

Journal Pre-proof

Comprehensive characterization of Cysteine-rich protein-coding genes of *Giardia lamblia* and their role during antigenic variation

Macarena Rodríguez-Walker, Cecilia R. Molina, Lucas A. Luján, Alicia Saura, Jon Jerlström-Hultqvist, Staffan G. Svärd, Elmer A. Fernández, Hugo D. Luján



PII: S0888-7543(22)00207-5

DOI: <https://doi.org/10.1016/j.ygeno.2022.110462>

Reference: YGENO 110462

To appear in: *Genomics*

Received date: 29 June 2022

Revised date: 15 August 2022

Accepted date: 17 August 2022

Please cite this article as: M. Rodríguez-Walker, C.R. Molina, L.A. Luján, et al., Comprehensive characterization of Cysteine-rich protein-coding genes of *Giardia lamblia* and their role during antigenic variation, *Genomics* (2022), <https://doi.org/10.1016/j.ygeno.2022.110462>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc.

Comprehensive characterization of Cysteine-rich protein-coding genes of *Giardia lamblia* and their role during antigenic variation

Macarena Rodríguez-Walker^{1,2†}, Cecilia R. Molina^{1,3†}, Lucas A. Luján^{1,3}, Alicia Saura^{1,3}, Jon Jerlström-Hultqvist⁴, Staffan G. Svärd⁴, Elmer A. Fernández^{1,2‡} and Hugo D. Luján^{1,3‡*}

¹Centro de Investigación y Desarrollo en Inmunología y Enfermedades Infecciosas (CIDIE), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)/Universidad Católica de Córdoba (UCC). Córdoba, Argentina.

²Facultad de Ingeniería, Universidad Católica de Córdoba (UCC). Córdoba, Argentina.

³Facultad de Ciencias de la Salud, Universidad Católica de Córdoba (UCC). Córdoba, Argentina.

⁴Department of Cell and Molecular Biology, BMC, Uppsala University, Uppsala, Sweden.

[†]These authors contributed equally to this work.

[‡]Co-senior authors.

*Correspondence and requests for materials should be addressed to H.D. Luján

(hlujan@ucc.edu.ar)

Keywords: Protozoa; variant surface antigens; antigenic switching; genome annotation; gene expression; immune evasion; parasite.

Abbreviations: Cys (cysteine), VSP (Variant-specific Surface Protein), HCMP (High Cysteine Membrane Protein), HCP (High Cysteine Protein), CRMP (Cysteine-rich Membrane Protein), SCRCP (Secretory Cysteine-rich Protein), TMD (Transmembrane Domain), CT (Cytoplasmic Tail), CDS (protein-coding genes), ORF (Open reading frame), PAS (poly(A) signal sites), APA (alternative polyadenylation), HCNCP (High Cysteine Non-variant Cyst protein), aa (amino acids), nt (nucleotides), ED (Ectodomain), EGF (Epidermal Growth Factor), INR (Initiator Element), DGE (Differential Gene Expression).

Highlights:

- Three different families of Cysteine-rich proteins are encoded in the *Giardia* genome.
- Variant-specific Surface Proteins undergo mutually exclusive changes in expression.
- Cysteine-rich Membrane proteins may protect the parasite during antigenic variation and host-parasite interactions.
- The VSP repertoire arose from retrotransposition, duplications and divergence.
- CRMP and SCRCP genes originated from VSP genes.

Abstract

Giardia lamblia encodes several families of cysteine-rich proteins, including the Variant-specific Surface Proteins (VSPs) involved in the process of antigenic variation. Their characteristics, definition and relationships are still controversial. An exhaustive analysis of the Cys-rich families including organization, features, evolution and levels of expression was performed, by combining pattern searches and predictions with massive sequencing techniques. Thus a new classification for Cys-rich proteins, genes and pseudogenes that better describes their involvement in *Giardia*'s biology is presented. Moreover, three novel characteristics exclusive to the VSP genes, comprising an Initiator element/Kozak-like sequence, an extended polyadenylation signal and a unique pattern of mutually exclusive transcript accumulation is presented as well as the finding that High Cysteine Membrane Proteins, upregulated under stress, may protect the parasite during VSP switching. These results allow better interpretation of previous reports providing the basis for further studies of the biology of this early-branching eukaryote.

1. Introduction

Giardia lamblia (syn., *G. intestinalis*, *G. duodenalis*) is a flagellated protozoan that inhabits the upper small intestine of many vertebrates. Infections are initiated by ingestion of cysts, followed by excystation and colonization of small intestine by the trophozoites. When trophozoites descend through the intestine, they differentiate into cysts, which are shed with the feces [1].

Giardia undergoes antigenic variation, a mechanism that allows the parasite to switch the

expression of its variant surface antigens, causing chronic and/or recurrent infections [2]. Trophozoites are covered with a single member of a family of Variant-specific Surface Proteins (VSPs), which are the main antigens recognized by the hosts [3]. Switching in expression of VSPs allows the parasite to evade the host immune response [1,4,5]. Both post-transcriptional and epigenetic mechanisms have been reported to control the expression of a unique VSP [6,7].

VSPs are integral membrane proteins having a signal peptide (SP) and a conserved C-terminal region comprising a single transmembrane domain (TMD) and a short cytoplasmic tail (CT) of only five amino acids (CRGKA) [8]. The extracellular portion of VSPs is of variable length and rich in cysteine (Cys), mainly as CXXC motifs. VSPs are resistant to proteolytic digestion and extreme pH and protect the parasite under the harsh conditions of the upper small intestine [9]. Furthermore VSPs have been reported to coordinate metals such as iron and zinc [10].

Other Cys-rich proteins have been described in *G. lamblia*: High Cysteine Membrane Proteins (HCMPs), High Cysteine Proteins (HCPs), High Cysteine Non-variant Cyst protein (HCNCp) and Tenascin-like Proteins (TLPs) [11–15]. HCMPs are Cys-rich proteins with a predicted TMD near the C-terminal end and a CT different from that of the VSPs [11]. In contrast, HCPs lack the TMD. TLPs were reported function as virulence factors [14], and the single HCNCp was described as a novel invariant HCMP specifically expressed during cyst formation [11]. However, there is no consensus for these proteins regarding their subcellular localizations, functions and relationship with VSPs. Moreover, it is unknown if all are actually transcribed and translated and if they undergo antigenic variation.

Currently there is no clear agreement regarding how many VSP genes exist in the genome

of *G. lamblia*. This characterization was limited because the first reference genome was of low quality and coverage [15]. Besides, the presence of several homologous gene families with highly repetitive sequence elements hindered the generation of an accurate genome assembly.

From the previous genome (GL2.0), several attempts have been made to characterize and quantify the VSP repertoire with dissimilar and inconclusive results accentuated by the lack of an established criterion to define a VSP. For example, in an earlier report between 235 and 275 VSP genes were estimated [16]. In a subsequent analysis, 228 complete VSPs and 75 partial VSPs were reported, comprising a repertoire of approximately 303 VSPs [16]. Besides, the presence of only 73 putative VSPs was also suggested [17].

Recently, a new version of the *Giardia* genome at 200x coverage and sequenced using PacBio long-reads and Illumina short-reads combined with structural mapping was published [18]. This new genome (GL2.1), assembled into five chromosomes, is better annotated both structurally and functionally. It is 12.6 Mpb in size and contains approximately 4,900 protein coding genes (CDS) and 320 pseudogenes. Since this new reference genome (RefSeq accession: GCF_000002435.2) differs substantially from the previous release, most of the transcriptomic and proteomic data produced in the past need to be re-evaluated.

The GL2.1 genome enables a more accurate analysis of the Cys-rich gene families, their organization, characteristic features, chromosomal localization and potential evolution. Here, different aspects of Cys-rich proteins of *G. lamblia* were exhaustively characterized, allowing us to propose a new classification and nomenclature for these genes and pseudogenes, thus contributing to a better understanding of the function of these protein families and their involvement during antigenic variation.

2. Results and Discussion

2.1. Cysteine-rich protein-coding genes and pseudogenes in the *G. lamblia* genome

Initially, the nucleotide (nt) and amino acid (aa) sequences of the 4,966 CDSs of the *G. lamblia* RefSeq genome were collected for analysis. A limit of $\geq 4\%$ of cysteine content and at least two of any of the characteristic CXC, CXXC and/or CXXXC motifs was established for a protein to be Cys-rich [3]. Since the major interest was in the VSPs and related families, a protein BLAST search for all CDS with similarity (E value <0.0001) to previously annotated VSPs (n=133), HCMPs (n=104), HCPs (n=11) or TLPs (n=11) was performed with the aim of finding any other genes belonging to these families but that might have been misannotated. This produced 273 CDS meeting these requirements.

Likewise, the nt sequences of the 320 pseudogenes from the RefSeq database were collected and a BLASTx search for all pseudogenes with similarity (E value <0.0001) to previously annotated Cys-rich protein genes was performed. This returned 217 pseudogenes with more than 4% of Cys content and at least two motifs in one of the three translated frames, yielding 490 sequences of putative Cys-rich genes and pseudogenes.

Then, a simple decision tree was used to classify these genes and pseudogenes ([Fig. 1](#)). First, a search for open reading frames (ORFs) from all the collected sequences was performed. If no ORF was found, the sequences were grouped as “**pseudogenes type I**” otherwise the presence of a predicted signal peptide (SP) was determined using Phobius [19], SignalP-5.0 [20] and TOPCONS [21]. At least two programs had to predict a SP to confirm the presence of this feature. It is worth mentioning that the search for an SP was

done not only in the annotated first methionine, but also in up- and downstream ones. Since the SP is key for trafficking and its deletion may result in loss of function, partial Cys-rich proteins without a proper SP were grouped as “**pseudogenes type II**”. If a SP was predicted, Phobius [19], TMHMM [22] and TOPCONS [21] were used to infer transmembrane regions. At least two programs had to find a positive TMD prediction to consider the presence of this feature, and the transmembrane region was extracted using the sequence ranges predicted by Phobius. If missing a TMD, the sequences were grouped as “**Secretory Cys-Rich Proteins (SCRPs)**”. Interestingly, all the Cys-rich proteins with TMDs were found to have a single TMD located at the C-terminal end of the proteins. Then, for those sequences with a TMD, the presence of the CRGKA cytoplasmic tail (CT) typical of the VSPs was determined. If found, the sequences were termed “**VSPs**”, otherwise, as “**Cys-Rich Membrane Proteins (CRMPs)**”.

Then, all type I and II pseudogenes were divided into pseudo VSP (pVSP), pseudo CRMP (pCRMP) or pseudo SCRPs (pSCRPs) depending on their similarity (highest score and the lowest E value) to each Cys-rich gene group using BLASTx against a local database constructed with the VSP, CRMP and SCRPs sequences.

Table 1 shows the number of sequences included in each group after applying this decision tree and the assigned names to which they were previously annotated. It was found that not only VSPs, HCMPs, HCPs and TLPs, but also other Cys-rich proteins were also included. These ORFs were formerly annotated as neurogenic locus notch proteins, neurogenic locus notch-like proteins, CXC-rich protein, EGF-like domain-containing proteins, uncharacterized and hypothetical proteins.

Of all identified Cys-rich proteins, 55% corresponded to VSPs, 38% to CRMPs and 7% to SCRPs, making up 4.9% of the total CDS and 4.6% of the entire genome. Moreover, Cys-

rich protein-coding genes plus pseudogenes contributes to almost 10% of the *Giardia* genome. A detailed information of all gene and protein IDs, annotated names, nucleotide and protein sequences, as well as the predicted TMD and SP for each VSP, CRMP and SCRPs are presented in [Table S1](#).

2.2. Transmembrane domain, cytoplasmic tail and signal peptide

A total of 136 VSPs were found where 133 were already named VSP, but the remaining 3 genes were misannotated as High Cysteine Membrane Protein (GL50803_0050332), pVSP (GL50803_0050390) and uncharacterized protein (GL50803_00114674). In the case of pVSP GL50803_0050390, the SP was found in the second Met in frame.

All 136 VSPs have a single, highly conserved TMD of ~25 aa located near the C-terminal end [19]. Of their 25 aa, 11 are conserved in all VSPs and are positioned at the beginning and at the end of the predicted TMD ([Fig. 2](#)). Besides, there is a conserved 13 aa stretch upstream the TMD that includes a C-xxx-G motif (consensus GGSTNKSSGLSTG; see below). On the other hand, the SP of the VSPs are not as conserved as the TMD, but still show a degree of similarity ([Fig. 2](#)).

