Bioinformatic insights to the ESAG5 and GRESAG5 gene families in kinetoplastid parasites

Amy R. Barker, Bill Wickstead, Eva Gluenz, Keith Gull*

Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

A R T I C L E   I N F O

Article history:
Received 13 May 2008
Accepted 1 August 2008
Available online 19 August 2008

Keywords:
Trypanosoma brucei
Expression site
VSG
ESAG5
Sequence analysis

A B S T R A C T

Trypanosoma brucei, the causative agent of African sleeping sickness, evades the immune response by expressing a coat of variant surface glycoprotein (VSG). VSG is expressed from a single telomeric expression site (ES), along with a number of expression site associated genes (ESAGs). Thus far, the function of most ESAGs is unknown. One ES contains the serum resistance associated gene (SRA), which confers resistance to trypanosome lytic factor in T. b. rhodesiense. Only three other ESAGs – 5, 6 and 7 – are present in this ES. ESAGs 6 and 7 encode a heterodimeric transferrin receptor, but the function of ESAG5 has not been identified. We present here a bioinformatic analysis of ESAG5 and distinguish between T. brucei-specific ESAGs and Genes Related to ESAG5 (GRESAG5s), which occur outside of ESs in chromosomal-internal contexts. Further, a genome-wide survey of these genes across kinetoplastids identifies a family of GRESAG5s in a number of species. Analysis of phylogenetic relationships indicates that this family may have evolved from a single ancestral copy. Predicted properties of (GR)ESAG5 proteins indicate a glycosylated protein containing either a signal peptide or transmembrane domain. Further analysis indicates a possible relationship to the lipid transfer/lipopolysaccharide-binding family which includes the bactericidal/permeability increasing (BPI) protein. Together, these results provide insights into the structure and evolution of an important extended gene family, and present a number of testable hypotheses which will aid in elucidating the function of ESAG5.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Trypanosoma brucei is the causative agent of nagana in cattle and, for T. b. rhodesiense and T. b. gambiense, human trypanosomiasis. The parasite is able to evade the host immune system by expressing a surface coat of a single variant surface glycoprotein (VSG [1]), which is switched at a rate of between 10−2 and 10−7 cell−1 generation−1 with culture-adapted strains exhibiting the lower switch rates [2]. Cells which have switched VSG escape the adaptive immune response directed against the previous coat and produce a new wave of infection. Since T. brucei is estimated to have a repertoire of over 1000 VSG genes [3], infection can be sustained over very long periods.

VSG is only expressed from telomeric expression sites (ESs). There are around 20 ESs, but only one is fully active at any one time. The VSG coat is changed either by transcriptional switching from the active expression site to a previously silent one, or by replacing the VSG in the active site with another from elsewhere in the genome [4–7]. Each ES also contains a number of expression site-associated genes (ESAGs). There are currently 12 known different ESAGs (including SRA); the function of most remains elusive [8,9]. The best characterised are ESAGs 6 and 7 which encode a heterodimeric transferrin receptor [10–12]. ESs are transcribed only in bloodstream form T. brucei; they remain silent in procyclic forms [13]. Other kinetoplastids do not use this mechanism of antigenic variation and therefore do not have ESs in their genomes.

Recent work has identified that slight variations in the sequences of ESAGs 6 and 7 from different ESs affects the affinity of the receptor for the transferrin of different species in the host range. Growth of trypanosomes in the presence of transferrin from another species can cause the parasite to switch ES [14]. It has therefore been hypothesised that ESAGs allow the parasite to adapt to different organisms in its wide host range [15,16]. These variations have also been theorised to allow efficient uptake of transferrin in the presence of host antibodies to
ESAG6/7 [14], though it now appears that under physiological conditions this may be unlikely [17,18]. It has been shown that certain ESs, even when expressing different VSGs, can cause growth rate differences in serum from different species [17], suggesting some possible role of ESs in host adaptation, though this role remains elusive.

Most ESs contain up to 11 ESAGs, but one critical ES – the SRA ES or R-ES – is truncated [19]. SRA is an ESAG found only in a single ES of T. b. rhodesiense and confers resistance to a trypanolytic factor present in human serum associated with high-density lipoprotein (HDL) particles [20,21]. Therefore this ES must be expressed during human infection in order for the parasite to survive. It contains only three ESAGs—5, 6 and 7 [19]. While the function of ESAGs 6 and 7 is known, that of ESAG5 remains to be elucidated. As these ESAGs are particularly implicated in human infection, their function may be essential to our understanding of the human disease.

