP-chimeras and MASP-related molecules within the pool of previously annotated MASPs and, more crucially, to accurately identify 77 previously unrecognized canonical MASP.

The robustness of our algorithm allowed for the exploration and classification of the MASP repertoire in different strains, representative of *T. cruzi* genetic diversity. As previously reported, hybrid DTUs displayed a higher overall dosage of MASP (i.e. canonical, chimeric and related) compared to ancestral strains TcI, TcII, TcIII and TcIV [8, 9, 11, 37, 38]. Nevertheless, when normalized by genome size, hybrid DTUs (except Tula cl4) had lower MASP densities compared to ancestral lineages. On this basis, the observed differences in genome size (which is larger in hybrid strains) are likely associated with the sequencing technologies used and, mostly with the level of resolution between the haplotypes, which is higher in hybrid strains. In line with this, the nearly complete and phase-assembled genome of Dm25 (TcI) presented a genome size well within the range of hybrid strains (Table 2).

TcI and TcII exhibited higher MASP/MASP-related ratios and a greater abundance of MASP-chimeras compared to other DTUs, suggesting positive selection to maintain functional MASP genes within these groups. In contrast, TcIII and TcIV strains exhibited the lowest MASP/MASP-related ratios and a very modest content of MASP-chimeras. A notable compartmentalisation was observed in the TcIV strain, which displayed the highest MASP-related density and, consequently, the lowest MASP/MASP-related ratio. Additional isolates from TcIV are needed to draw definite conclusions. With the exception of Tula cl4, hybrid strains (bearing TcII- and

TcIII-like haplotypes) displayed intermediate densities of MASP-chimeras and intermediate MASP/MASP-related ratios, suggesting that their genomes compiled the MASP allele repertoire from both parental strains.

Although several attempts to quantify and compare MASP sequences among strains/DTUs have been published [8, 9, 11, 35, 37], a recent read-based methodology aimed at estimating the variability and copy numbers of the multigene-repetitive families MASP, TS and TcMUC, represents, to our knowledge, the most robust effort to date [38]. However, it must be noted that several differences exist between this methodology and our pipeline. The read-based method relies exclusively on Illumina reads, whereas our algorithm accepts protein sequences as input and is easily adaptable for long-read NGS genomic sequences by a pre-processing step. The most remarkable difference lies in the outputs: while the read-based approach brings an estimation of the quantity and provides a global idea of the variability of the MASP repertoire in a strain (understood as the number of clusters in which the ~30 nt k-mers can be grouped), our algorithm delivers a detailed count of the classified molecules in distinct categories: 'MASP', 'MASP-chimera', and 'MASP-related', along with a list of the hits belonging to each category. Consequently, the read-based tool can only be interpreted comparatively against other strains, whereas our output is easily interpretable and facilitates the development of further functional studies on the identified molecules.

While significant progress has been made in characterizing the MASP family, our study underscores the ongoing nature of this research and the necessity for continued refinement of bioinformatics tools for multigene family annotation. Overall, we presented an automated algorithm developed for the identification, classification, and annotation of MASPs directly from *T. cruzi* genomic datasets that can be adaptable for the analysis of other multigene families. This tool facilitated the discovery of several novel MASP and MASP-chimeric molecules that had been previously overlooked due to limitations in the different methodologies used for genome annotation. Moreover, this gene annotation-independent strategy lays a solid foundation for comprehensive comparative and evolutionary genomics research on this important pathogen, paving the way towards a better understanding of *T. cruzi* variability.

