Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure

Lucio Marcello and J. David Barry

*Genome Res.* 2007 17: 1344-1352; originally published online Jul 25, 2007;
Access the most recent version at doi:10.1101/gr.6421207

**Supplementary data**

"Supplemental Research Data"

[http://www.genome.org/cgi/content/full/gr.6421207/DC1](http://www.genome.org/cgi/content/full/gr.6421207/DC1)

**References**

This article cites 50 articles, 13 of which can be accessed free at:

[http://www.genome.org/cgi/content/full/17/9/1344#References](http://www.genome.org/cgi/content/full/17/9/1344#References)

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

---

Notes

---

To subscribe to *Genome Research* go to:

[http://www.genome.org/subscriptions/](http://www.genome.org/subscriptions/)

© 2007 Cold Spring Harbor Laboratory Press
Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure

Lucio Marcello and J. David Barry

*Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow Biomedical Research Centre, Glasgow G12 8TA, United Kingdom*

*Trypanosoma brucei* evades host acquired immunity through differential activation of its large archive of silent variant surface glycoprotein (VSG) genes, most of which are pseudogenes in subtelomeric arrays. We have analyzed 940 VSGs, representing one half to two thirds of the arrays. Sequence types A and B of the VSG N-terminal domains were confirmed, while type C was found to be a constituent of type A. Two new C-terminal domain types were found. Nearly all combinations of domain types occurred, with some bias to particular combinations. One-third of encoded N-terminal domains, but only 13% of C-terminal domains, are intact, indicating a particular need for silent VSGs to gain a functional C-terminal domain to be expressed. About 60% of VSGs are unique, the rest occurring in subfamilies of two to four close homologs (≥50%–52% peptide identity). We found a subset of VSG-related genes, differing from VSGs in genomic environment and expression patterns, and predict they have distinct function. Almost all (92%) full-length array VSGs have the partially conserved flanks associated with the duplication mechanism that activates silent genes, and these sequences have also contributed to archive evolution, mediating most of the conversions of segments, containing ≥1 VSG, within and between arrays. During infection, intact array genes became activated by duplication after two weeks, and mosaic VSGs assembled from pseudogenes became expressed by week three and predominated by week four. The small subfamily structure of the archive appears to be fundamental in providing the interacting donors for mosaic formation.

This can also help assessing whether some sequences have evolved to neofunction, as often occurs for duplicated genes (Lynch and Katju 2004; Aury et al. 2006). Antigenic variation in *T. brucei* relies on an archive of silent VSG genes (Barry et al. 2005; Berriman et al. 2005; Taylor and Rudenko 2006), analogous to systems in other bacterial and protozoan pathogens (Barry et al. 2003). VSG genes occur at the subtelomeres of all chromosome classes in the *T. brucei* genome, which comprises 11 conventional diploid “megabase” chromosomes, an estimated 100 minichromosomes, and several intermediate chromosomes (Melville et al. 1998). Archival VSGs are effectively the only genes on minichromosomes, immediately adjacent to the telomere (Williams et al. 1982), and another, large, set is tandemly arrayed in megabase subtelomeres (Berriman et al. 2005; Callejas et al. 2006). Each trypanosome expresses only one VSG, and antigen switching replaces the expressed VSG by another. During infection, VSGs are transcribed from specialized loci, bloodstream expression sites (BESs), which are terminal on megabase and intermediate chromosomes (Becker et al. 2004). In switching, archive VSGs are duplicated into a BES, replacing the VSG previously resident there. Antigenic variation produces complex mixtures of variant subpopulations, which nevertheless display order in expression (Lythgoe et al. 2007). Order is thought to be important for prolongation of infection, by ensuring new variants appear as a trickle and by limiting to a sublethal level the number of extant variants. The different locus types occupied by archive VSGs help determine order, a key factor being the inherent activation probability of each VSG (Pays 1989;
Morrison et al. 2005). Thus, telomere-proximal genes, such as those on minichromosomes, are expressed early, array VSGs are expressed later, and two late variants have been observed to express mosaic VSGs, assembled from fragments of pseudogenes (Thom et al. 1990; Barbet and Kamper 1993). Very little is known about how or when mosaic genes are assembled, or their prominence in antigenic variation. It has been proposed recently that antigenic variation might be far more extensive than suspected, due to the large size of the archive and the combinatorial nature of mosaic formation, and that mosaic formation is the predominant switch mode in natural infections (Barry et al. 2005; Marcello and Barry 2007). Although subtelomeres are preferential sites for the evolution of eukaryotic multigene families, little is known about mechanisms involved, and the VSG family could be an informative paradigm.

