Genomic organization and expression profile of the mucin-associated surface protein (*masp*) family of the human pathogen *Trypanosoma cruzi*

Daniella C. Bartholomeu[1,*], Gustavo C. Cerqueira[2,3,4], Ana Carolina A. Leão[1], Wanderson D. daRocha[5], Fabiano S. Pais[5], Camila Macedo[2,4], Appolinaire Djikeng[6], Santuza M. R. Teixeira[5] and Najib M. El-Sayed[2,3,4,*]

[1]Department of Parasitology, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil, [2]Department of Cell Biology and Molecular Genetics, [3]Center for Bioinformatics and Computational Biology, [4]Maryland Pathogen Research Institute, University of Maryland, College Park, MD 20742, USA, [5]Department of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil and [6]J. Craig Venter Institute, Rockville, MD 20850, USA

ABSTRACT

A novel large multigene family was recently identified in the human pathogen *Trypanosoma cruzi*, causative agent of Chagas disease, and corresponds to ~6% of the parasite diploid genome. The predicted gene products, mucin-associated surface proteins (MASPs), are characterized by highly conserved N- and C-terminal domains and a strikingly variable and repetitive central region. We report here an analysis of the genomic organization and expression profile of *masp* genes. *Masps* are not randomly distributed throughout the genome but instead are clustered with genes encoding mucin and other surface protein families. *Masp* transcripts vary in size, are preferentially expressed during the trypomastigote stage and contain highly conserved 5′ and 3′ untranslated regions. A sequence analysis of a trypomastigote cDNA library reveals the expression of multiple *masp* variants with a bias towards a particular *masp* subgroup. Immunofluorescence assays using antibodies generated against a MASP peptide reveals that the expression of particular MASPs at the cell membrane is limited to subsets of the parasite population. Western blots of phosphatidylinositol-specific phospholipase C (PI-PLC)-treated parasites suggest that MASP may be GPI-anchored and shed into the medium culture, thus contributing to the large repertoire of parasite polypeptides that are exposed to the host immune system.

INTRODUCTION

Trypanosoma cruzi is an important human pathogen and the etiological agent of Chagas disease. It is estimated that there are 15–18 million infected people primarily in Central and South America. *Trypanosoma cruzi* trypomastigotes are transmitted typically from a reduviid bug to the mammalian host through the vector feces during the insect bite, but also by ingestion of contaminated food, following blood transfusion or organ donation. Trypomastigotes can invade several types of host cells where they differentiate intracellularly into replicative amastigotes. Amastigotes develop into non-dividing trypomastigotes which are released extracellularly upon cell disruption and can initiate another round of host cell infection. They can also infect a reduviid vector during feeding, within which they differentiate into replicative epimastigotes. Acute Chagas disease follows initial infection with *T. cruzi* and is characterized by high blood parasitaemia and broad tissue parasitism. This phase is usually a mild self-limited systemic illness involving fever and malaise, and in most cases, is not specifically diagnosed. Chronic Chagas disease appears years later and may lead to cardiomyopathy, megaesophagus and/or megacolon. Currently, there are no vaccines available. The drugs used in the treatment are toxic and effective only during the acute phase of the disease (1).

*To whom correspondence should be addressed. Tel: +1 301 4052999; Fax: +1 301 838 0208; Email: elsayed@umd.edu
Correspondence may also be addressed to Daniella C. Bartholomeu. Tel: +5531 3409 2825; Fax: +5531 3409 2970; Email: daniella@icb.ufmg.br
Present address:
Wanderson D. daRocha, Department of Biochemistry and Molecular Biology, Federal University of Paraná, Brazil.

The first draft of the *T. cruzi* genome was published (2) along with the complete genome sequences of two related trypanosomatid human pathogens, *T. brucei* (3) and *Leishmania major* (4). A comparative analysis of gene content of the three parasites has allowed the identification of a conserved core of ∼6200 genes and numerous species-specific genes (5). On one hand, these analyses provided the foundation for the development of unexplored chemotherapeutic approaches against these parasites, such as drugs that could be designed against conserved core processes and potentially useful against all three organisms. On the other hand, the characterization of species-specific genes is allowing us to better understand the distinct nature of the disease they cause and may help develop more specific interventions for treatment and prevention.

The *T. cruzi* diploid genome size is ∼100 Mb with an estimated haploid gene number of 12 000 genes in the CL-Brener strain (2). Compared to *T. brucei* and *L. major*, a remarkable feature of the *T. cruzi* genome is the massive expansion of surface protein gene families, which include the previously characterized gp85/trans-sialidase (TS)-like superfamily, mucins and the metalloprotease gp63. A major finding of the *T. cruzi* genome project was the discovery of a ∼1400-member gene family encoding the novel mucin-associated surface protein (MASP). Despite its large size, no member of the *masp* family has been characterized to date. We report here on the genome organization and expression profile of the family. Gene families encoding MASP and other surface protein gene are clustered in *T. cruzi*-specific regions of the genome which fall outside of the regions of synteny (i.e. where gene order is conserved) between the three trypanosomatids. Within these large clusters, *masp* genes and pseudogenes are preferentially located downstream of mucin TcMUCII. MASP members contain N- and C-terminal conserved domains that encode a putative signal peptide and a GPI-anchor addition site. The central region is variable both in length and in sequence and contains a large repertoire of repetitive motifs. In contrast to the highly heterogeneous coding region, *masp* mRNAs have conserved 5′ and 3′ untranslated regions (UTRs). Western blots of phosphatidylinositol specific phospholipase C (PI-PLC)-treated parasites suggest that MASP is GPI-anchored and is preferentially expressed during the trypomastigote (bloodstream) stage. Interestingly, despite the large number of genes, an examination of the expression profile reveals that a subset of *masp* members is preferentially expressed in a parasite population. This is the first detailed analysis of the *masp* gene family of *T. cruzi*.

MATERIALS AND METHODS

In silico analysis of MASP sequences

The figures depicting the genome organization of MASP family were generated by in-house PERL (Practical Extraction and Report Language) scripts taking advantage of the Bio::Graphics module, part of Bioperl toolkit (http://www.bioperl.org). The frequency distribution of genes in the vicinity of *masp* loci was computed by PERL and AWK scripts from the data stored in our local *T. cruzi* database and the result exported to excel to generate the graphs. To identify MASP conserved regions, the coding and flanking regions of the 771 MASP genes annotated in the *T. cruzi* genome containing both the N- and C-terminal conserved domains were aligned using the ClustalW algorithm (6). A consensus sequence was generated and the percentage identity for each position across the entire sequence was calculated using an in-house PERL script. The cDNA sequences were analyzed using Phred-Phrap-Consed package (7,8). To identify the MASP expressed members, the cDNA sequences were searched against a local database of the *T. cruzi* coding sequences and contigs using the BLASTN algorithm (9). MASP repetitive motifs were identified using the algorithm SSRIT (Simple Sequence Repeats Identification Tool) (10). *N*- and *O*-glycosylation and phosphorylation predictions were performed using the algorithms NetNGlyc 1.0, NetOGlyc 3.1 (11) and NetPhos 3.1 (12), respectively. Predictions for signal peptide and GPI-anchor addition sites were performed using SignalP (13) and GPI-SOM (14) algorithms, respectively. Multidimensional scaling was performed using XGobi/XGvis (15). The dataset consisted of a pairwise distance matrix of the first 60 nt encoding the N-terminal and the last 60 nt encoding the C-terminal of MASP (771 genes), serine-alanine-proline rich protein (SAP) (35 genes), TcSMUGS (eight genes), TcSMUGL (seven genes), TcMUC I (53 genes), TcMUC II (550 genes) and TcMUC III (two genes). The pairwise distance matrix was computed using Kimura-2p (16) implementation of MEGA2 (17).

Parasites

Epimastigote forms of CL-Brener strain of *T. cruzi* were maintained in logarithmic growth phase at 28°C in supplemented liver digested-neutralized tryptose (LDNT) medium (18) containing 10% fetal bovine serum. Amastigote and trypomastigote forms were obtained from infected Vero cells grown at 37°C or 34°C, respectively and in 5% CO_2 in RPMI medium supplemented with 5% or 1% of fetal bovine serum, respectively. Only preparations of the amastigote and trypomastigote forms having >90% purity were used.