Conclusions

This study addresses the challenges of capturing MASP diversity in *T. cruzi*. By developing a new bioinformatic pipeline and methodology, we created a flexible algorithm that allowed us to identify and classify the full repertoire of MASP sequences, including MASP genes, ´MASP-related' molecules, and, for the first time ´MASP-chimeras´, which have not been annotated in *T. cruzi* genomes until now. This achievement significantly enriched the dataset available for future comparative genomic studies. The high concordance rate achieved when validating our algorithm against a rigorously annotated dataset demonstrates its accuracy and potential applicability. Implementation of this pipeline to a set of *T. cruzi* genomes allowed us to analyze the repertoire of MASP across strains and to find previously unannotated sequences. Although our findings may be influenced by the precision of genome sequencing and assembly, they indicate that the diversity of the MASP repertoire (including genes, pseudogenes, and chimeras) is specific to each strain. Our work lays the groundwork for post-genomic studies in *T. cruzi*, offering valuable insights into the evolution and genetic landscape of the MASP family.

Abbreviations

DTU	Discrete typing unit
MASP	Mucin-associated surface protein from T. cruzi
NGS	Next generation sequencing
HMMER	Hidden Markov models
TcMUC	*T. cruzi* mucins expressed on the surface of mammal-dwelling stages
TS	*Trans*-sialidases

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11384-5.

Additional Table 1: MASP molecules used to generate the database

Additional Table 2: 'Novel' MASPs found in ***T. cruzi*** Dm28c

Additional File 1: TcMUC molecules used to generate the database

Additional File 2: TS molecules used to generate the database

Additional Figure 1: MASP database curation. (**A**) Editing of MASP sequences annotated with a premature translation initiation site. Above, Alignment of sequences corresponding to the conserved N-terminal region of MASPs before (left) and after (right) reassignment of the translation initiation codon. Below, Outputs of SignalP analysis for the pre- and post-edition version of the sequence C3747_8g499-t42_1-p1 are shown in pink and light blue, respectively. (**B**) Left, MASPs showing divergent N- or C-terminal sequences are boxed on the upper and bottom panels, respectively. Right, Divergent sequences were compared by BLASTp to the ***T. cruzi*** protein database and the ***E*** value of the best two hits is shown. (**C**) Left, Length analysis of MASP sequences from Brazil A4 and TCC before (pink) and after (light blue) manual curation. Right, Abundance (in percentage) per residue of MASP sequences curated from TCC (pink), Brazil A4 (light blue) and YC6 (purple). In both panels, each point represents a single sequence; and the mean ± SD for each population is indicated. **Additional Figure 2**: Generation of MASP molecular signatures. Above, Multiple alignments of MASP sequences from our curated database (***n*** = 1,249). For each position, the degree of conservation is indicated with a color scale (from blue to white) and by consensus bars (color scale from yellow to brown) placed at the bottom of the alignment. Gaps in the alignment are shown in gray. Below, Consensus sequences corresponding to MASP terminal regions (23 aa from the N-terminus and 35 aa from the C-terminus) are shown as ***WebLogo*** graphics derived from the alignment. For N-terminal sequences, the predicted initial Met is denoted as residue 1 whereas for C-terminal sequences aa positions are indicated with negative numbers and the last residue before the STOP codon is denoted as 0. **Additional Figure 3**: Generation of TcMUC molecular signatures. Consensus sequences corresponding to TcMUC terminal regions (30 aa from the N-terminus and 40 aa from the C-terminus) are shown as ***WebLogo*** graphics derived from the alignment. The number of sequences supporting each logo is indicated and conserved motifs chosen for RegEx are shown in yellow. The molecular signature mostly associated with chimerisation events is indicated in the red box. For N-terminal sequences, the predicted initial

Met is denoted as residue 1 whereas for C-terminal sequences aa positions are indicated with negative numbers and the last aa before the STOP codon is denoted as residue 0. **Additional Figure 4**: Generation of TS molecular signatures. Consensus sequences corresponding to TS terminal regions (30 aa from the N-terminus and 40 aa from the C-terminus) are shown as ***WebLogo*** graphics derived from the alignment. The number of sequences supporting each logo is indicated and conserved motifs chosen for RegEx are shown in yellow. The molecular signature mostly associated with chimerisation events is indicated in the red box. For N-terminal sequences, the predicted initial Met is denoted as residue 1 whereas for C-terminal sequences aa positions are indicated with negative numbers and the last aa before the STOP codon is denoted as residue 0

Acknowledgements
We are indebted to Florencia Diaz-Viraqué (Institut Pasteur de Montevideo, Uruguay) for providing the list of Dm28c MASP pseudogenes and Gonzalo Greif (Institut Pasteur de Montevideo, Uruguay) for sequencing and providing the complete genomes of the strains Merjo (TcIII), MT3663 (TcIII), Jose Julio (TcIV) and BolFc10A (TcV).