Many of the features of antigenic variation, from molecules to populations, are intimately linked; features at the molecular level are selected by what occurs at the host and parasite population levels, and in turn contribute to the molecular and population phenotypes (Marcello and Barry 2007). A major challenge is to reveal the functional nature of these links. With the recent near-completion of the genome sequence of *T. brucei* revealing the array archive, here we analyze bioinformatically and experimentally the nearly 1000 available VSG gene sequences (Berriman et al. 2005), to address the questions outlined above.

**Results and Discussion**

**Archive composition and coverage**

We analyzed 940 VSGs in the main contigs of chromosomes (chr) 1–11 (GENEdb v4), in the putative chr 8 homolog left array and in unordered contigs of chr 9, 10, and 11; all are retrievable from the VSGdb database (http://www.vsgdb.org). The archive is larger than this, as sequencing and assembly are incomplete and VSG array haplotypes on chromosome homologs remain to be characterized. The assembled diploid megabase chromosome core genome is ~44 Mb and pulsed field gel electrophoretic analysis (Melville et al. 1998) suggests these chromosomes in the genome strain total 53.5 Mb. Allowing for ~1.5 Mb comprising the five-to-nine bloodstream expression sites (M. Becker and E. Louis, pers. comm.) and other repetitive subtelomeric regions and telomeres, the remaining ~8 Mb difference could be VSG arrays. As the arrays appear to be haploid, and with the observed average gene density of one VSG every 5 kb, it is predicted conservatively that there are ~1600 array VSGs in the diploid genome. The 940 genes analyzed here therefore represent between half and two thirds of the array archive.

**The VSG protein family**

There are four categories of VSG (Berriman et al. 2005): “functional” (encodes all recognizable features of known functional VSGs), “atypical” (complete genes possibly encoding proteins with inconsistent VSG folding or post-translational modification), “pseudogene” (with frameshifts and/or in-frame stop codons), and “incomplete” (Berriman et al. 2005). Most of the 940 silent VSGs are “pseudogenes” (611, 65%) or “incomplete” (197, 21%), only a small percentage being “functional” (43, 4.5%) or “atypical” (89, 9.5%). Designation as “atypical” is necessarily conservative, 75% having small deviations from known GPI anchor signal sequences; increased knowledge of these signals possibly would lead to some being reclassified as “functional.” As the two VSG domains differ in structure, function, and evolution and are subject to independent recombination, we consider them separately.

**N-terminal domain**

In all the 771 full-length N-terminal domains, no novel types were found. Domain types A and B (Carrington et al. 1991), hereafter referred to as nA and nB, are about equal in number, while type nC forms a small cluster within nA (Fig. 1). For each N-terminal type, about one in three domains is theoretically “functional” (Supplemental Fig. S1), suggesting that the basis for a high proportion (94.6%) of the 743 full-length genes being degenerate is C termini being defective, through presence of stop codon(s), frameshift(s), or disruption of cysteine pattern. This means that one third of N-terminal domains possibly can be utilized directly, by combining with a functional C-terminal domain from, for example, the VSG already present in the expression site. More detailed analysis of 361 nA and 29 nC domains revealed several distinct clusters, distinguished by significant differences in the number and spacing of conserved cysteines (Fig. 2A). Group 1 is the most common (44%) and has been characterized biochemically (Carrington et al. 1991). As nC resembles group 4, we propose that it be reclassified into nA. In 335 nB sequences analyzed, the cysteine pattern is much more conserved, displaying only slight variation in inter-cysteine distance (Fig. 2A). It is important now to determine more structures, one reason being to enable identification of possible epitope-bearing substructures.