Generation of antisera

A MASP peptide was selected based on MEME-identified motifs (2) using the following criteria: (i) the most conserved motifs present in the predicted protein; (ii) predicted peptide immunogenicity; (iii) specificity of peptide to MASP. Using the above criteria, we have synthesized peptide 7 (Cys-QHQQHEH-(P/S)-AENGEESAKDK) (Pacific Immunology Corp., San Diego, CA). A cysteine residue was added to the N-terminal of the peptide to allow conjugation to the keyhole limpet hemocyanin (KLH) carrier protein. To generate affinity purified anti-MASP antibodies, two rabbits were immunized with 100 µg of the degenerate peptide. MASP antibodies present in the antiserum were purified by affinity

chromatography using CNBr-activated sepharose coupled to purified MASP peptide.

Southern and northern blot analyses

Total RNA was isolated from parasite cultures using the RNeasy Kit (Qiagen). For northern blot analysis, $10 \mu g$ of total RNA were separated in 1.2% agarose/MOPS/ formaldehyde gel. The RNA was blotted onto a Hybond N^+ membrane (Amersham-Pharmacia Biotech) and fixed through UV irradiation using UV linker 2400 (Stratagene). The MASP probes were labeled with $[\alpha\text{-}^{32}P]dCTP$ using the Megaprime DNA-labeling protocol from Amersham-Pharmacia Biotech. The membranes were hybridized in a 50% formamide buffer for 18 h at $42°C$ as previously described (19), washed twice with $2\times$ SSC/0.1% SDS at $65°C$ for 30 min each, and exposed to X-ray films (Kodak). The *T. cruzi* chromosomes blots were kindly provided by Dr Andersson (Karolinska Institute) using the pulse-field gel electrophoresis conditions previously described (20). The hybridization of the chromosome blots was performed as previously described (21).

Construction and screening of a trypomastigote cDNA library

A unidirectional trypomastigote expression library of the *T. cruzi* CL Brener strain was constructed in the lambda-ZAPII vector using procedures provided by the supplier of the kits for cDNA synthesis and *in vitro* phage packing (Stratagene). It was generated from poly-A^+ RNA isolated from tissue culture purified trypomastigotes obtained from Vero host cell. The primary library, which has the titer of 1.6×10^6 pfu/ml, was amplified to a titer of 10^9/ml. A total of ~15 000 clones were screened with a ^{32}P-labeled fragment derived from the MASP 3′ UTR conserved region as previously described (19). Plaques yielding positive signals in the first screen were purified by successive rounds of screening until isolated positive clones were obtained.

cDNA characterization

Recombinant phagemids were excised by co-infecting *Escherichia coli* XL-1 Blue cells with the selected λZAPII phage and R408 helper phage as described by the commercial supplier of the phage (Stratagene). Plasmid DNAs were prepared using standard protocols (22) and the cDNA inserts sequenced. The nucleotide sequence for the cDNAs clones has been deposited in the GenBank database under GenBank accession number EU825796–EU825847.

Protein extracts preparation and western blot

Total parasite proteins were extracted from epimastigotes, amastigotes and trypomastigotes by directly suspending 10^8 cells in 100 ml of Laemmli's sample buffer. Volumes corresponding to $10 \mu g$ of total parasite protein extracts were loaded onto a 12.5% SDS–PAGE gel. For western blot analyses, total protein extracts were transferred to polivinylidene difluoride (PVDF) membranes after electrophoresis (Bio Rad trans-blot transfer medium), stained with Ponceau S (Sigma) and blocked by incubation with 5.0% non-fat dry milk. MASP proteins were detected on western blots using affinity purified anti-MASP antibody diluted at 1:300 in PBS 0.1% Tween-20 and 1.0% non-fat dry milk. The blots were revealed with peroxidase conjugated anti-rabbit IgG at a dilution of 1:5000 and ECL Plus (GE HealthCare).

Immunofluorescence assay

The cells were fixed in 2% paraformaldehyde/PBS and kept at $4°C$. To proceed with the IFA, the cells were washed twice in PBS/50 mM NH_4Cl, then once in PBS, spotted on a clean glass slide and allowed to dry overnight. The fixed cells were re-hydrated with PBS and blocked for 1 h with $30 \mu l$ 5% fetal bovine serum in PBS. Slides were incubated with the affinity-purified anti-MASP antibody or pre-immune serum for 30 min with a 1:250 dilution. The slides were washed five times with blocking solution and incubated with the FITC-conjugated anti-rabbit IgG antibody (Molecular Probes) at the 1:500 dilution for 30 min and then with DAPI at 1:1000 dilution for 1 min. After washing with the blocking solution, the slides were mounted and examined by microscopy.