Author contributions
Conceptualization: VB, CAB. Design: VB, CAB. Data curation: AACD, VB. Formal analysis: AACD, VB, CAB. Methodology: AACD, VB, CAB. Funding acquisition and resources: CR, LB, CAB, VB. Writing - original draft: AACD, VB, CABWriting - review and editing: AACD, VB, CAB, LB, CR. Final approval of the version to be submitted: All the co-authors. All authors read and approved the final manuscipt.

Funding
AACD holds a CONICET fellowship, whereas CAB and VB are career investigators from the same institution. AACD also received a CIN fellowship and a UNU-Biolac fellowship. This investigation received financial support from the ANPCyT (PICT-2017-3908 and PICT-2021-I-A-00284 to CAB and PICT-2020-2396 to VB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability
All data generated and analysed during this study are included in this published article and its supplementary information files.AVAILABILITY AND REQUIREMENTS Project name: Disruptomics-MASP. Project home page: https://github.com/BuscagliaLab/Disruptomics-MASP. Operating system(s): Linux and MacOS.Programming language: Python. Other requirements: Python 3 or higher version. License: BSD 3-Clause License.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Consent for publication
Not applicable.

Availability and requirements
Project name: Disruptomics-MASP.
Project home page: https://github.com/BuscagliaLab/Disruptomics-MASP.
Operating system(s): Linux and MacOS.
Programming Language: Python.
Other requirements: Python 3 or higher version.
License: BSD 3-Clause License.

Author details
[1]Instituto de Investigaciones Biotecnológicas (IIBio), Universidad Nacional de San Martín (UNSAM), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Av. 25 de Mayo y Francia, Campus UNSAM, B1650HMP San Martín, Buenos Aires, Argentina
[2]Escuela de Bio y Nanotecnologías (EByN), UNSAM, Buenos Aires, Argentina
[3]Laboratorio de Interacciones Hospedero-Patógeno, Unidad de Biología Molecular, Institut Pasteur de Montevideo, Montevideo, Uruguay
[4]Laboratorio de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
[5]Unidad Académica de Bioquímica, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay

Received: 11 December 2024 / Accepted: 19 February 2025
Published online: 24 February 2025