Of the 94 CRMPs, 8 were previously annotated as HCMPs, 1 as High Cysteine non-variant Cyst Protein (GL50803_0040376), 1 as CXC-rich protein (GL50803_0014225), 1 as Tenascin-like protein (GL50803_0094510), and 3 as hypothetical proteins (GL50803_008733, GL50803_0015450 and GL50803_00113565). All of them have a predicted single C-terminal TMD. However, while VSPs have a highly conserved TMD, the TMDs of CRMPs are much more variable ([Fig. 2](#)). Examples of TMDs of CRMPs identical, similar and dissimilar to TMDs of VSPs are shown in [Table S2](#).

In contrast to the conserved CT of VSPs, those of CRMPs are highly variable, with their

length ranging from 1 to 60 aa. Notably, one CRMP (GL50803_00101589) with an identical TMD to that of the VSPs has an YRGKA CT. Due to its characteristics, GL50803_00101589 appears as a true VSP that has suffered a point mutation in its tail (see below). No other CRMP has a C-terminal tail of 5 aa with a mismatch of only one aa relative to the CRGKA. Only 3 additional CRMPs have a CT of 5 aa that differs from the CRGKA at two or three positions. Other CRMPs have completely different tails, in both length and sequence, although all comply with the positive-inside rule, which postulates the preferential occurrence of positively charged residues at the cytoplasmic edge of TMDs [23]. This is further evidence that the CT of both VSPs and CRMPs localize at the cytoplasmic side of cellular membranes.

Of the 104 previously annotated HCMPs, only 38 remained in the CRMP group. Former HCMPs are now distributed across the defined groups. Fourteen of them fell within the pseudogenes Type I, 1 within the SCRPs (GL50803_0029147), 1 is a VSP (GL50803_0050332), and 88 are CRMPs.

All Cys-rich proteins with a predicted SP but lacking TMD and CT were named SCRPs. Only 16 proteins fell into this group (Table S1). The SCRPs are composed mainly of the formerly named TLPs and HCPs. Of all previously annotated TLP and HCPs (11 and 8 in RefSeq genome), only 3 fell within this group. The remaining ones do not have a predicted SP and they were grouped together with pseudogenes type II. Moreover, although the 3 HCPs that remain in this group have a predicted SP, they might also be pseudogenes. This assumption is supported by the fact that for some of them it is possible to find a TMD in a different frame, for example in GL50803_00115202. Of all previously annotated TLPs, 2 fell within the pseudogenes type II group, 8 are SCRPs, and only 1 (GL50803_0094510) has a complete ORF with SP, TMD and CT (CRMP).

2.3. Length and predicted molecular weight

VSPs length ranges from 127 to 2293 aa, with a median of 645, equivalent to ~74 kDa. The predicted VSPs mass is between 15 and 260 kDa which is consistent with previous estimations indicating that the molecular mass of VSPs varies from 20 to 200 kDa [3]. CRMPs amino acid length ranges from 190 to 2539 aa, with a median of 824.5 equivalent to ~95 kDa. Lastly, the SCRPs amino acid length ranges from 123 to 1093, with a median of 538 which is equivalent to ~62 kDa (Fig. 2 and Table S1). It is worth to mention that pseudogene sequences were not taken into account in the calculation of protein lengths.

2.4. Predicted subcellular localization

DeepLoc-1.0 was used [24], which is a eukaryotic protein subcellular localization predictor that applies neural networks trained on Uniprot proteins with experimental evidence of subcellular localization. All 136 VSPs and the 94 CRMP were predicted by DeepLoc-1.0 [24] to localize at the plasma membrane with high confidence. SCRPs with SP but without TMD were predicted as extracellular. A representative example of the output of DeepLoc-1.0 for each group is shown in Fig. 2. TMD lengths are signatures for subcellular locations in eukaryotic cells [25,26]. Notably, the TMD of all VSPs, the single CXC-rich membrane protein (GL50803_000014225), and a few CRMPs possess a TMD containing extended GAS_{right} (G-xxx-G-xxx-G) and small(V/L/I)-xxx-small(V/L/I) motifs, which are known to facilitate TMD di- and oligomerization [27,28]. The conserved TMD of VSPs, the TMD of a VSP-like CRMP and a CRMP with a highly divergent TMD were analyzed to determine their oligomerization capability [29,30]. Only the Cys-rich membrane proteins having the consensus TMD sequence of VSPs were predicted to form oligomers (Fig. 3). Since TMD

length and sequence are key features of proteins prone to localize into liquid-ordered microdomains [31–33], all VSPs may be present in lipid raft-like structures.

2.5. Cysteine content and characteristic motifs

The percentage of Cys ranges from 7.06% to 15.46% among all VSPs, with a median of 11.89%, which is 5-fold higher than the median Cys content of all *Giardia* CDS. The percentage of Cys in CRMPs ranges from 4.93% to 15.63%, which is similar to the values found in the VSPs.

VSPs are known to have multiple CXXC motifs on their ectodomain (ED) [3,16]. In these proteins, the minimum number of CXXC motifs is 2 and the maximum 70, with more than half of the VSPs having between 20 and 30 CXXC, and the number of these motifs depended almost linearly of the length of the VSP ED (Fig. 4). Of the 136 VSPs, 36 have tandem repeats in their ED [34-36]. Interestingly, similar repeats are found in several VSPs but no duplicated VSPs have repeats. Examples of VSPs with tandem repeats are shown in Table S3 and all VSPs having repeats are listed in Table S4. Remarkably, no tandem repeats were found in any CRMPs.

Regarding the identity of the X amino acids in CXXC motifs, it was previously established that X could be any amino acid [8,16]. However, it was found that X could be any aa except C and W. However, there is a preference for T, K, A, S and E in both positions (Fig. 4). The CXXC motif is used by many enzymes to catalyze the formation, isomerization and reduction of disulfide bonds. Thus, the XX dipeptide located between the cysteines may be controlling the redox properties of these molecules [37,38].

Similar properties have been described for the CXC and CXXXC motifs [37,38]. Of all VSPs, 90 VSPs do not have any CXC motifs and 28 only have one (Fig. 4). Therefore, the

CXC motif is rare in VSPs, occurring only in one third of the proteins. Regarding the CXXXC motifs, most VSPs have fewer than 10 and there are 7 VSPs that do not have any. Collectively, X in CXXXC can be any amino acid except C and W, just like X in the CXXC motif (Fig. 4), suggesting that both CXC and CXXXC motifs have evolved from the most ubiquitous CXXC motifs of the VSPs.

All 94 CRMPs have CXXC motifs, but not all of them have CXC (59 of 94) or CXXXC (84 of 94) motifs (Fig. 4). Conversely, most SCRPs (12 of 16) have a marked majority of CXC motifs over CXXC or CXXXC, being this a real difference from the VSPs and CRMPs. All 16 SCRPs have at least 1 CXXC motif, but not all of them have CXXXC motifs (8 of 16) (Fig. 4).

Besides, since CXXC motifs in a VSP have been reported to coordinate Fe^{2+} or Zn^{2+} [10], DiANNA software was used to determine not only the capability of all these motifs to bind metals but also the possibility to form disulfide bonds [39]. All VSPs and most CRMPs were predicted to bind metals and to form intra- and inter-disulfide bonds (Table S5). It is likely that these characteristics favor the formation of the dense coat observed on the trophozoite [40], which is capable of resisting the action of proteases present in the upper small intestine [41]. Conversely, most SCRPs seem not to bind metals but can form disulfide bonds.

Previous reports stated that a GGCY motif is characteristic of VSPs and its mutations result in loss of VSP surface localization [3,16]. This motif is present in 117 of the 136 complete VSPs, with 68 VSPs having more than one. Moreover, GGCY motif is also found in 47 out of 94 CRMP and in 3 SCRPs. It would be interesting to test experimentally if the GGCY motif serves an important purpose in these proteins.

Since no structure of the Cys-rich proteins from *Giardia* has been reported, all Cys-rich

proteins were then analyzed with SMART [42]. Regarding VSPs, no characteristic motif was found present in all 136 sequences. In fact, most sequences (74/136) had no matches in the SMART database. The remaining sequences presented at least one of the following motifs: Furin-like, EGF, and EGF-like or Pfam VSP domains, which contain CXXC motifs. Surprisingly, the Pfam VSP domain (PF03302) was found only in few VSPs (19/136). Similar results were found in the CRMPs and in the SCRPs ([Fig. S1](#)). Since all Cys-rich families present the same motifs, it is a clear indication that all these sequences have a common origin. To determine the relationships between the amino acid sequences of VSPs and CRMP, a phylogenetic analysis was performed [43-46]. Notably, VSPs grouped together, whereas CRMP segregated into two well-separated clusters, one including more divergent CRMP and the other including CRMPs that share some similarity with the VSPs ([Fig. S2](#)).

2.6. Chromosomal location and duplicated pairs

All VSP and CRMP genes are distributed in noncontiguous locations on all five chromosomes ([Fig. 2](#)) and the longer the chromosome, the more Cys-rich genes it encodes ([Table S1](#)). No complete VSP or CRMP coding genes were found in the 30 still unplaced scaffolds.

Regions around VSP genes tend to be gene-poor [18]. The intergenic regions of the *G. lamblia* genome show a median length of 79 nt, indicating that the genome is highly compact and that most intergenic regions are very short. In contrast, the VSP intergenic regions were longer (median of 1280), whereas those of CRMP (median 363) tended to be larger than the average, but shorter than those of VSPs. Regarding these regions, it was earlier observed, using a sliding window GC content analysis that VSP genes have a higher

GC content than the rest of the genome [46]. A re-analysis of upstream regions of the Cys-rich genes showed that the average GC content in the 2000 bp upstream region is 53.9% for VSPs and 50.6% for CRMPs. This should be compared to the average GC content of all CDS, which is 45.2%. Thus, these “promoter” regions seemed larger and with a GC content slightly higher than of any other gene.

Although it was previously reported that VSP1267 (GL50803_00112208) was duplicated [47], duplicated VSP were not reported in the GL2.0 genome. Recently, Xu *et al.* found that among their 133 identified VSPs, 38 were in duplicated pairs [18]. Here, 40 VSPs were confirmed to be in duplicated pairs (in nt sequence as well as in aa sequence). Thirteen of them have tail-to-tail orientation ($\rightarrow \leftarrow$), and 7 have head-to-head orientation ($\leftarrow \rightarrow$). Such pairs were found in all chromosomes but non-identical VSP genes were found in different chromosomes (Table S1). Among them, 11 pairs were well annotated, with one member of the pair having a “d” added after the locus tag prefix GL50803_00 (e.g., GL50803_00112208 and GL50803_00d112208; VSP1267). However, the remaining 9 pairs were not explicitly indicated as duplicated, since their locus tags were different (e.g., GL50803_00117472 and GL50803_00117473). Therefore, to improve their annotation, these 9 pairs should be renamed. Furthermore, two VSPs identified as GL50803_00101765 and GL50803_00d101765 were wrongly annotated as duplicates, since they have different sequences (Table S1).

As VSP genes, there are also duplicated CRMP pairs, although not as many. Only 6 CRMPs were identified in duplicated pairs. Of the 3 pairs, 1 has head-to-tail ($\leftarrow \leftarrow$), and 2 have head-to-head orientations ($\leftarrow \rightarrow$). Furthermore, it was also noted that two CRMPs identified as GL50803_0026981 and GL50803_00d26981 are annotated as duplicates, but they have different sequences. In fact, GL50803_00d26981 is identical to

GL50803_00112673; therefore, they are both the duplicated pair (Table S1). SCRPs genes are also distributed in noncontiguous locations on chromosomes (Fig. 2), but they have no duplicated pairs.

