DNA damage drives antigen diversification through mosaic VSG formation in *Trypanosoma brucei*

Jaclyn E. Smith[1], Kevin J. Wang[1], Jill M.C. Hakim[1], Alexander K. Beaver[1,2], Aishwarya Magesh[1], Shane D. Gilligan-Steinberg[1,3], Jessica Zheng[1], Jaime So[1], Erin M. Kennedy[1], Bailin Zhang[1,4], Elgin Henry Akin[1], Lusajo Mwakibete[1], Monica R. Mugnier[1]*

[1]W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America

[2]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America

[3]Department of Bioengineering, University of Washington, Seattle, Washington, United States of America

[4]Scripps Research Department of Integrative Structural and Computational Biology, La Jolla, San Diego, California, United States of America

*Corresponding author
Email: mmugnie1@jhu.edu (MRM)

Summary

Antigenic variation, using large genomic repertoires of antigen-encoding genes, allows pathogens to evade host antibody. Many pathogens, including the African trypanosome *Trypanosoma brucei,* extend their antigenic repertoire through genomic diversification. While evidence suggests that *T. brucei* relies heavily on the generation of new variant surface glycoprotein (VSG) genes to maintain a chronic infection, a lack of experimentally tractable tools for studying this process has obscured its underlying mechanisms. Here, we present a highly sensitive targeted RNA sequencing approach for measuring VSG diversification. Using this method, we demonstrate that a Cas9-induced DNA double-strand break within the VSG coding sequence can induce VSG recombination with patterns identical to those observed during infection. These newly generated VSGs are antigenically distinct from parental clones and thus capable of facilitating immune evasion. Together, these results provide insight into the mechanisms of VSG diversification and an experimental framework for studying the evolution of antigen repertoires in pathogenic microbes.

Introduction

Pathogen survival in a host depends upon effective and continuous immune evasion. Several bacteria and eukaryotic pathogens have adopted strategies of antigenic variation to evade host immunity, a process in which they continuously alter antigenic surface proteins to escape the host's adaptive immune response. The African trypanosome *Trypanosoma brucei*, a unicellular eukaryotic parasite and causative agent of human and animal African trypanosomiasis, uses an especially sophisticated system of antigenic variation. The parasite, which remains extracellular throughout infection and thus faces a perpetual onslaught of host antibody, periodically "switches" expression of a surface coat consisting of 10^7 copies of a single, immunogenic protein known as the Variant Surface Glycoprotein (VSG). This process allows parasites to escape host antibody and maintain a chronic infection.

Although the *T. brucei* VSG repertoire contains thousands of VSGs, it may not be sufficient to maintain a chronic infection through VSG switching alone. During an infection, each *T. brucei* parasite expresses a single VSG at a time from one of ~15 telomeric Bloodstream Expression Sites (the "active" BES)[1]. The remaining thousands of VSG-encoding genes are stored in other expression sites, subtelomeric arrays, and minichromosomes, all of which remain transcriptionally silenced[2]. The parasite switches its VSG either by transcriptional activation of a silent BES (*in situ* switching) or through a gene conversion event in which a new VSG is copied into the active expression site. While gene conversion-based switching allows for the activation of VSGs outside of a BES, analysis of the *T. brucei* genome has shown that only ~20% of the VSGs in the parasite genome are full-length genes encoding a functional VSG protein. The remaining ~80% of VSGs in the parasite genome consist of pseudogenes or gene fragments[2] and cannot immediately be used for immune evasion through *in situ* or gene conversion switching. Moreover, the number of VSGs expressed at a single timepoint during experimental infection sometimes exceeds the total number of intact VSGs in the parasite genome[3,4], further indicating that the repertoire of intact VSGs is insufficient for the antigenic diversity required to maintain a chronic infection.

Evidence suggests that *T. brucei* deals with this shortage of antigens through diversification of the VSG repertoire. Many mouse studies have shown that novel VSGs, generated during infection, predominate at later stages of infection[3,5,6], while analysis of parasites from natural human infections revealed expressed VSGs that were nearly completely absent from the genomes of contemporary field isolates[7]. These observations suggest that the generation of new VSGs is likely to be a central feature of *T. brucei* antigenic variation.

There are two mechanisms thought to be responsible for extending the VSG repertoire: mosaic formation and *de novo* point mutation. Mosaic VSGs form when two or more VSG genes combine through segmental gene conversion to form a novel VSG. This mechanism allows parasites to access pseudogenes and VSG fragments within the repertoire. While mosaic VSGs have been observed in the literature, they are often found under extreme experimental conditions[8,9] or late during infection[3,5,10,11] making it difficult

to discern how exactly they arose. VSGs also appear to occasionally acquire *de novo* point mutations though these can be difficult to distinguish from a small gene conversion event using a similar donor VSG. The origins of these mutations also remain mysterious, though newly acquired mutations seem rare[9,12–14]. Ultimately, *de novo* mutation of VSGs would allow parasites to generate new VSG sequences regardless of the contents of their repertoire, further amplifying diversity.

Despite its clear importance, the mechanisms driving VSG diversification, whether by mutation or recombination, remain poorly understood. It is plausible that DNA damage and repair may play a role in either mechanism, as expressed VSG genes sit between two highly repetitive and damage-prone stretches of DNA, the conserved 70bp repeat and the telomere[2,15]. DNA breaks are frequently observed within the 70bp repeats at the active expression site[16,17] and near telomeres within silent BESs[17]. Experimental evidence implicates DNA damage in VSG switching more generally, as a DNA double-strand break induced upstream of the VSG induces a switch[16,17], but it is unknown whether DNA damage could also generate new VSGs.

There is a dearth of experimental evidence for the mechanisms driving VSG diversification because there has been no controlled and reproducible *in vitro* system for studying the process. Instead, researchers have had to rely on observation alone, characterizing the sequences of expressed VSGs that were isolated by chance. This approach has been crucial for generating hypotheses but is insufficient for deciphering mechanisms driving the process. Here, we have developed a comprehensive toolkit for the controlled and reproducible study of the diversification of individual VSGs. Using a highly sensitive barcode-based targeted RNA-seq approach, we show that DNA double-strand breaks can trigger the formation of mosaic VSGs that are identical to those observed *in vivo* during infection. Homology appears to drive donor VSG selection, with microhomology between the parent and donor VSG flanking the site of recombination in almost every event. Finally, we observe that break-induced diversification is most efficient in the portion of the VSG gene encoding the exposed top lobe of the VSG N-terminal domain. We suggest that this could represent a potential hypervariable region within the VSG, facilitating diversification in the region of the VSG protein where host antibody is most likely to bind.

Results

DNA double-strand breaks trigger the formation of mosaic VSGs

A number of studies have suggested that VSG diversification occurs, at least some of the time, within the active expression site[3,13], with the sequence of the actively expressed VSG being altered through recombination and/or mutation. For this reason, we focused our analyses on diversification of the actively expressed VSG. Because a double-strand DNA break upstream of the VSG within the active BES is known to induce switching[16,17], we hypothesized that a break within the VSG coding sequence might result in mosaic VSG formation. To investigate this, we engineered tetracycline-inducible Cas9 expressing EATRO1125 parasites, which express the VSG AnTat1.1, and induced breaks across the AnTat1.1 coding sequence using a set of guide RNAs.

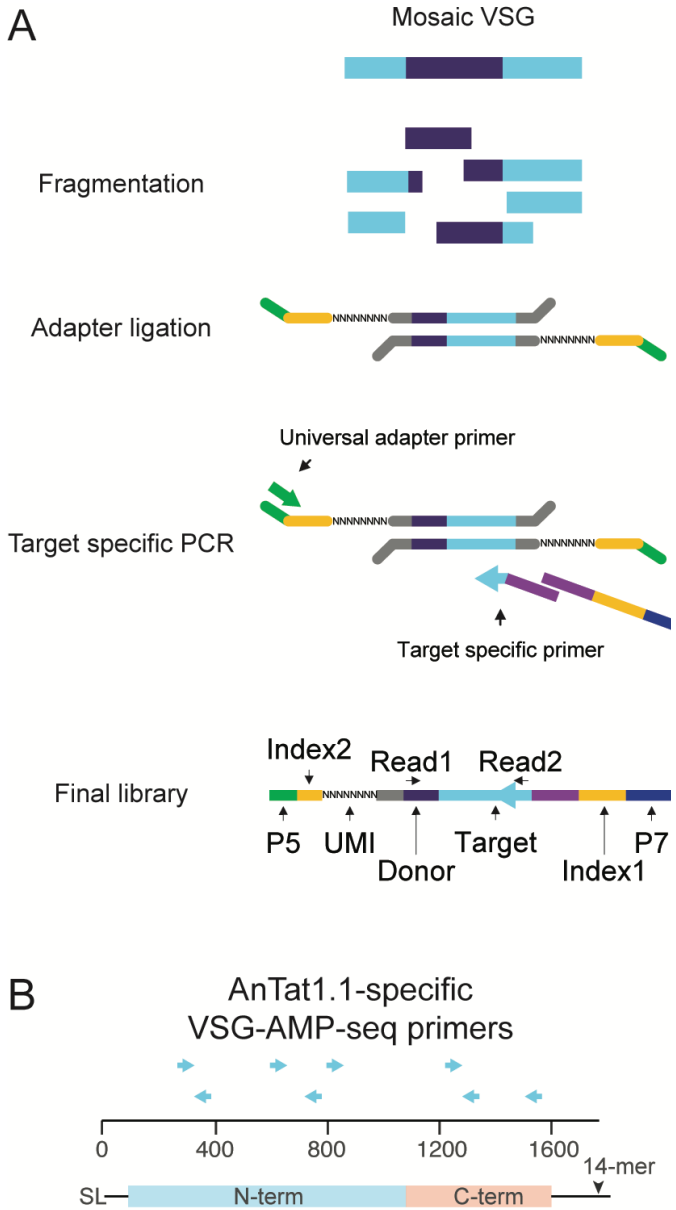


Figure 1. VSG-AMP-seq is a tool to detect diversification within a VSG of interest
A) Schematic of the library prep for VSG-AMP-seq. Figure adapted from [51]. B) Locations of target specific primers for VSG AnTat1.1. SL = 5′ splice leader cap, 14-mer = 3′ sequence conserved in all VSG transcripts

To evaluate potential recombination outcomes, we developed VSG Anchored Multiplex PCR Sequencing (VSG-AMP-seq), a technique that overcomes previous obstacles to studying VSG diversification by providing both high-throughput and highly accurate sequences (1A). Our approach uses long unique molecular indexes (UMIs) to generate high-confidence, accurate consensus sequences[18]. Consolidating reads into consensus sequences allows errors like PCR chimeras, which occur during later cycles of PCR and therefore represent a minority of sequences in a consensus group[19], to be eliminated while true events are retained. Libraries are prepared by fragmenting VSG-specific cDNA. Fragments are then end-repaired, A-tailed and ligated to universal adapters containing a

166 25bp unique molecular index (1A). A VSG target of interest is selected, and a series of
167 staggered primers are designed to cover the length of the target VSG's coding sequence
168 (1B). By pairing target-specific primers with a universal reverse adapter primer, target
169 VSG fragments are amplified within a sample regardless of their identity. Mosaic reads
170 are defined as those for whom a portion of the read matches the target and the remainder
171 matches another VSG (the "donor VSG") within the VSG repertoire (VSGnome) of the
172 strain being studied[2]. Due to the selective, target-specific amplification, this method can
173 sensitively detect thousands of rare diversification events, even from mixed samples
174 where diversified VSGs are a minority of the population.
175
176 To induce breaks across AnTat1.1, we induced Cas9 expression for 24 hours and
177 then transfected in DNA amplicons containing a T7 promoter and a guide RNA targeting
178 various regions throughout the AnTat1.1 coding sequence (cut positions, relative to the
179 5′ end of the VSG transcript: 243, 369, 694, 894, 978, and 1459) to induce breaks in the
180 VSG. Parasites were collected two days after transfection of the guide, and mosaic
181 derivatives of AnTat1.1 were analyzed by VSG-AMP-Seq (2A; Supplementary Figures
182 2A, 2B & 3C). We detected thousands of recombination events from two independently
183 generated Cas9 clones (C1 = 5956, C2 = 4488).
184
185 Our analysis revealed diverse mosaic recombination events centered around each
186 break site. Such events were virtually absent from the negative (no guide) control,
187 suggesting this process does not occur at a high rate in the absence of a trigger or in the
188 presence of Cas9 alone. Notably, as the DNA breaks progressed further away from the
189 center of the AnTat1.1 coding sequence, the frequency of mosaic formation decreased
190 dramatically (2B). This did not appear to be related to the guide sequences (Supplemental
191 Figure 2C-G) or to guide cutting efficiency (Supplemental Figure 2H). To ensure that the
192 observed mosaic events represented mosaic VSGs that were truly expressed by the
193 parasite, and not technical artifacts or VSGs incapable of being stably expressed by *T.*
194 *brucei*, we also obtained individual clones of parasites expressing AnTat1.1-derived
195 mosaic VSGs. Using parasite lines that stably express a VSG-targeted guide RNA after
196 7 days of Cas9 expression (2C; Supplemental Figure 2A), we isolated 26 mosaic-
197 expressing clones, 4 from guide 243 and 22 from guide 694, from 6 parental cell lines.
198 These parasites expressed VSGs containing recombination events identical to those
199 observed using VSG-AMP-Seq. These results suggest that DNA damage within the active
200 VSG can trigger the formation of mosaic VSGs, with recombination events centered
201 around the site of DNA damage.
202
203 **Homology drives mosaic VSG formation**
204
205 Analysis of the mosaic recombination events detected by VSG-AMP-seq revealed
206 an important role for sequence homology in the formation of mosaic VSGs. Almost all
207 (C1= 99.71%, C2 = 99.46%) recombination events occurred within a region of shared
208 sequence between AnTat1.1 and each donor VSG, with an average length of ~9 bps
209 (average C1 = 9.13 bp, average C2 = 9.44 bp, median = 6 bp) (2F; Supplemental Figure
210 1A). Within the isolated mosaic clones, we observe short insertions (< 200 bp, average =
211 47 bp, median = 23 bp) predominating among the events (2D). Although read lengths

limited our ability to detect larger insertions with VSG-AMP-seq, approximately 55-60% of the mosaic recombination reads detected by VSG-AMP-seq contained the same short insertions (average C1 = 46 bp, average C2 = 45.8 bp) (Supplemental Figure 3B).

Notably, only a small number of donor VSGs were used for the majority of recombination events. Upon closer inspection, these donors are members of a 6-VSG family that contains AnTat1.1 and represent the only sequences within the EATRO1125 VSGnome with significant homology to this VSG (2E). N-terminal recombination events appear restricted to just these family members (C1 = 100%, C2 = 99.87%). Together, these results suggest that VSG sequence homology influences the outcome of VSG recombination after a DNA break.