The recent release of the T. brucei genome has proven an invaluable tool, and the telomere-sequencing project for the Lister 427 strain [22] has also begun to release data. This means that for the first time, a complete or near-complete repertoire of ESAG sequences is available. We have used this background to analyse ESAG5 sequences using a detailed bioinformatic approach, leading to a number of testable hypotheses which can be used to clarify function. We have classified a number of ESAG5-like genes found in both ES and chromosomal–internal loci in T. brucei which fall into two distinct families. Syntenic homologues to these chromosome–internal genes can be found in other trypanosomatids, despite the lack of ESs in these organisms. Although no ESAG5-like genes were found outside of the kinetoplastids, a number of predictions indicate that ESAG5 may be a distant relative of the lipid binding/transfer protein superfamily in mammals [23], which includes bactericidal/permeability-increasing protein (BPI) and lipopolysaccharide binding protein (LBP).

2. Materials and methods

2.1. Identification of ESAG5 and related genes

The GeneDB databases for T. brucei, T. b. gambiense, T. congoense, T. vivax, T. cruzi, L. major, L. infantum and L. braziliensis were used to identify ESAG5 genes by annotation and BLAST similarity searches. Sequence and annotation data were produced by the Trypanosoma brucei Sequencing Group at the Sanger Institute and TIGR Sequencing Centres and can be obtained from http://www.genedb.org. Potential homologues to T. brucei genes were confirmed by reciprocal BLAST and sequence alignments produced using MAFFT [24–26]. Consensus quality scoring was carried out using CLUSTALX [27].

Iterative searches were carried out using both position–specific iterated (PSI)-BLAST (NCBI [28]) and hidden Markov models (HMM). PSI-BLAST searches were carried out against the full NCBI non-redundant protein database using an e-value threshold of 0.005.

For iterative HMM searches, protein sequences annotated as ESAG5 from the T. brucei genome were aligned with MAFFFT and manually edited. The alignment was then used to generate a hidden Markov model using Hmmer [29]. This model was used to search the NCBI non-redundant protein database. The resultant protein hits with an e-value below 5 were integrated into a new alignment, and columns containing more than 50% gaps were removed to prevent species– or clade–specific insertions biasing the results. This alignment was used to generate a new model and a further round of searches was carried out. Searching was terminated when no further hits below an e-value of 5 could be identified.

2.2. Analysis of ESAG5 proteins

A number of freely available programs were used to analyse ESAG5 proteins for motifs, domains or similarities to known proteins, including InterProScan [30] which searches a range of databases including PROSITE [31] and Pfam [32]. The servers SignalP 3.0 [33], TMHMM [34] and NetNGlyc [35] were used for protein property analysis, available from CBS at http://www.cbs.dtu.dk/Phobius [36,37] was used to further analyse protein sequences.

2.3. Structural prediction of ESAG5

T. brucei ESAG5 protein sequences were submitted to both 3D-PSSM and Phyre [38] to predict structural similarities to known proteins. Analysis of the same sequences via PSI-PRED http://bioinf.cs.ucl.ac.uk/psipred/ [39,40] produced similar results. Both secondary structure prediction and fold recognition was carried out with these programs.

2.4. Homology modelling of ESAG5 proteins

Predicted 3D models of all T. brucei GRESAG5 proteins and ESAG5 proteins from the T3 and 121 ESs were generated by the SWISS-MODEL server [41] in reference to the crystal structure of human BPI [42,43]. The model was based on a MAFFT alignment of ESAG5s with a number of representative members of the BPI/LBP superfamily including BPIs, LBPs, phospholipid transfer proteins (PLTPs), cholesterol ester transfer proteins (CETPs) and BPI-like proteins from a number of organisms including mammals, fish and plants.

2.5. Phylogenetic analysis

Sequences identified as belonging to the ESAG5 family were aligned with MAFFT version 5,662 [24] adopting the E-INS-i strategy [25]. Alignments were manually edited, and regions containing more than 50% gaps, or those where low sequence conservation precluded accurate alignment, were removed. These edited alignments were used to infer a minimum evolution phylogram using PAUP*4b10 [44] (heuristic search with tree-bisection and reconnection). For nucleotide-based inference, a general time reversible model was used with gamma-distributed rate variation. Support for the inferred topology was estimated from 100 bootstrap replicate inferences with resampling using the above method and for maximum parsimony.

3. Results

3.1. Identification of ESAG5 genes in Trypanosoma brucei

Genes found in the trypanosome ESs often have non-ES paralogues at other loci, including chromosome–internal sites [45–47]. Here, we use the term ESAG5 to refer only to sequences located within an ES. All other homologues are considered Gene Related to ESAG5 (GRESAG5s), according to precedent [46].