Apart from the VSP and CRMP genes, there are not many identical duplicated genes in the *G. lamblia* genome. Only 2.4% of the 4,966 CDS are duplicated. Consequently, the analysis of the local genomic context turns crucial to understand the expansion of the VSP gene family (Fig. S3). It was observed that some VSP genes are partially duplicated and that they have similar sequences in their intergenic region. Besides, many VSP pairs contain an annotated pseudogene between the members of the pair. When the intergenic regions of the duplicated pairs were analyzed, results showed that when pairs are oriented tail-to-tail, it was frequent to find pseudogenes and when pairs are oriented head-to-head, it was common to find retrotransposon remnant sequences in between (Table S6). For example, the region between the duplicated VSP genes GL50803_00119706 and GL50803_00119707, which are located tail to tail, has 3,301 nt with the presence of two LINE retrotransposons [48]. Moreover, almost identical intergenic sequences are located between duplicated VSP genes and have similarity with other VSP genes, such as the intergenic region between GL50803_00d112208 and GL50803_00112208 (head to head) that is highly conserved with the region located between the duplicated VSPs GL50803_00d115797 and GL50803_00115797. On the other hand, there are genomic regions where VSP, CRMP and pseudo Cys-rich genes are concentrated. For example, VSP GL50803_00113024 and VSP GL50803_0050225 have an identity of 98% in 85% of their sequence; they are located head to head and separated by 21,223 nt that includes two duplicated pVSPs and two duplicated CRMPs (Table S6).

The 7000-nt long 5'-upstream region of VSP1267 has homologous sequences in most of

the 5'-upstream regions of head-to-head duplicated VSP/VSP and VSP/CRMP pairs. CENSOR analysis of these regions identified sequences derived from the retrotransposons BEL, Gypsy and Mariner/Tc1 [49]. Since relics of transposable elements were also found in the flanking regions of tail-to-tail duplicated VSP genes, these findings suggest that VSP genes appeared earlier in evolution. Then, they expanded by initial duplications driven by transposition that, after suffering subsequent mutations, deletions and insertions, generated not only different VSPs, but also multiple CRMPs and SCRPs.

2.7. 5'-UTR and Initiator element (Inr)

For the appropriate expression of a eukaryotic gene, most organisms use two ubiquitous sequence motifs, the TATA box and the Initiator element (Inr) [50-52]. However, the representative TATA box (TATAWAV) is not present in the 5' upstream sequences of the *Giardia* genes, including the VSP genes (Fig. S4A). The alignment of the 200 nt upstream of all VSP genes shows the absence of any known sequence or consensus motif common to all of them, as previously observed [16,53]. However, there seems to be groups of VSP genes that share highly similar sequences (Fig. S4A), which group together in the tree depicted in Fig. 4B.

On the other hand, the Inr is the simplest promoter that can endorse transcription initiation without a TATA box. The Inr is a 17-nt sequence found at the transcription start site of eukaryotic genes [51]. It was described that some VSP genes have the DNA consensus sequence PyAATGTT, where Py is C or T and ATG is the start codon. Transient transfection studies showed that this consensus sequence is required for efficient expression of firefly luciferase from a VSP promoter region [16]. Therefore, this consensus region surrounding the start codon was named initiator element (Inr). By using the previous

version of *G. lamblia* genome, only a few VSP genes containing this sequence were described. Consequently, if Inr was found, VSPs were annotated as “VSP with INR”; and just as “VSP” otherwise. Previously, only 23 of the 136 VSP identified here were annotated as VSPs with INR. Initially, 132 of the 136 VSP genes were found to contain this sequence. A closer inspection of the remaining sequences revealed that a pair of duplicated VSPs (GL50803_00137722 and GL50803_00137723) and VSP GL50803_0050390 (formerly named pVSP) contained that sequence in the next in frame ATG codon, indicating that these 3 sequences have a wrongly annotated start codon. Making this correction, all but one (135 out of 136) are now VSPs with Inr (Fig. 2). Thus, the hypothesis that the Inr sequence is required for efficient expression of a subset of VSP genes is refused and all VSP genes should be annotated as VSP with Inr or simply VSP. Interestingly, the Inr element seems to be almost exclusive to VSP genes, since it was not found in other coding genes, with only very few exceptions (Table S7). The Inr was not found in any of the remaining Cys-rich protein-coding genes, except in CRMP GL50803_00101589, which is the one having the YRGKA CT and a TMD identical to that of VSPs. No characteristic motifs were found in the 5' untranslated region of CRMPs and SCRPs.

Due to the shortness of the *Giardia*'s Inr sequence as well as its localization, it may well function as a Kozak-like sequence [52] almost exclusive to *Giardia* VSPs. Therefore, it is not clear whether the PyAATGTT sequence works as an initiator element for DNA transcription or as a Kozak-like sequence in VSP mRNAs that may facilitate the highly efficient translation of members of this gene family [54,55]. Further studies are needed to clarify this issue.

2.8. 3'-UTR and Polyadenylation sites – Alternative Polyadenylation (APA)

It was previously stated that the poly(A) signal sites (PAS) of *Giardia* genes was AGRAAA, where R is a purine [47,53]. Recently, it was verified on a genome-wide scale that *G. lamblia* uses an AGURAA PAS [56]. AGUAAA and AGUGAA were present in 45% and 15% of protein coding genes, respectively. In contrast, AAUAAA was rare, occurring in only 5% of genes. The most frequent PAS, AGUAAA, differs from the metazoan AAUAAA motif by only a single nucleotide. Moreover, these PAS were found to be depleted in coding regions while occasionally overlap with stop codons, like in the murine parasite *G. muris* [57]. The 3' UTR lengths generated by that approach had a median of 59 nt, which is consistent with the idea that 3' UTRs of *G. lamblia* are unusually short [47,56]. Then, the presence of PAS in the 3' end of all Cys-rich genes was determined showing that all 136 VSPs have an extended PAS with the sequence ACUUAGRUAGURAAAYRY (R= purines and Y=pyrimidines) (Fig. S5A), placed at 6 nt (median) from the stop codon (Fig. S5B). This extended PAS is not present in most of the CRMPs. Thus, by searching for an extended PAS in the 3'UTR of all *Giardia* genes, it was found that the motif is intimately associated with VSPs and pVSPs. Manual curation of the detected motifs revealed that 323 of 327 hits are associated with VSPs and pVSPs with the remaining 4 cases being found in VSP-like CRMPs (Table S8).

For previously annotated HCMPs, the length of the 3' UTR has been determined to be variable, ranging from only a few nt up to thousands [47, 56]. Consistent with this finding, only 67 of 94 CRMPs have AGURAA in the 2000 nt downstream of the stop codon. The same was true for SCRPs, only 11 of 16 have AGURAA in the 2000 nt downstream of the stop codon. The remaining CRMPs and SCRPs might have the PAS located farther away or might have another signal. Incidentally, another explanation could be that a strict PAS may not be required for proper 3'-end formation [47]. Moreover, the presence of alternative

polyadenylation (APA) was analyzed and evidence that some VSP genes show APA was found. For example, VSP GL50803_004059 has two AGUAAA PAS in its 2000-nt downstream region. The PAS are located 12 and 1274 nt downstream of the stop codon. Interestingly, when a single cell RNA-seq experiment was reanalyzed (see below), in some cells this VSP was found to use the first PAS, producing a transcript with a very short 3'-UTR ([Fig. S6A](#)), whereas in other cells this transcript was found to have a significantly longer 3'-UTR and the second PAS was found nearby ([Fig. S6B](#)). This was also seen with other VSPs, such as GL50803_00113797. This is, to our knowledge, the first description that APA might occur in the VSP gene family.

2.9. Cys-rich Pseudogenes

All type I and II pseudogenes were divided into pVSP, pCRMP or pSCRIP based on their similarity to each group. It was found that 96% of type I pseudogenes are similar to VSPs and only 4% are similar to CRMP. In the case of type II cys-rich pseudogenes, 64% of them are similar to VSP, 26% to CRMP and 10% to SCRIP. These numbers indicate that in the *G. lamblia* genome there is a vast number of pVSP, representing ~89% of total Cys-rich pseudogenes, regardless of the type. Interestingly, even though the remaining ~11% of the sequences BLASTed first with a different Cys-rich protein than a VSP, they also BLASTed with VSP with lower query coverage or score. This might indicate that the differences between VSPs and CRMPs or SCRIPs in their ectodomains are not as significant as previously speculated. In contrast to the abundance of pVSPs in the genome, pCRMP and pSCRIP are scarcer, representing only ~9% and ~2% of total Cys-rich pseudogenes (type I and II), respectively.

Evidence of pVSPs is shown in [Table S9](#). For example, pVSP GL50803_0050472 has

multiple internal stop codons, and a closer inspection revealed that the putative coding sequence is split across all three frames, which can be explained by insertions and deletions. Most of the pVSPs without an ORF and without a proper start codon are incomplete and lack the 5' portion, as shown for pVSP GL50803_0050404 ([Table S10](#)). On the other hand, pVSP GL50803_0050265 has a complete ORF starting with a Met and ending in CRGKA, but no SP could be predicted in any Met in frame. Some pVSPs truncated at the C-terminus were also found, and pVSP GL50803_0050028 is an example of this group. The presence of these numerous partial and remnant VSP genes raises questions about their function.

2.10. Transcriptomic and proteomic analyzes

Because VSPs and CRMP share some characteristics but only the VSPs have been involved in antigenic variation, transcriptomic and proteomic analysis were performed. Since most of the early RNA-seq experiments used non-clonal parasites as well as the previous version of the *Giardia* genome as reference, a clear expression profile for VSPs and related genes was not achieved. Thus, to minimize the heterogeneity of expressed VSPs in culture, clonal trophozoite populations from the isolate WB “clone” C6 were generated by limiting dilutions. Sequencing libraries were prepared according to a strand-specific protocol using rRNA-depleted total RNA from 2 different clones, one expressing the VSP417 (formerly named TSA417 [58]; GL50803_00113797), and the other expressing a VSP duplicated pair (VSP1267; GL50803_00112208 and GL50803_00d112208 [59]). The expression of these VSPs on the cell surface was determined by immunofluorescence assays using a specific mAbs [6] and called clone 1 and clone 2, respectively. Three replicates were generated for clone 1 and one for clone 2.

The highest transcribed/accumulated transcript in clone 1 was VSP417 when using unique mapped (UM) reads ([Fig. 5A, left panel](#)). In clone 2, the VSP1267 was not detected as major transcripts, because it belongs to a duplicated pair ([Fig. 5A, left panel](#)). However, since many VSP genes are duplicated, multi-mapping reads were used, setting the maximum number of loci that a read is allowed to map to 2 (2M). Using this approach, it was possible to incorporate ~3.35% extra reads mapping to the genome and the pair GL50803_00112208/GL50803_00d112208 resulted in the most accumulated transcripts in clone 2 ([Fig. 5A, right panel](#)). At this point, it cannot be discriminated whether duplicated VSPs are expressed together or only one member of the pair is transcribed; however, whichever the case, the VSP that is expressed on the surface of the parasites correlates with the transcript that accumulates the most in each clone. Transcript per million (TPM) values for each transcript in each sample can be found in [Table S11](#).

Focusing on VSP, CRMP and SCRIP transcripts yielded that VSPs are transcribed and accumulated at different levels. However, it is clear that the VSP that is translated into the protein detected on the trophozoite membrane is the one that accumulates the most and its abundance differs radically from the remaining transcripts of this gene family ([Fig. 5B](#)). In clone 1, more than 90% of the total VSP gene transcript level is monopolized by VSP417. In the case of clone 2, VSP1267 monopolized 87.2% of the total VSP gene transcripts (~43.6% of each member of the pair) ([Fig. 5C](#)). This indicates that clones expressing a single surface antigen transcribe several VSP genes [6], but only accumulate in abundance transcripts encoding the VSP that is translated and expressed on the parasite surface.

Similar analysis was done with individual CRMP and SCRIP transcripts. CRMPs and SCRIPs are also transcribed and accumulated at different levels. However, no single transcript differs radically from the rest in terms of its TPM ([Fig. 5B](#)). Unlike VSPs, where

there is always one gene that accounts for nearly 90% of all VSP transcripts in clonal populations, in CRMPs and SCRPs none exceeds ~15% and ~21% of the total, respectively, similar to any other *Giardia* gene (Fig. 5C).

To identify differentially expressed genes (DEGs) that might be related to the process of antigenic variation triggered by antibodies, biological replicates of clone1 were incubated for 4 h with low concentrations of mAb 7C2 to induce VSP switching and RNA sequencing was performed [60]. Surprisingly, only 7 DEGs (FDR<0.05) were found between trophozoites before and after treatment (Table S12). The DEGs encode 6 CRMP (GL50803_00137727, GL50803_00112135, GL50803_0025816, GL50803_00113531, GL50803_00115066, and GL50803_00114930), which were all upregulated in antibody-treated clones. The level of upregulation was extremely small, with the highest logFC found being 1. These small logFCs were previously reported, although in experiments not related to antigenic variation [12].