RESULTS

Genomic context of the masp genes

We have previously reported the discovery of the MASP proteins encoded by the second largest multigene family in the *T. cruzi* genome (2). Of the 1377 *masp* genes identified, 814 are full-length and 563 are partial genes located on contig ends and/or pseudogenes. Of the 814 full-length genes, 771 encode proteins containing both N- and C-terminal MASP conserved domains and 43 are chimeric sequences containing the N- or the C-terminal MASP conserved domain. Genes for the MASP and mucin families are interspersed within large non-tandem arrays of genes which encode other surface proteins including mucin TcMUC, gp85/TS-like, GP-63, serine-alanine-proline rich protein (SAP) and dispersed gene family 1 (DGF-1) (Figure 1). Those same regions are enriched with *T. cruzi* retroelements, such as L1Tc, NARTc, DIRE, VIPER and SIRE and members of the retrotransposon hot spot (RHS) family (2); (23). Because a large number of those genes are located in close proximity of *T. cruzi* mucins (TcMUC), we have named the family of encoded proteins 'mucin-associated surface proteins' (MASPs). Here we have conducted a through inspection of the genes in the vicinity of all *masp* intact genes and pseudogenes. We observe a strong bias for the occurrence of TcMUCII upstream of *masp* intact genes or pseudogenes (Figure 2A and B). The frequency of downstream genes differs for *masp* intact and pseudogenes. Genes more frequently found downstream of intact *masp* copies correspond to those encoding hypothetical proteins (Figure 2A) whereas those more frequently located downstream of *masp* pseudogenes correspond to members of the gp85/TS-like superfamily (Figure 2B).
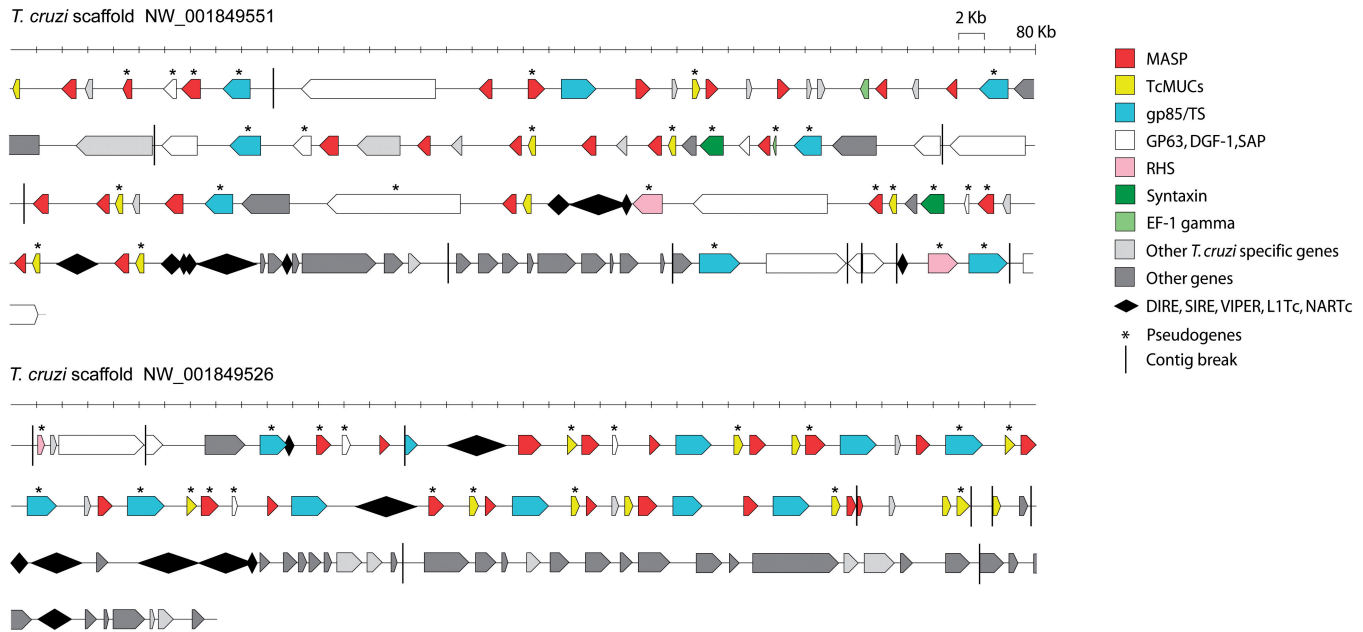
Figure 1. Genomic organization of the *masp* family. Two selected *T. cruzi* genomic scaffolds depicting typical regions containing *masp* genes and pseudogenes within arrays of surface protein-encoding genes. The colored arrows indicate the position and coding strand of *T. cruzi* genes and other genomic elements: TcMUCs, *T. cruzi* TcMUC mucins; TS, gp85/trans-sialidase-like; GP63, metalloprotease GP63; DGF-1, dispersed gene family 1; SAP, serine- alanine- and proline-rich protein; RHS, retrotransposon hot-spot protein; EF-1 gamma, translation elongation factor 1-γ. The category 'other *T. cruzi* specific genes' corresponds to *T. cruzi* genes not present in *T. brucei* or *L. major* genomes and which do not have a dedicated category in the legend. The category 'other genes' corresponds to *T. cruzi* genes also present in at least one of the other two Trypanosomatids. The scaffold identifier is the Genbank accession number.
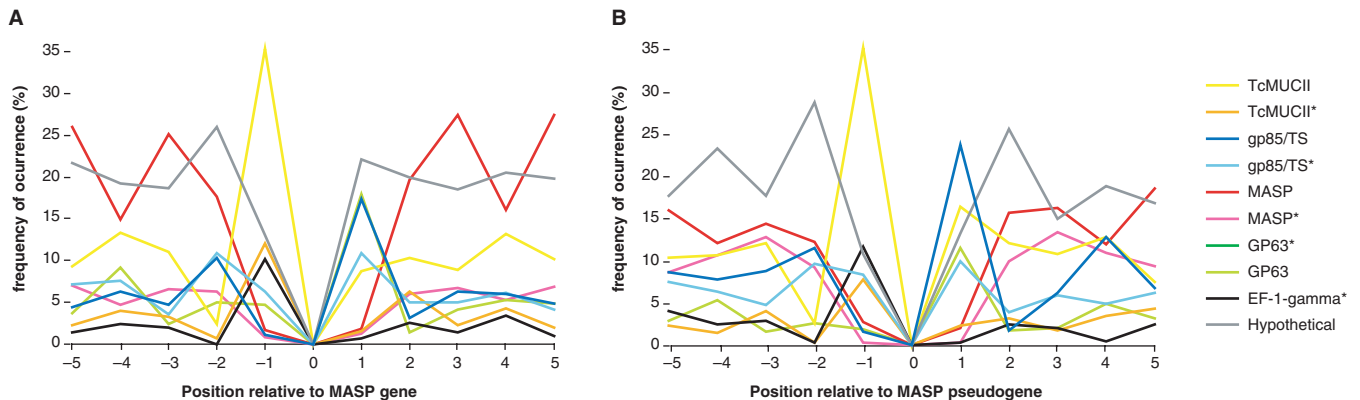

Figure 2. Genes in the vicinity of MASP. Frequency of occurrence of genes up to five positions upstream and downstream to *masp* intact genes (**A**) and pseudogenes (**B**). The values in the graph represent the frequency distribution of gene families independently computed at each ordinal position in reference to a *masp* gene. The analysis was restricted to five genes upstream and five genes downstream from *masp*. The relative positions between *masp* and other genes were inferred by loci coordinates along the genomic contig. Only gene families presenting a frequency higher than 10% are shown.

The *masp* genes are distributed over 210 scaffolds ranging in length from 5.3 to 799 kb. A 796-kb scaffold 1047053516725 (Genbank accession NW_001849575, Supplementary Figure S1) contains the largest number of *masp* genes, 79 in total. None of the intact *masp* genes were associated with the 47 subtelomeric regions that are represented in the assembled genome data (2). *T. cruzi* subtelomeric regions were defined here as sequences that extend from the telomeric hexamer repeats to the first two-way or three-way cluster of orthologous genes (COGs) across the *T. cruzi*, *T. brucei* and *L. major* genomes. In the *T. cruzi* genome, these telomeric regions are enriched with RHS genes (119 copies), gp85/trans-sia-lidases (84 copies), retroelements (81 copies that include SIRE, VIPER, DIRE, L1Tc and NARTc), dispersed gene family 1 (46 copies), and hypothetical proteins (42 copies). The only four *masp* genes found in those regions include three pseudogenes and a MASP/SAP chimeric sequence. Although the *T. cruzi* chromosome structure has not been completely elucidated and only about half of the *T. cruzi*
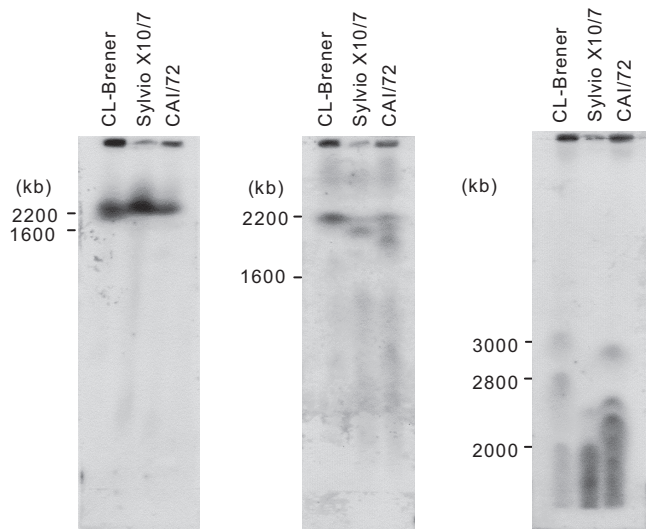
Figure 3. PFGE mapping of the *masp* genes onto *T. cruzi* karyotype. Three panels represent distinct pulsed-field gel conditions optimized to separate small (left); medium (center) and large chromosomes (right). The blots were hybridized with a probe derived from masp 3′ UTR conserved region.

telomeres are represented in the assembled genome, the available data suggests that the large clusters of surface protein-encoding genes reside internally in the chromosomes. More specifically, those clusters are located in the *T. cruzi*-specific regions of the genome which fall outside of the regions of synteny between the three trypanosomatids (Supplementary Figure S1).

To identify the chromosomal localization of *masp* genes, we performed Southern blot experiments on filters kindly provided by Dr Björn Andersson (Karolinska Institute). *T. cruzi* chromosomes were separated by pulsed-field gel electrophoresis under three distinct conditions, each optimized to separate small (1.1–0.3 Mb), medium (2.2–0.6 Mb) or large (3.5–2.0 Mb) chromosomes from three *T. cruzi* strains: CL-Brener, Sylvio X10/7 and CAI/72 (20). Two probes derived from the C-terminal domain and 3′ UTR region of *masp* were used since they represent the longest conserved regions among all members of this family (see below). Similar hybridization patterns were obtained with both probes and with the three *T. cruzi* strains. *Masp* genes appear to locate predominantly in large chromosomes (Figure 3).