In order to identify ESAG5 genes, we conducted a survey of the T. brucei genome sequences using as query sequences the known ESAG5 sequences of the 221, VO2 and BN-2 expression sites [48]. At the time of this study we identified 20 ESAG5 genes in GenDB, all from the 427 strain (Table 1); 3 previously published [48] plus 17 from the Sanger Institute telomere–sequencing project. As sequencing is ongoing for telomeres, the full repertoire of ESAG5 genes is unknown. Analysis of the gene sequences available indicates 13 non-identical sequences including three pseudogenes matching the sequences of the ψESAG5s from the 221 and VO2 expression
Table 1
ESAG5 genes in expression sites

<table>
<thead>
<tr>
<th>ESAG5 GeneDB ID</th>
<th>VSG in ES (clone ref)</th>
<th>VSG (MITat)</th>
<th>VSG (systematic name) a</th>
</tr>
</thead>
<tbody>
<tr>
<td>H25N7.16 b</td>
<td>221</td>
<td>MITat1.2</td>
<td>Lister427-2</td>
</tr>
<tr>
<td>Tb427.BES40.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N19B2.175 b c</td>
<td>VO2</td>
<td>MITat1.9</td>
<td>Lister427-9</td>
</tr>
<tr>
<td>Tb427.BES129.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13J1.13 b</td>
<td></td>
<td>Not present d</td>
<td></td>
</tr>
<tr>
<td>Tb427.BES4.3</td>
<td>121</td>
<td>MITat1.6</td>
<td>Lister427-6</td>
</tr>
<tr>
<td>Tb427.BES15.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tb427.BES5.3</td>
<td>17.13</td>
<td>MITat1.13</td>
<td>Lister427-13</td>
</tr>
<tr>
<td>Tb427.BES28.5</td>
<td>17.7</td>
<td>MITat1.17</td>
<td>Lister427-17</td>
</tr>
<tr>
<td>Tb427.BES56.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tb427.BES65.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tb427.BES98.3</td>
<td>17.22</td>
<td>MITat1.18</td>
<td>Lister427-18</td>
</tr>
<tr>
<td>Tb427.BES122.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tb427.BES126.4</td>
<td>bR2</td>
<td>MITat1.11</td>
<td>Lister427-11</td>
</tr>
<tr>
<td>Tb427.BES134.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tb427.BES153.4</td>
<td>224</td>
<td>MITat1.3</td>
<td>Lister427-4</td>
</tr>
</tbody>
</table>

A list of all known ESAG5 sequences from expression sites, including the relevant VSG information where known. Shading indicates a truncated protein. Where genes are grouped, both sequence identity (across the whole ES) and VSG present in the expression site sequence suggest these may be duplicate database entries (duplicate clones from the same expression site). Where the VSG is listed as ‘uncharacterised’, a VSG was present in the sequence but was not 100% identical to any sequence in GenBank. Sequences where the VSG is marked ‘not present’ did not contain a VSG as part of the released sequence. All recently released expression site sequences [22] contain full-length ESAG5 genes, with the only predicted pseudogenes being the previously sequenced 221 and VO2 ESAG5 genes and the recently released Tb427.BES122.3. Tb427.BES134 is unusual as it appears to contain only ESAGs 7, 6, 5 and an ESAG3 pseudogene. *As there is as yet no unified VSG nomenclature, both the clone and MITat reference have been listed alongside the proposed systematic name as suggested by G. Cross (full list can be found at http://tryps.rockefeller.edu/Lister427vsgsummary.html).** These three sequences are those originally sequenced [48]. This ES contains ESAGs 7, 6, 5 and 8 (no VSG) all of which are identical to the VO2 expression site. The predicted ESAG5 ORF appears to have been misannotated. **This ES sequence contains ESAGs 7, 6, 5, 3, 4, 8, 3, 2, 11 and 1 all identical to the VO2 expression site. The predicted ESAG5 ORF appears to have been misannotated.**

A tree of ESAG5 sequences was constructed to assess the relationships between genes, and in particular to analyse the location of the ESAG5 sequence from the T. b. rhodesiense R-ES which contains the SRA gene (Fig. 1). ESAG5 genes from the 427 strain fall roughly into two groups with Tb427.BES4.3 (ES 121) and the R-ES ESAG5 sequences not falling directly into either group. In the case of the R-ES sequence, it should be noted that this is the only gene not sequenced from the 427 strain and as such represents a very different isolate with a different culture history which may have been under very different selective pressure. The 121 ES containing Tb427.BES4.3 is known to be capable of expression in a genetically engineered laboratory strain [49]. The VO2 and 221 site ESAG5 pseudogenes (ESs both shown to be expressed in the List427 strain) occur within the same group, and are very closely related.

At present there is no available sequence data on ESs from other African trypanosomes. Other kinetoplastid parasites such as Leishmania do not contain ES structures.