Although coming from a single clone the triplicates have a vast intrinsic variation, clearly evidencing the stochastic nature of gene expression in *Giardia*. Stochastic gene expression has a significant effect on the biology of microorganisms since genetically identical clones may diverge phenotypically to adapt to fluctuating environmental conditions [61].

Besides the present study, only one report [12] performed RNA-seq of a clonal population. Authors were interested in assessing the effects of iron on the total *G. lamblia* transcriptome. The raw data from 4 SRA randomly chosen samples from those experiments were analyzed using the described pipeline. These samples were called Clon3-rep1 to Clon3-rep4. In Table S13, the SRA accession numbers are indicated. Unlike our RNA-seq, these libraries were prepared including a poly(A) selection step. Similar to our results, this clone highly accumulates the transcript corresponding to a single VSP, in this case

GL50803_00113450, which monopolizes between 74.48% and 78.79% of the total VSP gene transcripts. Again, this characteristic of a single transcript accumulated above others is not observed for any member of the CRMP or SCRPF families (Fig. 6A).

In another set of RNA-Seq experiments, but using non-clonal populations, Peirasmaki *et al.* [12] used an *in vitro* model of interaction of trophozoites with intestinal epithelial cells to identify genes that might correlate with the establishment of infection and disease induction. The raw data of 4 randomly chosen samples from SRA were collected and analyzed (named Population1-rep1 to Population1-rep4 and their SRA accession numbers are indicated in Table S13). In these populations, no single VSP dominated most of the transcripts (Fig. 6B). In fact, two VSP transcripts corresponding to GL50803_0041472 and GL50803_0040591 have only ~35% of the total VSP transcripts each, indicating that this population comprises two different VSP clones.

Furthermore, by using raw single cell RNA-seq data (SRA SRR9222552-SRR9222606) from experiments where the authors tested different modifications a single-cell RNA sequencing protocol originally developed for mammalian cells to establish a cost-efficient workflow for protists [62], the preponderance of VSP transcript accumulation was also verified. They generated 55 transcriptomes of single *G. lamblia* cells. Unfortunately, the authors used the GL2.0 genome for analysis. Analyzed using our own pipeline and the new reference genome, 30 cells accumulated VSP417 as the major transcript (the same VSP found in our clone 1), 11 cells accumulated VSP GL50803_0040591 and the rest accumulated other VSPs or pairs of duplicates. Six representative samples called SC1 to SC6 were selected for a closer inspection (Table S13). SC1 and SC2 accumulated VSP417 and this VSP monopolized 87.49% and 86.18% of the total VSP gene transcripts, whereas SC3 and SC4 accumulated VSPA8 and this transcript monopolized 84.72% and 82.23%,

respectively. SC5 accumulated VSP GL50803_00137618 with 74.70% of total VSP transcripts in that cell (Fig. 6C). These samples were chosen to show that despite coming from the WB “clone” C6 (ATCC[®] 50803), different cells accumulate a VSP transcript that is not always the same. On the other hand, SC6 does not accumulate any VSP with more than 17% of the total VSP transcripts using the UM strategy. Therefore, for this cell transcriptome, the 2M strategy was used and, indeed, SC6 accumulated the VSP pair GL50803_0050229/GL50803_00d50229, which contributed ~4.3% each to the total VSP transcripts in that cell (Fig. 6C). From these analyzes it can be concluded that only in bulk RNA-seq of well-defined clones and in single cell RNA-seq experiments it is possible to determine the VSP that will be translated and incorporated into the cell membrane of individual trophozoites.

Regarding proteomic data, earlier studies performed on the previous version of the *Giardia* genome and with parasite populations showed translational evidence for all groups of Cys-rich encoding genes [14,63–66]. Therefore, the raw data from mass spectrometry experiments were downloaded from the PRIDE repository and re-analyzed with the MaxQuant software using the specific parameters described in each of these reports, but using the proteins fasta file of the GL2.1 genome. The following PRIDE projects were used: PXD017597, PXD004398, PXD000452, PXD002398, PXD007183 and PXD022565. Additionally, two recent proteomic studies that use the updated version of the *Giardia* genome [67,68]. The results showed evidence of translation for only a subset of VSP proteins with unique peptides (29/136), which is in agreement with the results of the transcriptional analysis. The proteins with evidence of translation can be found in [Table S14](#). Interestingly, some peptides detected in those analyzes were shared between VSPs and CRMPs, while others were shared between CRMPs and SCRPs, strongly suggesting that

SCRPs may derive from CRMPs and these from VSPs.

3. Conclusions

Giardia colonizes the lumen of the small intestine, where digestion of nutrients takes place. How *Giardia* can survive in this harsh environment has been suggested to depend on the resistance of the VSPs to proteolytic digestion, properties that are provided by the CXXC motifs and their capability to bind metals [9,41]. Additionally, VSPs are involved in the process of antigenic variation, making the parasite capable of avoiding the constant immune pressure generated by their hosts by switching the expression of antigenically different VSPs [69]. Other Cys-rich proteins are also encoded in the *Giardia*'s genome, but their function is controversial.

Here, by performing an exhaustive analysis of these groups of proteins, we defined a new classification that better describes their characteristics: VSPs, CRMPs, and SCRPs. This much simpler classification is supported not only by the analysis of their sequences, but also by their patterns of expression both at the transcriptional and post-transcriptional levels.

Despite their sequence variability, VSPs have strikingly constant features that are not shared by any other Cys-rich family. The presence of a CRGKA cytoplasmic tail at the C-terminal end of the protein, an extended PAS and the Inr element /Kozak-like sequence at the 5'-end of the gene seems to be the defining feature of the VSPs, in addition to the presence of an SP, multiple CXXC motifs in the ED and a highly conserved and almost exclusive TMD. Interestingly, we also found that the extended PAS is associated with most

pVSP genes and some VSPs show APA. Besides, the variability in the features of CRMPs (no Inr and no CRGKA CT) and their similarity to the VSPs in their ED suggest that these proteins accomplish a function analogous to that of VSPs in protecting the parasite surface from digestion. However, only the VSPs accumulate a unique transcript at high levels. This is not the case for the CRMP and SCRIP.

CRMPs seemed up regulated during host-parasite interactions *in vivo* [70], during parasite-epithelial cell interactions *in vitro* [12], during encystation [71] and after confronting trophozoites antibodies against their cognate VSP (this work). It seems that some relative CRMP transcript levels increase in response to stress factors. On the other hand, SCRIP may function as virulence factors, as suggested [14], or might stabilize VSP and CRMP on the membrane of the trophozoites.

In sum, our results not only provide original observations regarding the Cyst-rich proteins of *Giardia* but also these findings pave the way to decipher in detail the involvement of these proteins on the biology of this important human pathogen.

4. Materials and Methods

4.1. Sources of genomic, transcriptomic and proteomic sequences

Genomic, transcriptomic and proteomic data were obtained from RefSeq (GCF_000002435.2), the sequence read archive SRA and PRIDE proteomics repository, respectively. Genomic and transcriptomic data were visualized using the IGV_2.8.13 software [72].

4.2. Computational characterization of cysteine rich proteins

The presence of a predicted SP was determined using Phobius [19], SignalP-5.0 [20] and TOPCONS [21]. Phobius [19], TMHMM 2.0 [22] and TOPCONS [21] were used to infer transmembrane regions. The subcellular localization of all Cys-rich proteins was predicted by DeepLoc 1.0 software [24]. All software were used with default parameters.

4.3. Sequence analysis

For sequence analysis, R (v4.0.3) were used with multiple packages like Biostrings, seqinr and stringr. Different parameters were searched using R for both nucleotide and amino acid sequences. Some of these parameters were the length of the protein, number of CXXC, CXC and CXXXC motifs, presence of PAS and of Inr. Statistical analyzes were also performed in R. The amino acid sequences of the proteins rich in cysteines were analyzed using DiANNA Analysis 1.1 web server [39] to look for cysteines that coordinate metals such as Fe²⁺ and Zn²⁺. Conserved motifs in VSP were searched with MEME Suite 5.3.3 [35]. The classic mode was selected in the motif discovery mode. Zero or one occurrence per sequence was selected and the software was set to search for 10 motifs of up to 70 characters. No other parameter was modified. The XSTREME software of the MEME suite [35] was used to scan the 136 VSP 3'UTR regions (stop codon + 50 bp) of VSP genes for the presence of conserved motifs using standard settings. The top-matching MEME output was used to extract a sequence logo of the motif and the position. The Fimo software of the MEME suite was used to scan the 3'UTR region (stop codon + 50 bp) of all genes (including pseudogenes) in the *Giardia* genome using the extended PAS. The 381 putative hits reported at a p-value of 1.0⁻⁰⁵. The hits were manually curated, with 327 cases remaining after the hits showing poor motif conservation in otherwise high conserved

positions, shifted placement, or wrong orientation were excluded. Using the manual analysis the cutoff representing a genuine hit had a p-value of 1.41^{-07} .

4.4. VSP and CRMP duplicated pairs and intergenic regions

To study the genomic context of the duplicated VSP and CRMP gene pairs, the specific intergenic region of each gene pair was analyzed and BLASTn was done using the BioProject 1439: *Giardia intestinalis* strain WB C6.

4.5. In silico VSP TMD oligomerization models

Oligomerization capability of the TMD was analyzed with TMHOP [73] and PREDIMMER [29] and visualized with VMD [74]. Alignments were performed with Clustal Omega [75] with default parameters and visualized with WebLogo [76].

4.6. Parasite culture

Giardia lamblia assemblage A1 isolate WB (ATCC[®] 50803) was cultured in borosilicate glass tubes containing TYI S-33 medium supplemented with 0.5 mg/ml of bovine bile (Sigma-Aldrich, Cat. #B5883) and 10% adult bovine serum (Natocor), as described [6]. Clones expressing different surface antigens were obtained by limiting dilution in 96-well culture plates placed in anaerobiosis chambers (Anaerogen[®] Compact, Thermo Scientific[®] Oxoid[®], Cat. # AN0010C) at 37 °C during 5 days; positive clones were then selected using specific anti-VSP mAb by immunofluorescence assays (IFA). Reactive clones were then expanded in culture medium overnight and tested for homogeneity before use (Fig. S7).

4.7. Transcriptomic analysis

To minimize the heterogeneity of expressed VSPs in culture, clonal trophozoite populations were generated from the original WB C6 “clone”. Total RNA was purified from clonal population by Trizol extraction according to the manufacturer’s protocol. The experiment had 7 RNA samples collected from 2 clones, 6 samples from clone 1 expressing VSP GL50803_00113797 (3 for control and 3 for antibody treatment) and 1 sample from clone 2 expressing the VSP pair GL50803_00112208 and GL50803_00d112208. Ribosomal RNA was depleted in all samples and strand-specific Illumina sequencing libraries were constructed. RNA sequencing was performed as a paid service by Genehub.com. The libraries were sequenced as 2x150 bp paired-end reads on an Illumina MiSeq instrument, and 3.1-4.3 million reads were subsequently generated for each library. Raw sequence reads from Illumina were submitted to the NCBI Sequence Read Archive, with SRA accession numbers SRR17933280 to SRR17933282 and SRR17933276 for control samples, and SRR17933277 to SRR17933279 for antibody-treated samples. The sequences can also be accessed from the BioProject PRJNA804420. Genome index was generated using STARv2.5.4b with default parameters except for `--genomeSAindexNbases`. According to the STAR manual, for small genomes, the parameter `--genomeSAindexNbases` must be scaled down, with a typical value of $\min(14, \log_2(\text{GenomeLength})/2 - 1)$. Given that *G. lamblia* genome is 12.6 Mb in length, this parameter was set to 11. Quality control of raw reads was performed using FastQC v0.11.5. If needed, trimming of low quality base calls (below 30) and adapter sequences was performed with TrimGalore v0.6.4 and Cutadapt v1.15, in order to improve mapping efficiency. Sequences shorter than 70 bp were removed. Reads that passed quality criteria were aligned to the latest reference genome using STAR v2.5.4b with default parameters except for: (1) `--alignIntronMax` (maximum intron size), which was set to 1 because, with

few exceptions, *G. lamblia* genes lack introns, and (2) `--outFilterMultimapNmax` (max number of multiple alignments allowed for a read), which was set to 2 to account for duplicated genes. On average, ~80% of all sequenced reads could be mapped to a single location in the genome. After mapping, reads were assigned to genomic features using FeatureCounts v2.0.0. Multi-overlapping and reads not overlapping with any features in the annotation file were not counted. Since many VSP genes are duplicated, two strategies were used. In strategy UM, multimapping reads were not counted, whereas in strategy 2M, multimapping reads were fractionally counted (each alignment carrying 1/2 count, since the parameter `--outFilterMultimapNmax` was previously set to 2 in the alignment step). Sense and antisense counts for each gene in each library were then normalized to transcript per million (TPM). Commands used are indicated in Supplemental Methods file 1. The same procedure used for analyzing our own samples was used for analyzing several transcriptomic datasets deposited in SRA. Differential expression analysis was done using edgeR v3.28 package. Counts-per-million were calculated for each gene to standardize for differences in library-size. Genes that do not have a worthwhile number of counts in any sample were filtered out of downstream analyzes. Therefore, filtering was carried out to retain genes with a baseline expression level of at least 10 cpm in 3 or more samples. For each data set, TMM normalization was then applied to account for the compositional biases. All plots, images and statistical analyzes were performed in R (v4.0.3) and GraphPad Prism (v9.00), and were later compiled in Adobe Illustrator.