**C-terminal domain**

Amongst the 651 sequences analyzed, more structural diversity and complexity (presence of one or two subdomains, presence of indels) were detected for the C-terminal domain than for the N-terminal domain. The C-terminal domain is described by the presence of one or two four-cysteine subdomains, the spacing of cysteines, and the type of GPI anchor signal sequence. Two new domain types were found: c5 (one or two four-cysteine subdomains) and c6 (two such subdomains). c1 is the most abundant (235), followed by c2 (186), then c3 (152), c4, c5, and c6 are much less common (10, 30, and 38, respectively). Only 13% (85/651) of encoded C-terminal domains are predicted to be func-

Figure 1. Diversity of VSG N-terminal domain. Neighbor-joining tree based on ClustalX-generated multiple sequence alignment of predicted peptide sequences of 725 sequences. The three domain types are colored individually (by HyperTree) as shown in the key. g represents each group depicted in Fig. 2A.
tional, ranging from 23% (43/186) of C2 to 7% of C1 (Supplemental Fig. S1B). A notable difference from N-terminal domains is a higher incidence of likely indels, making cysteine spacing much more variable and hampering identification of subgroups. Phylogenetic analysis (data not shown) suggests that there is extensive mixing of domain types in the region of the first four cysteines, but much less so toward the carboxy end of the domain, where the pattern of the last four cysteines and the sequence of the GPI anchor signal are more type-specific (Fig. 2B). The solved structure of a C2 domain reveals a novel, compact motif spanning the four cysteines, flanked by disorder (at least in solution) (Chattopadhyay et al. 2005). Our observation of extensive mixing of the first four-cysteine region (see also Fig. 2B) is compatible with its observed greater sequence conservation and provides the context for the documented case of recombination between sequences encoding the third and fourth cysteines of a C1 and a C2 domain (Hutchinson et al. 2003). It thus appears that the 5'

region encoding the C-terminal domain provides a relative hotspot for recombination, yielding expressed interdomain hybrids (Pays et al. 1985). Such modular use of domains appears to have led to a high level of degeneracy of C-terminal domains.

**Full-length VSGs**

Full-length genes and pseudogenes were analyzed phylogenetically and for domain combinations. Whereas global phylogeny of predicted full-length VSGs matches that produced by aligning N-terminal domains alone (Supplemental Fig. S2), analysis of domain combinations revealed bias in combination of domain types, toward nAc2, nBc1, nBc3, and nBc6 (Table 1). However, the propensity for formation of hybrid genes during duplicative activation, allied with the requirement for most genes to gain a functional C-terminal encoding sequence, means that these biases may not be reflected in what occurs in expressed VSGs. Analysis of cDNAs shows no apparent restriction on combination of N- and C-terminal domains, and the same N-terminal domain can be expressed with different C-terminal types (Hutchinson et al. 2003). This promiscuity in domain combinations underlines the importance of the shared four-cysteine motif in the C-terminal domain in promoting N-terminal domain recombination at the DNA level.

**VSG-related proteins**

Thirty one genes, termed VSG-related (VR) (Berriman et al. 2005), cluster separately from VSGs (Fig. 3) and have several features suggesting they are not authentic VSGs. All VR N-terminal domains form a discrete subgroup of nB, but the C-terminal domain is distinct from all of the VSG C-terminal domains. The VR family is as divergent as the much larger VSG archive but, unlike the VSGs, appears not to have assortment between putative N- and C-terminal domain types, suggesting a different mode of evolution. Only seven of 31 (22%) VRs are pseudogenes, in contrast to the ≥86% in the VSG archive. VRs are organized as single genes or in small arrays at strand switch regions in chromosome cores, rather than in subtelomeric VSG arrays. Over the examined 300 bp, VR flanks cluster separately from those of VSGs, and VRs do
not have the upstream 70-bp repeats common to most VSGs. Features of VSG and VR genes are compared in Table 2.

If VRs are functionally distinct from VSGs, they might be predicted to be conserved between strains, unlike VSGs (Van Meirvenne et al. 1977; Frasch et al. 1982). PCR analysis revealed that about half of the 10 tested VR sequences are present in genomic DNA from three other T. brucei strains (Fig. 4A). To test whether VRs have different expression control from that of VSGs, which yield mRNA in bloodstream stage trypanosomes in a clonetype pattern, but no mRNA in the procyclic stage parasite from the tsetse midgut, we applied the RT-PCR. Using the same primers as for genomic DNA analysis, we found at least eight of the 10 VRs, including four of the five ubiquitous genes (see Fig. 4A), expressed as RNA in both stages of the genome strain (Fig. 4B). We next tested whether VR expression is not clonally restricted in bloodstream trypanosomes, through analysis of VR expression in a T. brucei Lister 427 strain in which the HYG gene is in tight linkage with the 221 VSG (McCulloch et al. 1997). RT-PCR analysis of this strain, grown under hygromycin selection, showed transcripts of all three tested VR genes (VR2, VR4, and VR1S) and of 221 VSG, but not of five other VR genes to which 221 commonly switches (Fig. 4C). We conclude that these VRs are coexpressed with the 221 VSG. Overall, the expression patterns of VRs are inconsistent with VSG function and are consistent with novel function.