MASP protein features

An alignment of the amino-acid sequences of 771 full-length MASP proteins revealed highly conserved N- and C-terminal domains that were predicted to encode a signal peptide and a GPI-anchor addition site, respectively (Figure 4A), suggesting a surface location in the parasite. The central region varies both in sequence (Figure 4A) and in length (ranging from 176 to 645 amino-acid residues) (Supplementary Figure S2) and often contains repetitive motifs.

To further investigate the repertoire of repetitive motifs in the family, all 944 MASP proteins encoded by intact genes were analyzed. The sequences corresponding to the signal peptide and the GPI-anchor addition site were removed, so that only the central region of each protein, which is likely to be exposed at the parasite surface, was analyzed. The following repetitions were searched: (i) single amino acids repeated at least five times; (ii) di-peptides repeated at least four times; (iii) tri-peptides repeated at least three times and (iv) tetra- to dodeca-peptides repeated at least twice. From the total of 944 sequences, 406 proteins were found to contain at least one repetitive motif. A total of 679 repetitions were identified in the dataset (Supplementary Table S1). Single residue repetitions are the most common, and those consisting of glutamic acid are the most frequent corresponding to ~27% of the total of the identified repetitive motifs.

A large number of *T. cruzi* surface proteins are heavily glycosylated. Analyses of full-length MASP proteins revealed at least four potential *O*-glycosylation sites per sequence, 70% of which correspond to threonines. Also, a total of 2000 predicted *N*-glycosylation sites were predicted in 710 out of 771 MASP intact proteins. MASP may also be subjected to phosphorylation, since 11–59 potential phosphorylation sites were predicted in the same MASP family subset analyzed.

A search of all MASP proteins against the Pfam database revealed only a few weak hits, 18 in total, to a Mucin-like glycoprotein Pfam domain (PF01456) found in mucin proteins (Supplementary Table S2). A close inspection of these 18 members revealed that in all but two cases the predicted proteins correspond to chimeras containing the N-terminal domain characteristic of TcMUC family and the MASP C-terminal conserved domain. Alignments of the PF01456 domain matches (data not shown) indicate this domain is degenerated in the MASP family. This is reflected by less significant *P*-values when compared to matches from the TcMUC family (Supplementary Table S2). It is also worth noting that the consensus N- and C-terminal regions of TcMUC and MASP gene families show some level of identity at the nucleotide level, with the N-terminal domains sharing 57% identity and the C-terminal domain 38% (data not shown). In addition, threonine residues located near the C-terminal of TcMUC proteins are also found in the consensus MASP sequence (Figure 4A). Despite these similarities, sequence comparison of the N- and C-terminal conserved regions from all MASP and TcMUC members against each other confirms the unique identity of each family. A multidimensional scaling plot (Figure 4B) which represents the visual schema of the distance matrix reveals two main clusters, one representing the MASP and the other one the TcMUC sequences. The amino-acid composition of these families is also quite distinct (Supplementary Table S3).

Analysis of conserved flanking regions and masp mRNA levels during the *T. cruzi* life cycle

An alignment of the 5′ and 3′ flanking regions from all 771 *masp* intact genes revealed that both regions are highly conserved (Figure 5A). Based on sequence similarity with two expressed sequence tags (ESTs) derived from an
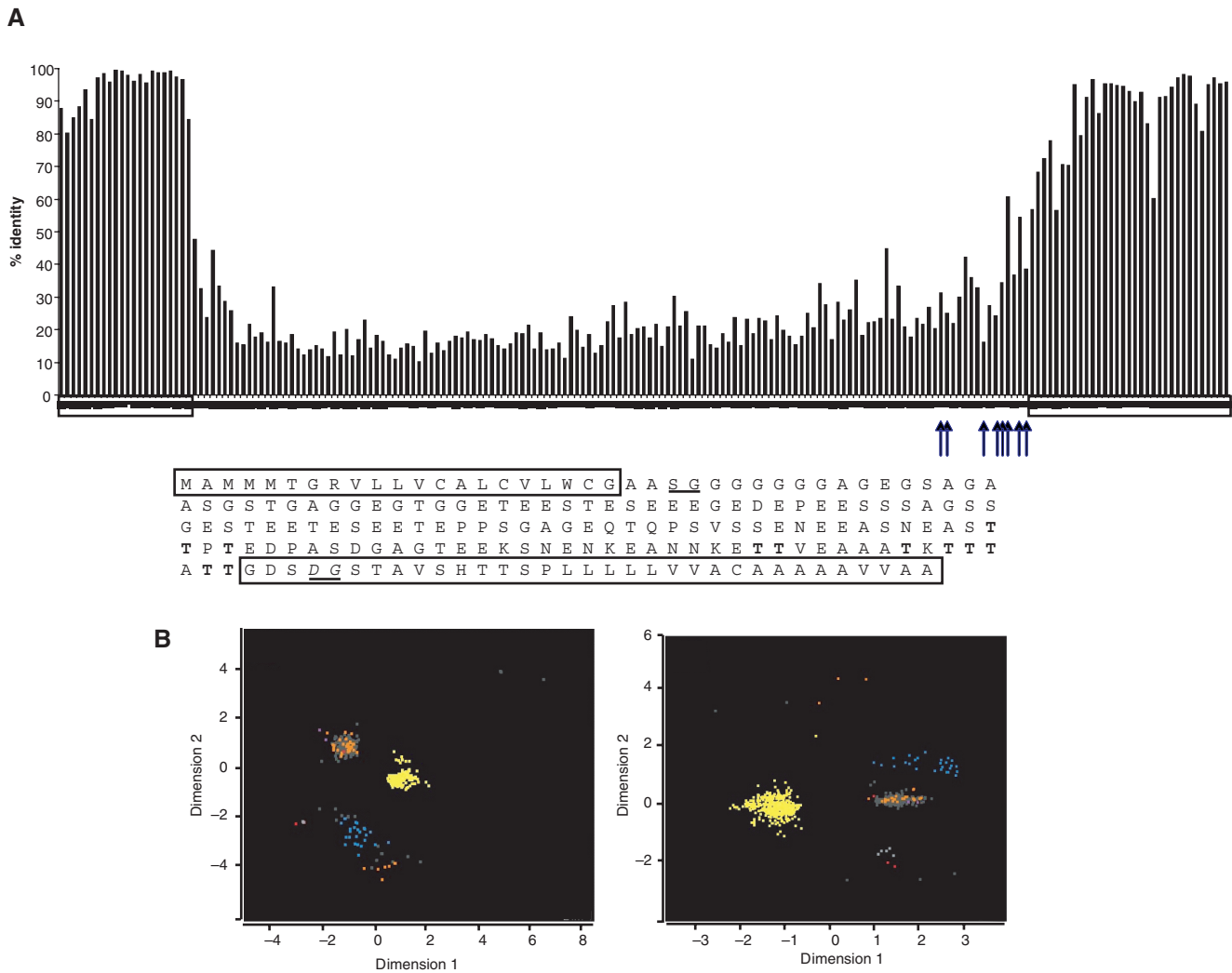
A

M A M M M T G R V L L V C A L C V L W C G A A SG G G G G G G A G E G S A G A
A S G S T G A G G E G T G G E T E E S T E S E E E G E D E P E E S S S A G S S
G E S T E E T E S E E T E P P S G A G E Q T Q P S V S S E N E E A S N E A S **T**
T P **T** E D P A S D G A G T E E K S N E N K E A N N K E **T T** V E A A A **T** K **T T T**
A **T T** G D S _D G_ S T A V S H T T S P L L L L L V V A C A A A A A V V A A

B

Figure 4. MASP protein features. (**A**) MASP protein sequence. All 771 MASP proteins containing both N- and C-terminal conserved regions were aligned using ClustalW (6). A consensus sequence was generated and a percentage identity score was computed for each position across the entire sequence. The conserved N- and C-terminal domains encoding the predicted signal peptide and predicted GPI anchor addition site are boxed. The amino acids surrounding the putative signal peptide addition site are underlined, while those surrounding the predicted GPI-addition site are italicized and underlined. The arrows indicate threonine residues close to the mature C-terminal region. (**B**) Multidimensional scaling plot representing the matrix of distance among the sequences derived from the N- (right panel) and C-terminal (left panel) conserved domains from MASP (yellow), TcMUCI (dark gray), TcMUCII (orange), TcMUCIII (pink), TcSMUGS (light gray), TcSMUGL (red) and SAP genes (turquoise).

amastigote cDNA library (accession number CB923887 and S.M.R. Teixeira, unpublished data) we were able to map the polyadenylation addition site (Figure 5A) and to verify that the 3′ flanking conserved region is part of the *masp* 3′ UTR. This conserved region has been previously described as the repetitive TcIRE element (24), which in fact corresponds to the reverse complement of the MASP C-terminal region and the 3′ UTR conserved sequences. Consistent with our *masp* chromosomal location results, hybridization of PFGE blots with the TcIRE probe showed that this sequence is located in high molecular weight chromosome bands (24).