To determine the outcome following DNA damage of a VSG when there are no homologous donors available, we used the same Cas9 system in the commonly used Lister427 *T. brucei* line to cut the actively expressed VSG, VSG-2, which lacks homologous family members. After inserting stably expressed guides targeting VSG-2 into the genome, we induced Cas9 expression for 7 days. Parasite clones isolated after a break had all switched VSG expression from VSG-2, with no evidence of mosaic recombination (Supplemental Figure 3A; cut position 707, n = 4; cut position 1082, n = 3). This further supports the hypothesis that homology between donor VSGs and the region surrounding the DNA break drives mosaic formation.

We sought to determine what proportion of the VSG repertoire contains VSGs that are members of VSG families and thus capable of diversifying through break-induced mosaic formation. We defined the phylogenetic distance between VSGs within each of the EATRO and Lister427 strains using a generalized time reversible model and found that the proportions of the repertoire within families in each strain were remarkably similar. In both strains, ~75% of the known VSG sequences are within a family and likely capable of diversifying through the mechanisms described here (2G).

242

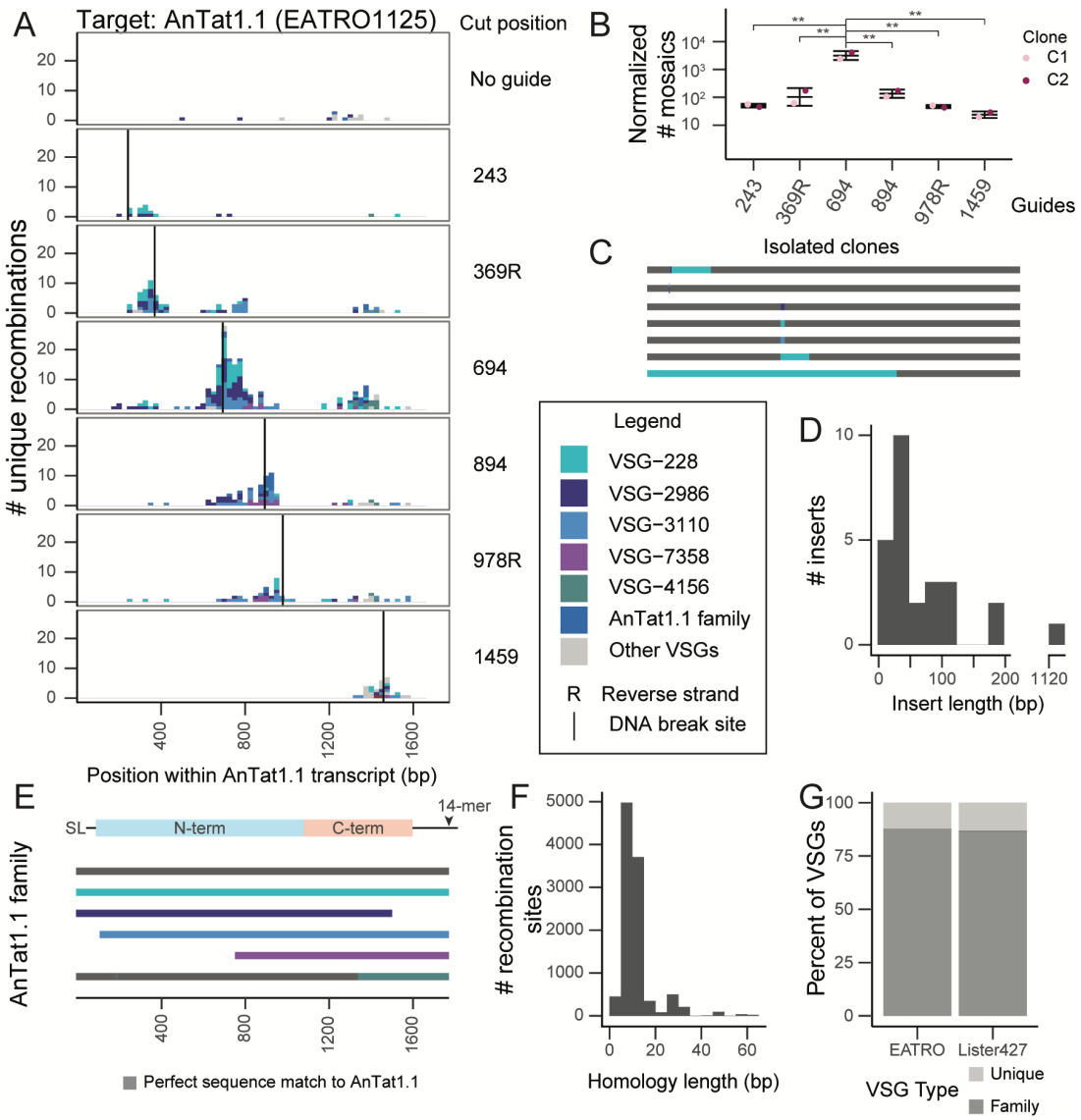


Figure 2. DNA double-strand breaks trigger mosaic VSG formation if homology is available
A) A histogram of unique recombination events identified along the AnTat1.1 transcript. The Cas9 DNA break site is indicated by a vertical line. Sections of the histogram are colored to indicate donor VSG identified. R indicates a guide that binds to the reverse strand. The midpoint of the perfect homology between AnTat1.1 and the donor VSG at the recombination site is plotted. If a donor VSG could not be unambiguously identified, the average recombination position was plotted. B) Quantification of mosaic recombination events induced by DNA breaks. The number of recombination events detected within 250bp up or downstream of the cut site was normalized to the number of total unanchored reads aligning within that region compared to the unanchored read count from the region with the smallest coverage to control for sequencing depth. (n=2, two independent clones) Statistical significance was determined with a one-way ANOVA with post-hoc Tukey HSD (** $p < 0.01$) C) Schematics of mosaic VSGs from clones isolated after DNA breaks within AnTat1.1. Representative sequences shown. D) A histogram of donor VSG insertion lengths identified in all mosaic VSG clones isolated. The insert length only includes newly inserted sequence and does not include recombination sites. E) A schematic of the AnTat1.1 family aligned to the AnTat1.1 transcript. Gray sequences are a perfect match to AnTat1.1. F) A histogram of the length of the shared identity between AnTat1.1 and the donor VSG at the recombination sites. G) Quantification of types of VSGs within the VSGnomes from EATRO1125 and Lister427 parasites. The Lister427 VSGnome has 12 VSGs which are duplicated but lack diversification. SL = 5′ splice leader cap, 14-mer = 3′ sequence conserved in all VSG transcripts

Mosaic formation relies on a template for break-induced VSG diversification

Although most AnTat1.1-derived mosaics were identical to the putative donor VSG, many of these insertions altered only a few bps in AnTat1.1. We thus reasoned that it was possible that these events were *de novo* mutations created during DNA repair that happened to match the putative genome-encoded donor VSG. To test this possibility, we engineered dox-inducible Cas9-expressing Lister427 parasites and replaced VSG-2 with AnTat1.1 at the active expression site, BES1. Lister 427 has a different repertoire of VSGs compared to EATRO1125[2,20]. While AnTat1.1 is not endogenously present, there are four VSGs nearly identical to the AnTat1.1 family member VSG-2986 (99.5%, 98%, 97.5%, and 95.5% nucleotide sequence identity) (3A) that could serve as donor VSGs after a break in AnTat1.1. We detected hundreds of recombination events in two independent clones (clone 1 = 398, clone 2 = 833; 3C). Upon cutting, all mosaic recombination events detected utilized donor VSGs exclusively found in the Lister427 genome (3B), indicating that the repair events that generate mosaic VSGs require a template.

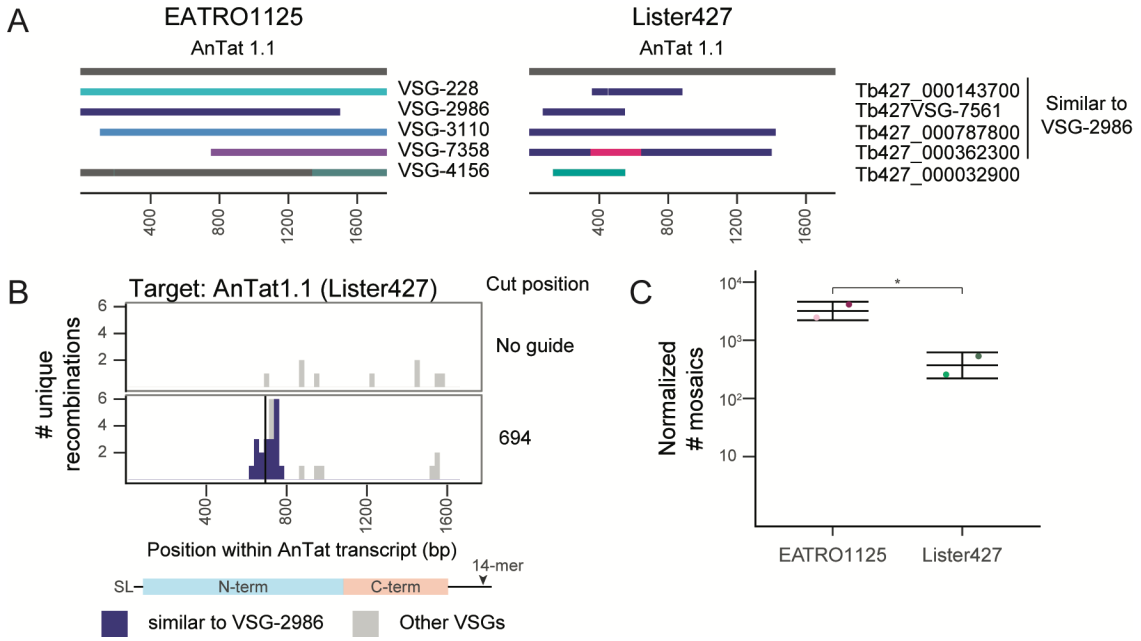


Figure 3. Nucleotide changes observed in mosaic VSGs arise from templated insertions
A) Schematics of the EATRO1125 AnTat1.1 VSG family and the VSGs similar to AnTat1.1 found within Lister427. Each is aligned to the AnTat1.1 transcript. For regions that are identical between two donor VSGs, regions are shown in the same color. B) A histogram of unique recombination events identified within Lister427 expressing AnTat1.1 parasites after a cut at position 694 along the AnTat1.1 transcript. The midpoint of the perfect homology between AnTat1.1 and the donor VSG at the recombination site is plotted. All AnTat1.1 family members except VSG-2986 were included as potential identifiable donor VSGs (all Lister VSGs + VSG-228, VSG-3110, VSG-7358, and VSG-4156). C) Quantification of mosaic recombination events induced by Cas9 at position 694. EATRO data is from 2A and 2B. The number of recombination events detected within 250bp up or downstream of the cut site was normalized to the number of total unanchored reads aligning within that region compared to the unanchored read count from the region with the smallest coverage from the EATRO cut sites to control for sequencing depth. SL = 5′ splice leader cap, 14-mer = 3′ sequence conserved in all VSG transcripts

260 **Mosaic VSGs generated *in vitro* are identical to a subset of mosaic VSGs formed *in***
261 ***vivo***
262
263 To investigate whether the mosaic VSGs we detected after break induction *in vitro*
264 reflected the events that occur naturally *in vivo*, we performed VSG-AMP-seq on parasites
265 isolated from wildtype mouse infections, on day 15 post-infection, when AnTat1.1 has
266 been mostly eliminated from the blood (Supplemental Figure 4B & 4C). We observed a
267 strong C-terminal bias to all recombination events (316 recombination events across 7
268 mice). Very few sequences aligned to AnTat1.1 within the N-terminus, suggesting that
269 most of these mosaic VSGs were complete replacements of the VSG N-terminal domain
270 (4A; Supplemental Figure 4A). These donor VSGs shared significantly less homology
271 with AnTat1.1 than the AnTat1.1 family members, usually only sharing spans of 100bps
272 or less of imperfect homology within the C-terminal region of the VSGs. Nevertheless,
273 there was still a short (average = 12 bp) span of perfect identity between AnTat1.1 and
274 the donor at the recombination site (4E). Since VSGs expressed by parasites at the
275 second peak of parasitemia are antigenically distinct from AnTat1.1, we reasoned these
276 N-terminal replacement VSGs ensured complete immune evasion since only a small
277 portion of AnTat1.1 was retained.
278
279 We wondered whether this restriction within mouse infections reflected a
280 mechanistic bias towards recombination within the VSG C-terminal domain or the effect
281 of host antibody selection, with only the most evasive recombination events surviving. To
282 investigate these possibilities, we repeated the experiment in μMT mice[21], which do not
283 have B-cells and therefore do not generate antibodies. We again analyzed day 15 post-
284 infection, though AnTat1.1 is never cleared from these infections and parasitemia remains
285 high throughout (Supplemental Figure 4B & 4C). In μMT mice, mosaic recombination
286 events span the full length of the VSG, and many are identical to mosaic VSGs observed
287 *in vitro* (4A; Supplemental Figure 4A) (2144 recombination events across 7 mice). *In vivo,*
288 we observe short insertions (average = 41 bp) using donors homologous to AnTat1.1
289 flanked by short (average = 13 bp, median = 9 bp) regions of perfect identity, matching
290 the patterns we observed *in vitro* (4E & 4F). Interestingly, as in the wildtype infections,
291 there are an additional subset of recombination events within the C-terminus which are
292 largely absent *in vitro* and use a much more diverse set of donors (4B). These data
293 demonstrate that our *in vitro* system recapitulates a subset of *in vivo* recombination
294 events, but there may be other pathways facilitating recombination *in vivo* in addition to
295 those that can be triggered by a single, blunt DNA double-strand break.
296
297 **Mosaic VSGs form in extravascular spaces during infection**
298
299 While AnTat1.1 is eliminated from the blood in wildtype mice, it persists within
300 tissues until at least day 14[4]. Given that parasite clearance is delayed in extravascular
301 spaces, we wondered if mosaic formation may occur more readily in this parasite niche.
302 To investigate this, we analyzed all assembled VSGs from a previous study of
303 extravascular parasite populations[4]. Again, we found AnTat1.1 mosaics present in tissues
304 identical to those observed in both μMT mice and *in vitro* (4C & 4G). Moreover, we
305 observe that mosaic derivates of AnTat1.1 increase over time within tissue spaces during

infection (4D). Together, these data suggest that mosaic VSGs form preferentially within extravascular spaces, possibly due to the slower VSG-specific parasite clearance in these spaces.