Despite this, there are no specific sequence features conserved amongst the predicted VR proteins. Several functional trypanosomal surface proteins homologous to VSGs have been proposed to form the distinctive coiled-coil “VSG fold” (Carrington and Boothroyd 1996). One is the heterodimeric transferrin receptor, both subunits of which resemble VSGs, although it is unknown whether they evolved from, or share an ancestor with, VSGs (Salmon et al. 1997). A clearer case for evolution from the VSG has been argued for the serum resistance protein (Campillo and Carrington 2003). Although the VR putative products may have evolved from VSGs in T. brucei, they could have evolved from an earlier precursor, as they have an N-terminal domain, which is the only type demonstrated for VSGs of Trypanosoma congolense (Bussler et al. 1998).

**VSG gene array architecture and evolution**

All of the VSGs analyzed here are located in subtelomeric arrays, with a sole exception. Tb927.5.3990 is interstitial, has none of the upstream 70-bp imperfect repeats characteristic of VSGs, and has a nonconserved C-terminal domain. Within the arrays, as elsewhere in the genome, the basic VSG unit is a 3.4 kb cassette, flanked upstream by the 70-bp repeats and downstream by the partially conserved region from within the 3′ end of the coding sequence to elements in the 3′ untranslated region (UTR) (Michels et al. 1983). The 70-bp repeats form the upstream boundary of the duplication events of VSG switching. Some 92% of full-length array VSGs (687/743) are flanked upstream by at least one such repeat, 74% having one copy, 14% having two, and 5% having between three and 15 copies. The degeneracy amongst these repeats is much greater than previously recognized, especially in the case of section 3 of the repeat (Aline et al. 1985), which sometimes is absent or incomplete.

Our initial global analysis of annotated VSGs showed almost complete dispersion of closest homologs across the arrays (Berman et al. 2005), in keeping with earlier analysis of one array gene in different trypanosome strains (Bernards et al. 1986), which had led to a model involving deletion and duplication of VSG cassettes. We now confirm and extend this model, through analysis of recent duplications in the genome. For pairs of N-terminal domain-encoding sequences >90% nucleotide-identical, we located the physical extent of each putative gene duplication event, through pairwise alignment of their flanking regions. Scrutiny of several long (up to 30 kb) gene conversion tracts revealed localized disruption of homology, often coincident with 70-bp repeats, which in many cases displayed repeat contractions and expansions. This, and subsequent possible conversion events, complicated the identification of precise ends of conversion events. Over the 38 conversion events analyzed, nucleotide identity between putative conversion tracts ranged between 74% and 99%. The shortest putative conversion tract was 1.6 kb long, and the longest contained nine VSGs. In 89% of cases, at least one of the ends corresponded to either 70-bp repeats or the C-terminal end region; 52% of cases involved either

![Figure 3. Tree of amino acid alignment of putative VSG-related proteins. The tree compares predicted products of 35 full-length, functional VSGs with those of 29 VR genes. The VR set falls into two clusters, denoted by 1 and 2.](image-url)

**Table 2. Comparison of VSGs and VSG-related proteins**

<table>
<thead>
<tr>
<th></th>
<th>VSG</th>
<th>VSG-related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of genes</td>
<td>940</td>
<td>31</td>
</tr>
<tr>
<td>No. of pseudogenes</td>
<td>808</td>
<td>7</td>
</tr>
<tr>
<td>70-bp-repeat upstream</td>
<td>Most</td>
<td>None</td>
</tr>
<tr>
<td>In arrays with 70-bp repeats</td>
<td>All but 1 (Tb927.5.3990)</td>
<td>None</td>
</tr>
<tr>
<td>Length (amino acids)</td>
<td>474–540</td>
<td>340–490</td>
</tr>
<tr>
<td>C domain types</td>
<td>A, B, C</td>
<td>8 type clusters 1, 2, 3</td>
</tr>
<tr>
<td>GPI signal sequence</td>
<td>Present</td>
<td>3 clusters, cluster 2 with cysteines</td>
</tr>
<tr>
<td>3′ UTR</td>
<td>Two main subtypes clustering closely</td>
<td>Absent or not conserved; possibly present in cluster 1 and 2</td>
</tr>
<tr>
<td>5′ UTR</td>
<td>Diverged, slight overlap with VR</td>
<td>Cluster separately from VSGs</td>
</tr>
</tbody>
</table>