Because the 3′ UTR is the largest and the most conserved region of *masp* transcripts, we selected this sequence to use as a probe in northern blots. As shown in Figure 5B, *masp* is preferentially expressed in trypomastigotes with low levels of expression also detected in

amastigotes and epimastigotes. A similar pattern of hybridization was obtained using a probe derived from the C-terminal conserved region (data not shown). This differential expression is consistent with the results of a *T. cruzi* proteome study in which nine MASP peptide fragments (out of 5792 total unique peptides) were detected exclusively in trypomastigotes (25). Interestingly, despite the fact that the length of *masp* coding regions is quite variable (Supplementary Figure S2), a predominant hybridization band of ~1.4 kb was detected.

MASP expression profile and mRNA structure

To investigate the repertoire of *masp* expression in trypomastigotes, we constructed and probed a trypomastigote cDNA library with the *masp* 3′ UTR conserved region. A total of ~15 000 phages were screened yielding a total
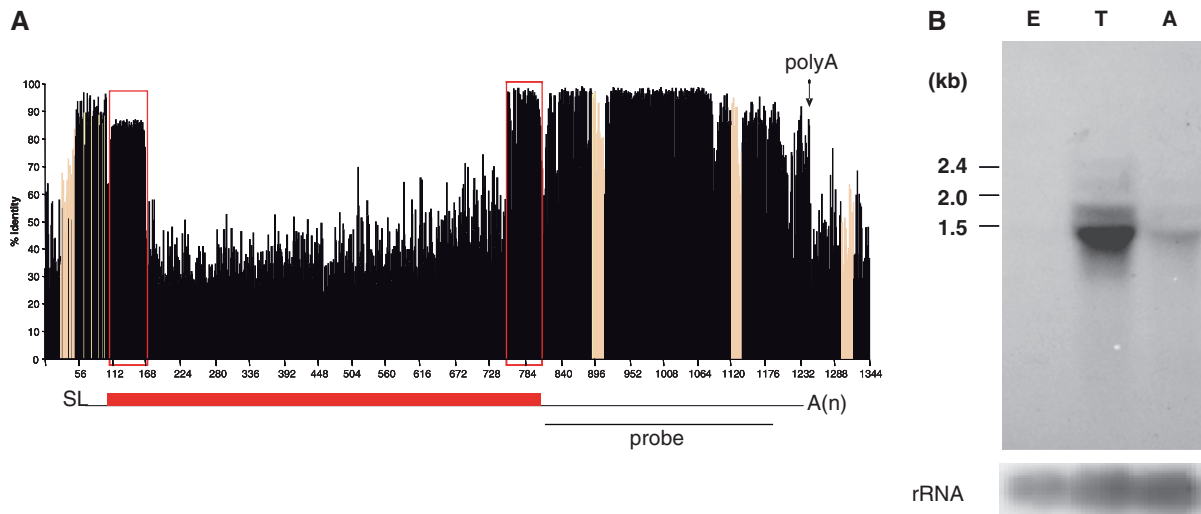
Figure 5. MASP conserved domains and mRNA expression during the *T. cruzi* life cycle. (**A**) The coding and flanking regions of the 771 MASP members containing both the N- and C-terminal conserved domains were aligned as in Figure 4. The mRNA structure is depicted below the graph with the coding region represented by a red rectangle and the UTRs by thin lines. The polyadenylation site is indicated with an arrow. (**B**) Northern blot containing total RNA (10 μg) isolated from epimastigote (E), trypomastigote (T) and amastigote (A) forms was probed with the first 330 nt of the 3′ UTR. The blot was stripped and re-probed with a fragment containing part of the *T. cruzi* 24Sα rRNA gene.

of ∼600 positive clones (∼4%). A total of 52 independent cDNA clones were sequenced from both ends and searched against the *T. cruzi* coding sequence database. The *T. cruzi* genes corresponding to the best hit of each cDNA clone are shown in the Supplementary Table S4. The corresponding gene could be identified for 36 of the cDNA clones. A total of 12 cDNAs could not be unambiguously associated to a specific *masp* gene, since their sequences covered only the C-terminal conserved domain and the downstream conserved 3′UTR. Four clone sequences were not related to *masp*.

The analysis of the cDNA sequences revealed that 24 clones correspond to full length *masp* genes or *masp* pseudogenes whose N- and C-terminal domains could be reconstituted. Another 13 clones could not be matched to a specific *masp* gene because the cDNA sequence did not cover the central variable region of the corresponding gene or because the 5′ region is partial due to contig ends. Taken together, this suggests that 46–71% of the 52 sequenced cDNA match full length *masp* genes or pseudogenes whose N- and C-terminal conserved regions could be reconstituted. Since ∼4% of the trypomastigote cDNA clones hybridized with the *masp* probe and assuming that 46–71% of the positives clones are in fact full-length *masp* sequences, we estimate that *masp* non-chimeric genes corresponds to around 1.8–2.8% of the trypomastigote transcripts.

We have previously subjected all MASP predicted proteins containing both N- and C-terminal conserved domains to the MEME algorithm to identify conserved motifs in the family (2). MASP members sharing the same combination of motifs were then clustered, yielding a total of 135 MASP subgroups (2) and Supplementary File 1). As shown in Supplementary Table S4, 7 out of 20 cDNAs (35%) that match full-length *masp* genes belong to a single MASP subgroup (S008). The remaining

13 cDNAs are derived from genes that belong to distinct subgroups. The cDNA sequences derived from S008 MASP subgroup are however not identical and are derived from distinct genes. In summary, our data indicates an expression bias towards MASP subgroup S008 in the trypomastigote population used to construct the cDNA library.

Another interesting observation was the significant number of cDNA clones containing chimeric MASPs. A total of 11 clones encoded 5′ segments of hypothetical proteins, TcMUCII or TS-like protein and 3′ segments of MASP C-terminal domain and 3′ UTR. One of the chimeric sequences (Supplementary Figure S3) corresponds to the TcMUCII-MASP gene Tc00.1047053508219.50 present in contig 7643 (Genbank acc. #AAHK 01000303). This chimera contains the first 126 nt of TcMUCII that encode a signal peptide and 20 amino acids of the putative mature protein (the best non-chimeric TcMUCII hit is Tc00.1047053508873.240). The remaining sequence corresponds to MASP coding sequence (the best non-chimeric MASP hit is Tc00.1047053504239.30) followed by the conserved MASP 3′ UTR. This chimeric cDNA is colinear with the genome sequence, ruling out an artifact of the genome assembly or cDNA library construction. In addition, two individual whole genome shotgun reads, TCHKU21TF and TCJTB48TR, match the cDNA perfectly (Supplementary Figure S3). Also, this expressed chimera sequence codes for potential signal peptide and GPI-anchor addition site and therefore contains *a priori* the required signals for a surface localization.