References

1. Buscaglia CA, Kissinger JC, Agüero F. Neglected tropical diseases in the Post-Genomic era. Trends Genet TIG. 2015;31(10):539–55.
2. Bern C. Chagas' disease. N Engl J Med. 2015;373(5):456–66.
3. Magalhães LMD, Gollob KJ, Zingales B, Dutra WO. Pathogen diversity, immunity, and the fate of infections: lessons learned from trypanosoma Cruzi human-host interactions. Lancet Microbe. 2022;3(9):e711–22.
4. Messenger LA, Miles MA, 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. 2015;13(8):995–1029.
5. Marcili A, Lima L, Cavazzana M, Junqueira ACV, Veludo HH, Maia Da Silva F, 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. 2009;136(6):641–55.
6. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et al. The genome sequence of trypanosoma Cruzi, etiologic agent of Chagas disease. Science. 2005;309(5733):409–15.
7. Weatherly DB, Boehlke C, Tarleton RL. Chromosome level assembly of the hybrid trypanosoma Cruzi genome. BMC Genomics. 2009;10:255.
8. Berná L, Rodriguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G et al. Expanding an expanded genome: long-read sequencing of trypanosoma Cruzi. Microb Genomics. 2018;4(5).
9. Callejas-Hernández F, Rastrojo A, Poveda C, Gironès N, Fresno M. Genomic assemblies of newly sequenced trypanosoma Cruzi strains reveal new genomic expansion and greater complexity. Sci Rep. 2018;8(1):14631.
10. Díaz-Viraqué F, Pita S, Greif G, de Souza R, de CM, Iraola G, Robello C. Nanopore sequencing significantly improves genome assembly of the protozoan parasite trypanosoma Cruzi. Genome Biol Evol. 2019;11(7):1952–7.
11. Wang W, Peng D, Baptista RP, Li Y, Kissinger JC, Tarleton RL. Strain-specific genome evolution in trypanosoma Cruzi, the agent of Chagas disease. PLoS Pathog. 2021;17(1):e1009254.
12. Hoyos Sanchez MC, Ospina Zapata HS, Suarez BD, Ospina C, Barbosa HJ, Carranza Martinez JC, et al. A phased genome assembly of a Colombian trypanosoma Cruzi TcI strain and the evolution of gene families. Sci Rep. 2024;14(1):2054.
13. Talavera-López C, Messenger LA, Lewis MD, Yeo M, Reis-Cunha JL, Matos GM, et al. Repeat-Driven generation of antigenic diversity in a major human pathogen, trypanosoma Cruzi. Front Cell Infect Microbiol. 2021;11:614665.
14. De Pablos LM, Osuna A. Multigene families in trypanosoma Cruzi and their role in infectivity. Infect Immun. 2012;80(7):2258–64.
15. Bartholomeu DC, Cerqueira GC, Leão ACA, daRocha WD, Pais FS, Macedo C, et al. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen trypanosoma Cruzi. Nucleic Acids Res. 2009;37(10):3407–17.
16. Campetella O, Buscaglia CA, Mucci J, Leguizamón MS. Parasite-host glycan interactions during trypanosoma Cruzi infection: trans-Sialidase rides the show. Biochim Biophys Acta Mol Basis Dis. 2020;1866(5):165692.
17. Bernardo WP, Souza RT, Costa-Martins AG, Ferreira ER, Mortara RA, Teixeira MMG, et al. Genomic organization and generation of genetic variability in the RHS (Retrotransposon hot Spot) protein multigene family in trypanosoma Cruzi. Genes. 2020;11(9):1085.
18. Díaz-Viraqué F, Chiribao ML, Libisch MG, Robello C. Genome-wide chromatin interaction map for trypanosoma Cruzi. Nat Microbiol. 2023;8(11):2103–14.
19. Lima ARJ, de Araujo CB, Bispo S, Patané J, Silber AM, Elias MC, et al. Nucleosome landscape reflects phenotypic differences in trypanosoma Cruzi life forms. PLoS Pathog. 2021;17(1):e1009272.
20. Comprehensive Analysis of Nascent Transcriptome Reveals Diverse Transcriptional Profiles Across the Trypanosoma cruzi Genome Underlining the Regulatory. Role of Genome Organization, Chromatin Status, and Cis-Acting