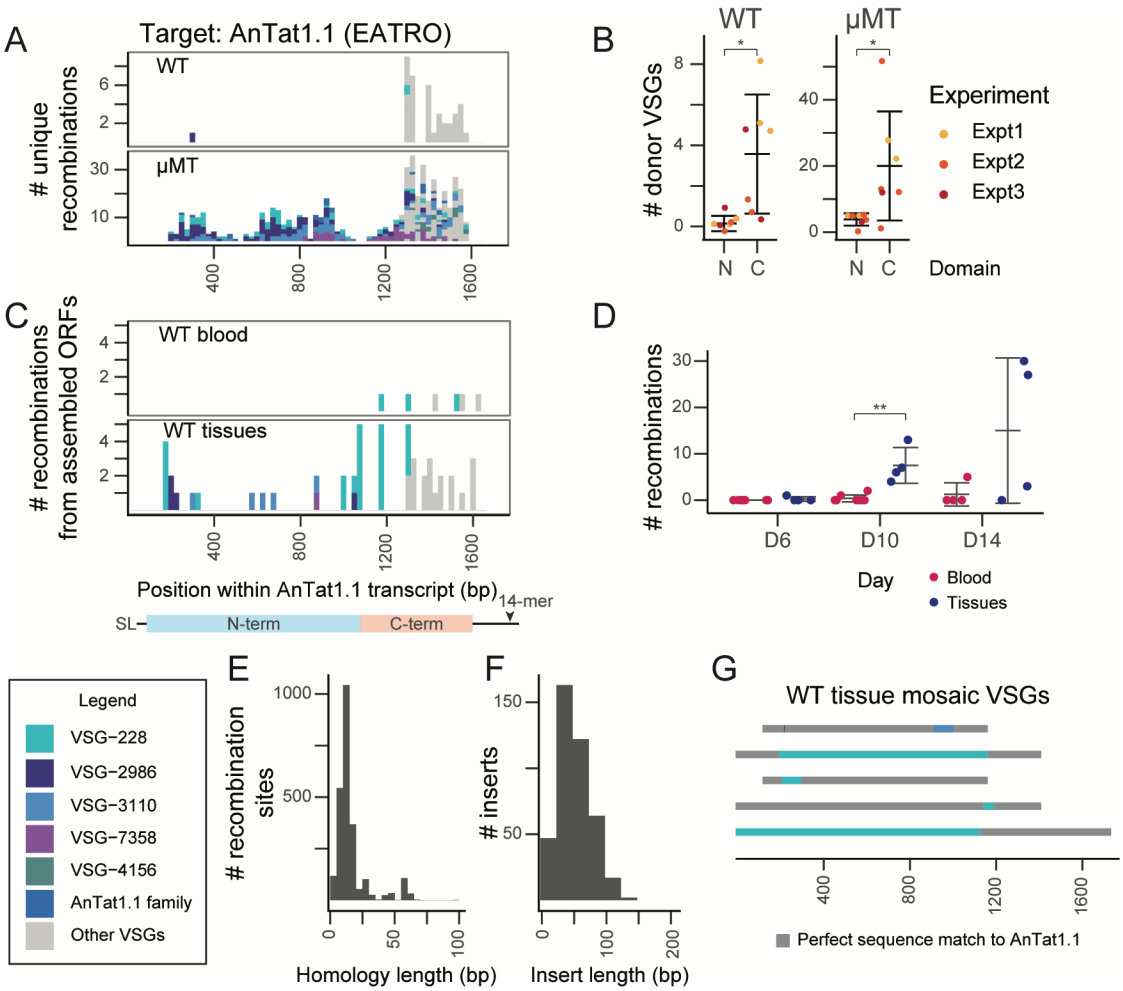


Figure 4. AnTat1.1 mosaic VSGs form *in vivo* and reside within tissues of WT mice
A) A histogram of unique recombination events identified in all mice in wildtype or μMT mice from Day 15 post-infection. Recombination events found in multiple mice are represented once. (WT n=7; μMT n=7, from 3 independent experiments) The midpoint of the perfect homology between AnTat1.1 and the donor VSG at the recombination site is plotted. If a donor VSG could not be unambiguously identified, the average recombination position was plotted. B) Quantification of the number of donor VSGs utilized in mosaic recombination events within the N and C terminal domain of AnTat1.1 in wildtype and μMT mice from A). Statistical significance determined by a pairwise Wilcoxon test (*p<0.05) C) A histogram of the mosaic recombination events identified from VSG ORFs assembled in Beaver et al. All recombination events identified are shown at all time points sampled during infection: D6, D10, and D14. If a mosaic recombination event was identified in more than one tissue within the same mouse, it was counted once (n=12, 4 mice per tissue time point). D) Quantification of the mosaic recombination events detected within wildtype mouse blood or tissue. Statistical significance determined by pairwise Wilcoxon within each timepoint. (**p<0.01) E) A histogram of the length of the shared identity between AnTat1.1 and the donor VSG at the recombination sites. F) A histogram of donor VSG insertion lengths identified within individual reads. The insert length only includes newly inserted sequence and does not include recombination sites. G) Representative schematics of mosaic VSGs from assembled ORFs in Beaver et al. SL = 5′ splice leader cap, 14-mer = 3′ sequence conserved in all VSG transcripts

309 **Small changes in the VSG sequence provide substantial antibody evasion**
310
311 Many AnTat1.1-derived mosaic VSGs differ from their original sequence by only a
312 few amino acids. To determine how these changes impacted the antigenic character of
313 the mosaic VSGs, we analyzed live parasites expressing AnTat1.1 mosaic derivatives by
314 flow cytometry using a potent rabbit anti AnTat1.1 polyclonal antibody raised against
315 purified AnTat1.1 protein (5A). Many of the mutations block about 50% of antibody
316 binding. Interestingly, three VSG-2986 mosaic clones were isolated with increasing
317 insertion length (5B). The smallest insertion (18bps, with 3 a.a. changes, Mosaic 2),
318 appears to account for most of the antigenic change, with only small decreases in binding
319 associated with each additional length of insertion. We modeled the structure of these
320 mosaics using ColabFold[22] and their general structures matched AnTat1.1 (5C), except
321 at the disordered top of the N-terminal lobe. A careful examination of the locations of the
322 mutations on AnTat1.1 show that mutations from Mosaic 2 which have the largest effect
323 on antibody binding are at the apex of the structure while other alterations can be found
324 on the side of the monomer (5D). These results demonstrate that, although the insertions
325 characteristic of mosaic recombination are quite short, even these small changes to the
326 VSG can confer large consequences for host antibody binding depending on their
327 position.
328

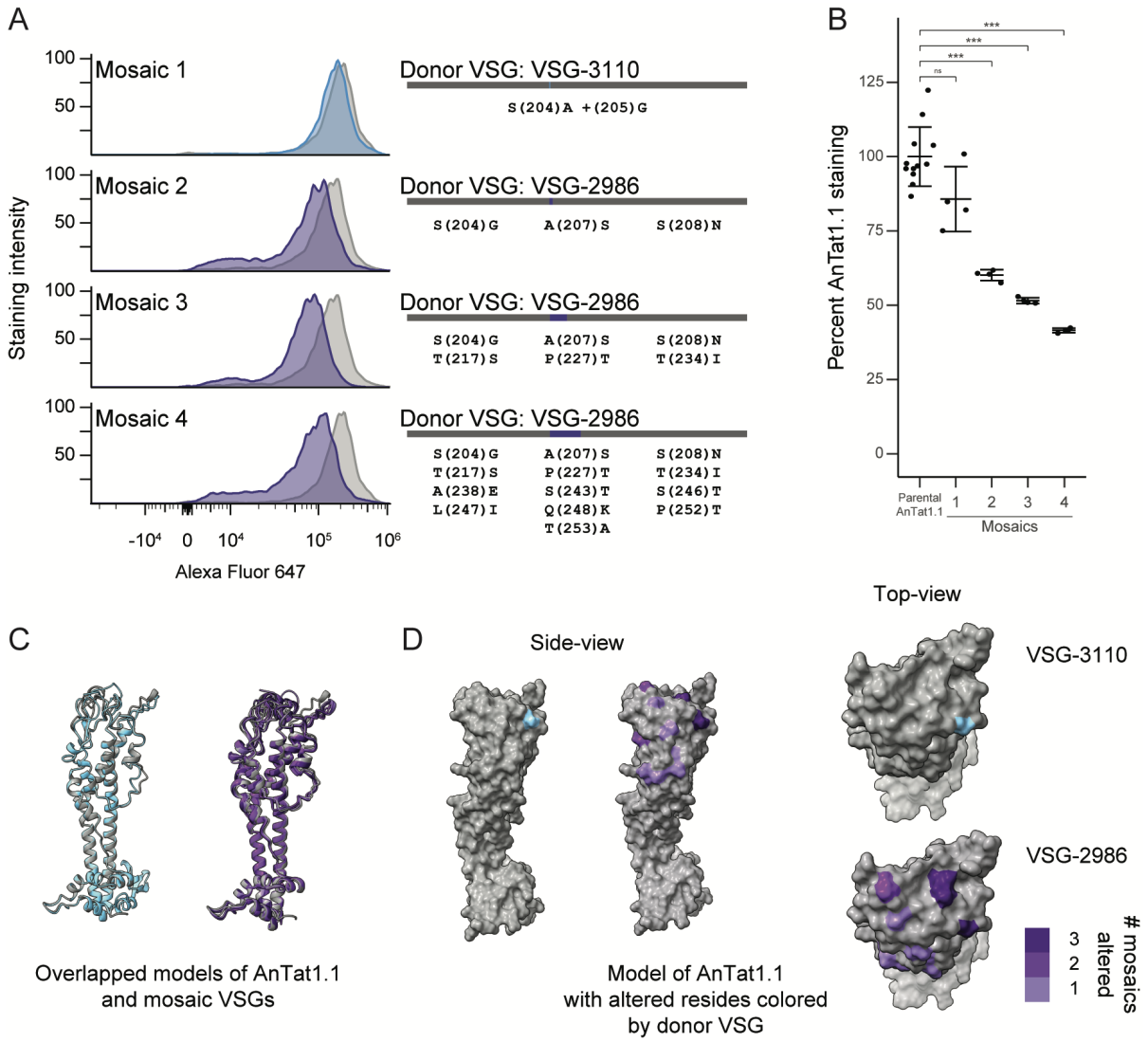


Figure 5. Mutations within VSG hypervariable region shift antibody binding
A) A histogram showing the Alexa Fluor 647 staining intensity for parental controls (gray) and mosaic clones colored by donor VSG. (n= 4, from 3 independent parental lineages derived from 2 Cas9 clones) Schematics showing the mosaic VSGs are to the right. Donor VSG and amino acid substitutions are specified. B) Quantification of staining intensity changes for individual clones, based on median staining intensity. The median Alexa Fluor 647 intensities were normalized to the average of the parental clone. Statistical significance was determined based on a one-way ANOVA with a post-hoc Tukey HSD. (***p<0.001) C) Overlapping ribbon structures of AnTat1.1 and Mosaic 1 in blue and AnTat1.1 and Mosaic 2, Mosaic 3, and Mosaic 4 in purple as predicted by ColabFold. D) A space filling model of AnTat1.1 highlighting the changed residues within the monomer. VSG-3110 in blue and VSG-2986 in purple.

Discussion

Chronic *T. brucei* infection relies on the generation of new VSGs, with novel VSGs dominating late stages of infection[3,5,6]. However, the new VSGs generated during infection are complex, possibly arising from several intermediate events, and selected within a host environment with immunity against numerous previously expressed VSGs. The inherent complexity of chronic infection thus obscures the underlying biological principles driving VSG diversification. Here, we demonstrate that VSG diversification can be induced *in vitro* using Cas9-mediated double-strand DNA breaks within the VSG coding sequence, reproducibly generating mosaic VSGs that faithfully recapitulate mosaics formed naturally *in vivo*. By selecting just one VSG and looking at thousands of recombination outcomes, we have defined patterns characteristic of mosaic recombination. Mosaic VSGs typically form through short, templated insertions, and homology drives this process, restricting donor VSGs within the N-terminus to those from a set of closely related family members. Finally, we demonstrate that mosaic VSGs provide substantial immune evasion, particularly when these changes occur at the top of the VSG N-terminus, which may reflect a hypervariable region within the VSG.

We have shown that DNA breaks, previously shown to drive VSG switching[16,17], can also result in VSG diversification through mosaic VSG formation. While the exact DNA repair mechanism that generates mosaics is unclear, the patterns we observe—short templated-insertions relying on short stretches of sequence homology—suggest a mechanism similar to microhomology-mediated end joining (MMEJ). MMEJ is known to function in *T. brucei* and is typically RAD51-independent[23]. In line with this, there is evidence for a RAD51-independent pathway capable of driving antigenic variation in the parasite[24]. If an MMEJ-mediated small insertion occurs following a blunt, double-strand break, other types of breaks, like 5′ or 3′ staggered double-strand breaks, may result in alternative repair outcomes[25]. In B-cells, which use DNA breaks to undergo class switching and somatic hypermutation, the type of DNA break appears to dictate repair pathway choice[26,27]. It is thus possible that the N-terminal replacements we observe *in vivo*, but not *in vitro*, originate from multiple breaks, or staggered breaks with single-strand overhangs, that are repaired through a pathway distinct from the one driving repair of Cas9-induced breaks *in vitro*.

While our data suggest most (75%) VSGs in the archive are capable of diversification through the mechanism we have described, this is not the case for all VSGs. Indeed, we find that mosaic recombination occurs only if a homologous donor VSG is available. Most studies of antigenic variation in *T. brucei* have focused on VSG-2, a unique VSG that does not readily diversify; this may explain, at least in part, why mosaic recombination has been observed so rarely *in vitro* until now. The proportion of VSGs within families is surprisingly similar between the EATRO1125 and Lister427 strains. Perhaps this ratio has evolved to provide a balance between diversification events, which may not always provide full immune evasion, and switching, which is more likely to completely escape pre-existing immunity.

While the members of the AnTat1.1 VSG provide numerous possibilities for

recombination and the generation of new VSGs, we found that the center portion of the gene appears to be especially successful at generating recombinants. We considered that this might be attributable to available donor sequences or guide efficiency, but there was nothing apparently unique about this site compared to other break locations. Protein modeling of AnTat1.1, however, revealed that this region encodes the top of the N-terminal lobe of the VSG, a site directly exposed to host antibody. For AnTat1.1 and many similar VSGs[2] (A2-type), this is an unstructured region[28], and indeed the apex of most crystalized VSG appears unstructured despite a surprising variety of underlying structures[28]. We propose that this disordered region may have evolved to be more tolerant to amino acid changes, allowing diversification within the part of the protein most likely to facilitate immune evasion. While many recombination events may occur after a DNA break, we hypothesize that only a few, particularly those within disordered regions, maintain VSG structural integrity. Since expression of the VSG is essential for cell cycle progression[29], it is possible that many newly formed mosaic VSGs lead to cell death for the parasites, allowing only a subset of the recombination events to survive and resulting in an apparent increase in recombination within the regions of the VSG where mutation is most tolerated.

Although mosaic insertion events are often small (average ~46 bp) and can result in few amino acid changes, these small changes, especially within the top of the N-terminal lobe, confer substantial immune evasion. Just three amino acid substitutions, from an 18 bp templated replacement, block approximately 50% antibody binding from a potent polyclonal antiserum. While we have shown that these recombination events can result in dramatic antigenic changes, many are also likely to result in incomplete immune evasion. Perhaps as a result, mosaic VSGs derived from AnTat1.1 in WT blood appear to be mostly N-terminal replacements, where the AnTat1.1 N-terminal domain is completely replaced. A wider diversity of AnTat1.1-derived mosaics can be observed within the tissues, however, where AnTat1.1-expressing parasites linger due to differences in immune pressure[4]. In μMT mice, where AnTat1.1 expressing parasites can linger indefinitely, this observation is further amplified. We thus hypothesize that generating new VSGs is an iterative process, where a series of recombination events progressively shifts the character of a VSG until an immunologically distinct variant is formed. The relatively “protected” tissue spaces may serve as a site for this iterative process to occur. Another intriguing possibility is that partial immune evasion is advantageous for the parasite: a recent study suggested that exposure to sublethal antibody concentrations can trigger a VSG switch in *T. brucei*[30]. Perhaps parasites expressing these partially evasive variants are more prone to switching to a completely new VSG.