Analysis of cDNA sequences containing the spliced leader and/or poly-A tail allowed us to map the *masp* untranslated regions. As previously predicted from the alignment of all flanking sequences of the *masp* coding regions (Figure 5), both 5′ and 3′ UTRs are very well conserved among the cDNA clones. The 5′ UTRs are
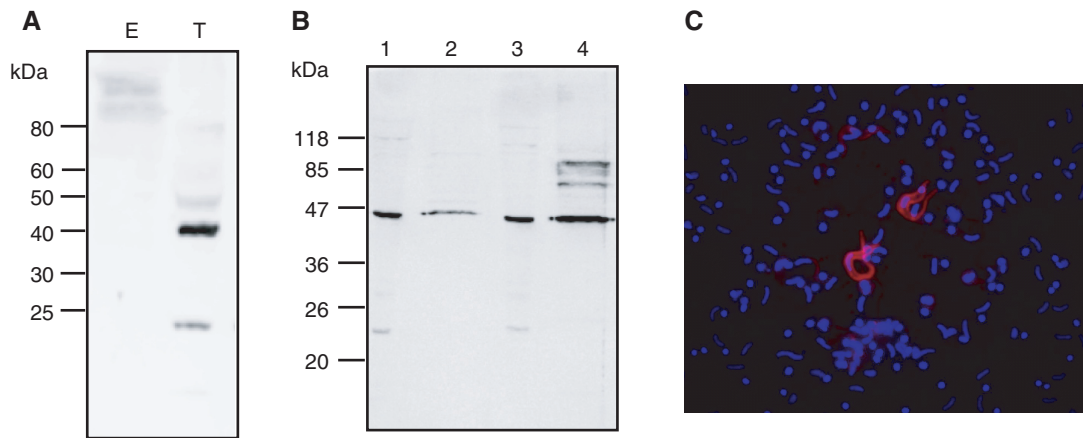
Figure 6. MASP protein expression and cellular localization. Western blot analysis of total protein extracts (10 μg) from the epimastigote (E) and trypomastigote (T) forms using anti-MASP peptide 7 rabbit antisera (A). (**B**) PI-PLC treatment of parasites. Trypomastigotes were treated or not with 1U PI-PLC for 40 min at 37°C and after centrifugation, the pellet and the filtrated supernatant were analyzed by western blot as in (A). 1, pellet of non-treated parasites; 2, supernatant of non-treated parasites; 3, pellet of treated parasites; 4, supernatant of treated parasites. (**C**) IFA assay of paraformaldehyde fixed trypomastigotes incubated with the affinity-purified anti-MASP antibody. Parasite DNA is stained with DAPI (blue).

short (50 nt on average) while the 3′ UTRs range from 361 to 498 nt in size.

MASP protein expression and cellular localization

To analyze the expression of MASP throughout the parasite life cycle, protein extracts from the epimastigote and trypomastigote stages were separated on SDS–PAGE followed by immunoblotting using anti-MASP antibodies. The antibodies were raised against a MASP peptide (peptide 7) which, according to the MEME analysis (2), is found in 109 members. As shown in Figure 6A, anti-MASP antibodies reacted with trypomastigote proteins of around 45 kDa. Light additional bands are likely due to a non-specific recognition as they do not appear consistently and reproducibly in independent experiments. Pre-immune sera do not react with the parasite protein extracts (data not shown).

To determine the cellular localization of MASP proteins, non-permeabilized trypomastigotes were analyzed by immunofluorescence using anti-MASP antibodies. As shown in Figure 6C the entire cell surface of trypomastigotes expressing MASPs containing the peptide 7 was labeled. Only a small proportion (∼5%) of the parasites, however, reacted with this antibody, indicating that the expression of peptide 7-containing MASP is limited to a subset of the population.

To confirm the *in silico* prediction that suggested MASP is GPI-anchored, trypomastigotes were treated with phosphatidylinositol-specific phospholipase C (PI-PLC) and the supernatant and the pellet fractions were analyzed by western blot. Proteins in the expected range (∼45 kDa) were detected by anti-MASP antibodies in the supernatant of the PI-PLC-treated trypomastigotes (Figure 6B, lane 4) suggesting that MASP is a GPI-anchored surface protein. The supernatant of non-treated parasites also reacted with the anti-MASP antibodies (Figure 6C, lane 2), suggesting that at least some of the MASP family members are shed into the culture by trypomastigotes.

DISCUSSION

The *T. cruzi* surface protein-encoding genes are often clustered into large haploid and heterogeneous arrays that can be as large as 600 kb (2). Here we show that the occurrence of genes in the vicinity of masps is not random, with a strong bias towards the occurrence of TcMUCII genes upstream of *masp* genes and a frequent occurrence of members of the gp85/TS-like superfamily downstream of *masp* intact genes as well as pseudogenes (Figure 2). Interestingly relatively few head-to-tail arrays of *masp* genes were observed in the genome (Figure 2). We speculate this genomic organization with *masp* interspersed with other surface protein gene families may play a contributing factor in maintaining diversity within the masp family by avoiding sequence homogenization. We also show that clusters of surface protein genes containing members of the *masp* gene family are preferentially associated with large *T. cruzi* chromosomes (Figure 3). The same pattern of chromosome distribution has been reported for TcMUC (26), the *T. cruzi* specific SAP family (27) and the satellite SAT 195 (28,29), which are all associated with the same arrays (2,27,30). Although not all subtelomeric regions could be assembled with confidence in the CL-Brener genome (2), the available data clearly demonstrates that the clusters of surface proteins containing *masp* genes and other large gene families such as TcMUC and SAP are internal in the chromosomes at regions of synteny breaks with *T. brucei* and *L. major*. This is in contrast to other protozoans such as *T. brucei* and *P. falciparum* where variable surface protein genes (vsg and var, respectively) are often subtelomeric in the megachromosomes (3,31). This is also the case for all chromosomes of *T. brucei*, including mini- (50–150 kb) and intermediate (200–900 kb) chromosomes, which

contain telomere-linked vsgs (3,32). The clustering of *T. cruzi* surface protein genes internally within the chromosomes may correlate with the absence of telomere-associated classical antigenic variation in this parasite. Instead, it has been demonstrated that different members of *T. cruzi* surface protein families such as mucin TcMUC and gp85/TS family are co-expressed (25,33). Our survey of expressed *masp* genes in the trypomastigote revealed that many are co-expressed in a parasite population. The expression profile is not random, however, since one MASP subgroup is over-represented (Supplementary Table S4). Our IFA results (Figure 6C) also reveal that MASP expression is heterogeneous in the trypomastigote culture analyzed since only ∼5% of the parasites express MASP members containing the peptide 7, which is present in 14% of all possible MASP proteins encoded by intact genes. Whether an individual trypomastigote expresses a single masp gene, co-expresses different members of the masp family or none at all remains to be investigated.

In addition to the large repertoire of polypeptides derived from the various polymorphic surface protein families, *T. cruzi* combines sequences from different gene families to increase diversity. *T. cruzi* mosaic genes formed by group II and III members of the gp85/TS superfamily (34) and containing the N- or C-terminal conserved domain of MASP combined with the N- or C-terminal domain of TcMUC, SAP or gp85/TS superfamily have been identified (2). As part of this study, we identified two cDNAs that contain masp sequences associated with segments of TcMUCII or TS-like genes (Supplementary Table S4 and Supplementary Figure S3), indicating that at least some of the *T. cruzi* predicted mosaic genes are expressed. It is possible that these mosaic sequences have originated by means of segmental gene conversion, favored by the clustering of surface protein genes. Recombination events within and between genes families can generate an unlimited repertoire of parasite protein exposed to the host. In fact, several lines of evidence suggest that those areas containing arrays of surface proteins are subject to intense rearrangements: (i) the average length of the directional gene clusters in these *T. cruzi* specific regions is much smaller than for syntenic regions with *T. brucei* and *L. major* genomes; (ii) the presence of large number of pseudogenes (∼42% of all predicted gp85/ts, masp, mucins, retrotransposon hot spot, dispersed gene family-1 and metalloprotease gp63 genes), (iii) the abundance of retroelements and (iv) syntenic break between the two CL-Brener haplotypes.

It has been proposed that sequence polymorphism of *T. cruzi* surface proteins might represent an evasion mechanism to drive the immune system into a series of spurious and inefficient activations of naïve CD8 + T cells, resulting in their delayed generation of protective immune responses in the initial phase of the infection (35). Also, Pitcovsky *et al.* (36) proposed that trans-sialidase family variability might contribute to simultaneous presence of B-cell-related epitopes during the infection resulting in a series of non-neutralizing antibody response.