4.8. Proteomic analysis

Data was processed using MaxQuant version 2.0.1.0 with the parameters indicated in the corresponding PRIDE projects. In general, however, the following parameters were used:

Enzyme, Trypsin/P; Database, NCBI *Giardia lamblia* RefSeq 2.1 proteome (concatenated forward and reverse plus common contaminants); Fixed modification, Carbamidomethyl (C); Variable modifications, Oxidation (M), Acetyl (N-term), Pyro-Glu (N-term Q), Deamidation (N/Q); Mass values, Monoisotopic; Peptide Mass Tolerance, 10 ppm; Fragment Mass Tolerance, 0.02 Da; Max Missed Cleavages, 2. Data were filtered using 1% protein and peptide FDR and requiring at least two unique peptides per protein.

Journal Pre-proof

Acknowledgments

Authors thank Dr. Albano H. Tenaglia for preparing the clones for transcriptomic analysis. This work was supported by grants FONCYT (PICT-13469, PICT-2703, PICT-E 0234, and PICT-2116), CONICET (D4408), and UCC (80020150200144CC) of Argentina to H.D.L.

Author contributions

MRW and CRM performed most of the sequences analyzes. MRW contributed the transcriptomic analysis. LAL contributed the proteomic analyzes and the immunofluorescence assays. MRW, CRM and LAL verified each other results. AS generated monoclonal antibodies used for cloning. EAF supervised the bioinformatic analyzes. SGS and JJH contributed with PAS analysis. HDL conceived the project and designed the experiments. All authors analyzed the data. MRW, CRM, EAF, SGS and HDL wrote the paper. All authors read and commented on the manuscript.

References

1. Luján HD. Mechanisms of adaptation in the intestinal parasite *Giardia lamblia*. *Essays Biochem.* 2011; 51: 177–91. DOI: [10.1042/bse0510177](https://doi.org/10.1042/bse0510177)
2. Deitsch KW, Lukehart SA, Stringer JR. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Microbiol.* 2009; 7: 493–503. DOI: [10.1038/nrmicro2145](https://doi.org/10.1038/nrmicro2145)
3. Nash TE. Surface antigenic variation in *Giardia lamblia*. *Mol Microbiol.* 2002; 45:585-90. DOI: [10.1046/j.1365-2958.2002.03029.x](https://doi.org/10.1046/j.1365-2958.2002.03029.x)
4. Gargantini PR, Serradell MC, Ríos DN, Tenaglia AH, Luján HD. Antigenic variation in the intestinal parasite *Giardia lamblia*. *Curr Opin Microbiol.* 2016; 32: 52–58. DOI: [10.1016/j.mib.2016.04.017](https://doi.org/10.1016/j.mib.2016.04.017)
5. Prucca CG, Rivero FD, Luján HD. Regulation of antigenic variation in *Giardia lamblia*. *Annu Rev Microbiol.* 2011; 65: 611–30. DOI: [10.1146/annurev-micro-090110-102940](https://doi.org/10.1146/annurev-micro-090110-102940)
6. Prucca CG, Slavin I, Quiroga R, Elías EV, Rivero FD, Saura A, et al. Antigenic variation in *Giardia lamblia* is regulated by RNA interference. *Nature.* 2008; 456: 750–4. DOI: [10.1038/nature07585](https://doi.org/10.1038/nature07585)
7. Carranza PG, Gargantini PR, Prucca CG, Torri A, Saura A, Svärd S, et al. Specific histone modifications play critical roles in the control of encystation and antigenic variation in the early-branching eukaryote *Giardia lamblia*. *Int J Biochem Cell Biol.* 2016; 81: 32–43. DOI: [10.1016/j.biocel.2016.10.010](https://doi.org/10.1016/j.biocel.2016.10.010)
8. Adam RD, Nash TE. Antigenic Variation of the VSP Genes of *Giardia lamblia*. In: *The Pathogenic Enteric Protozoa: Giardia, Entamoeba, Cryptosporidium and Cyclospora.* Sterling CR, Adam RD, editors. Kluwer Academic Publishers; 2004. pp. 59–73. DOI: [10.1007/1-4020-7878-1_5](https://doi.org/10.1007/1-4020-7878-1_5)
9. Serradell MC, Rupil LL, Martínez PA, Prucca CG, Carranza PG, Saura A, et al. Efficient oral vaccination by bioengineering virus-like particles with protozoan surface proteins. *Nat Commun.* 2019; 10: 361. DOI: [10.1038/s41467-018-08265-9](https://doi.org/10.1038/s41467-018-08265-9)
10. Luján HD, Mowatt MR, Wu JJ, Lu Y, Lees A, Chance MR, et al. Purification of a variant-specific surface protein of *Giardia lamblia* and characterization of its metal-binding properties. *J Biol Chem.* 1995; 270: 13807–13. DOI: [10.1074/jbc.270.23.13807](https://doi.org/10.1074/jbc.270.23.13807)
11. Davids BJ, Reiner DS, Birkeland SR, Preheim SP, Cipriano MJ, et al. A New Family of Giardial Cysteine Rich Non-VSP Protein Genes and a Novel Cyst Protein. *PLoS One.* 2006, 1: e44. DOI: [10.1371/journal.pone.0000044](https://doi.org/10.1371/journal.pone.0000044)
12. Peirasmaki D, Ma'ayeh SY, Xu F, Ferella M, Campos S, Liu J, Svärd SG. High Cysteine Membrane Proteins (HCMPs) are up-regulated during *Giardia*-host cell interactions. *Front Genet.* 2020; 11: 913. DOI: [10.3389/fgene.2020.00913](https://doi.org/10.3389/fgene.2020.00913)
13. Emery, S., Mirzaei, M., Vuong, D, Pascovici D, Chick JM, Lacey E, Hayneset PA, et al. Induction of virulence factors in *Giardia duodenalis* independent of host attachment. *Sci Rep.* 2016; 6: 20765. DOI: [10.1038/srep20765](https://doi.org/10.1038/srep20765)
14. Dubourg A, Xia D, Winpenny JP, Al Naimi S, Bouzid M, Sexton DW, et al. *Giardia* secretome highlights secreted tenascins as a key component of pathogenesis. *Gigascience.* 2018; 7: 1–13. DOI: [10.1093/gigascience/giy003](https://doi.org/10.1093/gigascience/giy003)
15. Morrison HG, McArthur MG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, et al. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science.* 2007; 317: 1921–6. DOI: [10.1126/science.1143837](https://doi.org/10.1126/science.1143837)

16. Adam RD, Nigam A, Seshadri V, Martens CA, Farneth GA, Morrison HG, et al. The *Giardia lamblia* vsp gene repertoire: characteristics, genomic organization, and evolution. BMC Genomics. 2010; 11: 424. DOI: [10.1186/1471-2164-11-424](https://doi.org/10.1186/1471-2164-11-424)
17. Li W, Saraiya AA, Wang CC. Experimental verification of the identity of variant-specific surface proteins in *Giardia lamblia* trophozoites. mBio; 2013. pp. 00321–13. DOI: [10.1128/mBio.00321-13](https://doi.org/10.1128/mBio.00321-13)
18. Xu F, Jex A, Svärd SG. A chromosome-scale reference genome for *Giardia intestinalis* WB. Sci Data. 2020; 7: 38. DOI: [10.1038/s41597-020-0377-y](https://doi.org/10.1038/s41597-020-0377-y)
19. Käll L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007; 35: 429–32. DOI: [10.1093/nar/gkm256](https://doi.org/10.1093/nar/gkm256)
20. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunet S, et al, SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019; 37: 420–423. DOI: [10.1038/s41587-019-0036-z](https://doi.org/10.1038/s41587-019-0036-z)
21. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 2015; 43: 401–7. DOI: [10.1093/nar/gkv485](https://doi.org/10.1093/nar/gkv485)
22. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001; 305: 567–80. DOI: [10.1006/jmbi.2000.4315](https://doi.org/10.1006/jmbi.2000.4315)
23. Baker JA, Wong WC, Eisenhower B, Warwicker J, Eisenhaber F. Charged residues next to transmembrane regions revisited: “Positive-inside rule” is complemented by the “negative inside depletion/outside enrichment rule”. BMC Biol. 2017; 15: 66. DOI: [10.1186/s12915-017-0404-4](https://doi.org/10.1186/s12915-017-0404-4)
24. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017; 33: 3387–3395. DOI: [10.1093/bioinformatics/btx548](https://doi.org/10.1093/bioinformatics/btx548)
25. Singh S, Mittal A. Transmembrane domain lengths serve as signatures of organismal complexity and viral transport mechanisms. Sci Rep. 2016; 6: 22352. DOI: [10.1038/srep22352](https://doi.org/10.1038/srep22352)
26. Quiroga R, Trenchi A, Colizález Montoro A, Valdez Taubas J, Maccioni HJF. Short transmembrane domains with high-volume exoplasmic halves determine retention of Type II membrane proteins in the Golgi complex. J Cell Sci. 2013; 126: 5344–9. DOI: [10.1242/jcs.130658](https://doi.org/10.1242/jcs.130658)
27. Mueller BK, Subramaniam S, Senes A. A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical α -H hydrogen bonds. PNAS. 2014; 111: E888–E895. DOI: [10.1073/pnas.1319944111](https://doi.org/10.1073/pnas.1319944111)
28. Russ WP, Engelman DM. The GxxxG motif: a framework for transmembrane helix-helix association. J Mol Biol. 2000; 296: 911–919. DOI: [10.1006/jmbi.1999.3489](https://doi.org/10.1006/jmbi.1999.3489)
29. Polyansky AA, Chugunov AO, Volynsky PE, Krylov NA, Nolde DE, Efremov RG. PREDDIMER: a web server for prediction of transmembrane helical dimers. Bioinformatics. 2014; 30: 889–90. DOI: [10.1093/bioinformatics/btt645](https://doi.org/10.1093/bioinformatics/btt645)
30. Weinstein JY, Elazar A, Fleishman SJ. A lipophilicity-based energy function for membrane-protein modelling and design. PLOS Comp Biol. 2019; 15: e1007318. DOI: [10.1371/journal.pcbi.1007318](https://doi.org/10.1371/journal.pcbi.1007318)
31. Brown DA. Lipid rafts, detergent-resistant membranes, and raft targeting signals. Physiology (Bethesda). 2006; 21: 430–439. DOI: [10.1152/physiol.00032.2006](https://doi.org/10.1152/physiol.00032.2006)