3.2. Identification of GRESAG5 genes in kinetoplastids

A number of ESAGs including ESAGs 2 and 4 have been shown to have homologues outside of ESs [46,47]. Our survey of the trypanosomatid genome sequences available from GeneDB also identified ESAG5-like sequences outside of ESs, the results of which are shown in Table 2. These GRESAG5s were found in all kineto-
plastid parasites for which data is available, although numbers of genes vary between organisms. *GRESAG5s* can be grouped according to syntenic chromosomal location. Groups are indicated in rows in Table 2. The genome of *T. brucei* contains seven *GRESAG5s* in chromosomal-internal positions, on chromosomes 2, 4, 5, 7 and 9. The genes on chromosomes 2 and 4 are situated within directional gene clusters syntenic with other kinetoplastid parasites. Those on chromosomes 5 and 9 are located in what appear to be subtelomeric regions, whereas *GRESAG5* on chromosome 7 occurs at a strand-switch region.

Table 3 shows the identity and similarity of a number of (GR)ESAG5 proteins, demonstrating the low identity of GRESAG5s (15–40%) compared with the higher identity of ESAG5s from expression sites. It should be noted that despite low identity, similarity levels are much higher (40–63%). To clarify relationships between GRESAG5 proteins within and between species, a phylogram containing both ESAG5 and GRESAG5 sequences was constructed (Fig. 2). This phylogram demonstrates that orthologous proteins are also syntenic.

*Leishmania* ssp. and *T. cruzi* have only a single *GRESAG5*, syntenic to the *GRESAG5* found on *T. brucei* chromosome 4. Only this gene is conserved in all kinetoplastids (though in *T. congoense* it appears to be a pseudogene) and as such appears to be the ancestral gene. All of the African trypanosomes exhibit additional *GRESAG5* gene expansions, which appear to be independent in each organism (for instance the chromosome 9 *GRESAG5* in *T. brucei*, the chromosome 4 *GRESAG5* in *T. vivax*). In the case of *T. congoense*, expansion appears to have involved a gene at a position not found in *T. brucei*. The analysis presented here used contiguous sequence assemblies from the current genome release. It should be noted that due to the preliminary nature of this data, more GRESAG5 genes may be identified in these species in the future.

In *T. brucei*, the ES sequences share a common ancestor with the sequence found on chromosome 7 (Tb927.7.6860) which is the most closely related chromosome-internal copy. Interestingly, a non-syntenic gene located on *T. b. gambiense* chromosome 11 (Tbgamb.38351) appears more closely related in sequence to the *T. b. brucei* chromosome 7 gene, while the syntenic copy found on *T. b. gambiense* chromosome 7 (Tbgamb.16066) is more closely related to the ES genes. This implies some level of exchange between ESAGs and GRESAGs.
Table 2
Non-telomeric GRESAG5 genes and their conservation in trypanosomatids

<table>
<thead>
<tr>
<th>T. brucei Chromosome</th>
<th>T. b. gambiense</th>
<th>T. congolense</th>
<th>T. vivax</th>
<th>T. cruzi</th>
<th>L. major</th>
<th>L. infantum</th>
<th>L. braziliensis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Tb927.2.1920</td>
<td>Tb927.1102</td>
<td>Tb927.5.284</td>
<td>Tb927.5.284</td>
<td>Tb927.5.284</td>
<td>Tb927.5.284</td>
<td>Tb927.5.284</td>
</tr>
<tr>
<td>4</td>
<td>Tb927.4.810</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
</tr>
<tr>
<td>5</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
<td>Tb927.5.340</td>
</tr>
<tr>
<td>7</td>
<td>Tb927.7.6860</td>
<td>Tb927.6.666</td>
<td>Tb927.6.666</td>
<td>Tb927.6.666</td>
<td>Tb927.6.666</td>
<td>Tb927.6.666</td>
<td>Tb927.6.666</td>
</tr>
<tr>
<td>9</td>
<td>Tb927.4.2210</td>
<td>Tb927.4.2210</td>
<td>Tb927.4.2210</td>
<td>Tb927.4.2210</td>
<td>Tb927.4.2210</td>
<td>Tb927.4.2210</td>
<td>Tb927.4.2210</td>
</tr>
</tbody>
</table>

GRESAG5-like sequences (named GRESAG5, Gene Related to ESAG5) were identified by BLAST. No BLAST hits were found outside of trypanosomatids. BLAST hits were manually checked and sorted into groups according to chromosomal location and synteny, and these groups are presented in rows. Partial sequences are indicated in italics. In some cases there are multiple sequences which appear to be the same gene; these are indicated in brackets. Genes below the bold line do not show any synteny with T. brucei GRESAG5s. Grey shading indicates a pseudogene. The sequences from the preliminary genomes of T. congolense and T. vivax has been taken from contig data. A number of additional reads identified by BLAST have not been included due to insufficient length, but may form additional ESAG5 genes. The two T. vivax tvv1015 genes are adjacent, and may represent a gene duplication event. The T. congolense congo1403h1.1,1.q1k 3 and congo1549b02.p1k 4 encode the N- and C-terminal portions of an ESAG5 pseudogene interrupted by multiple frameshifts. They have been included for the sake of completeness.