**Trypanosome VSG gene silent archive**

Genome Research 1347

www.genome.org
of these sequences at both ends. To dissect the evolution on a finer scale and assess the incidence of relatively recent duplication events, we queried geneDB for the number of duplicates (defined as ≥75% nucleotide identity) of each of 31 VSGs in the chr 8 left subtelomere array (chr 8L). Five of the 31 genes are within a family of five or more members (Fig. 5). Again, duplications in general are delimited by the conserved cassette flanks and often have involved single genes, although some events will be concealed by subsequent incidents. It appears now that the cassette flanks driving VSG duplication are important not only for antigen switching, but also in array expansion and diversification.

A revised model for array evolution can now be suggested: Shorter rearrangements, involving deletions and conversions of VSG cassettes, promote gene duplication and family divergence, while longer-range contractions and expansions are effected via the non-LTR retrotransposon INGI, which is found at most strand switches in VSG arrays (Berriman et al. 2005). The passive and active routes by which such transposons reshape genomes (Kazazian 2004) could contribute extensively to the large differences in archive size between trypanosome strains and, within a strain, between arrays and between chromosome homologs. Gene conversion clearly is a major mechanism in array evolution, and we detected no polarity in the duplication map (Fig. 5), which argues against a role for break-induced replication (Fischer et al. 2006). Subtelomeres of other organisms, including Kluyveromyces lactis (Fairhead and Dujon 2006) and Homo sapiens (Linardopoulou et al. 2005), are patchworks with few tandem duplications and so resemble the VSG array subtelomere regions. A reason for this common architectural theme might be evident from studies of Saccharomyces cerevisiae, where experimental selection for duplication yielded many cases of translocation, an event shown to be most likely to lead to fixation (Koszul et al. 2004, 2006), presumably because tandem duplication creates excessive, local genome instability. The extensive traffic between trypanosome subtelomeres, including gene duplication, is followed by rapid sequence diversification, which presumably limits which VSGs can readily interact in homologous recombination within their coding sequence. As we show below, that is central to ordered VSG expression after the initial phase of infection.

**VSG gene subfamilies**

As we have proposed that mosaic VSG expression is a major mechanism in antigenic variation (Marcello and Barry 2007) and mosaic formation appears to depend on highly related donor partners, we investigated whether the archive contains VSG sequence subfamilies. We aligned N-terminal domains to avoid the complication of higher similarity amongst C-terminal sequences, and because the variable domain is more directly related to mosaic gene formation, as we show below. Systematic ClustalW pairwise analysis of 361 type nA domain sequences revealed an outlier group of pairs with >51% peptide identity. The outlier group (41.8% of genes) comprises small subfamilies, ranging in size from two to six members (Fig. 6, upper panel), with a minimum pairwise nucleotide identity of 67% (average for the least-score 10 pairs is 73.5%). Almost exclusively, the subfamilies are pairs or triplets, in the ratio 3:1 (Fig. 6, upper panel). The largest subfamily, of six members, is abnormal, as four unusually form a directional gene cluster demarking the core of chr 4 from the right-side VSG array. The subfamily of four illustrates the dispersion of subfamily members: They are in arrays on different subtelomeres. For type nB VSGs, comparison of 345 N-terminal domains, using an observed minimum cutoff value of 53% peptide identity, yielded similar data (Fig. 6, lower panel). Again, the largest subfamily is anomalous, consisting of an array of VR genes. Across the whole archive of an estimated 1500 VSGs, and taking account of the proportion of sequences that are full-length or encode N-terminal fragments, it might be extrapolated that ~550 VSGs are in high-identity subfamilies, comprising ~180 pairs, ~50 triplets, and ~10 quadruplets.