32. Diaz-Rohrer BB, Levental KR, Simons K, Levental I. Membrane raft association is a determinant of plasma membrane localization. *PNAS*. 2014; 111: 8500–8505. DOI: [10.1073/pnas.1404582111](https://doi.org/10.1073/pnas.1404582111)
33. Simons K, Ikonen E. Functional rafts in cell membranes. *Nature*. 1997; 387: 569–72. DOI: [10.1038/42408](https://doi.org/10.1038/42408)
34. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*. 1999; 27: 573–580. DOI: [10.1093/nar/27.2.573](https://doi.org/10.1093/nar/27.2.573)
35. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009; 202–8. DOI: [10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335)
36. Madeira F. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. 2019; 47: 636–641. DOI: [10.1093/nar/gkz268](https://doi.org/10.1093/nar/gkz268)
37. Quan S, Schneider I, Jonathan Pan J, Von Hacht A, Bardwell JCA. The CXXC motif is more than a redox rheostat. *J Biol Chem*. 2007; 282: 28825–28833. DOI: [10.1074/jbc.M705291200](https://doi.org/10.1074/jbc.M705291200)
38. Schultz LW, Chivers PT, Raines RT. The CXXC motif: crystal structure of an active-site variant of *Escherichia coli* thioredoxin. *Acta Crystallogr D Biol Crystallogr*. 1999; 55: 1533–8. DOI: [10.1107/s0907444999008756](https://doi.org/10.1107/s0907444999008756)
39. Ferrè F, Clote P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res*. 2005; 33: 230–2. DOI: [10.1093/nar/gki412](https://doi.org/10.1093/nar/gki412)
40. Pimenta PF, da Silva PP, Nash T. Variant surface antigens of *Giardia lamblia* are associated with the presence of a thick cell coat: thin section and label fracture immunocytochemistry survey. *Infect Immun*. 1991; 59: 3989–96. DOI: [10.1128/iai.59.11.3989-3996.1991](https://doi.org/10.1128/iai.59.11.3989-3996.1991)
41. Rupil LL, Serradell MC, Luján HL. Using protozoan surface proteins for effective oral vaccination. *Trends Parasitol*. 2019; 36: 7–10. DOI: [10.1016/j.pt.2019.07.004](https://doi.org/10.1016/j.pt.2019.07.004)
42. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Research*. 2021; 49: D458–D460. DOI: [10.1093/nar/gkaa937](https://doi.org/10.1093/nar/gkaa937)
43. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre² web portal for protein modeling, prediction and analysis. *Nat Protoc*. 2015; 10: 845–58. DOI: [10.1038/nprot.2015.052](https://doi.org/10.1038/nprot.2015.052)
44. Katoh K, Rozewicz J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 2019; 20: 1160–1166. DOI: [10.1093/bib/bbx108](https://doi.org/10.1093/bib/bbx108)
45. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014; 30: 1312–3. DOI: [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033)
46. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021; 49: 293–296. DOI: [10.1093/nar/gkab301](https://doi.org/10.1093/nar/gkab301)
47. Franzén O, Jerlström-Hultqvist J, Einarsson E, Ankarklev J, Ferella M, Andersson B, Svärd SG. Transcriptome profiling of *Giardia intestinalis* using strand-specific RNA-seq. *PLoS Comput Biol*. 2013; 9: 1003000. DOI: [10.1371/journal.pcbi.1003000](https://doi.org/10.1371/journal.pcbi.1003000)
48. Grandi FC, An W. Non-LTR retrotransposons and microsatellites: Partners in genomic variation. *Mob Genet Elements*. 2013; 3: e25674. DOI: [10.4161/mge.25674](https://doi.org/10.4161/mge.25674)

49. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007; 8: 973–982. DOI: [10.1038/nrg2165](https://doi.org/10.1038/nrg2165)
50. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. *Annu Rev Biochem.* 2003; 72: 449–79. DOI: [10.1146/annurev.biochem.72.121801.161520](https://doi.org/10.1146/annurev.biochem.72.121801.161520)
51. Smale ST, Baltimore D. The “initiator” as a transcription control element. *Cell.* 1989; 57: 103–13. DOI: [10.1016/0092-8674\(89\)90176-1](https://doi.org/10.1016/0092-8674(89)90176-1)
52. Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene.* 2002; 299: 1–34. DOI: [10.1016/s0378-1119\(02\)01056-9](https://doi.org/10.1016/s0378-1119(02)01056-9)
53. Adam RD. Biology of *Giardia lamblia*. *Clin Microbiol Rev.* 2001; 14: 447–475. DOI: [10.1128/CMR.14.3.447-475.2001](https://doi.org/10.1128/CMR.14.3.447-475.2001)
54. Yee J, Nash TE. Transient transfection and expression of firefly luciferase in *Giardia lamblia*. *Proc Natl Acad Sci U S A.* 1995; 92: 5615–5619. DOI: [10.1073/pnas.92.12.5615](https://doi.org/10.1073/pnas.92.12.5615)
55. Singer SM, Yee J, Nash TE. Episomal and integrated maintenance of foreign DNA in *Giardia lamblia*. *Mol Biochem Parasitol.* 1998;92: 59–69. DOI: [10.1016/s0166-6851\(97\)00225-9](https://doi.org/10.1016/s0166-6851(97)00225-9)
56. Bilodeau DY, Sheridan RM, Balan B, Jex AR, Rissland OS. Precise gene models using long-read sequencing reveal a unique poly(A) signal in *Giardia lamblia*. 2021. DOI: [10.1261/rna.078793.121](https://doi.org/10.1261/rna.078793.121)
57. Onsbring H, Tice AK, Barton BT, Brown MV, Ettema TJG. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on *Giardia intestinalis* cells. *BMC Genomics.* 2020; 21: 448. DOI: [10.1186/s12864-020-06858-7](https://doi.org/10.1186/s12864-020-06858-7)
58. Gillin FD, Hagblom P, Harwood J, Alejo SB, Reiner DS, McCaffery M, et al. Isolation and expression of the gene for a major surface protein of *Giardia lamblia*. *PNAS.* 1990; 87: 4463–4467. DOI: [10.1073/pnas.87.12.4463](https://doi.org/10.1073/pnas.87.12.4463)
59. Mowatt MR, Aggarwal A, Nash TE. Carboxy-terminal sequence conservation among variant-specific surface proteins of *Giardia lamblia*. *Mol Biochem Parasitol.* 1991; 49: 215–27. DOI: [10.1016/0166-6351\(91\)90065-e](https://doi.org/10.1016/0166-6351(91)90065-e)
60. Tenaglia AH, Luján LA, Píros DN, Midlej V, Iribarren PA, Molina CR, et al. Antibodies to protozoan variable surface antigens induce antigenic variation. *BioRxiv.* 2022. Preprint at: <https://www.biorxiv.org/content/10.1101/2022.06.21.497077v1>
61. McAdams HH, Arkin A. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA.* 1997; 94: 814–819. DOI: [10.1073/pnas.94.3.814](https://doi.org/10.1073/pnas.94.3.814)
62. Xu F, Jiménez-González A, Einarsson E, Ástvaldsson Á, Peirasmaki D, Eckmann L, et al. The compact genome of *Giardia muris* reveals important steps in the evolution of intestinal protozoan parasites. *Microb Genom.* 2020; 6(8):mgen000402. DOI: [10.1099/mgen.0.000402](https://doi.org/10.1099/mgen.0.000402)
63. Müller J, Braga S, Uldry A-N, Heller M, Müller N. Comparative proteomics of three *Giardia lamblia* strains: investigation of antigenic variation in the post-genomic era. *Parasitology.* 2020; 147: 1008–1018. DOI: [10.1017/S0031182020000657](https://doi.org/10.1017/S0031182020000657)
64. Emery SJ, Baker L, Ansell BRE, Mirzaei M, Haynes PA, McConville MJ, et al. Differential protein expression and post-translational modifications in metronidazole-resistant *Giardia duodenalis*. *Gigascience.* 2018; 7. DOI: [10.1093/gigascience/giy024](https://doi.org/10.1093/gigascience/giy024)
65. Emery SJ, Sluyter S, Haynes PA. Proteomic analysis in *Giardia duodenalis* yields insights into strain virulence and antigenic variation. *Proteomics.* 2014; 14: 2523–34. DOI: [10.1002/pmic.201400144](https://doi.org/10.1002/pmic.201400144)

66. Heller M, Braga S, Müller N, Müller J. Transfection with plasmid causing stable expression of a foreign gene affects general proteome pattern in *Giardia lamblia* trophozoites. *Front Cell Infect Microbiol.* 2020; 10: 602756. DOI: [10.3389/fcimb.2020.602756](https://doi.org/10.3389/fcimb.2020.602756)
67. Zhao P, Cao L, Wang X, Dong J, Zhang N, Li, X, et al. Extracellular vesicles secreted by *Giardia duodenalis* regulate host cell innate immunity via TLR2 and NLRP3 inflammasome signaling pathways. *PLoS Negl Trop Dis.* 2021; 15: 4 e0009304. DOI: [10.1371/journal.pntd.0009304](https://doi.org/10.1371/journal.pntd.0009304)
68. Krakovka S, Ribacke U, Miyamoto Y, Eckmann L, Svärd SG. Characterization of metronidazole-Rresistant *Giardia intestinalis* lines by comparative transcriptomics and proteomics. *Front Microbiol.* 2022; 13:834008. DOI: [10.3389/fmicb.2022.834008](https://doi.org/10.3389/fmicb.2022.834008)
69. Nash TE. Antigenic variation in *Giardia lamblia* and the host's immune response. Liew FY, Vickerman K, editors. *Phil Trans R Soc Lond B.* 1997; 352: 1369–1375. DOI: [10.1098/rstb.1997.0122](https://doi.org/10.1098/rstb.1997.0122)
70. Pham JK, Nosala C, Scott EY, Nguyen KF, Hagen KD, Stancovich HN, Dawson SC. Transcriptomic profiling of high-density *Giardia* foci encysting in the murine proximal Intestine. *Front Cell Infect Microbiol.* 2017; 7: 227. DOI: [10.3389/fcimb.2017.00227](https://doi.org/10.3389/fcimb.2017.00227)
71. Einarsson E, Troell K, Hoepfner MP, Grabherr M, Ribacke U, Svärd SG. Coordinated changes in gene expression throughout encystation of *Giardia intestinalis*. *PLoS Negl Trop Dis.* 2016; 10: 0004571. DOI: [10.1371/journal.pntd.0004571](https://doi.org/10.1371/journal.pntd.0004571)
72. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol.* 2011; 29: 24–6. DOI: [10.1038/nbt.1754](https://doi.org/10.1038/nbt.1754)
73. Xiao Y, Zeng B, Berner N, Frishman L, Langosch D, Teese MG. Experimental determination and data-driven prediction of homotypic transmembrane domain interfaces. *Comput Struct Biotechnol J.* 2020; 18: 3230–3242. DOI: [10.1016/j.csbj.2020.09.035](https://doi.org/10.1016/j.csbj.2020.09.035)
74. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph.* 1996; 14: 33–38, 27–28. DOI: [10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
75. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011; 7: 539. DOI: [10.1038/msb.2011.75](https://doi.org/10.1038/msb.2011.75)
76. Crooks GE, Hon C, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004; 14: 1188–1190. DOI: [10.1101/gr.849004](https://doi.org/10.1101/gr.849004)

Main figures and Tables with legends (in order of appearance in the text of the manuscript)

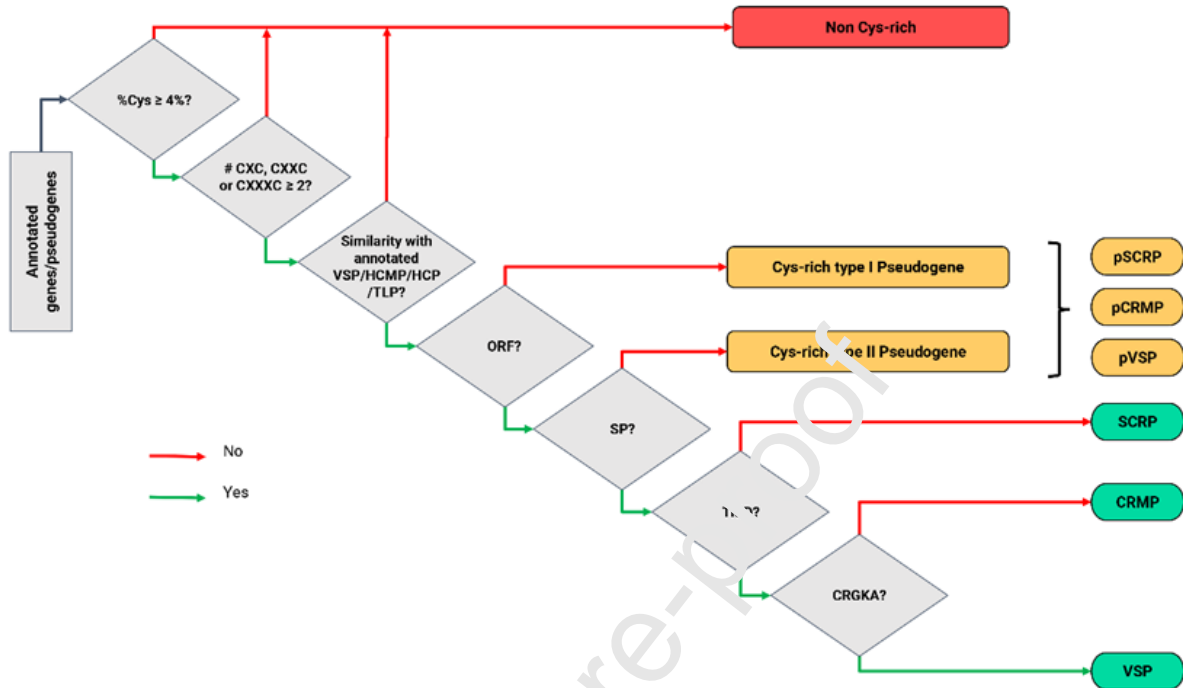
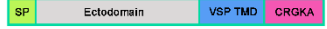
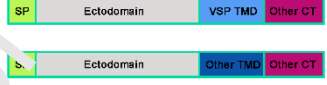


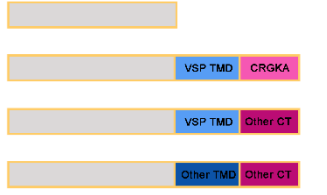


Fig. 1. Decision tree used for classification of Cys-rich genes and pseudogenes. Initially, the nucleotide and amino acid sequences of genes and pseudogenes deposited in *G. lamblia* RefSeq database were collected. Three requirements were considered for classifying a gene/pseudogene as Cys-rich: (1) $\geq 4\%$ of cysteine content, (2) at least 2 of any of the CXC, CXXC and/or CXXXC motifs in one of the 3 translated frames, and (3) sequence similarity to previously annotated VSPs, HCMPs, HCPs or TLPs. Next, a search for open reading frames (ORFs) was performed and sequences were grouped as described in the text.