We analysed the GRESAG5 repertoire of the Lister 427 strain, for which genomic data from chromosome-internal regions is not currently available, by PCR and sequencing. We were able to specifically amplify from 427 DNA fragments from all GRESAG5s except the chromosome 7 copy (data not shown). However, this gene could be amplified from 927 genomic DNA with the same primer sets. In addition, as only one of the chromosome 9 GRESAG5s (Tb09.244.2120) has been integrated into the current genome assembly while the others remain as additional unordered contigs, we confirmed the presence of three distinct DNA sequences similar to the chromosome 9 GRESAG5s in the 427 strain. These are likely to form part of a subtelomeric array.

3.3. Predicting properties of ESAG5/GRESAG5 proteins

After analyses of the (GR)ESAG5 gene family found in kinetoplastids, we asked whether any related genes could be found in other organisms. Simple BLAST similarity searches failed to find likely homologues in non-kinetoplastids. However searches for more distantly related sequences using both PSI-BLAST and iterative hidden Markov models identified (GR)ESAG5 as being possible members of the human lipid binding/lipid transfer superfamily [50]. This family includes the BPI, LBP, PLTP and CETP among many others. So far, BPI superfamily members have been described only in metazoans and higher plants. Confirming the classification, the reciprocal PSI-BLAST analysis starting from the human BPI proteins also identifies (GR)ESAG5 as superfamily members. This classification is specific, as only BPI superfamily members and (GR)ESAG5s were identified. Moreover, the BPI superfamily genome assignments on the Superfamily database (SF55394) includes GRESAG5 proteins from T. brucei, T. cruzi and Leishmania. The (GR)ESAG5 proteins and the BPI family members are of similar sizes (~480 aa) and, while no highly conserved regions can be identified, the proteins can be aligned without significant stretches of insertions or deletions.

Further analysis identified a stretch of hydrophobic amino acids at the N-terminus of all (GR)ESAG5s predicted to be either a signal peptide (SignalP) or transmembrane domain (TMHMM), though strength of prediction for each varies between proteins with some appearing more likely to have a signal peptide than transmembrane domain (and vice versa). Use of the Phobius server, which is reported to have higher accuracy at discerning between the two [37], indicates a signal peptide is more likely. There are at least three very strongly predicted N-glycosylation sites in every ESAG5 protein (NetNGlyc), though only one (found at ~70 aa) is conserved at the same position in all ESAG5 sequences.

Table 3
Identity and similarity of (GR)ESAG5 proteins

<table>
<thead>
<tr>
<th>T. brucei</th>
<th>Tb427.BESS5.3</th>
<th>Tb927.2.1920</th>
<th>Tb927.4.810</th>
<th>Tb927.5.340</th>
<th>Tb927.7.6860</th>
<th>Tb09.244.2120</th>
<th>Tc00.1047053508257.220</th>
</tr>
</thead>
<tbody>
<tr>
<td>13J3.12</td>
<td>83 (88)</td>
<td>24 (48)</td>
<td>25 (49)</td>
<td>25 (49)</td>
<td>28 (51)</td>
<td>22 (46)</td>
<td>26 (52)</td>
</tr>
<tr>
<td>Tb427.BESS5.3</td>
<td>83 (88)</td>
<td>24 (48)</td>
<td>25 (49)</td>
<td>25 (49)</td>
<td>28 (51)</td>
<td>22 (46)</td>
<td>26 (52)</td>
</tr>
<tr>
<td>Tb927.2.1920</td>
<td>20 (43)</td>
<td>19 (42)</td>
<td>25 (49)</td>
<td>25 (49)</td>
<td>28 (51)</td>
<td>22 (46)</td>
<td>26 (52)</td>
</tr>
<tr>
<td>Tb927.4.810</td>
<td>23 (50)</td>
<td>23 (49)</td>
<td>39 (61)</td>
<td>26 (51)</td>
<td>26 (52)</td>
<td>26 (52)</td>
<td>26 (52)</td>
</tr>
<tr>
<td>Tb927.5.340</td>
<td>70 (80)</td>
<td>69 (79)</td>
<td>28 (51)</td>
<td>22 (46)</td>
<td>22 (46)</td>
<td>22 (45)</td>
<td>18 (41)</td>
</tr>
<tr>
<td>Tb927.7.6860</td>
<td>26 (52)</td>
<td>27 (52)</td>
<td>40 (62)</td>
<td>26 (51)</td>
<td>40 (63)</td>
<td>29 (56)</td>
<td>29 (56)</td>
</tr>
<tr>
<td>Tb09.244.2120</td>
<td>18 (42)</td>
<td>18 (43)</td>
<td>24 (46)</td>
<td>21 (46)</td>
<td>22 (45)</td>
<td>18 (41)</td>
<td>24 (48)</td>
</tr>
<tr>
<td>Tc00.1047053508257.220</td>
<td>16 (40)</td>
<td>15 (40)</td>
<td>18 (43)</td>
<td>19 (48)</td>
<td>19 (45)</td>
<td>15 (40)</td>
<td>18 (47)</td>
</tr>
<tr>
<td>LmjF34.3930</td>
<td>18 (43)</td>
<td>18 (43)</td>
<td>18 (43)</td>
<td>18 (43)</td>
<td>18 (43)</td>
<td>18 (43)</td>
<td>18 (43)</td>
</tr>
</tbody>
</table>