The MASP protein resembles other *T. cruzi* surface proteins with N- and C-terminal conserved regions that encode putative signal peptide and GPI-anchor addition site, respectively (Figure 4A). We showed in this study that MASP may be GPI-anchored at the surface of the trypomastigote and shed (at least some of the members) into the culture medium (Figure 6B). We have also shown that MASP central region is highly variable in length and in sequence, with frequent repeated motifs (Figure 4, Supplementary Figure S2 and Supplementary Table S1). It has been recognized that the expression of repetitive proteins is a common phenomenon among parasites that persistently evade immune destruction in their hosts (37). In fact, *T. cruzi* infection induces a strong antibody response against many proteins containing tandem repeats (38–43). No experimental evidence, however, supports the concept that repetitive antigens are involved in a specific mechanism to subvert host immune responses. Other functions have also been attributed to *T. cruzi* tandem amino-acid repeats. In one study, repetitive motifs derived from *trans*-sialidase and SAPA (shed-acute phase antigen) were shown to delay trans-sialidase clearance in the blood (44).

As for other *T. cruzi* surface proteins, MASP may also be subjected to post-translational modifications since a large number of *N*- and *O*-glycosylation sites as well as phosphorylation sites were predicted in MASP proteins. Proteomic analyses have confirmed that at least some MASP members are subject to *N*-glycosylation (45). The glycosylation status of MASP proteins may, in part, explain why this large family has escaped detection for so long. Studies using library screening approaches with antisera from chagasic patients have typically relied on bacterial expression systems that lack the glycosylation machinery.

We have demonstrated that MASP conserved coding domains display limited, yet detectable, nucleic-acid sequence identity with the TcMUC gene family. In addition, a few MASP members contain a degenerate version of Pfam domain PF01456, characteristic of the mucin proteins, including those of *T. cruzi* (TcMUC) (Supplementary Table S2). No other Pfam domain was identified in the MASP proteins. MASP may have originated from an ancestral TcMUC member, followed by divergence, extensive gene duplications and sequence diversification to generate the actual MASP repertoire.

In addition to the conservation of sequences coding for the signal peptide and the GPI-anchor addition site, striking identity levels were observed in the 5′ and 3′ UTRs of the MASP transcripts (Figure 5A). Those UTRs may contribute to the differential abundance of MASPs throughout the *T. cruzi* life cycle. Northern blots reveal that MASP transcripts are preferentially expressed during the trypomastigote stage (Figure 5B). It is well established that trypanosomatids rely primarily on post-transcriptional mechanisms for control of gene expression. One of the most well known mechanisms responsible for developmental regulation of gene expression in trypanosomatids involves the presence of regulatory sequences within untranslated regions that modulate transcript stability and/or translation efficiency (19,46–50). Another possible explanation for the conservation of MASP untranslated regions may be related to the generation of sequence diversity within the family. An alternative model of recombination was recently proposed as a way to generate

diversity in the surface protein MSP2 involved in antigenic variation in *Anaplasma marginale*, a bacterial pathogen (51). According to this model, one recombination site occurs in the conserved flanking domains and the other within the hypervariable region. There is no apparent minimum region of sequence identity required for recombination in the hypervariable region; however, this is only possible due to the strong anchoring of the conserved ends holding the two recombining molecules in close proximity (51). It is noteworthy that MASP proteins display a mosaic structure with different members having a distinct combination of similar motifs (2).

The massive expansion of *masp* genes in the *T. cruzi* genome suggests this family may be critical for parasite survival. Its pattern of expression in the surface of the circulating trypomastigote forms, shedding properties, extensive sequence variability and repetitive nature suggests MASP may be involved in important parasite-host cell interactions. The ability of *T. cruzi* to infect and replicate within a variety of cell types is an essential feature of the parasite's survival strategy. Genetic polymorphism of *T. cruzi* surface glycoproteins has been hypothesized to be an important factor that may contribute to this phenomenon (52,53), although no clear association between a specific profile of parasite surface molecules and the ability of the parasite to infect a given host cell has been established. We speculate that the highly variable MASP central region may contribute to a large repertoire of peptides that may interact with different receptors from a variety of host cell types. Alternatively, the strong pressure for diversification may be imposed by the host immune system. In this regards, it will be interestingly to investigate whether MASP elicit the immune system especially during the acute phase of the infection when a large number of trypomastigotes are circulating.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Urbina,J.A. and Docampo,R. (2003) Specific chemotherapy of Chagas disease: controversies and advances. *Trends Parasitol.*, **19**, 495–501.
2. El-Sayed,N.M., Myler,P.J., Bartholomeu,D.C., Nilsson,D., Aggarwal,G., Tran,A.N., Ghedin,E., Worthey,E.A., Delcher,A.L., Blandin,G. *et al.* (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science*, **309**, 409–415.
3. Berriman,M., Ghedin,E., Hertz-Fowler,C., Blandin,G., Renauld,H., Bartholomeu,D.C., Lennard,N.J., Caler,E., Hamlin,N.E., Haas,B. *et al.* (2005) The genome of the African trypanosome Trypanosoma brucei. *Science*, **309**, 416–422.
4. Ivens,A.C., Peacock,C.S., Worthey,E.A., Murphy,L., Aggarwal,G., Berriman,M., Sisk,E., Rajandream,M.A., Adlem,E., Aert,R. *et al.* (2005) The genome of the kinetoplastid parasite, Leishmania major. *Science*, **309**, 436–442.
5. El-Sayed,N.M., Myler,P.J., Blandin,G., Berriman,M., Crabtree,J., Aggarwal,G., Caler,E., Renauld,H., Worthey,E.A., Hertz-Fowler,C. *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science*, **309**, 404–409.
6. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
7. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
8. Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Temnykh,S., DeClerck,G., Lukashova,A., Lipovich,L., Cartinhour,S. and McCouch,S. (2001) Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.*, **11**, 1441–1452.
11. Julenius,K., Molgaard,A., Gupta,R. and Brunak,S. (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, **15**, 153–164.
12. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
13. Bendtsen,J.D., Nielsen,H., von,H.G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
14. Fankhauser,N. and Maser,P. (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics*, **21**, 1846–1852.
15. Swayne,D.F., Cook,D. and Buja,A. (1998) XGobi: Interactive dynamic data visualization in the X window system. *J. Comput. Graph. Stat.*, **7**, 113–130.
16. Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.
17. Kumar,S., Tamura,K., Jakobsen,I.B. and Nei,M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.
18. Kirchhoff,L.V., Engel,J.C., Dvorak,J.A. and Sher,A. (1984) Strains and clones of Trypanosoma cruzi differ in their expression of a surface antigen identified by a monoclonal antibody. *Mol. Biochem. Parasitol.*, **11**, 81–89.
19. Bartholomeu,D.C., Silva,R.A., Galvao,L.M., El-Sayed,N.M., Donelson,J.E. and Teixeira,S.M. (2002) Trypanosoma cruzi: RNA

structure and post-transcriptional control of tubulin gene expression. *Exp. Parasitol.*, **102**, 123–133.

20. Branche,C., Ochaya,S., Aslund,L. and Andersson,B. (2006) Comparative karyotyping as a tool for genome structure analysis of Trypanosoma cruzi. *Mol. Biochem. Parasitol.*, **147**, 30–38.

21. Bartholomeu,D.C., Batista,J.A., Vainstein,M.H., Lima,B.D. and de Sa,M.C. (2001) Molecular cloning and characterization of a gene encoding the 29-kDa proteasome subunit from Trypanosoma cruzi. *Mol. Genet. Genomics*, **265**, 986–992.

22. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press. New York, USA.

23. Bringaud,F., Bartholomeu,D.C., Blandin,G., Delcher,A., Baltz,T., El-Sayed,N.M. and Ghedin,E. (2006) The Trypanosoma cruzi L1Tc and NARTc non-LTR retrotransposons show relative site specificity for insertion. *Mol. Biol. Evol.*, **23**, 411–420.