In addition to shedding light on the selective pressures imposed upon mosaic variants, our mouse data also demonstrate that our *in vitro* model of antigenic diversification recapitulates VSG diversification *in vivo*. The recombination events we detect in μMT mice represent the full breadth of possible recombination events which AnTat1.1 can form, and these are largely represented within our *in vitro* break-induced mosaic VSG populations. We observe short insertions flanked by short regions of identity in both contexts, but the dominance of the 694 break position we observe *in vitro* is not

422 reflected in our *in vivo* analysis. This could be for a variety of reasons, including non-
423 uniform break locations along the VSG *in vivo* or a variety of DNA break types driving a
424 different pattern of outcomes. Intriguingly, in the absence of an exogenous DNA break,
425 we do not observe any diversification *in vitro*, while diversification occurs continuously in
426 the μMT context. This suggests that some aspect of the host environment other than
427 antibody pressure induces diversification *in vivo.* It remains to be investigated whether an
428 *in vivo* cue for diversification exists.
429
430 The diversification mechanisms we have described may also be at play in other
431 gene families where diversity is critical. Many pathogens express variable antigens, with
432 the antigenic repertoire stored within subtelomeric regions of the genome where
433 continuous DNA damage might facilitate diversification. Mosaic antigen genes have been
434 described in plasmodium var genes[31–34] and giardia variant-specific surface proteins
435 (VSPs)[35,36], some of which resemble the small insertion mosaics we have observed here.
436 Alternative end joining pathways appear to be intact in these organisms[37], and may
437 facilitate mosaic formation. Many alternative forms of DNA repair are also conserved in
438 mammals[38]; it is thus plausible that large gene families like olfactory receptors[39,40] and
439 protocadherins[39], which depend upon diversity and are known to swap sequences
440 through gene conversion events[41–44], may also be diversifying through similar
441 mechanisms. A rigorous targeted approach, like we have used here, may be required to
442 determine the role this mechanism plays in diversification of other gene families.
443
444 Here we answer long-standing questions about how the *T. brucei* VSG repertoire
445 evolved to generate new antigens critical for maintaining a chronic infection. Our high-
446 throughput, sensitive technique has enabled the characterization of thousands of mosaic
447 diversification events, orders of magnitude more than was previously feasible. This
448 analysis revealed a pattern of short, homology-driven, and templated insertions around
449 DNA break sites that can shift the antigenic character of the VSG. More broadly, this study
450 provides an experimental framework for the hypothesis-driven exploration of antigen
451 diversification in *T. brucei* as well as in other pathogenic microbes.
452

Methods

Parasites:

Pleiomorphic EATRO1125 AnTat1.1 90-13 *T. brucei* parasites were maintained in HMI-9 media with 10% heat inactivated FBS and 10% Serum Plus. Parasites were passaged when they reached approximately $5*10^5$ cells/mL. Monomorphic Single Marker Lister427 VSG221 TetR T7RNAP bloodstream form (NR42011; LOT: 61775530)[45]. *T. brucei* were maintained in HMI-9 up to $1*10^6$ parasites/mL. VSG221 has since been renamed to VSG-2.

Plasmids:

Plasmids were synthesized with Gibson Assembly using the following templates, when specified. Gibson Assembly master mix was custom made[46]. Whole plasmid sequencing was performed by Plasmidsaurus using Oxford Nanopore Technology with custom analysis and annotation.

pLEW100v5-BSD-FLAG-La-Cas9 was synthesized from pLEW100v5-BSD and pRPaCas9[47]. A *T. brucei* codon optimized FLAG tag was added to the N-terminus of Cas9. pLEW100V5-BSD was a gift from George Cross (Addgene plasmid # 27658 ; http://n2t.net/addgene:27658 ; RRID:Addgene_27658). pRPaCa9 was a gift from David Horn (Addgene plasmid # 111819 ; http://n2t.net/addgene:111819 ; RRID:Addgene_111819).

pHH-HYG-AnTat1.1 and pHH-HYG-VSG-228 were synthesized from pHH-HYG-VSG-3-S317A.[48] This was a gift from Joey Verdi. AnTat1.1 sequence was obtained from AnTat1.1 specific cDNA from mouse infection D6 cloned into a pMiniT vector with the PCR Cloning Kit (NEB, E1202S). VSG-228 was partially amplified from VSG PCR (see below) derived from AnTat1.1 depleted D16 μMT parasites from blood. The partial fragment was amplified with an AnTat1.1 family specific forward primer (5' -ACTACACCCACAACAAGCTCTA-3') and a pan VSG reverse primer which binds to a conserved 14bp region of the 3' UTR (5'-GATTTAGGTGACACTATAGTGTTAAAATATATC-3') with AmpliTaq Gold (Applied Biosystems, 4398881) (anneal & extension 60C 1m, 35 cycles).The resulting PCR was cloned into the pMiniT backbone and the remainder of VSG-228 was *de novo* synthesized with Gibson Assembly.

The pT7sgRNA plasmids were obtained according to Rico et al[47]. Briefly, sticky-end annealed sgRNAs were ligated into a BbsI-HF (NEB, R0539S) digestion of pT7sgRNA. Supplemental Table 1 contains the sequences for the annealed guides. pT7sgRNA was a gift from David Horn. (Addgene plasmid # 111820 ; http://n2t.net/addgene:111820 ; RRID:Addgene_111820)

Transgenic parasites:

To obtain all transgenic parasites, 5 million parasites were electroporated with 5-10ug of digested plasmid with an AMAXA Nucleofector II using X-001 in Human T-cell Nucleofector Solution (Lonza VPA-1002).

For tetracycline-inducible Cas9 parasites, EATRO1125 or Lister427 parasites were electroporated with pLEW100v5-BSD-FLAG-La-Cas9, digested with NotI-HF (NEB, R3189S). Parasites were immediately selected in 5ug/mL Blasticidin (Thermo Scientific, R21001) and maintained in selection unless otherwise specified.

To obtain AnTat1.1 and VSG-228 expressing Lister427 parasites, Lister427 parasites were electroporated with pHH-HYG-VSG plasmids digested with BamHI-HF (NEB, R3136S). After 16-24 hours recovery, 25ug/mL of hygromycin (Fisher Scientific, J67371-XF) was added to the culture. After obtaining colonies 5-7 days later, selection was reduced to 5ug/mL and maintained unless otherwise specified.

Lister427 parasites with pLEW100v5-BSD-FLAG-La-Cas9 inserted were additionally electroporated with pHH-HYG-VSG plasmids expressing Antat1.1. Clones were obtained as above.

To obtain constitutive guide expressing parasite lines, parasites were electroporated with pT7sgRNA guide containing plasmids digested with NotI-HF[47]. After 16-24 hours recovery, single colonies of parasites were selected with 2ug/mL phleomycin (Sigma-Aldrich, SML3001). Parasites were maintained without phleomycin selection.

T7-Guide Synthesis & Purification:

DNA fragments containing a T7 promoter and guide RNA sequence were synthesized as described previously[49]. Briefly, T7-guide primers and a G00 primer were amplified with Phusion polymerase for 35 cycles (annealing temp 60C, extension 5s). PCR products from 12 identical PCRs were pooled and purified via ethanol precipitation.

Cas9 Transient Electroporations:

Approximately 24 hours prior to electroporation, in a single flask, 1 million parasites in 12mLs of media per sample was seeded and induced with 1ug/mL doxycycline (Millipore Sigma, D9891-1G). Only blasticidin selection (Cas9) was maintained and for Lister427 parasites expressing AnTat1.1, hygromycin (active VSG expression) was removed at this stage. Either 8mLs (Lister427) or 10mLs (EATRO1125) of parasites was spun down, media removed, and parasites were resuspended in 100uL Human T-cell Nucleofector Solution. Each sample was electroporated using the X-001 program on the AMAXA Nucleofector II with approximately 1-1.5ug of purified T7-guide in a volume less than 10uL or a sample without any DNA as a negative control. Parasites were moved into 5mLs HMI-9 in 6-well plates to recover for 30 minutes then moved into 20mLs total in flasks to recover overnight. About 24 hours after electroporation, parasites were counted and split. For EATRO1125, 12 million cells (or all the cells if there were fewer than 12million) were seeded into 60mLs total with blasticidin. For Lister427, 2 million cells were seeded into 20mLs total with blasticidin. At the 48 hour time point, parasites were counted, collected, and stored in TRIzol (Invitrogen, 15596026) for subsequent RNA extraction.

Isolation of mosaic expressing parasites:

Multiple clones of parasites expressing T7sgRNA with guides were derived from two distinct inducible Cas9-expressing parasite clones. Dilutions of parasites were plated in 96-well plates in 5ug/mL blasticidin and 1ug/mL doxycycline. Individual clones were visible after 7-14 days. Upon isolation of a parasite clone, all drugs were removed. Parasites were analyzed by flow cytometry. VSG sequences for clones were determined

545 from extracted RNA. cDNA was synthesized using the Superscript III Reverse
546 Transcriptase (Invitrogen, 18080051) and a VSG-specific primer which binds to a
547 conserved 14-bp sequence within the 3' UTR. (5'-GTGTTAAAATATATC-3'). 2uL of
548 RNase-treated cDNA was amplified for 35 cycles with VSG specific primers: a spliced-
549 leader (5'-ACAGTTTCTGTACTATATTG-3') and SP6-VSG 14-mer (5'-
550 GATTTAGGTGACACTATAGTGTTAAAATATATC-3') using Phusion polymerase
551 (Thermo Fisher, F530L) (annealing temp 55C, extension 45s). PCR products were
552 cleaned using the Monarch PCR & DNA cleanup kit (NEB, T1030L). VSG sequences
553 were determined by amplicon nanopore sequencing performed by Plasmidsaurus using
554 Oxford Nanopore Technology with custom analysis and annotation or Sanger sequencing
555 with the sequencing primer (5'- AGAGAATACTAAGCTAGTTGGC-3') performed by
556 Azenta Life Sciences.
557
558 **Mouse Infections:**
559 C57BL/6 mice and μMT^{-} (B6.129S2-Ighm^{tm1Cgn}/J)[21] mice were delivered at 8
560 weeks old from Jackson Labs. Approximately 5 EATRO1125 parasites were injected into
561 mice intravenously through the tail vein. These parasites express either AnTat1.1 or VSG-
562 421. Starting at D4 post-infection, parasitemia was monitored within mice every two days
563 via tail bleed. Blood was harvested at D6 post-infection through a submandibular bleed.
564 An additional gel pack and in-cage food pellets were provided to mice during recovery
565 since it helped μMT mice survive the blood loss. At D15, post-infection, mice were
566 humanely euthanized with ketamine injection and blood was harvested through cardiac
567 puncture. Blood was stored in TRIzol LS (Invitrogen, 10296028) for RNA extraction.
568
569 **RNA preparation:**
570 Parasites were stored in TRIzol prior to RNA extraction. Blood with parasites was
571 stored in TRIzol LS. RNA was extracted via phenol/chloroform extraction according to the
572 manufacturers protocol. Purified RNA was DNase treated with Turbo Dnase (Thermo
573 Fisher, AM2239) and purified with 1.8X Mag-Bind TotalPure NGS Beads (Omega Bio-
574 tek, M1378-01). Verification of effective DNase treatment was performed via PCR of
575 Hygromycin (EATRO1125 only) (Fwd: 5'-ACAGCGGTCATTGACTGGAG-3'; Rev: 5'-
576 ATTTGTGTACGCCCGACAGT-3', annealing temp 52C, extension 30s) or HSP70
577 (Lister427 & EATRO1125) (Fwd: 5'-AGAACACTATCAATGACCCCAAC-3'; Rev: 5'-
578 CCATGCCCTGGTACATCT-3', annealing temp 50C, extension 15s) genes for 30 cycles
579 using OneTaq DNA Polymerase (NEB, M0480L).
580
581 **VSG-seq:**
582 VSG-seq was performed as described elsewhere[3,4]. Briefly, cDNA was
583 synthesized using the Superscript III Reverse Transcriptase and a VSG-specific primer
584 which binds to a conserved 14-bp sequence within the 3' UTR. (5'-GTGTTAAAATATATC-
585 3'). cDNA was treated with Rnase A (Quiagen, 19101) and RNase H (Invitrogen,
586 18080051) for 30 minutes then purified with 1.8X Mag-Bind Total NGS Beads (Omega
587 Bio-tek, M1378-01). Purified cDNA was amplified for 25 cycles with VSG specific primers:
588 a spliced-leader (5'-ACAGTTTCTGTACTATATTG-3') and SP6-VSG 14-mer (5'-
589 GATTTAGGTGACACTATAGTGTTAAAATATATC-3') using Phusion polymerase
590 (Thermo Scientific, F530L) (annealing temp 55C, extension 45s). This PCR product was

prepared for sequencing using the Nextera XT DNA Sample Prep Kit (Illumina, FC-131-1096) according to the manufacturer's instructions. Libraries were sequenced with 100 bp single-end reads on a NovaSeq6000. Analysis was performed using the VSGSeqPipeline found at github.com/mugnierlab.

VSG-AMP-seq library preparation:

VSG-AMP-seq was based upon AMP-seq[50] and GUIDE-seq[51]. cDNA was synthesized from DNA-free RNA using Superscript III Reverse Transcriptase and a VSG-specific primer which binds to a conserved 14-bp sequence within the 3' UTR. (5'-GTGTTAAAATATATC-3'). Second Strand synthesis was performed with NEBNext mRNA Second Strand Synthesis Module (E6111L). The resulting double stranded cDNA was purified with purified with 1.8X Mag-Bind Total NGS Beads. cDNA was fragmented briefly with NEBNext dsDNA Fragmentase (M0348L) for ten minutes at 37C to obtain fragments of approximately 500bps. Fragments were purified with a double-sided Mag-bind bead cleanup. First, fragments were incubated with 0.5X beads, DNA bound to the beads were discarded and an equal volume of beads as before, then a 1X PEG concentration, was added to the fragments and the cleanup proceeded as normal. Fragments were end-repaired with Enzymatics low concentration end repair mix (Y9140-LC-L) and A-tailed with Recombinant Taq (Life Technologies, 100021276). Y-Adapters were pre-annealed by incubating a MiSeq Common Adapter with Adapters (A01-A10) containing 25 bp Unique Molecular Indexes (UMIs) and a barcode at 10uM at 95C for 1 sec, 60C for 1s, and slowly cooled to 4C. Adapters were ligated to the A-tailed fragments with T4 DNA Ligase (Enzymatics L6030-LC-L). The resulting fragments were cleaned up with 0.8X Mag-Bind beads. Two target specific PCRs were performed on the fragments with pools of target specific primers, one in the forward and one in the reverse direction. All primers are listed in Supplemental Table 1. Target primers have spacers of varying lengths to generate sequence diversity without the need for PhiX. Target specific primers are paired with P5_2 and a sample specific P primer (P701-P710) which contains a second barcode. Fragments are amplified with Platinum Taq (Life Technologies, 10966018) using the following program: 95C 5 mins, 35 cycles of 95C for 30s, 55C for 30s, and 72C for 30s, and 72 for 5 mins. The resulting products were cleaned up with 0.6X Mag-bind beads. Libraries were quantified using the Qubit dsDNA HS Kit (Life Technologies, Q32854) and run on a 1% agarose gel to determine average length. Libraries were sequenced on a MiSeq or for deep sequencing of μMT samples on a NovaSeq6000 with custom index1 and read2 primers using the following cycle conditions: "151|8|33|131" with the paired-end Nextera sequencing protocol.