GRESAG5 proteins from T. brucei were aligned with representative ESAG5 proteins (Tb427.BESS4.3 and 13J3.12) as well as the more diverse GRESAG5s from T. cruzi (Tc00.1047053508257.220) and L. major (LmjF34.3930). This alignment was then used to calculate identity and similarity scores between proteins. Similarity scores are indicated in brackets.
Fig. 2. Unrooted minimum evolution tree of ESAG5/GRESAG5 protein sequences in trypanosomatids. A minimum evolution tree was constructed in PAUP* from an alignment made in MAFFT and edited with Seaview. All identified GRESAG5 sequences were included (listed in Table 2), and the 10 non-redundant ESAG5 proteins. Groups have been highlighted based on both tree branching relationships and synteny of chromosomal location. Topology support from 100 bootstrap replicates is indicated as ME/MP where ME is minimum evolution and MP maximum parsimony. Where '-' is shown, the grouping is not supported in that method. All the African trypanosomatids (T. brucei, T. congolense and T. vivax) appear to have multiple GRESAG5s. Each of these organisms appears to have an expansion of a single gene group, though the group differs between organisms. Only a single group is represented in all trypanosomatids, suggesting that this group (corresponding to Tb927.4.810) may have been the ancestral gene.

analysed. No single glycosylation site is conserved across the GRESAG5s.

3.4. Structural properties of ESAG5/GRESAG5 proteins

The BPI superfamily is characterised by low sequence identity, but a conserved 3D structure [50]. For this reason, we used a variety of programs to predict secondary structures of (GR)ESAG5 proteins from T. brucei. The consensus between predictions was good. Interestingly, this secondary structure is highly similar to BPI (Fig. 3). The predictions have been aligned pairwise beginning at amino acid 1. Both show a similar pattern of α-helices, β-strands and coil regions, indicating that they may form a similar 3D structure.

3.5. Structural modelling and amino acid conservation

The ES ESAG5 proteins share a high level of sequence identity, with variation occurring throughout the protein. The CLUSTALX consensus quality score at each residue position of an alignment of non-redundant ESAGs was used as a measure of variation (Fig. 4). Sequence variations are not clustered in the primary sequence, though the C-terminal half of the protein exhibits more variability than the N-terminal half. Where variations are found, the quality
Fig. 3. Alignment of 2D structural prediction for ESAG5 and human BPI. The protein sequences of ESAG5 (1J3J.12) and the human BPI protein (ID 1BP1_A; the protein sequence from the crystal structure) were entered into the 2D structural prediction protein 3D-PSSM. A schematic representation of the results is shown to scale for both proteins matched pairwise by amino acid (with each starting at 1 aa). The domains predicted by TMHMM/SignalP and InterProScan are indicated above the ESAG5 protein. TM/SP is the transmembrane domain or signal peptide spanning approximately amino acids 7–27, as predicted by TMHMM (SignalP prediction = 1–26 aa). The two BPI domains predicted by InterProScan span amino acids 40–256 and 257–469 (though actually amino acid numbers vary between ESAG5 proteins). The box labelled SP is the 31aa BPI signal peptide, which is cleaved.

score is often high, indicating point mutations found in a single sequence only or that the substituted residue has similar biochemical properties.

Human BPI family members are secreted proteins able to bind lipid molecules. If ESAG5 is a distant relative, it may have similar properties. Therefore the location of sequence variations is potentially important; in a surface-exposed or secreted protein, variation would be expected in domains exposed to antibodies, whereas a family of lipid-binding molecules might show variation around binding sites to accommodate different substrate specificities.

As the high level of identity between ESAG5 proteins indicates they probably all form the same structure, and assuming the possible relationship with BPI to be true (while recognising the caveats with such approaches), we used the 3D structure of human BPI as a template on which to model a possible ESAG5 structure using SWISS-MODEL. To analyse possible effects of ESAG5 variation on this 3D model, positions where the CLUSTALX consensus quality score is 60% or lower were mapped on to this 3D model. Fig. 5 shows that the more variable positions appear throughout the molecule, but there is a bias toward clustering around the putative lipid binding pockets and the tips of the boomerang-shaped molecule—the sites which in BPI would bind phospholipids or interact with other molecules such as LPS.

