Genome-wide association of trypanosome infection status in the tsetse fly Glossina fuscipes, the major vector of African trypanosomiasis in Uganda

Norah Saarman (norah.saarman@usu.edu)
Utah State University
Jae Hak Son
Rutgers, The State University of New Jersey
Hongyu Zhao
Yale School of Public Health
Luciano Cosme
Yale University
Yong Kong
Yale School of Public Health
Mo Li
Yale School of Public Health
Shiyu Wang
Emory University
Brian Weiss
Yale School of Public Health
Richard Echodu
Gulu University
Robert Opiro
Gulu University
Serap Aksoy
Yale School of Public Health
Adalgisa Caccone
Yale University

Research Article

Keywords: ddRAD, trypanosomiasis, vector, genome wide association, GWAS, population genomics, Muller elements, chromosome arms, aneuploidy

Posted Date: September 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1685795/v1

Abstract

The primary vector of the trypanosome parasite causing human and animal African trypanosomiasis in Uganda is the riverine tsetse fly *Glossina fuscipes fuscipes* (*Gff*). We conducted a genome-wide association (GWA) analysis with field-caught *Gff*. To increase statistical power, we first improved the *Gff* genome assembly with whole genome 10X Chromium sequencing, used ddRAD-seq to identify autosomal versus sex-chromosomal regions of the genome with data from 96,965 SNPs, and conducted a GWA with a subset of 50,960 autosomal SNPs from 351 flies. Results assigned a full third of the genome to the sex chromosome, suggested possible sex-chromosome aneuploidy in *Gff*, and identified a single autosomal SNP to be highly associated with trypanosome infection. The top SNP was ~ 1200 bp upstream of the gene *lecithin cholesterol acyltransferase* (LCAT), an important component of the molecular pathway that initiates trypanosome lysis and protection in mammals. Results indicate that variation upstream of LCAT and/or linked genetic elements are associated with trypanosome infection susceptibility in *Gff*. This suggests that there may be naturally occurring genetic variation in *Gff* that can protect against trypanosome infection, thereby paving the way for targeted research into novel vector control strategies that can promote parasite resistance in natural populations.

Introduction

Human and animal African trypanosomiasis (HAT and AAT, respectively) limit livestock production and represent a significant public health constraint [1, 2]. These diseases are caused by protozoan parasites in the family Trypanosomatidae. The parasites are transmitted to humans and animals through the bite of an infected tsetse fly, which collectively inhabits about 10 million km^2 of land in sub-Saharan Africa. There are multiple forms of AAT, with the most important forms of the disease caused by *Trypanosoma congolense, T. vivax and T. brucei brucei*. In contrast, there are just two forms of HAT, the chronic form (caused by *T. b. gambiense*), found in west and central Africa, and the acute form (caused by *T. b. rhodesiense*), found in countries in eastern and southern Africa [3, 4]. Uganda is the only country presenting both forms of the human disease. In Uganda, up to 90% of HAT cases are transmitted by a single species of tsetse fly, *Glossina fuscipes fuscipes* (*Gff*) [5, 6], which lives in lowland forests and along waterways in the western and central regions of the country (Fig. 1).

Host-pathogen dynamics are among the strongest forces of selection [7–9]. The expectation of strong positive selection in genes involved in host-pathogen interaction, as well as experimental evidence of influence of genotype on susceptibility to infection [10–12] suggests that there may be adaptive responses in *Gff* to avoid infection with *Trypanosoma* parasites. If there are adaptive responses, we expect significant associations between single nucleotide polymorphisms (SNPs) and infection status. Indeed, previous studies have found significant genetic associations with *Trypanosoma* infection status in this species [13, 14]. Previous studies identified 56 candidate genes in the vicinity of the 18 regions associated with *Trypanosoma* infection status in *Gff*. These genes were involved in DNA regulation, neurophysiological functions, and immune responses [13].

Since the most recent GWA study in *Gff* was published [13], there have been advances in genome sequencing technology, knowledge of *Gff* genome architecture, and genome-wide population structure that, when accounted for, can improve accuracy of the GWA analysis. Advances in technology include the development of linked-read sequencing (i.e. 10X Chromium sequencing) with long-range contiguity. Having long-range contiguity allows assembly of longer stretches of DNA molecules especially in structurally complex genomic regions [15–17],

improving quality, scaffold length, and overall coverage of the genome assembly in non-model organisms such as *Gff*.

The *Gff* genome is ~ 530 Mb in length [18]. Since approximately a third of it is on the sex chromosome [19], this creates complications in GWA study design and interpretation [20] because of the different copy numbers in males *versus* females and the potential inactivation of large regions as a mechanism of dosage compensation. Population genomic analyses of genetic diversity, divergence, migration, introgression, and genetic clustering (i.e., model-based clustering and multidimensional summaries) indicated extensive genome-wide divergence among *Gff* populations in Uganda with three distinct genetic clusters (discrete ancestral populations) in the north, west, and south of the country [21]. Analysis also demonstrated complex patterns of introgression that were non-uniform across the genome [21], raising the possibility that population structure should also be accounted for non-uniformity across the genome in GWA analysis. These advances in knowledge suggest that an improved genome assembly with proper identification of the sex chromosomes and a method that explicitly accounts for population structure in a chromosome-specific manner can improve our ability to identify genetic elements associated with trypanosome infection.