Table 1. Groups of Cys-rich genes and pseudogenes defined in this work.

Group (total number)	Previously annotated as	Number	Scheme
VSP (136)	VSP	104	
	VSP with INR	23	
	VSP AS8	3	
	VSP AS12	1	
	VSP S8	1	
	Variant-specific surface protein	1	
	VSP4A1	1	
	pVSP	1	
	Uncharacterized protein	1	
	High cysteine membrane protein	1	
CRMP (94)	CXC-rich protein	1	
	High cysteine membrane protein	21	
	High cysteine membrane protein EGF-like	9	
	High cysteine membrane protein EGF-like	13	
	High cysteine membrane protein Group 1	10	
	High cysteine membrane protein Group 1	6	
	High cysteine membrane protein Group 2	5	
	High cysteine membrane protein Group 2	5	
	High cysteine membrane protein Group 3	5	
	High cysteine membrane protein Group 3	5	
	High cysteine membrane protein Group 4	11	
	High cysteine membrane protein Group 4	1	
	High cysteine membrane protein Group 5	3	
	High cysteine membrane protein Group 5	1	
	High cysteine membrane protein Group 6	1	
	High cysteine membrane protein TMK-like	1	
	High cysteine membrane protein VSP-like	1	
High cysteine non-variant cyst protein	1		
Hypothetical protein	1		
Tenascin-like protein	1		
SCRIP (16)	EGF-like domain-containing protein	1	
	High cysteine membrane protein	1	
	High cysteine protein	3	
	Hypothetical protein	1	
	Neurogenic locus Notch protein	1	
	Neurogenic locus notch-like protein	1	
	Tenascin-like protein	8	
Cys-rich type I pseudogenes (190)	pseudo High cysteine membrane protein	5	
	pseudo High cysteine membrane protein	1	
	EGF-like	1	
	pseudo High cysteine membrane protein Group 4	1	
	pseudo High cysteine protein pVSP	182	
Cys-rich type II pseudogenes (54)	High cysteine membrane protein	9	
	High cysteine membrane protein EGF-like	1	
	High cysteine membrane protein EGF-like	1	
	High cysteine membrane protein Group 1	2	
	High cysteine membrane protein Group 1	1	
	High cysteine membrane protein Group 2	1	
High cysteine membrane protein Group 2	7		
High cysteine membrane protein Group 2	2		

	4	1	
	pseudo High cysteine membrane protein	25	
	VSP-like	2	
	High cysteine protein	2	
	Hypothetical protein		
	Neurogenic locus notch-like protein		
	pVSP		
	Tenascin-like protein		
	Uncharacterized protein		

Each group has its own new name, the total number of sequences and their previous annotation. Diagrams correspond to the detected features. Diagrams with green and orange border represent genes and putative pseudogenes, respectively.

Journal Pre-proof

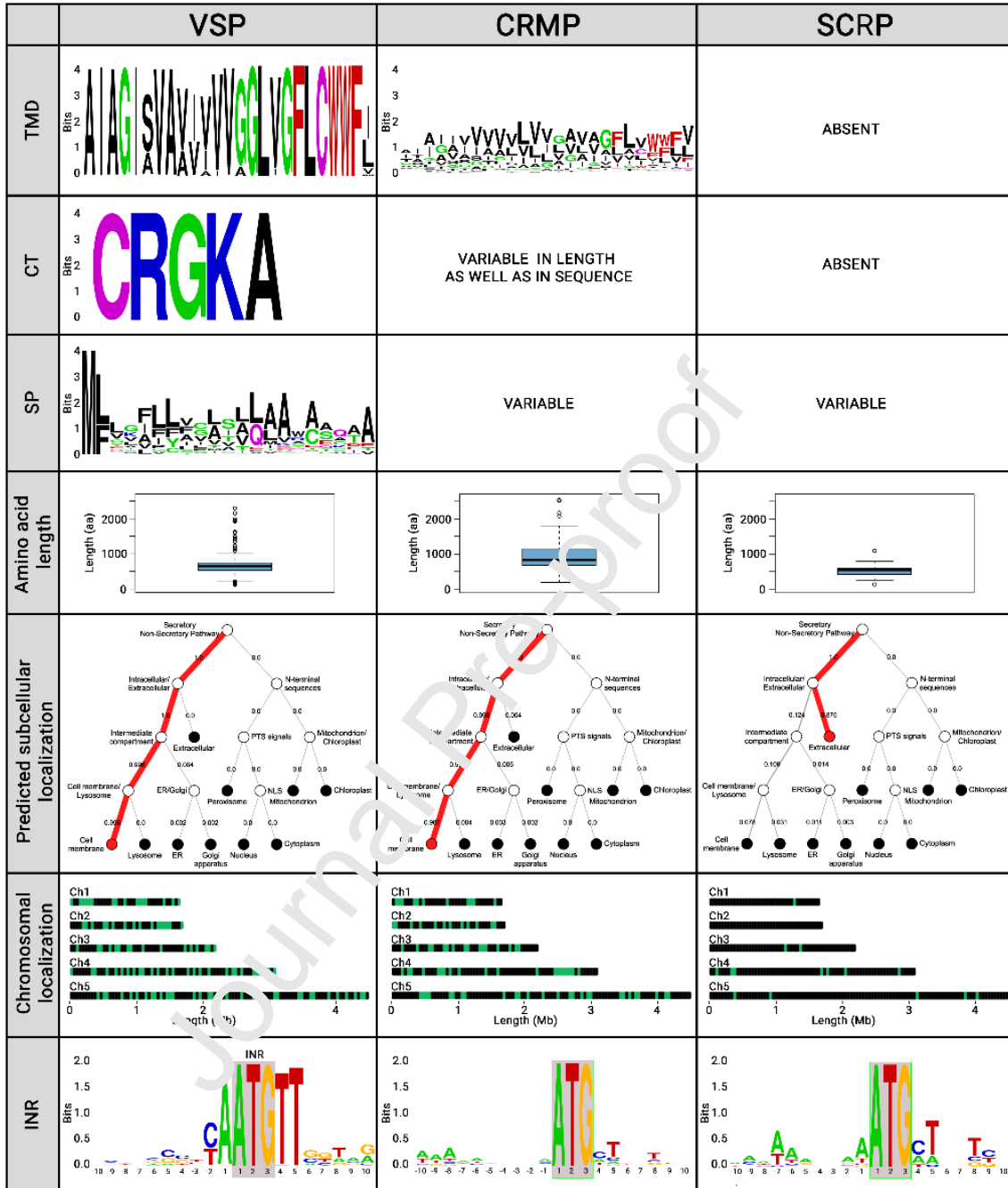


Fig. 2. Comparison among VSP, CRMP and SCRIP features. Rows (from top to bottom) indicate: Logo of the predicted transmembrane domain; cytoplasmic tail; signal peptide; boxplot of amino acid length; representative diagram outputted by DeepLoc-1.0 subcellular localization predictor; chromosomal distribution of Cys-rich genes (green bars indicate presence of one or more genes in a certain locus); and logo of -10 to +10 nt with respect to the first base of the translational start codon (ATG, boxed).

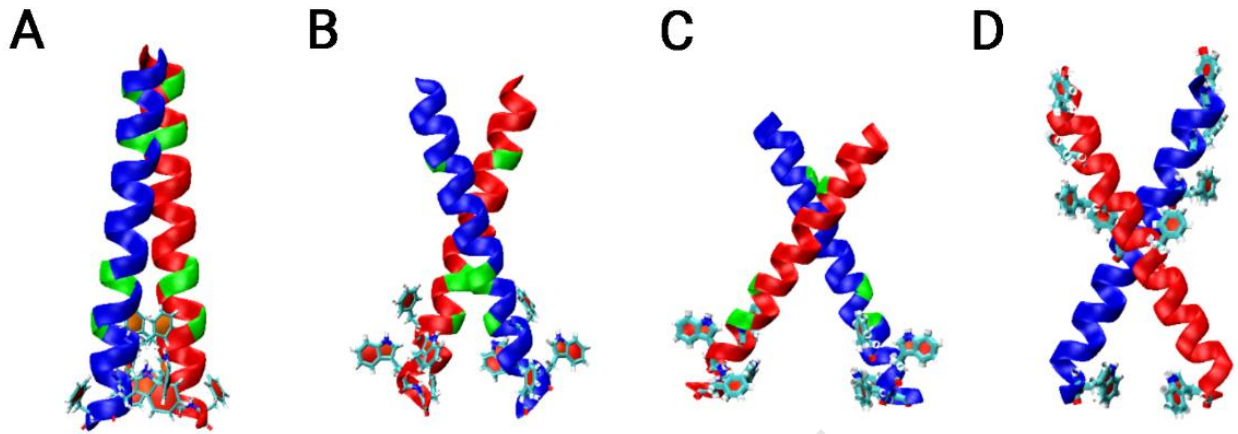


Fig. 3. Oligomerization capability of VSP and CRMP TMDs. Predicted oligomers by PREDDIMER. Glycine residues are shown in green. Aromatic amino acid residues are shown as ball and sticks. (A) VSP GL50803_00112757 (B) CRMP (VSP-like) GL50803_0010659 (C) CRMP GL50803_0014225 (D) CRMP GL50803_0094510. The characteristic feature of close angle crossing typical of oligomerized TMD structures is only found on the VSP TMD (A).