VSG-AMP-seq analysis:

Index sequences were added to the names of sequencing reads. FASTQ files were demultiplexed by target specific primer using cutadapt v=3.5[1] searching for multiple target-specific primers at the 5' end of read2 in paired-end mode using flags --action=retain --overlap 10. Then, reads were quality trimmed with trim_galore v0.6.4_dev (github.com/FelixKrueger/TrimGalore) where adapter sequences, if present, were removed with flags --dont_gzip --paired --trim1. Spacers were removed from read1 if present using cutadapt by searching for the target-specific primer sequence at the 3' end of the read with flags --action=retain --overlap 10. Dual-indexed barcodes were used to

637 demultiplex reads. Mismatch allowed was determined by the barcodes present, so that a
638 unique barcode could still be identified.
639
640 To consolidate reads, UMIs were extracted from large FASTQ files, split into
641 smaller files and grouped by 100% identity using cd-hit-est v4.8.1[53,54] with flags -c 1.0 -n
642 8 -M 16000 -d 0. UMIs were then grouped by 92% using flags -c 0.92 -n 8 -M 16000 -d 0.
643 Reads were sorted into consensus groups based on UMIs and a consensus sequence
644 was formed as the most popular base at each position of the read with a threshold quality
645 score of at least 2. If there was a tie, an N was used. Clusters with fewer than 3 reads
646 were removed.
647
648 The full sequence length of the AnTat1.1 transcript was determined by
649 Plasmidsaurus using Oxford Nanopore Technology with custom analysis and annotation
650 and the 5'-end was appended with alternative splicing information from control samples.
651 The positions were 0 indexed. Read2, the anchored read, was aligned to AnTat1.1. Read
652 pairs were removed if: the first 4/10 bases after the primer did not align to AnTat1.1, an
653 alternative VSG which contained the primer sequence was amplified, the wrong position
654 of AnTat1.1 was amplified by the anchor, if the reads were too short (<15 bps), contained
655 too many Ns (>5), if there was an inversion event or a duplication, or if the transcript was
656 alternatively spliced. Read1, the unanchored read, was also aligned to AnTat1.1 with a
657 mismatch of up to 1bp. If no alignment could be found, read1 was trimmed to remove
658 AnTat1.1 like sequences from the 5' and 3' ends, allowing up to 1 mismatch on each end
659 to be removed, leaving a fragment which can be used to search the VSGnome and
660 identify potential donor VSGs. A consensus sequence was generated from the reads, if
661 they overlapped, using their alignment positions within AnTa1.1. AnTat1.1, the consensus
662 read, and putative donor VSG were aligned, and recombination sites were identified. If
663 no consensus could be generated, read1 was used. Ambiguous recombination sites
664 where a single donor could not be identified were represented by the average position of
665 the potential recombination sites and identified as ambiguous.
666
667 **VSG Clustering and family identification:**
668 FASTA files of genes encoding VSGs from the Lister427 strain were derived from
669 TrypsRU (vsgs_tb427_all_atleast150aas_cds.txt, vsgs_tb427_
670 nodups_atleast250aas_cds.txt, vsgs_tb427_nodups_atleast250aas_cdsplusflanks.txt)
671 (George Cross) and TriTrypDB[55] (TriTrypDB-65 Lister Strain 427 2018)[20]. A combined
672 master FASTA file was standardized to include important attributes as part of the
673 sequence name (VSG name, GB Accession Code, Nucleotide Length, Status of
674 Sequencing, Chromosomal Location, Source, Parasite Line (Lister427), and any other
675 Extra Information) and the corresponding sequence. From the 2018 genome, the genes
676 encoding for VSGs were isolated by looking for the keywords, "variant" or "VSG" in the
677 description attribute. The genomic sequence for these VSG genes was extracted from
678 the Genome file and included 300 bp of flanking sequence up and downstream for further
679 analysis. Identical VSGs were clustered using cd-hit-est, with the following flags: -c 1.0 -
680 n 8 -M 16000 -d 0. A representative sequence, usually from the 2018 genome with the
681 flanking sequences, was selected for each VSG and was the longest sequence available.
682 Duplicate VSGs identified at two different chromosomal locations within the 2018 genome

683 were added to the file to represent all known Lister427 VSGs. AnTat1.1 was added to the
684 master Lister427 VSG file for further analysis.
685 EATRO1125 VSGs were obtained from TrypsRU (George Cross) and included 200
686 bp flanking sequences. (vsgs_tb1125_all_atleast150aas_cdsplusflanks.txt)
687
688 Multiple sequence alignment was performed using MAFFT[56], using the following
689 flags: --auto --reorder. Following multiple sequence alignment, a phylogenetic tree was
690 constructed using a generalized time reversible mode in FastTree[57,58] with the following
691 flags: -nt -gtr. Clustering of AnTat1.1 and its family members was used to identify a
692 threshold genetic distance to generate subsequent subtrees. Threshold distances of
693 0.24659 and 0.21677 were used for EATRO1125 and Lister427, respectively. Down
694 sampling and subtree generation of representative sequences was performed using
695 PARNAS (0.1.4)[59]. The command used for PARNAS included flag: --cover –radius
696 <threshold distance>. From the subtree generated, the number of unique VSGs and the
697 number of VSG families were identified. Duplicate VSGs from the Lister427 genome were
698 analyzed to determine if they were within a group that lacked any diversification.
699
700 **Western Blotting:**
701 5 million parasites were spun down and washed with 25C PBS. Then, pelleted
702 parasites were resuspended in 50uL RIPA buffer (50mM Tris, 150mM NaCl, 1% NP-40,
703 0.5% Sodium Deoxycholate, 0.1% SDS, pH = 7.4) + 2X Laemmli buffer. Lysed parasites
704 were boiled at 95C for 5 minutes. 5uL of lysates were separated on a Tris-glycine
705 polyacrylamide gel at 110V for 100 minutes in Tris/glycine running buffer (25mM Tris,
706 192mM glycine, 0.1% SDS, pH = 8.3). Proteins were then transferred onto a nitrocellulose
707 membrane using transfer buffer (25mM Tris, 192mM glycine, 20% methanol) overnight at
708 a 60mAmps per transfer box at 25C. For Cas9 (1:1000) and EF1α (1:1000), lysates were
709 separated on 10% polyacrylamide gels and transferred onto 0.45um membranes. For γ-
710 H2A (1:200), lysates were separated on a 15% polyacrylamide gel and transferred onto
711 a 0.2um membrane. After transfer, blots were blocked with 0.5% BSA in TBS with 0.05%
712 Triton X-100 (TBST, 20mM Tris, 150mM NaCl, pH = 7.6) for an hour at RT under constant
713 agitation. Blots were incubated with primary antibody (see above dilutions) for 2 hours at
714 25C. After 5 TBST washes, blots were stained for 1 hour with goat anti mouse (1:5000,
715 Cell Signaling, 7076S) or goat anti rabbit-HRP-conjugated secondary (1:5000, Cell
716 Signaling, 7074S). After another 5 washes, blots were incubated with ECL (Cytiva,
717 RPN2109) and film was exposed to blots in a dark room. Developed film was scanned
718 and images were processed with FIJI[60]. Primary antibodies used were m anti FLAG (M2
719 clone) (Millipore Sigma, F3165-1MG), m anti EF1α (CBP-KK1 clone) (Millipore Sigma,
720 05-235), and rb anti γ-H2A was a kind gift from Galadriel Hovel-miner based upon Glover
721 and Horn[61].
722
723 **Flow Cytometry:**
724 In 96 well plates, 200,000 parasites were stained with 1:20,000 rabbit anti
725 AnTat1.1 primary antibody (Jay Bangs) for ten minutes at 4C while shaking in PBS +
726 10mg/mL glucose. Parasites were washed once with 100uL PBS + glucose. Then,
727 parasites were stained with Alexa Fluor 647-conjugated goat anti rabbit IgG (H +L),
728 F(ab')2 Fragment (Cell Signaling, 4414S) at 1:1000 at 4C while shaking in PBS + glucose.

After washing again with 100uL PBS + glucose, parasites were resuspended in PBS + glucose + 1:20 Propidium Iodide (BD Biosciences, 556463) and analyzed on a Attune Nxt flow cytometer (Invitrogen). Data analysis was performed using FlowJo v10.

Analysis and Modeling of VSG N-terminal Domains:

Full length protein coding sequences from AnTat1.1 and its isolated mosaics were used for structural modeling. Only N-terminal domain sequences were used for protein structural prediction since this region of the VSG is the most well-defined experimentally. Signal peptides are cleaved from the mature VSG during processing, so we used SignalP 6.0[62] to predict and remove the Sec/SPI sequence (--organism eukarya, --mode fast) resulting in a FASTA file of mature proteins. To determine the coordinate of the N-terminal domain, we used a python analysis pipeline available at (https://github.com/mugnierlab/find_VSG_Ndomains). The script identifies the boundaries of the VSG N-terminal domain using the HMMscan function under HMMer version 3.1b2[63]. Query sequences are searched against an HMM profile containing 735 known N-terminal domain sequences from Cross et al[2]. and N-terminal domains defined by the largest envelope domain coordinate that meets E value threshold (1×10^{-5}, -domE 0.00001). The processed FASTA file containing only mature VSG N-terminal domain sequences was used as input for structural prediction of monomers using LocalColabFold[22] function colabfold_batch run using the following arguments: --amber, --templates, --num_recycle 3. The best ranked output model with the highest average predicted local distance test score (pLDDT), that is the highest confidence model, was used for downstream analyses. Model visualization and alignment were performed using with UCSF ChimeraX version 1.7.1[64], developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Acknowledgements

We would like to thank Jay Bangs for generating and sending us primary rabbit anti-AnTat1.1 polyclonal antibody. Thanks to Galadriel Hovel-miner for the primary rabbit anti-γ-H2A antibody. We appreciate members of the Mugnier lab for their thoughtful review of this manuscript, particularly Carolina Duque and Sneider Gutierrez. We also would like to thank Richard McCulloch for thoughtful feedback about this manuscript. We would like to thank David Mohr and the JHU GCRF Genomics Core for sequencing services. Thanks to Joey Verdi for pHH-HYG-VSG-3-S317A and advice about the generation of transgenic parasites. Data analysis was carried out at the Advanced Research Computing at Hopkins (ARCH) core facility (rockfish.jhu.edu), which is supported by the National Science Foundation (NSF) grant number OAC 1920103. This work was supported by the JHU Catalyst Award. JES, JMCH, AKB, JS, BZ, and MRM were supported by Office of the Director, NIH (DP5OD023065). JES and JS were additionally supported NIH training grant T32GM007 and JES was supported by the National Science Foundation (NSF) Graduate Research Fellowship under grant numbers 1232825 and 1746891. JMCH, AKB, JS and EHA were additionally supported by NIH T32AI007417.

775 **References**

776 1. Hertz-Fowler, C. *et al.* Telomeric Expression Sites Are Highly Conserved in
777 Trypanosoma brucei. *PLoS ONE* **3**, e3527 (2008).
778 2. Cross, G. A. M., Kim, H.-S. & Wickstead, B. Capturing the variant surface
779 glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427. *Mol.*
780 *Biochem. Parasitol.* **195**, 59–73 (2014).
781 3. Mugnier, M. R., Cross, G. A. M. & Papavasiliou, F. N. The in vivo dynamics of
782 antigenic variation in *Trypanosoma brucei*. *Science* **347**, 1470–1473 (2015).
783 4. Beaver, A. *et al.* Extravascular spaces are reservoirs of antigenic diversity in
784 Trypanosoma bruci infection. *bioRxiv* (2022)
785 doi:https://doi.org/10.1101/2022.06.27.497797.
786 5. Hall, J. P. J., Wang, H. & Barry, J. D. Mosaic VSGs and the Scale of Trypanosoma
787 brucei Antigenic Variation. *PLoS Pathog.* **9**, e1003502 (2013).
788 6. Roth, C., Bringaud, F., Layden, R. E., Baltz, T. & Eisen, H. Active late-appearing
789 variable surface antigen genes in Trypanosoma equiperdum are constructed entirely
790 from pseudogenes. *Proc. Natl. Acad. Sci.* **86**, 9375–9379 (1989).
791 7. So, J. *et al.* VSGs Expressed during Natural T. b. gambiense Infection Exhibit
792 Extensive Sequence Divergence and a Subspecies-Specific Bias towards Type B N-
793 Terminal Domains. *mBio* **13**, e02553-22 (2022).
794 8. Pays, E. *et al.* Gene conversion as a mechanism for antigenic variation in
795 Trypanosomes. *Cell* **34**, 371–381 (1983).
796 9. Baltz, T. *et al.* Exposed epitopes on a Trypanosoma equiperdum variant surface
797 glycoprotein altered by point mutations. *EMBO J.* **10**, 1653–1659 (1991).
798 10. Kamper, S. M. & Barbet, A. F. Surface epitope variation via mosaic gene formation
799 is potential key to long-term survival of Trypanosoma brucei. *Mol. Biochem.*
800 *Parasitol.* **53**, 33–44 (1992).
801 11. Thon, G., Baltz, T. & Eisen, H. Antigenic diversity by the recombination of
802 pseudogenes. *Genes Dev.* **3**, 1247–1254 (1989).
803 12. Lu, Y., Hall, T., Gay, L. S. & Donelson, J. E. Point mutations are associated with a
804 gene duplication leading to the bloodstream reexpression of a trypanosome
805 metacyclic VSG. *Cell* **72**, 397–406 (1993).
806 13. Rice-Ficht, A. C., Chen, K. K. & Donelson, J. E. Point mutations during generation of
807 expression-linked extra copy of trypanosome surface glycoprotein gene. *Nature* **298**,
808 676–679 (1982).
809 14. Lu, Y., Alarcon, C. M., Hall, T., Reddy, L. V. & Donelson, J. E. A strand bias occurs
810 in point mutations associated with variant surface glycoprotein gene conversion in
811 Trypanosoma rhodesiense. **14**, 10 (1994).
812 15. Berriman, M. *et al.* The architecture of variant surface glycoprotein gene expression
813 sites in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **122**, 131–140 (2002).
814 16. Boothroyd, C. E. *et al.* A yeast-endonuclease-generated DNA break induces
815 antigenic switching in Trypanosoma brucei. *Nature* **459**, 278–281 (2009).
816 17. Glover, L., Alsford, S. & Horn, D. DNA Break Site at Fragile Subtelomeres
817 Determines Probability and Mechanism of Antigenic Variation in African
818 Trypanosomes. *PLoS Pathog.* **9**, e1003260 (2013).