Elements| bioRxiv. [cited 2024 Sep 24]. Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/2024.04.16.589700v1?s=03
21. De Pablos LM, Osuna A. Conserved regions as markers of different patterns of expression and distribution of the mucin-associated surface proteins of trypanosoma Cruzi. Infect Immun. 2012;80(1):169–74.
22. dos Santos SL, Freitas LM, Lobo FP, Rodrigues-Luiz GF, Mendes TA, de Oliveira O. The MASP family of trypanosoma Cruzi: changes in gene expression and antigenic profile during the acute phase of experimental infection. PLoS Negl Trop Dis. 2012;6(8):e1779.
23. Oliveira AER, Grazielle-Silva V, Ferreira LRP, Teixeira SMR. Close encounters between trypanosoma Cruzi and the host mammalian cell: lessons from genome-wide expression studies. Genomics. 2020;112(1):990–7.
24. Seco-Hidalgo V, De Pablos LM, Osuna A. Transcriptional and phenotypical heterogeneity of trypanosoma Cruzi cell populations. Open Biol. 2015;5(12):150190.
25. Sabalette KB, Sotelo-Silveira JR, Smircich P, De Gaudenzi JG. RNA-Seq reveals that overexpression of TcUBP1 switches the gene expression pattern toward that of the infective form of trypanosoma Cruzi. J Biol Chem. 2023;299(5):104623.
26. Cánepa GE, Mesías AC, Yu H, Chen X, Buscaglia CA. Structural features affecting trafficking, processing, and secretion of trypanosoma Cruzi mucins. J Biol Chem. 2012;287(31):26365–76.
27. Burle-Caldas G, de Dos Santos A, de Castro NSA, Mugge JT, Grazielle-Silva FLB, Oliveira V. Disruption of active Trans-Sialidase genes impairs egress from mammalian host cells and generates highly attenuated trypanosoma Cruzi parasites. mBio. 2022;13(1):e0347821.
28. Atwood JA, Minning T, Ludolf F, Nuccio A, Weatherly DB, Alvarez-Manilla G, et al. Glycoproteomics of trypanosoma Cruzi trypomastigotes using subcellular fractionation, lectin affinity, and stable isotope labeling. J Proteome Res. 2006;5(12):3376–84.
29. Alves MJM, Kawahara R, Viner R, Colli W, Mattos EC, Thaysen-Andersen M, et al. Comprehensive glycoprofiling of the epimastigote and trypomastigote stages of trypanosoma Cruzi. J Proteom. 2017;151:182–92.
30. Bayer-Santos E, Aguilar-Bonavides C, Rodrigues SP, Cordero EM, Marques AF, Varela-Ramirez A, et al. Proteomic analysis of trypanosoma Cruzi secretome: characterization of two populations of extracellular vesicles and soluble proteins. J Proteome Res. 2013;12(2):883–97.
31. De Pablos LM, González GG, Solano Parada J, Seco Hidalgo V, Díaz Lozano IM, Gómez Samblás MM, et al. Differential expression and characterization of a member of the mucin-associated surface protein family secreted by trypanosoma Cruzi. Infect Immun. 2011;79(10):3993–4001.
32. Espinoza B, Martínez I, Martínez-Velasco ML, Rodríguez-Sosa M, González-Canto A, Vázquez-Mendoza A, et al. Role of a 49 kda trypanosoma Cruzi Mucin-Associated surface protein (MASP49) during the infection process and identification of a mammalian cell surface receptor. Pathog Basel Switz. 2023;12(1):105.
33. JACKSON AP. Genome evolution in trypanosomatid parasites. Parasitology. 2015;142(Suppl 1):S40–56.
34. Souza RT, Santos MRM, Lima FM, El-Sayed NM, Myler PJ, Ruiz JC, et al. New trypanosoma Cruzi repeated element that shows site specificity for insertion. Eukaryot Cell. 2007;6(7):1228–38.
35. Stoco P, Wagner G, Talavera-López C, Gerber A, Zaha A, Thompson C, et al. Genome of the avirulent Human-Infective Trypanosome—Trypanosoma rangeli. PLoS Negl Trop Dis. 2014;8:e3176.
36. Bradwell KR, Koparde VN, Matveyev AV, Serrano MG, Alves JMP, Parikh H, et al. Genomic comparison of trypanosoma Conorhini and trypanosoma rangeli to trypanosoma Cruzi strains of high and low virulence. BMC Genomics. 2018;19(1):770.
37. Franzén O, Ochaya S, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, et al. Shotgun sequencing analysis of trypanosoma Cruzi I Sylvio X10/1 and comparison with T. Cruzi VI CL Brener. PLoS Negl Trop Dis. 2011;5(3):e984.