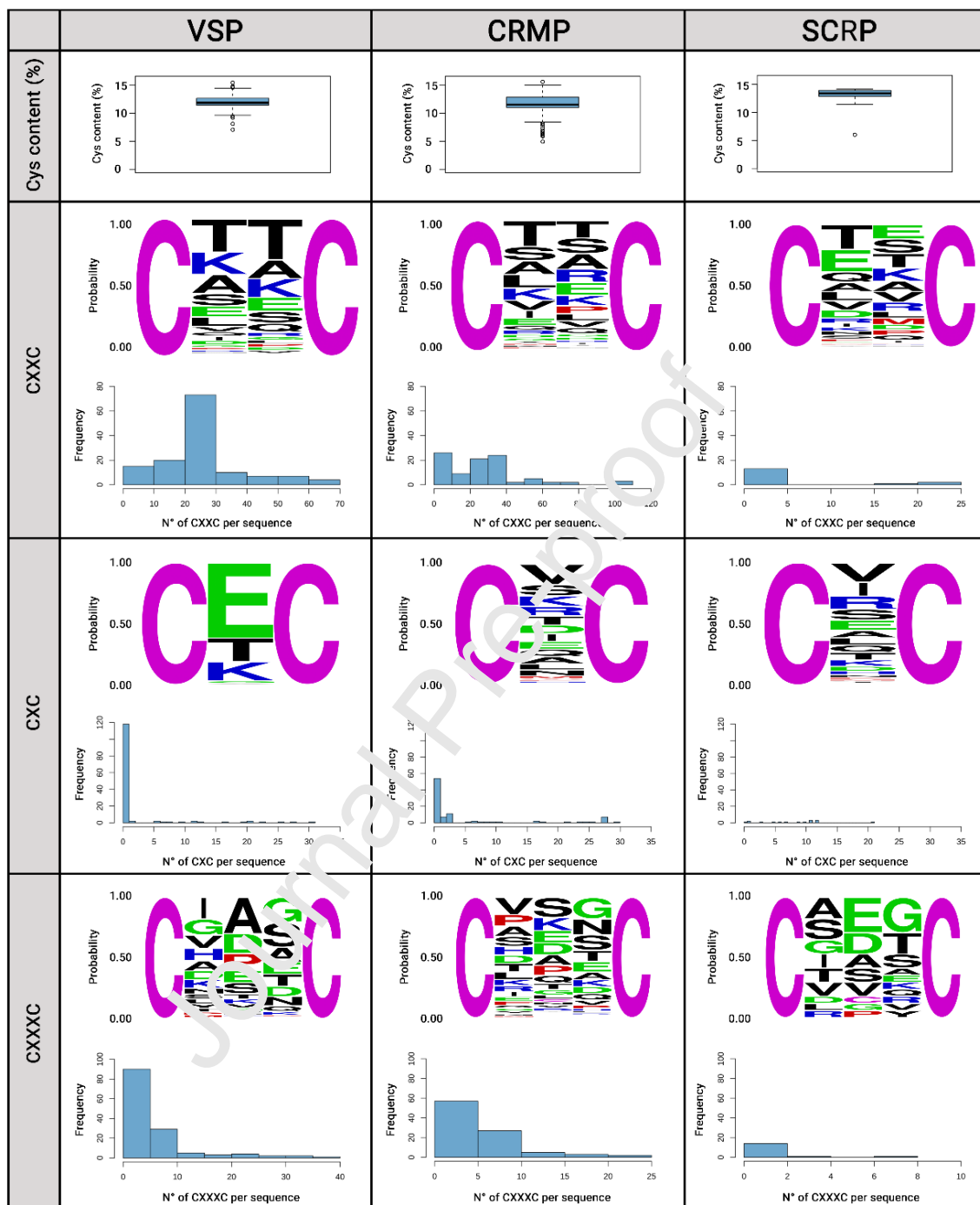


Fig. 4. Cysteine content, CXC, CXXC and CXXXC motifs found in Cys-rich proteins. The first row shows cysteine content as the percentage of total amino acids in each group of Cys-rich proteins. The second row shows a logo of CXXC motifs indicating the identity of the second and third amino acids. It also shows a histogram of the number of CXXC motifs per sequence. The third and fourth rows show the same as row two, but regarding CXC and CXXXC motifs, respectively.

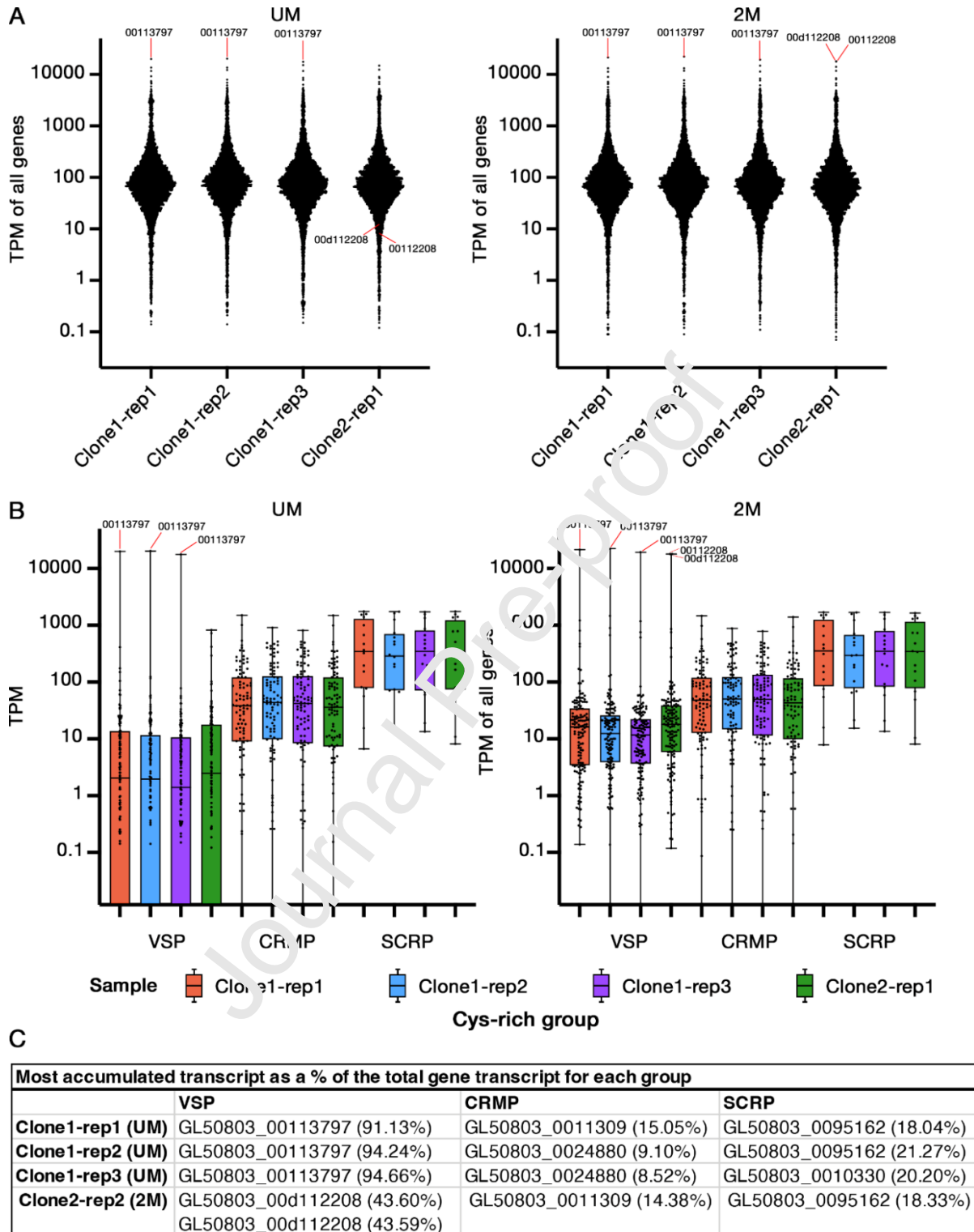


Fig. 5. Analysis of our own RNA-seq experiment. (A) Gene expression/transcript accumulation in TPM of all genes and pseudogenes. **(B)** Gene expression/transcript accumulation in TPM of genes corresponding to each Cys-rich family: CRMP, SCRIP and VSP. **(C)** Most accumulated transcript as a percentage (%) of the total gene transcript for each group, each group.

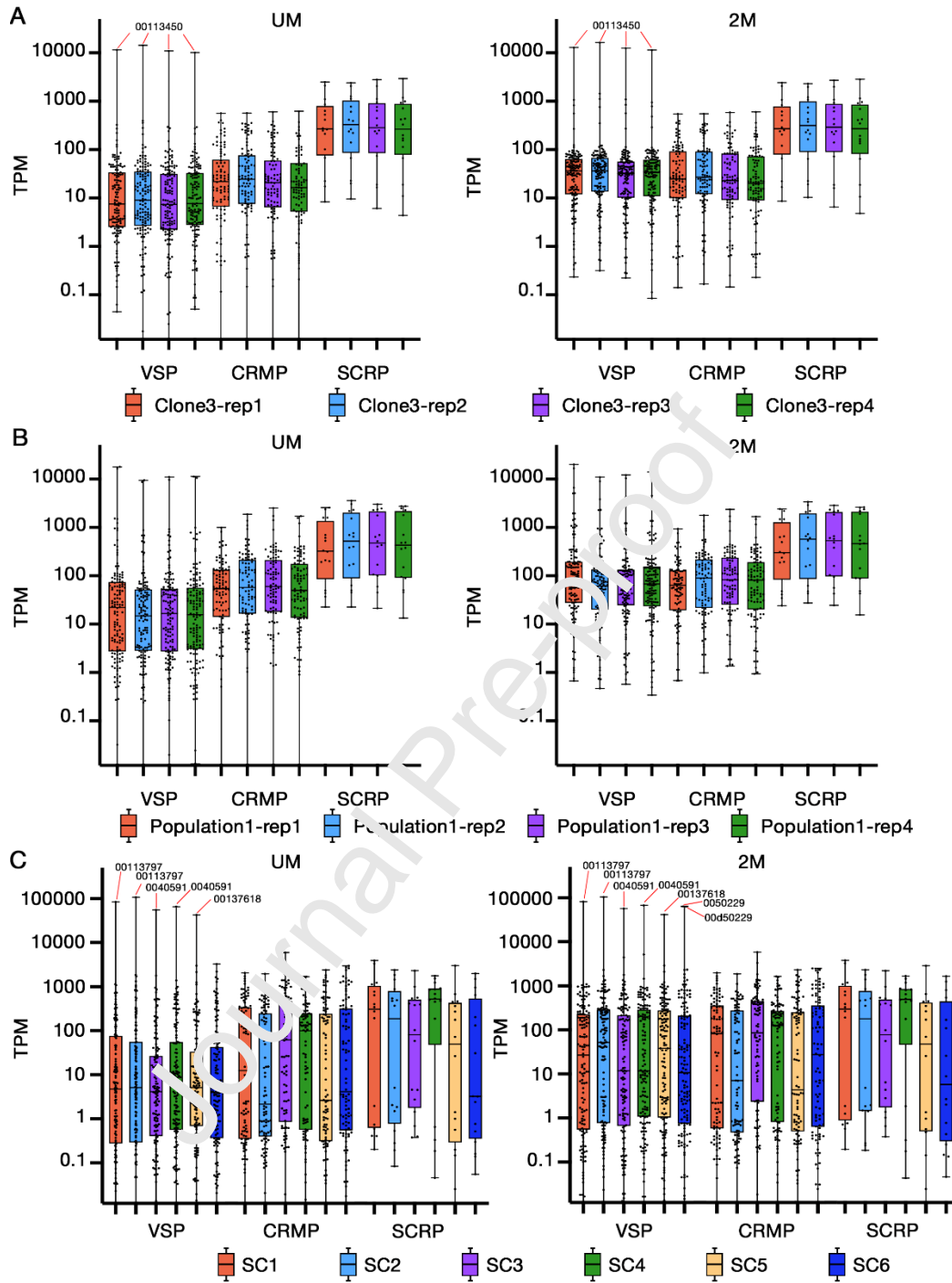


Fig. 6. Analysis of RNA-seq experiments from other authors. Gene expression/transcript accumulation of the Cys-rich genes in (A) the clone (clone3) used by Peirasmaki *et al.* [12]; (B) the non-clonal population (population1) used by Peirasmaki *et al.* [12]; and (C) the single cell (SC) experiment performed by Onsbring *et al.* [57]. Note that the characteristic of a single transcript accumulated radically above others is only observed in the VSP family (labeled with the corresponding gene ID).

Author contributions

MRW and CRM performed most of the sequences analyzes. MRW contributed the transcriptomic analysis. LAL contributed the proteomic analyzes and the immunofluorescence assays. MRW, CRM and LAL verified each other results. AS generated monoclonal antibodies used for cloning. EAF supervised the bioinformatic analyzes. SGS and JJH contributed with PAS analysis. HDL conceived the project and designed the experiments. All authors analyzed the data. MRW, CRM, EAF, SGS and HDL wrote the paper. All authors read and commented on the manuscript.

Conflict of Interest

Authors declare no conflict of interest.

Journal Pre-proof

Comprehensive characterization of Cysteine-rich protein-coding genes of *Giardia lamblia* and their role during antigenic variation

Highlights:

- Three different families of Cysteine-rich proteins are encoded in the *Giardia* genome.
- Variant-specific Surface Proteins undergo mutually exclusive changes in expression.
- Cysteine-rich Membrane Proteins may protect the parasite during antigenic variation and host-parasite interactions.
- The VSP repertoire arose from retrotransposition, duplications and divergence.

CRMP and SCRIP genes originated from VSP genes