18. Kou, R. *et al.* Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLOS ONE* **11**, e0146638 (2016).
19. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* **96**, 317–323 (2003).
20. Müller, L. S. M. *et al.* Genome organization and DNA accessibility control antigenic variation in trypanosomes. *Nature* **563**, 121–125 (2018).
21. Kitamura, D., Roes, J., Kühn, R. & Rajewsky, K. A B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin μ chain gene. *Nature* **350**, 423–426 (1991).
22. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
23. Glover, L., Jun, J. & Horn, D. Microhomology-mediated deletion and gene conversion in African trypanosomes. *Nucleic Acids Res.* **39**, 1372–1380 (2011).
24. McCulloch, R. & Barry, J. D. A role for RAD51 and homologous recombination in Trypanosoma brucei antigenic variation. *Genes Dev.* **13**, 2875–2888 (1999).
25. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).
26. Ling, A. K. *et al.* Double-stranded DNA break polarity skews repair pathway choice during intrachromosomal and interchromosomal recombination. *Proc. Natl. Acad. Sci.* **115**, 2800–2805 (2018).
27. So, C. C. & Martin, A. DSB structure impacts DNA recombination leading to class switching and chromosomal translocations in human B cells. *PLOS Genet.* **15**, e1008101 (2019).
28. Đaković, S. *et al.* A structural classification of the variant surface glycoproteins of the African trypanosome. *PLoS Negl. Trop. Dis.* **17**, e0011621 (2023).
29. Sheader, K. *et al.* Variant surface glycoprotein RNA interference triggers a precytokinesis cell cycle arrest in African trypanosomes. *Proc. Natl. Acad. Sci.* **102**, 8716–8721 (2005).
30. Tenaglia, A. H. *et al. Antibodies to Protozoan Variable Surface Antigens Induce Antigenic Variation*. http://biorxiv.org/lookup/doi/10.1101/2022.06.21.497077 (2022) doi:10.1101/2022.06.21.497077.
31. Zhang, X. *et al.* Rapid antigen diversification through mitotic recombination in the human malaria parasite Plasmodium falciparum. *PLOS Biol.* **17**, e3000271 (2019).
32. Bopp, S. E. R. *et al.* Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families. *PLoS Genet.* **9**, e1003293 (2013).
33. Claessens, A. *et al.* Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLoS Genet.* **10**, e1004812 (2014).
34. Bull, P. C. *et al. Plasmodium falciparum* antigenic variation. Mapping mosaic *var* gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol. Microbiol.* **68**, 1519–1534 (2008).
35. Yang, Y. & Adam, R. D. A group of Giardia lamblia variant-specific surface protein (VP) genes with nearly identical 5 ’ regions. *Mol. Biochem. Parasitol.* (1995).

36. Adam, R. D. *et al.* The Giardia lamblia vsp gene repertoire: characteristics, genomic organization, and evolution. *BMC Genomics* **11**, 424 (2010).
37. Nenarokova, A. *et al.* Causes and Effects of Loss of Classical Nonhomologous End Joining Pathway in Parasitic Eukaryotes. *mBio* **10**, e01541-19 (2019).
38. Schimmel, J., van Schendel, R., den Dunnen, J. T. & Tijsterman, M. Templated Insertions: A Smoking Gun for Polymerase Theta-Mediated End Joining. *Trends Genet.* **35**, 632–644 (2019).
39. Williams, D. L. *et al.* May the Odds Be Ever in Your Favor: Non-deterministic Mechanisms Diversifying Cell Surface Molecule Expression. *Front. Cell Dev. Biol.* **9**, 720798 (2022).
40. Zozulya, S., Echeverri, F. & Nguyen, T. The human olfactory receptor repertoire. *Genome Biol.* **2**, research0018.1 (2001).
41. Lai, P. C. *et al.* An olfactory receptor pseudogene whose function emerged in humans: a case study in the evolution of structure–function in GPCRs. *J. Struct. Funct. Genomics* **9**, 29–40 (2008).
42. Niimura, Y. & Nei, M. Evolution of olfactory receptor genes in the human genome. *Proc. Natl. Acad. Sci.* **100**, 12235–12240 (2003).
43. Newman, T. & Trask, B. J. Complex Evolution of 7E Olfactory Receptor Genes in Segmental Duplications. *Genome Res.* **13**, 781–793 (2003).
44. Noonan, J. P., Grimwood, J., Schmutz, J., Dickson, M. & Myers, R. M. Gene Conversion and the Evolution of Protocadherin Gene Cluster Diversity. *Genome Res.* **14**, 354–366 (2004).
45. Wirtz, E., Leal, S., Ochatt, C. & Cross, George A. M. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **99**, 89–101 (1999).
46. Isothermal Reaction (Gibson Assembly) Master Mix. *Cold Spring Harb. Protoc.* **2017**, pdb.rec090019 (2017).
47. Rico, E., Jeacock, L., Kovářová, J. & Horn, D. Inducible high-efficiency CRISPR-Cas9-targeted gene editing and precision base editing in African trypanosomes. *Sci. Rep.* **8**, 7960 (2018).
48. Gkeka, A. *et al. Immunodominant Surface Epitopes Power Immune Evasion in the African Trypanosome*. http://biorxiv.org/lookup/doi/10.1101/2021.07.20.453071 (2021) doi:10.1101/2021.07.20.453071.
49. Shaw, S., Knüsel, S., Hoenner, S. & Roditi, I. A transient CRISPR/Cas9 expression system for genome editing in Trypanosoma brucei. *BMC Res. Notes* **13**, 268 (2020).
50. Zheng, Z. *et al.* Anchored multiplex PCR for targeted next-generation sequencing. *Nat. Med.* **20**, 1479–1484 (2014).
51. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMB.journal* **17**,.
53. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
54. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

910 55. Amos, B. *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics
911 resource center. *Nucleic Acids Res.* **50**, D898–D911 (2022).
912 56. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version
913 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
914 57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum
915 Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* **26**, 1641–
916 1650 (2009).
917 58. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-
918 Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).
919 59. Markin, A. *et al.* PARNAS: Objectively Selecting the Most Representative Taxa on a
920 Phylogeny. *Syst. Biol.* **72**, 1052–1063 (2023).
921 60. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat.*
922 *Methods* **9**, 676–682 (2012).
923 61. Glover, L. & Horn, D. Trypanosomal histone γH2A and the DNA damage response.
924 *Mol. Biochem. Parasitol.* **183**, 78–83 (2012).
925 62. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein
926 language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
927 63. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195
928 (2011).
929 64. Meng, E. C. *et al.* UCSF CHIMERAX : Tools for structure building and analysis.
930 *Protein Sci.* **32**, e4792 (2023).
931
932

933 **Supplemental figures**

934

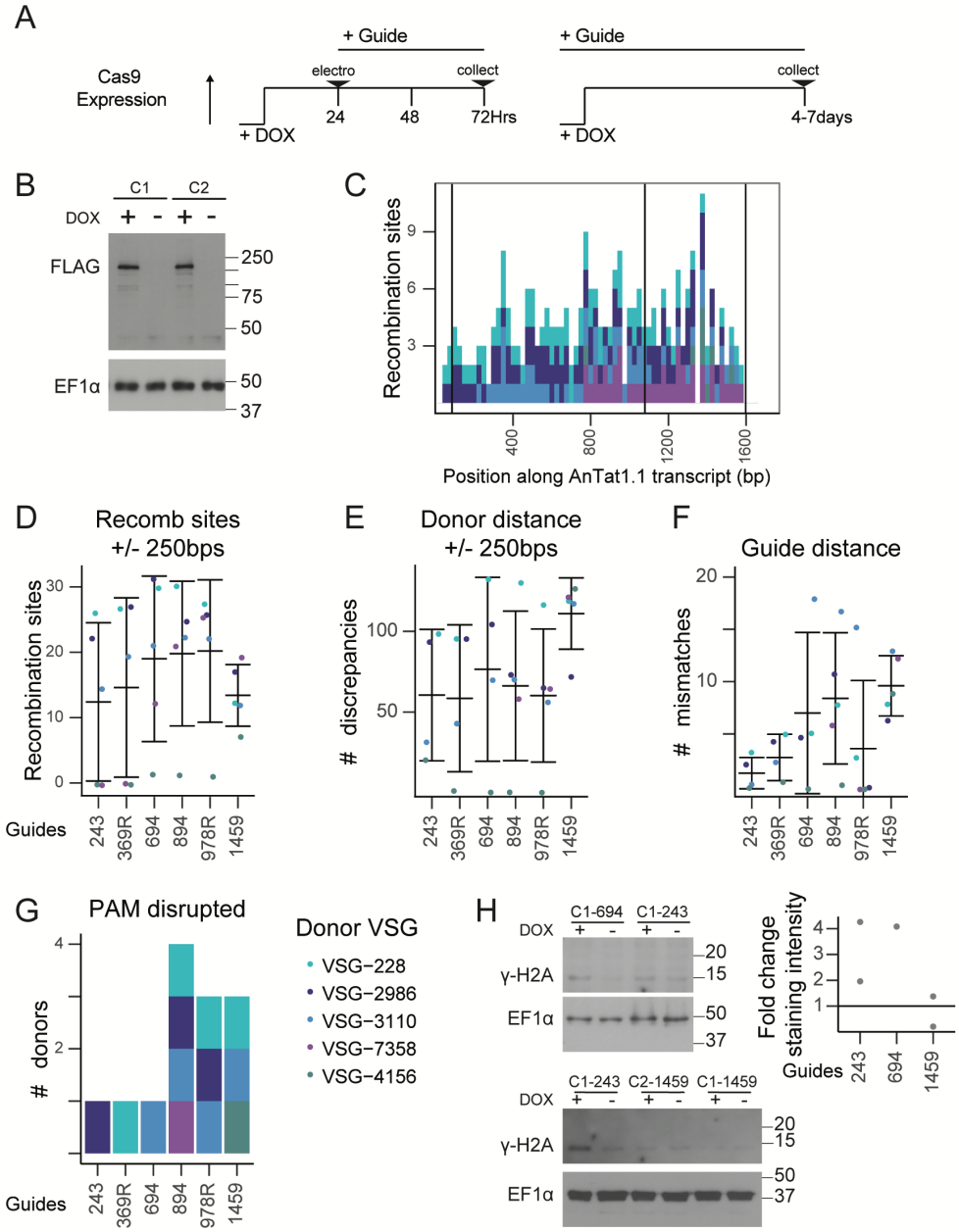
A

```
Target:AGAAGCAGCCGCGACACTGTTAATTTACGCCACGCACAAAATACAAGAC
Read:  AGAAGCAGCCGCGACACTGTTAATTTACGTCAGGATGAAAGTGGAAGCA
Donor: GCAAGAAGCCGGAACGCTGTTAATTTACGTCAGGATGAAAGTGGAAGCA
```

Supplemental Figure 1. Mosaic recombination site
A) A schematic of a recombination site. The region in yellow is the putative recombination site, shared by all 3 sequences.