Bringing together these results, it is clear that a wider family of related (GR)ESAG5 genes are present in a range of kinetoplastids. This family probably arose from a single gene in a common ancestor, and this gene remains conserved in every kinetoplastid organism for which sequence data is available. The proteins encoded by these genes have low sequence identity (15–40%) but higher similarity (40–65%) and are likely to form glycosylated proteins which are either secreted or membrane-bound. Sensitive iterative searches and structural predictions indicate a similarity to the lipid transfer/lipopolysaccharride-binding family (LT/LBP).

4. Discussion

Trypanosoma brucei ESAGs are genes co-expressed with VSG from the single active telomeric expression site. Because they are expressed in the host bloodstream, they may be relevant in infection and may be potential drug targets. One particular ES (that containing SRA) is implicated in all cases of East African human trypanosomiasis. Only three ESAGs are co-expressed with SRA [19], with ESAG5 the only one of unknown function.

With the recent release of genomic information from both the 927 and 427 strains of T. b. brucei, much more information about kinetoplastid genes is available. We have used this data for an in-depth bioinformatic analysis to assess diversity of ESAGS and ESAG5-like sequences in T. brucei and look for an indication of function; in addition, we aimed to consolidate the information available from the two strains and compare it with genome sequences available for other organisms. This has revealed that ESAG5 genes are part of a larger family within kinetoplastids, where we define an additional sub-family of GRESAG5 genes not previously described. Moreover, our analysis indicates that (GR)ESAG5s may belong to a much larger superfamily, that containing the bactericidal/permeability-increasing protein (BPI).
Fig. 5. Sites of common mutations in an ESAG5 model. An alignment of ESAG5 and BPI protein sequences was input to SWISS-MODEL and a homology model for ESAG5 (Tb427.BES5.3) generated by threading onto the BPI crystal structure. Sites of variation, as indicated by CLUSTALX q-score, were then mapped on to this 3D model (please refer to Fig. 4). (a) The side and top view of the ribbon diagram of ESAG5. A diagram of BPI is shown in (b) for comparison, from the same points of view. In both models, the two phospholipids bound by the BPI protein are shown in teal to indicate binding sites. Residues from the ESAG5 alignment which have a CLUSTALX q-score lower than 60 are indicated in orange. Those with a q-score lower than 40 are shown in red. The model indicates that while variations occur throughout the protein, a majority occur either on the tips of the molecule or around the binding pocket regions, where they may affect binding affinity or interactions.

We define ESAG5 sequences as those located in ESs only. These genes exhibit high levels of sequence identity and fall into at least two groups, though the full catalogue of genes has not yet been fully sequenced. The ESAG5 found in the R-ES of T. b. rhodesiense does not fall into either group; whether this is due to its implied role in human infection or the different history of the sequenced strain is unknown, but will be clarified as further genomic information emerges. Such data may reveal the R-ES ESAG5 sequence of T. b. rhodesiense to have been influenced by its expression solely in humans. The ESAG5 genes located in the 221 and VO2 expression
sites are very closely related, and both feature mutations which produce truncated proteins; however, these mutations are different in each protein and as such the two appear to have undergone separate pseudogenisation events. Since these two ESs are known to be expressed in culture and both contain ESAG5s predicted to be pseudogenes, this may indicate that ESAG5 is not required for parasite survival under culture conditions (or that it may even be disadvantageous). However, since low-level transcription of inactive expression sites is known to contribute to the pool of ESAG6 and 7 mRNA, it may be that similar transcription of ESAG5 from alternative loci is able to rescue pseudogenised ESAG5 in the active ES [51].

Further, we have identified a family of ESAG5-like genes located outside of ESs in T. brucei. In concordance with precedent [46], we refer to these as GRESAG5. One member of this family, orthologous to that found on T. brucei chromosome 4, is likely to be the ancestral copy and is conserved in all sequenced trypanosomatid genomes. All African trypanosomes exhibit independent expansions of GRESAG5s, implying that positive selection may be occurring. These GRESAG5s form a diverse family, with low levels of sequence identity both within and between organisms (<40% at protein level).

Three GRESAG5s found only in T. b. brucei and T. b. gambiens share a common ancestor with ESAG5s from ESs, and display peculiarities in their synteny. This has two possible explanations; firstly, that these genes represent an ESAG5 which recombined into an ancestral expression site or secondly, that an ESAG5 from an expression site has recombined into a non-telomeric location. Whether this situation occurs with other ESAGs is unknown; as further genomic data becomes available the situation may become clearer.