**Expression of array VSGs and formation of mosaic genes**

To address the significance, to antigenic variation, of mosaic gene expression and
we recovered only the VSGs due to the incomplete extent of genome sequencing. On day 3, it should be noted that our interpretations are partly speculative, depending in the sequential use of different donor types, as expected with the T. brucei genome strain 927 (Table 3). A pattern is evident in the sequential use of different donor types, as expected from previous studies (Pays 1989; Robinson et al. 1999), although it should be noted that our interpretations are partly speculative, due to the incomplete extent of genome sequencing. On day 3, we recovered only the VSG, GUTat 10.1, expressed by the infecting trypanosomes. For day 9 and in one case for day 14, no corresponding donor gene sequences are present in geneDB. Precedence suggests they arose by duplication of minichromosomal VSGs into the expression site (Liu et al. 1985; Robinson et al. 1999); the minichromosomal genes of TREU 927 have not been sequenced. Array donor genes became involved from day 14. Of the 21 distinct sequences, 14 could be attributed to array VSG donors, with 10 matching fully sequenced silent genes and four matching short, partial reads. Five of these 14 arose from “functional” array donors, although they had gained some or all C-terminal sequence from elsewhere. Two cDNAs (days 14, 24) arose from “atypical” genes, each with donation of a short stretch of sequence at the 3’ end. Mosaic genes were present as seven cDNAs from days 22 and 28 and became predominant, contributing 40% of recovered cDNAs by days 22–24 and 62.5% by day 28 (Table 3). Previously, mosaics were detected on days 23–35 in rabbits, but their prevalence was not investigated (Capbern et al. 1977; Marcello and Barry 2007). Notably, we found distinct sets of variants in different infections on day 28, which is consistent with the proposition that mosaic formation leads to distinct variant strings in different animals, due to antigenic variation being stochastic (rather than to host-related differences) (Barry et al. 2005; Marcello and Barry 2007).

The proposal that the number of 70-bp repeats upstream of the gene, and their degree of identity with those in the BES, help determine order in antigenic variation (Aline et al. 1985) is not supported by our data, as there is only minimal variation in 70-bp repeat number amongst silent cassettes. For the downstream end of duplications into the BES, our experimental analysis reveals that inheritance of at least some of the C-terminal encoding sequence from the previously expressed VSG is common. In the 14 cDNAs that arose from array VSGs, eight had gained a 3’ end (Table 3), in agreement with the observation that most array genes encode defective C-terminal domains.

The archive’s subfamily structure appears to be central to the subfamily structure, we sequenced 37 cDNAs, which corresponded to 21 different VSGs, isolated from 11 mice infected with the T. brucei genome strain 927 (Table 3). A pattern is evident in the sequential use of different donor types, as expected from previous studies (Pays 1989; Robinson et al. 1999), although it should be noted that our interpretations are partly speculative, due to the incomplete extent of genome sequencing. On day 3, we recovered only the VSG, GUTat 10.1, expressed by the infecting trypanosomes. For day 9 and in one case for day 14, no corresponding donor gene sequences are present in geneDB. Precedence suggests they arose by duplication of minichromosomal VSGs into the expression site (Liu et al. 1985; Robinson et al. 1999); the minichromosomal genes of TREU 927 have not been sequenced. Array donor genes became involved from day 14. Of the 21 distinct sequences, 14 could be attributed to array VSG donors, with 10 matching fully sequenced silent genes and four matching short, partial reads. Five of these 14 arose from “functional” array donors, although they had gained some or all C-terminal sequence from elsewhere. Two cDNAs (days 14, 24) arose from “atypical” genes, each with donation of a short stretch of sequence at the 3’ end. Mosaic genes were present as seven cDNAs from days 22 and 28 and became predominant, contributing 40% of recovered cDNAs by days 22–24 and 62.5% by day 28 (Table 3). Previously, mosaics were detected on days 23–35 in rabbits, but their prevalence was not investigated (Capbern et al. 1977; Marcello and Barry 2007). Notably, we found distinct sets of variants in different infections on day 28, which is consistent with the proposition that mosaic formation leads to distinct variant strings in different animals, due to antigenic variation being stochastic (rather than to host-related differences) (Barry et al. 2005; Marcello and Barry 2007).

The proposal that the number of 70-bp repeats upstream of the gene, and their degree of identity with those in the BES, help determine order in antigenic variation (Aline et al. 1985) is not supported by our data, as there is only minimal variation in 70-bp repeat number amongst silent cassettes. For the downstream end of duplications into the BES, our experimental analysis reveals that inheritance of at least some of the C-terminal encoding sequence from the previously expressed VSG is common. In the 14 cDNAs that arose from array VSGs, eight had gained a 3’ end (Table 3), in agreement with the observation that most array genes encode defective C-terminal domains.