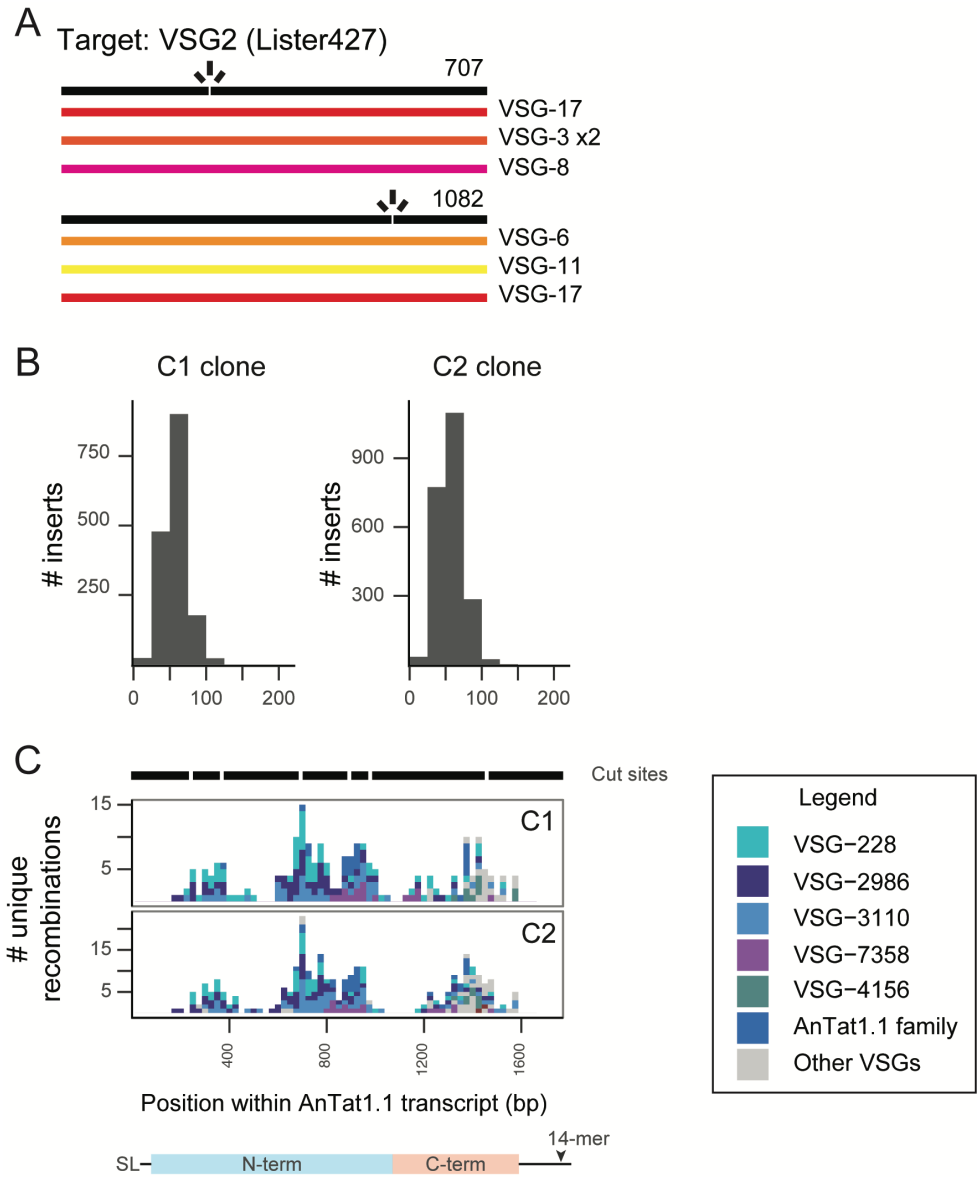
935

936



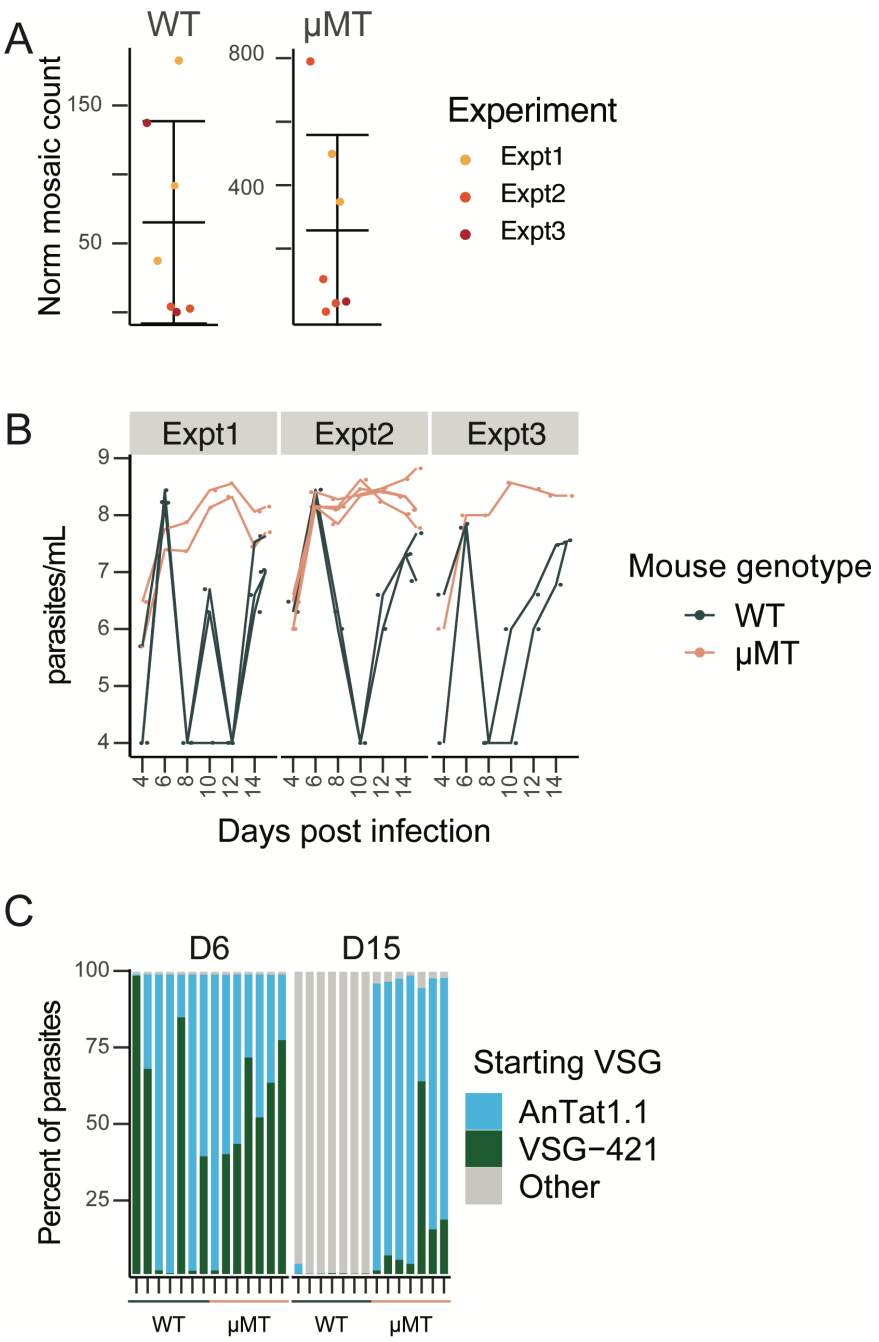
Supplemental Figure 2. Cas9 system and sgRNA design and analysis

A) A schematic of the Cas9 induction experiment. Transient electroporation shown on the left while constitutive guide expression to isolate mosaic clones shown at right. B) An immunoblot showing FLAG-tagged Cas9 induction 24 hours following doxycycline treatment (DOX). DMSO was used as a vehicle control. EF1α was used as a loading control. C) A histogram showing the midpoint of all possible recombination sites 5bps or longer between AnTat1.1 and its family members. D) Quantification of the number of recombination sites within 250bps up or downstream of the cut site for each donor VSG. E) Quantification of the Levenshtein distance between AnTat1.1 and family members. This includes mismatches, insertions and deletions. F) Quantification of the number of mismatches at the guide binding site between AnTat1.1 and family members. G) Histogram of which donor VSGs can disrupt the PAM when used to repair AnTat1.1. H) An immunoblot of γ-H2A from doxycycline induced clones and uninduced controls. DMSO was used as a vehicle control. Samples were normalized to EF1α loading and γ-H2A induction following doxycycline treatment after 24 hours was quantified at right.



Supplemental Figure 3. DNA breaks result in switchers if a homologous donor is not present
A) A schematic showing identified switchers following a Cas9-induced double strand break within VSG-2 at locations 707 and 1082. B) Histograms of donor VSG insertion lengths identified in all mosaic VSG reads from Figure 2A). The insert length only includes newly inserted sequence and does not include recombination sites. Clone C1 and C2 sequenced via VSG-AMP-seq are shown separately. The limit of detection for an insertion is approximately 200 bp. C) A histogram with a summary of the unique recombination sites found within guide-induced break for each clone. Cut sites are indicated above the histograms as gaps within the black line. The midpoint of the perfect homology between AnTat1.1 and the donor VSG at the recombination site is plotted. If a donor VSG could not be unambiguously identified, the average recombination position was plotted. The legend for the donor VSG colors is to the right. SL = 5′ splice leader cap, 14-mer = 3′ sequence conserved in all VSG transcripts

938
939



Supplemental Figure 4. Mouse infection extended data

A) Quantification of the number of recombination events detected per mouse. Each mouse was normalized to the number of total consolidated, aligned, and unanchored reads compared to the consolidated, aligned, unanchored read count for one mouse to control for sequencing depth. This was performed separately for each genotype. B) A time course of the parasitemia for the mouse infections. Blood was harvested every two days. C) Quantification of the percent of parasites expressing the starting VSG at D6 and D15 post-infection as quantified by VSG-seq.

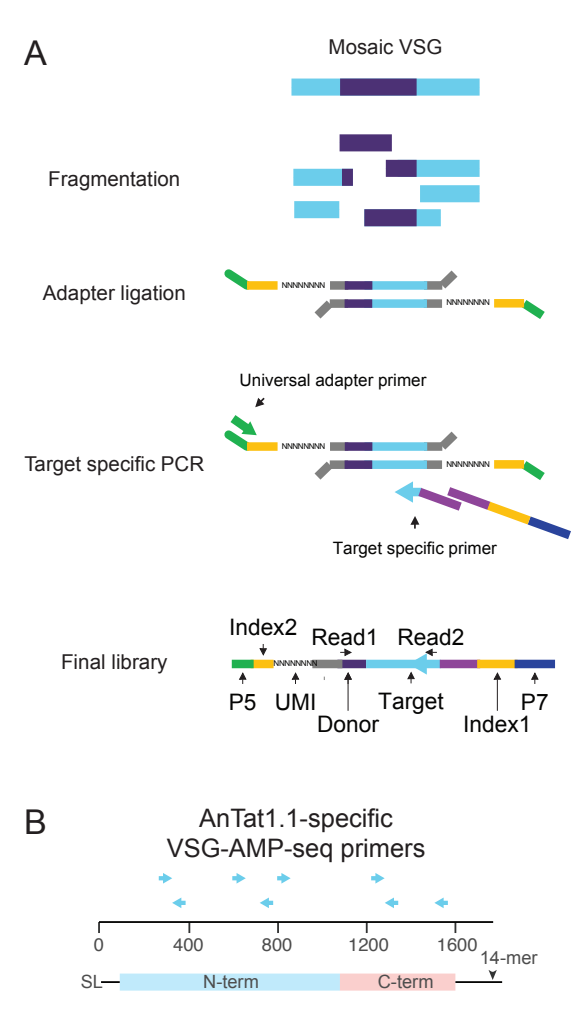
942 Supplemental Table 1

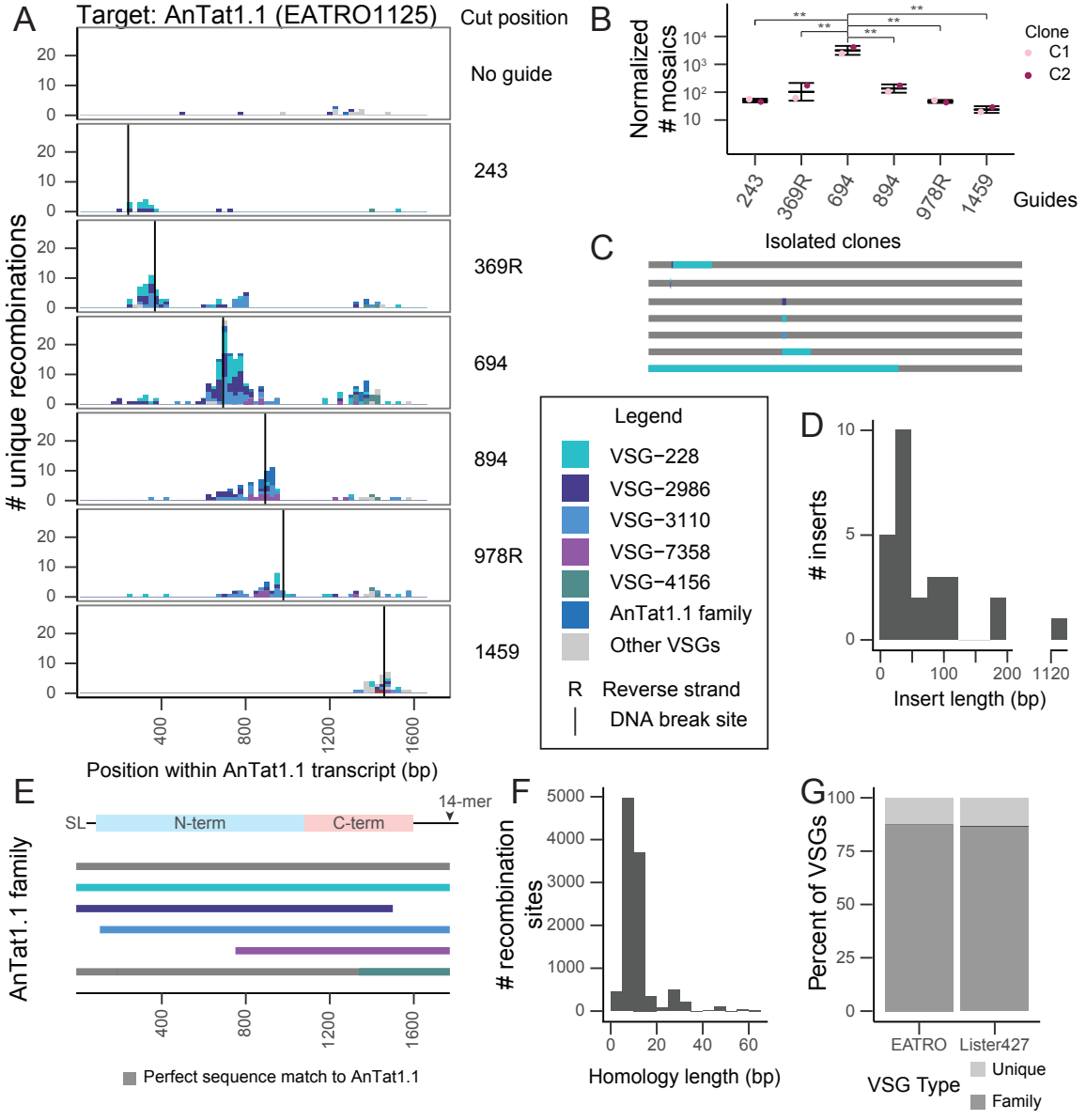
Guide Primers		
Primer Name	Sequence	
AnTat1.1_243	GAAATTAATACGACTCACTATAGGATTCAAAAACGGCCAAACGCCGTTTTAGAGCTAGAAATAGC	
AnTat1.1_369R	GAAATTAATACGACTCACTATAGGGGCGTAAATTAACAGTGTCGGTTTTAGAGCTAGAAATAGC	
AnTat1.1_694	GAAATTAATACGACTCACTATAGGAGTACAGACCCAGAAGCCAGGTTTTAGAGCTAGAAATAGC	
AnTat1.1_894	GAAATTAATACGACTCACTATAGGACGCCGGTGTCGCAGCTAAACGTTTTAGAGCTAGAAATAGC	
AnTat1.1_978R	GAAATTAATACGACTCACTATAGGGTCGTTGGCTGCTTGGAGTTGTTTTAGAGCTAGAAATAGC	
AnTat1.1_1459	GAAATTAATACGACTCACTATAGGACCAATCCAGAAAAGTGCAAGTTTTAGAGCTAGAAATAGC	
G00	AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC	
Annealed Guides for T7-sgRNA insertion		
Fragment Name	FWD	REV
AnTat1.1_243	AGGGATTCAAAAACGGCCAAACGCC	AAACGGCGTTTGGCCGTTTTTGAAT
AnTat1.1_694	AGGGAGTACAGACCCAGAAGCCAG	AAACCTGGCTTCTGGGTCTGTACT
VSG-2_707	AGGGACCAACGGCCTCGGCAAAAG	AAACCTTTTGCCGAGGCCGTTGGT
VSG-2_1082	AGGGCCAGTGGCGCAAAACCTGGT	AAACACCAGGTTTTGCGCCACTGG
DNase Verification Primers		
Primer Name	FWD	REV
HSP-70 (Tb927.11.11330)	AGAACACTATCAATGACCCCAAC	CCATGCCCTGGTACATCT
Hyg	ACAGCGGTCATTGACTGGAG	ATTTGTGTACGCCCGACAGT
VSG-AMP-seq Primers		
Primer Name	Sequence	
All-VSG-3'UTR	GTGTTAAAATATATC	
Y-adapter%	[Phos]GATCGGAAGAGC*C*A	
A01^	AATGATACGGCGACCACCGAGATCTACACTAGATCGC(N:25252525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGATC*T	
A03^	AATGATACGGCGACCACCGAGATCTACACTATCCTCT(N:25252525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA	