The (GR)ESAG5 proteins are diverse, with no conserved regions, and the protein sequences themselves reveal little about the function. The first 30 amino acids are hydrophobic and form either a signal peptide or transmembrane domain. As the features of such domains are very similar, in silico analysis cannot always differentiate between the two. A number of N-glycosylation sites are also predicted across the length of the protein, though no single one is found in every family member. These results indicate that ESAG5 encodes a glycosylated membrane-bound or secreted protein, as has been previously suggested [9].

The only domains predicted in a number of (GR)ESAG5 proteins are the BPI-like C- and/or N-terminal domains (bactericidal/permeability-increasing protein) [52,53]. This is supported by our data from sensitive iterative searches and secondary structure prediction. BPI is a member of the lipid transfer/lipopoly saccharide binding superfamily (LT/LBP [54,55]) which also contains LBP [56], PLTP [57,58] and CETP [59,60] as well as a number of recently discovered members [61]. These proteins have been identified only in metazoans and plants so far and have diverse functions; BPI and LBP form part of the innate immune response, binding bacterial LPS and affecting macrophages via transfer of LPS to CD14 [62,63]. In addition, BPI is able to exert a bactericidal effect on gram-negative bacteria involving disruption of the bacterial membrane [64]. CETP and PLTP are plasma lipid-transfer proteins associated with high-density lipoprotein remodelling and as such have mostly been characterised in the context of atherosclerosis [65], though these proteins too can bind LPS [66].

As LT/LBP family members appear to form a similar structure [42,50,67], SWISS-MODEL was used to thread (GR)ESAG5 protein sequences onto the BPI crystal structure [42,43]. All (GR)ESAG5 proteins were able to be threaded onto the BPI boomerang structure, though these theoretical models fall well beneath the 40% identity limit below which inaccuracies are known to occur [68] and as such should be treated with caution. The amino acid variation found between ESAGs clusters around the sites of the BPI lipid-binding pockets and tips of the classical BPI boomerang molecule. Variability at the tips and at the surface may also be indicative of antibody recognition causing a selective pressure on these accessible regions.

Given our conjecture that (GR)ESAG5 is related to this family, one can discuss a number of possible areas of function. Firstly, it may be involved in lipid/sterol binding or transfer; the BPI superfamily contains members which bind phospholipids, sterols and lipoproteins. T. brucei is known to require lipoproteins from the host [69,70] and obtain cholesterol via receptor-mediated endocytosis and lysosomal degradation of lipoprotein particles [69]. As (GR)ESAG5 is strongly predicted to be N-glycosylated, a property linked with the endocytic pathway in trypanosomes [71], ESAG5 could play a role in the uptake or metabolism of lipids. At present, few lipoprotein receptors have been characterised [72].

Secondly, LP/LBP and BPI are known to interact with the MD-2/TLR-4/CD14 receptor complex on macrophages to influence the inflammatory response, including up- or down-regulation of nitric oxide production [73]. The role of macrophages during T. brucei infection is known to be significant, though the mechanisms involved are not fully elucidated [74]. It is possible that trypanosome factors may be involved in modulating these cells.

Finally, (GR)ESAG5 may have functions in parasite–parasite communication or, for GRESAG5, in parasite–vector interactions. BPI is able to bind LPS and can disrupt the membranes of gram-negative bacteria [75], possibly providing a competitive advantage to T. brucei during invasion.

It should be noted that, while ESAG5 sequences are highly similar, GRESAG5s are quite diverse and therefore may form a family with related but separate functions. As T. brucei has a wide host range, the slight differences between ESAG5 proteins may produce a range of proteins with varying affinities, binding to molecules within different species. Some evidence of this has been suggested for ESAG6 and 7, where trypanosomes switch the active ES in response to a change of transferrin species [14,15], though the relevance in vivo remains unclear [17].

The results in this study provide a number of testable hypotheses as well as an analysis of currently available data on the ESAG5 proteins. ESAG5 may be, as postulated for other ESAGs, a surface or secreted protein; as such determining its glycosylation status and subcellular localisation will be important steps toward the elucidation of a function. We theorise that ESAG5 may function as a receptor and may, because of its position in the R-ES, be critical in human infection. Our analysis has revealed distinct differences between the ESAG5 and GRESAG5 gene sets. As GRESAG5 genes are conserved throughout the kinetoplastids, their function is not limited to the requirements of T. brucei; trypanosomatid organisms cause a range of diseases and occupy a number of different niches within the host, indicating a possible wider role for (GR)ESAG5 proteins.

Acknowledgements

This work was supported by grants from the Wellcome Trust and EP Abraham Trust. AB holds a studentship from the Biotechnology and Biological Sciences Research Council (BBSRC). KG is a Wellcome...
Trust Principal Research Fellow. We thank past and present members of our group for direct support of this work, in particular, S. Kelly for constructive methodological discussions and advice.

References