The archive’s subfamily structure appears to be central to

Figure 6. VSG families. Pairwise comparison within N-terminal domain types. Peptide sequences encoded by 361 type nA VSGs were all scored against each other in ClustalW multiple sequence alignment, and 345 type nB VSGs were compared in the same way (bottom chart). The number of families (left scale of Y-axis) of given size (X-axis) are depicted in black, and the percentage of genes (right scale of Y-axis) within each domain type is depicted in white. VR indicates the family of VSG-related genes.

Table 3. VSGs expressed in chronic infections

<table>
<thead>
<tr>
<th>Day</th>
<th>Mouse</th>
<th>Expressed VSG</th>
<th>Silent donor (geneDB)</th>
<th>Donor type</th>
<th>GenBank accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1, 2</td>
<td>Tb10.v4.0001</td>
<td>Tb10.v4.0001</td>
<td>Functional</td>
<td>EF568372</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>09-03-01</td>
<td>Unknown</td>
<td>Unknown</td>
<td>EF560194</td>
</tr>
<tr>
<td>9</td>
<td>3, 4</td>
<td>09-03-03</td>
<td>Unknown</td>
<td>Unknown</td>
<td>EF560195</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>09-03-04</td>
<td>Unknown</td>
<td>Unknown</td>
<td>EF560196</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>09-04-02</td>
<td>Unknown</td>
<td>Unknown</td>
<td>EF560197</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>14-05-01</td>
<td>Unknown</td>
<td>Inc*</td>
<td>EF560198</td>
</tr>
<tr>
<td>14</td>
<td>5, 6, 8</td>
<td>14-05-02</td>
<td>Tb09.v4.0077</td>
<td>Atypical + 3’ donor</td>
<td>EF560199</td>
</tr>
<tr>
<td>22</td>
<td>7</td>
<td>22-07-01</td>
<td>Tb09.244.1580</td>
<td>Functional</td>
<td>EF174487</td>
</tr>
<tr>
<td>22</td>
<td>7</td>
<td>22-07-03</td>
<td>Tb927.3.190</td>
<td>Mosaic</td>
<td>EF174488</td>
</tr>
<tr>
<td>22</td>
<td>7</td>
<td>22-07-04</td>
<td>Tb11.30.0005, Tb11.04.0021, Tb09.354.0090</td>
<td>Functional</td>
<td>EF560201</td>
</tr>
<tr>
<td>24</td>
<td>9</td>
<td>24-09-01</td>
<td>Tb09.v4.0102</td>
<td>Mosaic at C-terminal domain</td>
<td>EF174490</td>
</tr>
<tr>
<td>28</td>
<td>10</td>
<td>28-10-02</td>
<td>Tb11.09.0005, Tb11.13.0003</td>
<td>Atypical + 3’ donor</td>
<td>EF174492</td>
</tr>
<tr>
<td>28</td>
<td>10</td>
<td>28-10-03</td>
<td>Tb11.09.0005, Tb11.13.0003, AZ217061</td>
<td>Mosaic</td>
<td>EF174491</td>
</tr>
<tr>
<td>28</td>
<td>10</td>
<td>28-10-07</td>
<td>Tb11.v4.0027, Tryp_X1.934d12q1k.rev, Tryp_X334d06.p1c</td>
<td>Mosaic</td>
<td>EF174492</td>
</tr>
<tr>
<td>28</td>
<td>10</td>
<td>28-10-10</td>
<td>Tryp_X1.1058d01.q1k, Tb11.57.0032, Tb09.244.0090</td>
<td>Mosaic</td>
<td>EF174493</td>
</tr>
<tr>
<td>28</td>
<td>11</td>
<td>28-11-01</td>
<td>Contig 10023</td>
<td>Functional/hybrid</td>
<td>EF560202</td>
</tr>
<tr>
<td>28</td>
<td>11</td>
<td>28-11-02</td>
<td>Tb927.5.0580</td>
<td>Functional</td>
<td>EF560203</td>
</tr>
<tr>
<td>28</td>
<td>11</td>
<td>28-11-09</td>
<td>Tb927.7.6530, AL4S5630, AQ660659, AL474336</td>
<td>Mosaic</td>
<td>EF174493</td>
</tr>
</tbody>
</table>

VSGs corresponding to known archive genes are given the geneDB identifier of the donor; novel VSGs do not have a geneDB code.