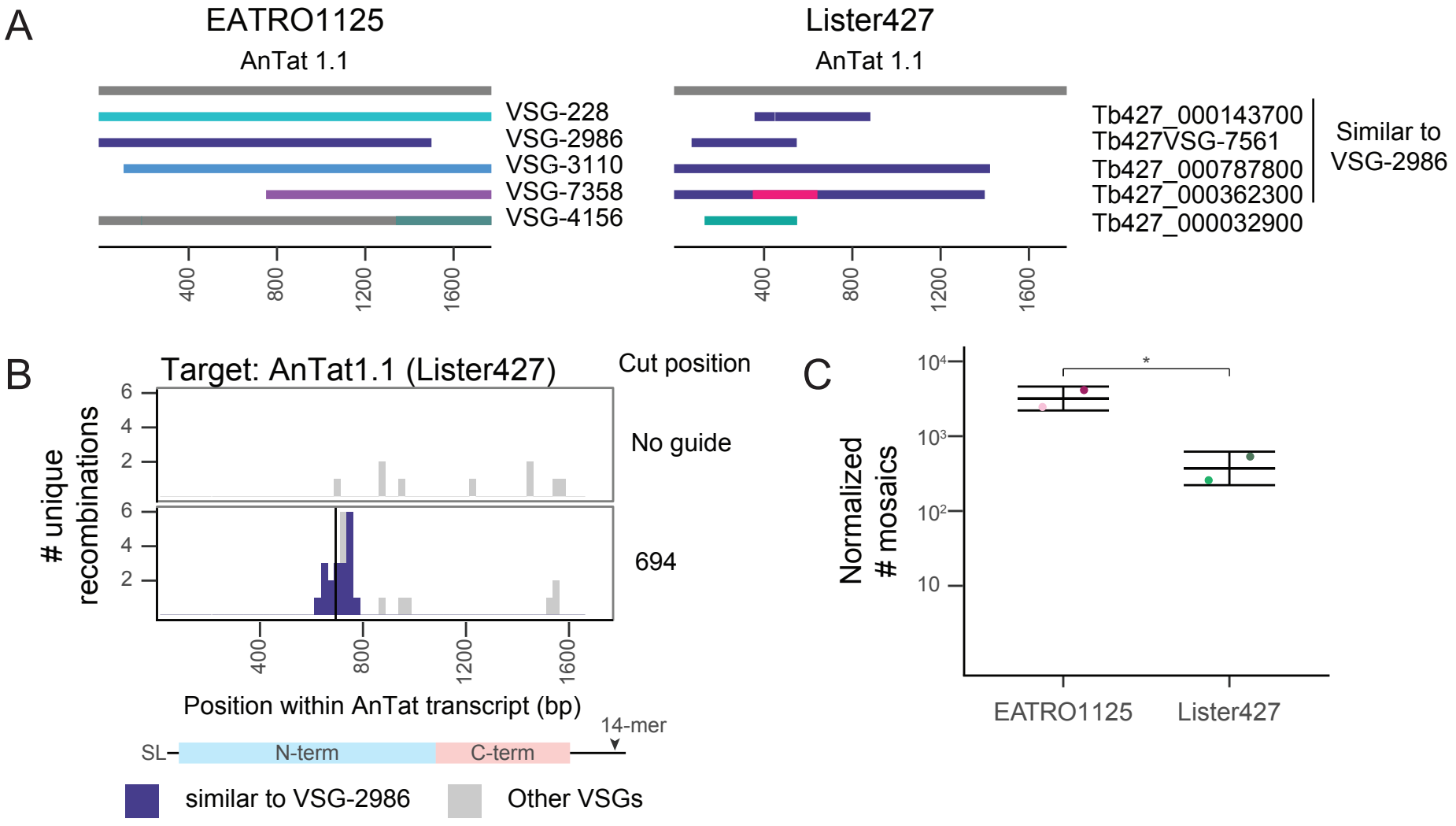
	TC*T
A04^	AATGATACGGCGACCACCGAGATCTACACAGAGTAGA(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
A05^	AATGATACGGCGACCACCGAGATCTACACGTAAGGAG(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
A06^	AATGATACGGCGACCACCGAGATCTACACACTGCATA(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
A07^	AATGATACGGCGACCACCGAGATCTACACAAGGAGTA(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
A08^	AATGATACGGCGACCACCGAGATCTACACCTAAGCCT(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
A09^	AATGATACGGCGACCACCGAGATCTACACGACATTGT(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
A10^	AATGATACGGCGACCACCGAGATCTACACACTGATGG(N:252 52525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)ACACTCTTTCCCTACACGACGCTCTTCCGA TC*T
P5_2	AATGATACGGCGACCACCGAGATCTACAC
P701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P708	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA
P709	CAAGCAGAAGACGGCATACGAGATATCACGGTGTGACTGGA GTCCTCTCTATGGGCAGTCGGTGA

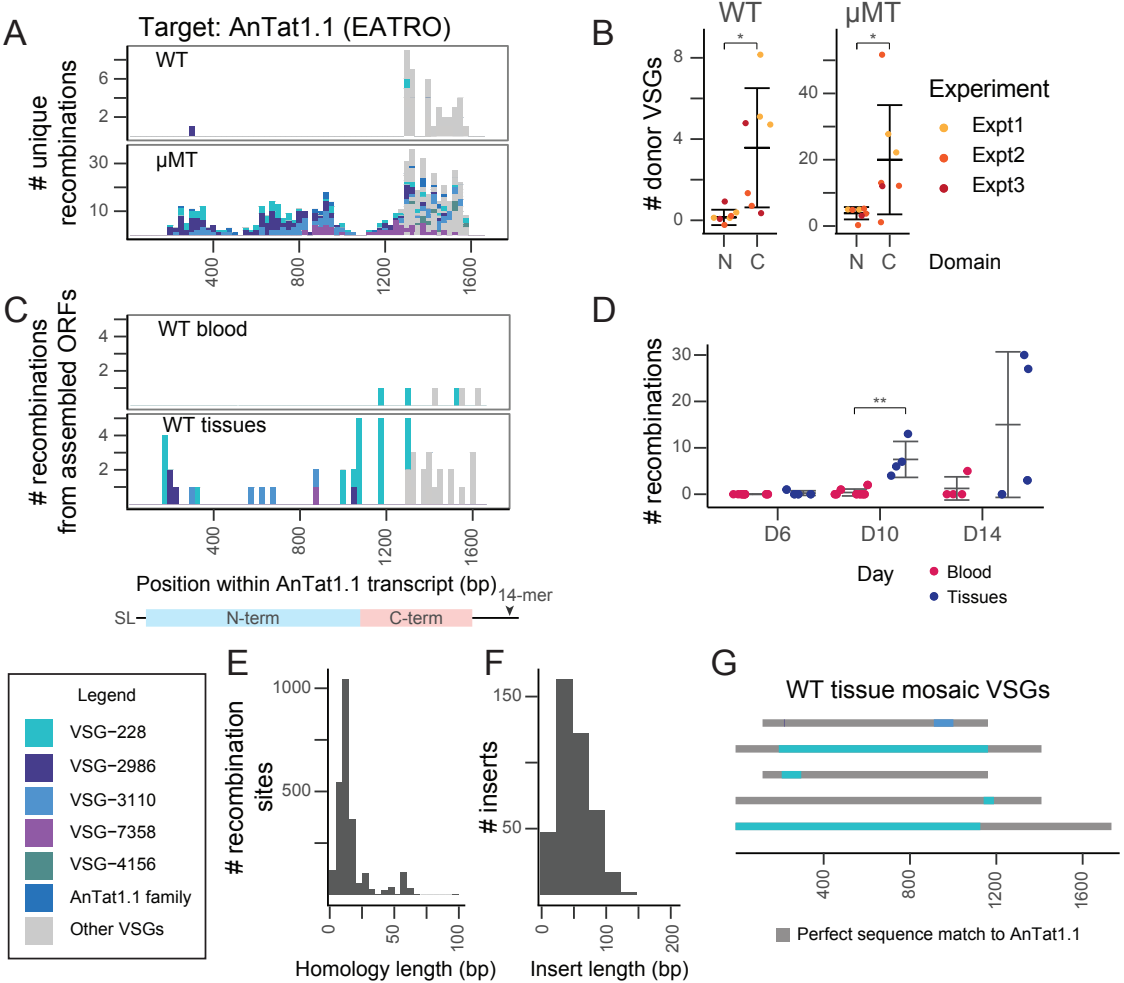
P710	CAAGCAGAAGACGGCATACGAGATCGATGTGCGTGACTGGAGTCCTCTCTATGGGCAGTCGGTGA
AnTat1.1_1F[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$CGCAAACACTACAACGAGCC
AnTat1.1_2F[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$CAGAATGCGACACGGAAAGC
AnTat1.1_3F[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$ACGCAGGCGGCTTCAAAACA
AnTat1.1_4F[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$AACAGCCGCAGCAACCAAAC
AnTat1.1_0R[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$GGCCACAAATGCGGCAGAAAC
AnTat1.1_1R[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$GCCATAAGCTGCGGTTTCGT
AnTat1.1_2R[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$GTTGTGTATGGTTAGCAGGC
AnTat1.1_3R[&]	CCTCTCTATGGGCAGTCGGTGAT$(N)_{0-7}$CTTGTATTTTGTGCGTGGCG
Index1[%]	ATCACCGACTGCCCATAGAGAGGACTCCAGTCAC
Read2[%]	GTGACTGGAGTCCTCTCTATGGGCAGTCGGTGAT
Miscellaneous Additional Primers	
Name	Sequence
SL-FWD	ACAGTTTCTGTACTATATTG
SP6-14mer-REV	GATTTAGGTGACACTATAGTGTTAAAATATATC
AnTat1.1 Sanger Sequencing Primer	AGAGAATACTAAGCTAGTTGGC
Pan AnTat1.1family FWD	ACTACACCCACAACAAGCTCTA

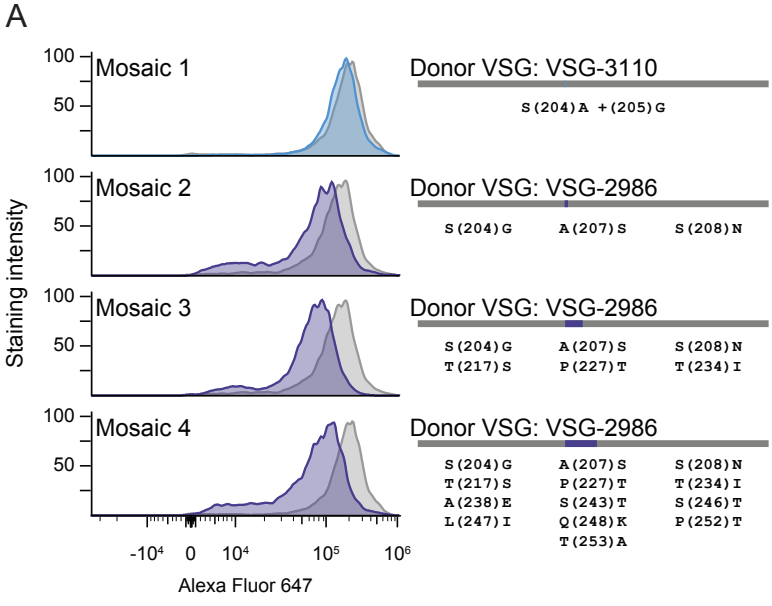
943 * = indicates a phosphorotiolate bond modification
944 [Phos] = 5’ phosphorylation
945 ^ = hand mixing
946 & = machine mixing
947 % = HPLC purification

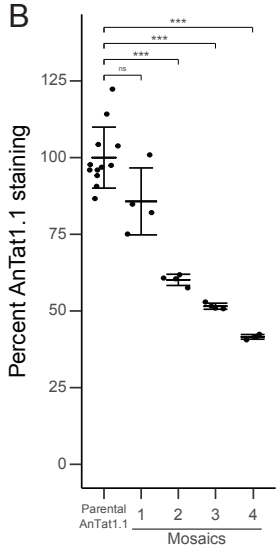

A
Mosaic VSG
Fragmentation
Adapter ligation
NNNNNNNN
NNNNNNNN
Universal adapter primer
Target specific PCR
NNNNNNNN
NNNNNNNN
Target specific primer
Final library
Index2
Read1
Read2
NNNNNNNN
P5
UMI
Donor
Target
Index1
P7
B
AnTat1.1-specific
VSG-AMP-seq primers
0
400
800
1200
1600
14-mer
SL
N-term
C-term

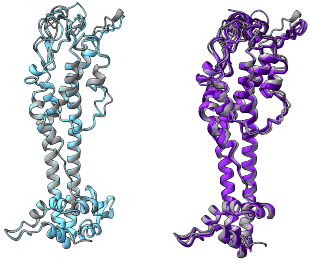

A
Target: AnTat1.1 (EATRO1125)
Cut position
No guide
243
369R
694
894
978R
1459
unique recombinations
Position within AnTat1.1 transcript (bp)
B
Normalized # mosaics
Clone
C1
C2
Guides
C
Isolated clones
Legend
VSG−228
VSG−2986
VSG−3110
VSG−7358
VSG−4156
AnTat1.1 family
Other VSGs
R Reverse strand
DNA break site
D
inserts
Insert length (bp)
E
AnTat1.1 family
SL
N-term
C-term
14-mer
Perfect sequence match to AnTat1.1
F
recombination sites
Homology length (bp)
G
Percent of VSGs
EATRO
Lister427
VSG Type
Unique
Family


A
EATRO1125
AnTat 1.1
VSG-228
VSG-2986
VSG-3110
VSG-7358
VSG-4156
400
800
1200
1600
Lister427
AnTat 1.1
Tb427_000143700
Tb427VSG-7561
Tb427_000787800
Tb427_000362300
Tb427_000032900
Similar to VSG-2986
400
800
1200
1600
B
Target: AnTat1.1 (Lister427)
Cut position
unique recombinations
6
4
2
No guide
694
400
800
1200
1600
Position within AnTat transcript (bp)
14-mer
SL
N-term
C-term
similar to VSG-2986
Other VSGs
C
Normalized # mosaics
10^4
10^3
10^2
10
*
EATRO1125
Lister427

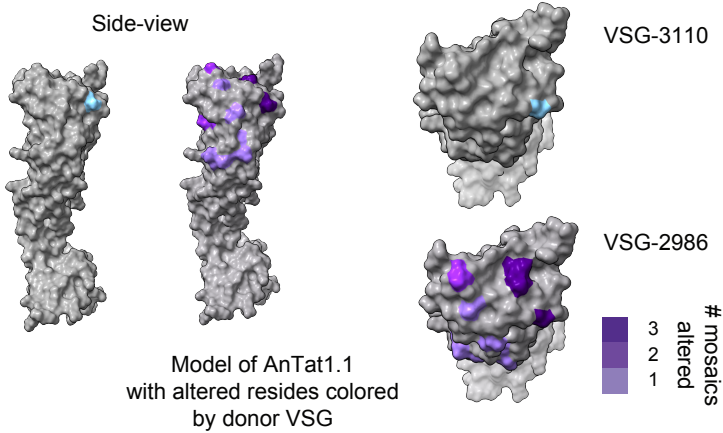

A
Target: AnTat1.1 (EATRO)
WT
μMT
unique recombinations
B
WT
μMT
donor VSGs
Experiment
Expt1
Expt2
Expt3
N
C
Domain
C
WT blood
WT tissues
recombinations from assembled ORFs
Position within AnTat1.1 transcript (bp)
SL
N-term
C-term
14-mer
D
recombinations
D6
D10
D14
Day
Blood
Tissues
Legend
VSG−228
VSG−2986
VSG−3110
VSG−7358
VSG−4156
AnTat1.1 family
Other VSGs
E
recombination sites
Homology length (bp)
F
inserts
Insert length (bp)
G
WT tissue mosaic VSGs
Perfect sequence match to AnTat1.1







Overlapped models of AnTat1.1 and mosaic VSGs



Model of AnTat1.1 with altered resides colored by donor VSG