24. Aguero,F., Verdun,R.E., Frasch,A.C. and Sanchez,D.O. (2000) A random sequencing approach for the analysis of the Trypanosoma cruzi genome: general structure, large gene and repetitive DNA families, and gene discovery. *Genome Res.*, **10**, 1996–2005.

25. Atwood,J.A. III, Weatherly,D.B., Minning,T.A., Bundy,B., Cavola,C., Opperdoes,F.R., Orlando,R. and Tarleton,R.L. (2005) The Trypanosoma cruzi proteome. *Science*, **309**, 473–476.

26. Di Noia,J.M., Sanchez,D.O. and Frasch,A.C. (1995) The protozoan Trypanosoma cruzi has a family of genes resembling the mucin genes of mammalian cells. *J. Biol. Chem.*, **270**, 24146–24149.

27. Baida,R.C., Santos,M.R., Carmo,M.S., Yoshida,N., Ferreira,D., Ferreira,A.T., El Sayed,N.M., Andersson,B. and Da Silveira,J.F. (2006) Molecular characterization of serine-, alanine-, and proline-rich proteins of Trypanosoma cruzi and their possible role in host cell infection. *Infect. Immun.*, **74**, 1537–1546.

28. Elias,M.C., Vargas,N.S., Zingales,B. and Schenkman,S. (2003) Organization of satellite DNA in the genome of Trypanosoma cruzi. *Mol. Biochem. Parasitol.*, **129**, 1–9.

29. Vargas,N., Pedroso,A. and Zingales,B. (2004) Chromosomal polymorphism, gene synteny and genome size in T. cruzi I and T. cruzi II groups. *Mol. Biochem. Parasitol.*, **138**, 131–141.

30. Martins,C., Baptista,C.S., Ienne,S., Cerqueira,G.C., Bartholomeu,D.C. and Zingales,B. (2008) Genomic organization and transcription analysis of the 195-bp satellite DNA in Trypanosoma cruzi. *Mol. Biochem. Parasitol.*, **160**, 60–64.

31. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.

32. El Sayed,N.M., Hegde,P., Quackenbush,J., Melville,S.E. and Donelson,J.E. (2000) The African trypanosome genome. *Int. J. Parasitol.*, **30**, 329–345.

33. Buscaglia,C.A., Campo,V.A., Di Noia,J.M., Torrecilhas,A.C., De Marchi,C.R., Ferguson,M.A., Frasch,A.C. and Almeida,I.C. (2004) The surface coat of the mammal-dwelling infective trypomastigote stage of Trypanosoma cruzi is formed by highly diverse immunogenic mucins. *J. Biol. Chem.*, **279**, 15860–15869.

34. Allen,C.L. and Kelly,J.M. (2001) Trypanosoma cruzi: mucin pseudogenes organized in a tandem array. *Exp. Parasitol.*, **97**, 173–177.

35. Tarleton,R.L. (2007) Immune system recognition of Trypanosoma cruzi. *Curr. Opin. Immunol.*, **19**, 430–434.

36. Pitcovsky,T.A., Buscaglia,C.A., Mucci,J. and Campetella,O. (2002) A functional network of intramolecular cross-reacting epitopes delays the elicitation of neutralizing antibodies to Trypanosoma cruzi trans-sialidase. *J. Infect. Dis.*, **186**, 397–404.

37. Kemp,D.J., Coppel,R.L. and Anders,R.F. (1987) Repetitive proteins and genes of malaria. *Annu. Rev. Microbiol.*, **41**, 181–208.

38. Hoft,D.F., Kim,K.S., Otsu,K., Moser,D.R., Yost,W.J., Blumin,J.H., Donelson,J.E. and Kirchhoff,L.V. (1989) Trypanosoma cruzi expresses diverse repetitive protein antigens. *Infect. Immun.*, **57**, 1959–1967.

39. Alvarez,P., Leguizamon,M.S., Buscaglia,C.A., Pitcovsky,T.A. and Campetella,O. (2001) Multiple overlapping epitopes in the repetitive unit of the shed acute-phase antigen from Trypanosoma cruzi enhance its immunogenic properties. *Infect. Immun.*, **69**, 7946–7949.

40. Buscaglia,C.A., Campetella,O., Leguizamon,M.S. and Frasch,A.C. (1998) The repetitive domain of Trypanosoma cruzi trans-sialidase enhances the immune response against the catalytic domain. *J. Infect. Dis.*, **177**, 431–436.

41. DaRocha,W.D., Bartholomeu,D.C., Macedo,C.D., Horta,M.F., Cunha-Neto,E., Donelson,J.E. and Teixeira,S.M. (2002) Characterization of cDNA clones encoding ribonucleoprotein antigens expressed in Trypanosoma cruzi amastigotes. *Parasitol. Res.*, **88**, 292–300.

42. Duncan,L.R., Gay,L.S. and Donelson,J.E. (1991) African trypanosomes express an immunogenic protein with a repeating epitope of 24 amino acids. *Mol. Biochem. Parasitol.*, **48**, 11–16.

43. Pais,F.S., DaRocha,W.D., Almeida,R.M., Leclercq,S.Y., Penido,M., Fragoso,S.P., Bartholomeu,D.C., Gazzinelli,R.T. and Teixeira,S.M. (2008) Molecular characterization of ribonucleoproteic antigens containing repeated amino acid sequences from *Trypanosoma cruzi*. *Microbes Infect.*, **10**, 716–725.

44. Buscaglia,C.A., Alfonso,J., Campetella,O. and Frasch,A.C. (1999) Tandem amino acid repeats from Trypanosoma cruzi shed antigens increase the half-life of proteins in blood. *Blood*, **93**, 2025–2032.

45. Atwood,J.A. III, Minning,T., Ludolf,F., Nuccio,A., Weatherly,D.B., varez-Manilla,G., Tarleton,R. and Orlando,R. (2006) Glycoproteomics of Trypanosoma cruzi trypomastigotes using subcellular fractionation, lectin affinity, and stable isotope labeling. *J. Proteome. Res.*, **5**, 3376–3384.

46. Abuin,G., Colli,W., de,S.W. and Alves,M.J. (1989) A surface antigen of Trypanosoma cruzi involved in cell invasion (Tc-85) is heterogeneous in expression and molecular constitution. *Mol. Biochem. Parasitol.*, **35**, 229–237.

47. Di Noia,J.M., D'Orso,I., Sanchez,D.O. and Frasch,A.C. (2000) AU-rich elements in the 3′-untranslated region of a new mucin-type gene family of Trypanosoma cruzi confers mRNA instability and modulates translation efficiency. *J. Biol. Chem.*, **275**, 10218–10227.

48. Nozaki,T. and Cross,G.A. (1995) Effects of 3′ untranslated and intergenic regions on gene expression in Trypanosoma cruzi. *Mol. Biochem. Parasitol.*, **75**, 55–67.

49. Teixeira,S.M., Kirchhoff,L.V. and Donelson,J.E. (1995) Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of Trypanosoma cruzi. *J. Biol. Chem.*, **270**, 22586–22594.

50. Weston,D., La Flamme,A.C. and Van Voorhis,W.C. (1999) Expression of Trypanosoma cruzi surface antigen FL-160 is controlled by elements in the 3′ untranslated, the 3′ intergenic, and the coding regions. *Mol. Biochem. Parasitol.*, **102**, 53–66.

51. Futse,J.E., Brayton,K.A., Knowles,D.P. Jr and Palmer,G.H. (2005) Structural basis for segmental gene conversion in generation of Anaplasma marginale outer membrane protein variants. *Mol. Microbiol.*, **57**, 212–221.

52. Macedo,A.M., Machado,C.R., Oliveira,R.P. and Pena,S.D. (2004) Trypanosoma cruzi: genetic structure of populations and relevance of genetic variability to the pathogenesis of chagas disease. *Mem. Inst. Oswaldo Cruz.*, **99**, 1–12.

53. Burleigh,B.A. and Woolsey,A.M. (2002) Cell signalling and Trypanosoma cruzi invasion. *Cell Microbiol.*, **4**, 701–711.