*Inc, only incomplete sequence obtained.
Figure 7. Structure and origin of putative mosaic cDNAs. (A) Mosaic structure. The seven detected mosaics for which significant donor information is available. Colors indicate patches matching different donors. The white regions in mosaic 28-10-10 indicate where no donor has been identified. (B) Origin and structure of N-terminal domain mosaic 22-07-02 from mouse 7 on day 22. The three putative silent donor genes (A, B, C) are shown below the cDNA, as thinner rectangles. Colors represent patches in common between donor and cDNA (positions of junctions in cDNA are averages where donors are identical in that region). The percentage nucleotide identity between silent donors is shown to the right. Vertical red bars depict frameshifts or stop codons in donor genes. The likely donors are: A, Tb11.v4.0074; B, Tb10.v4.0161; C, Tb11.14.0001. (C) Origin and structure of related mosaics coexisting in mouse 10 on day 28. Diagram is constructed as in B. For both mosaics, likely donors A and B are Tb11.09.0005 and Tb11.13.0003. For 28-10-03, likely donor C is read AZ217061. (D) Deduced pathway for development of related mosaics in mouse 10 on day 28.
panosome subtelomeres has been selected to achieve equilibrium between gene duplication and rapid diversification, satisfying needs for a subfamily structure conducive to expressed mosaic formation and for the production of novel silent VSGs.

Methods

Bioinformatic analysis of the VSG archive

The VSG annotation procedure has been described (Berriman et al. 2005). Multiple sequence alignments were conducted, via default settings, using ClustalX or ClustalW at EBI (http://www.ebi.ac.uk/clustalw/) or at the Pasteur Institute (http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html), the alignment was then edited manually with Bioedit (http://www.mibio.ncsu.edu/BioEdit/bioedit.html) and then a second alignment with the edited sequences was performed. A neighboring tree produced by ClustalX was then visualized and colored using the HyperTree software (Bingham and Sudarsanam 2000). GPI signal sequences were identified initially by manual comparison with known VSG signals, then through big-PI (http://mendel.imp.ac.at/gpi/gpi_server.html) and DGPI (http://129.194.185.165/dgpi/index_en.html) prediction programs.

Trypanosomes

Most analyses were performed on the genome sequencing strain, T. brucei TREU 9274/4 GuTat 10.1, either as bloodstream stage in mice or, for analysis of VR gene expression, as the procyclic stage in SDM79 medium (Brun and Schonenberger 1979). Some VR gene expression analysis was performed on T. brucei Lister 427 bloodstream stage line 3174.2, which has the hygromycin resistance gene HYG inserted upstream of the expressed 221 VSG. This line was grown in vitro in HMI-9 medium as described (McCulloch et al. 1997). Genomic DNA of T. brucei EATRO 795 and STIB 247 was also analyzed for presence of VR genes. For chronic infections with the genome strain, 11 mice were each infected with 80 µL of blood containing 1.3 × 10^7 trypanosomes/mL. In order to maximize recovery of parasite gDNA and RNA, each mouse was sampled only terminally. On days 3, 9, 14, 21 to maximize recovery of parasite gDNA and RNA, each mouse was each infected with 2 µL of each primer (1 µM final concentration), 0.2 µL Taq, 2 µL custom buffer mix (yielding final concentrations of 45 mM Tris-HCl at pH 8.8, 11 mM [NH₄]₂SO₄, 4.5 mM MgCl₂, 6.7 mM 2-mercaptoethanol, 4.4 µM EDTA at pH 8.0, 113 µg/mL BSA, 1 mM each of the four dNTPs), and made up to 20 µL with water. Reaction conditions were 30 cycles of 95°C for 50 sec, 55°C–63°C for 50 sec, and 65°C for 1 min per kb of expected product.

Acknowledgments

We thank colleagues at the Sanger Institute (Matt Berriman, Christiane Hertz-Fowler, Hubert Renauld) and The Institute for Genome Research (Najib El-Sayed, Gaelle Blandin) for extensive discussions during annotation of the VSG arrays. We thank the Wellcome Trust for funding; J.D.B. is a Wellcome Trust Principal Research Fellow.

References


17: 197

338: 221

517: 537

223: 355

10.1371/journal.pgen.0020032.eor.

14: 17.

Trypanosoma brucei equiperdum Parasitol. 127: 79–84.