38. Reis-Cunha JL, Coqueiro-Dos-Santos A, Pimenta-Carvalho SA, Marques LP, Rodrigues-Luiz GF, Baptista RP, et al. Accessing the variability of multi-copy genes in complex genomes using unassembled Next-Generation sequencing reads: the case of trypanosoma Cruzi multigene families. mBio. 2022;13(6):e0231922.
39. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010;38(Database issue):D457–462.
40. Gremme G, Steinbiss S, Kurtz S, GenomeTools:. A comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(03):645–56.
41. Higgins DG, Thompson JD, Gibson TJ. [22] Using CLUSTAL for multiple sequence alignments. In: Methods in Enzymology. Academic Press; 1996 [cited 2022 Sep 30]. pp. 383–402. (Computer Methods for Macromolecular Sequence Analysis; vol. 266). Available from: https://www.sciencedirect.com/science/article/pii/S0076687996660248
42. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinforma Oxf Engl. 2009;25(9):1189–91.
43. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012;28(4):464–9.
44. Nielsen H. Predicting secretory proteins with signalp. Methods Mol Biol Clifton NJ. 2017;1611:59–73.
45. Gíslason MH, Nielsen H, Almagro Armenteros JJ, Johansen AR. Prediction of GPI-anchored proteins with pointer neural networks. Curr Res Biotechnol. 2021;3:6–13.
46. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20(4):1160–6.
47. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2012;12.
48. RDP4. Detection and analysis of recombination patterns in virus genomes| Virus Evolution| Oxford Academic. [cited 2024 Sep 24]. Available from: https://academic.oup.com/ve/article/1/1/vev003/2568683
49. Larralde M, Zeller G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. Bioinformatics. 2023;39(5):btad214.
50. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
51. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet TIG. 2000;16(6):276–7.
52. Callejas-Hernández F, Gironès N, Fresno M. Genome sequence of trypanosoma Cruzi strain Bug2148. Genome Announc. 2018;6(3):e01497–17.
53. Berná L, Greif G, Pita S, Faral-Tello P, Díaz-Viraqué F, Souza RDCMD, et al. Maxicircle architecture and evolutionary insights into trypanosoma Cruzi complex. PLoS Negl Trop Dis. 2021;15(8):e0009719.
54. Barnabé C, Brenière SF, Santillán-Guayasamín S, Douzery EJP, Waleckx E. Revisiting gene typing and phylogeny of *Trypanosoma Cruzi* reference strains: comparison of the relevance of mitochondrial DNA, single-copy nuclear DNA, and the intergenic region of mini-exon gene. Infect Genet Evol. 2023;115:105504.
55. Romer G, Bracco LA, Ricci AD, Balouz V, Berná L, Villar JC, et al. Deep serological profiling of the trypanosoma Cruzi TSSA antigen reveals different epitopes and modes of recognition by Chagas disease patients. PLoS Negl Trop Dis. 2023;17(8):e0011542.
56. Majeau A, Murphy L, Herrera C, Dumonteil E. Assessing trypanosoma Cruzi parasite diversity through comparative genomics: implications for disease epidemiology and diagnostics. Pathogens. 2021;10(2):212.
57. Campetella O, Sánchez D, Cazzulo JJ, Frasch AC. A superfamily of trypanosoma Cruzi surface antigens. Parasitol Today Pers Ed. 1992;8(11):378–81.
58. Freitas LM, Santos SL, dos, 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.
59. Buscaglia CA, Campo VA, Frasch ACC, Di Noia JM. Trypanosoma Cruzi surface mucins: host-dependent coat diversity. Nat Rev Microbiol. 2006;4(3):229–36.
60. Masip YE, Caeiro LD, Cosenza M, Postan M, Molina G, Taboga O, et al. Vaccination with parasite-specific TcTASV proteins combined with Recombinant baculovirus as a delivery platform protects against acute and chronic trypanosoma Cruzi infection. Front Cell Infect Microbiol. 2024;14:1297321.
61. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome trypanosoma brucei. Science. 2005;309(5733):416–22.
62. Durante IM, Spina PEL, Carmona SJ, Agüero F, Buscaglia CA. High-resolution profiling of linear B-cell epitopes from mucin-associated surface proteins (MASPs) of trypanosoma Cruzi during human infections. PLoS Negl Trop Dis. 2017;11(9):e0005986.

Publisher's note