The goal of this study is to strengthen the identification of SNPs associated with trypanosome infection in *Gff* by making use of advances in sequencing technology and explicitly accounting for new knowledge of genomic architecture and population structure. To achieve this goal, we (i) conducted a whole-genome sequencing effort to improve the available genome assembly for *Gff*, (ii) used this new assembly and double digest restriction enzyme associated DNA sequencing [22] from 627 flies to identify autosomal *versus* sex-chromosomal regions of the genome and to identify population genomic structure, and (iii) conducted a GWA with 50,960 autosomal SNPs from 351 individuals of balanced infection status (174 infected and 177 control flies). Results from this study provide an improved genome assembly for *Gff*, assigned scaffolds to autosomal *versus* sex chromosome Muller elements (ME; chromosomal arms that are conserved across Diptera species), and identified a single SNP on autosomal Muller element B (chromosome L1) 1067 bp upstream of the gene *lecithin cholesterol acyltransferase* (LCAT) associated with trypanosome infection in *Gff.* These findings are important because they provide an improved genomic toolkit for future genomic studies in *Gff*, and identified candidate genes and gene-pathways important in trypanosome susceptibility in the obligate insect vectors of AAT and HAT.

Results

Whole-genome sequencing, assembly and annotation

The DNA extraction from a *Gff* pupae whole body used for 10X Chromium GenCode Technology (10x Genomics, USA) sequencing yielded DNA with average molecular weight > 30 kb and concentration > 600 ng/uL according to the pulse field gel run as part of the Yale Center for Genomic Analysis' standard quality control protocol. Sequencing returned 226.74 million long-range contiguous paired-end DNA sequences, with a length-weighted mean molecule length of 27,937.92 bases. Mean molecular length was shorter than estimated with the pulse field gel, indicating possible DNA damages (i.e. nicks, UV damage, etc). A total of 1.5% of sequences were identified as non-*Gff* (bacteria, viral, archaea, and cloning vector sequences) with Kraken 2 [23] and were removed. *De novo* assembly with the 10X Genomics supernova software package [24] yielded a total assembly of 390.09 Mb, with 7,986 scaffolds, an N50 scaffold size of 9.52 Mb, and 6 scaffolds > 10 Mb. This represents a

vast improvement over the existing *Gff* assembly (NCBI accession GCA_000671735.1), which had an N50 scaffold size of 0.6 Mb and zero scaffolds > 10 Mb (Table 1) [25].

Table 1
Assembly statistics and assignment of genes and scaffolds to Muller elements (ME) of the existing and new genome assemblies for *Gff*.

Statistic	Existing Assembly	New Assembly
Total coverage	52×	56x
Genome size (Mb)	375	390
Total no. of Scaffolds	2395	7986
>10 Mb	0	6
1–10 Mb	59	53
GC content (%)	34.00%	34.20%
N50 scaffold length (Mb)	0.56	9.6
L50 (rank of N50 scaffold)	178	8
Genes annotated	20138	18538
Genes with 1:1 orthologs	6487	6345
Scaffolds containing 1:1 orthologs	793	542
Scaffolds assigned to ME	764	169
Assembly length assigned to ME (Mb)	321	350
Total assigned to ME (%)	85.83%	89.74%

We transferred 18,538 gene annotations from the existing assembly onto the new assembly using the UCSC liftOver suite of programs [26]. Of these genes, 6,345 had 1:1 *Drosophila melanogaster* orthologs identified using OrthoDB [27]. Assignment to MEs (six chromosomal elements of Diptera species known as MEs A-F) were made following the method outlined [19], wherein scaffolds with the majority (> 50%) of genes with 1:1 orthologs on a single element were assigned that ME. With this method, 169 scaffolds and 350 Mb (89.74%) were assigned to MEs (Table 1; Supplementary Table S1 online), with 55.12% of the assembly was assigned to autosomal regions, and 34.66% was assigned to sex-chromosome associated regions (Supplementary Table S1 online).

ddRAD-seq population genomics

The ddRAD-seq protocol used yielded an average of 14.1 million reads and 80,256 ddRAD-tags per individual (Fig. 1, Table 2, Supplementary Table S2 online). For population genomics analysis, our genotyping and filtering protocol retained 92,720 SNPs from 627 flies originating from 33 geographically diverse sampling sites. These sampling sites spanned the three major genetic clusters of *Gff* found in Uganda, plus an outgroup (Fig. 1; Table 2; Supplementary Table S2 online) chosen based on mtDNA sequence data that indicated a more distant phylogenetic relationship [28].

Table 2

Sampling details for population genomics (PopGen) and genome-wide association (GWA) analysis for each genetic cluster. We report PopGen and GWA number of sampling sites, PopGen and GWA total sample size (N total), number of infected individuals in GWA (infected), number of uninfected control individuals in GWA (control), the average infection rate of the region based on field microscopy (regional infection rate), and the average per individual number of reads retained (reads) and number of ddRAD-tags sequenced and passing all quality control filters (tags).

Genetic Cluster	PopGen sites	GWA sites	PopGen N	GWA N	GWA Infected	GWA Control	Regional Infection Rate	Average Reads	Average Tags
North	19	8	291	166	76	79	2.80%	14874020	69857
West	3	1	201	190	89	89	11.91%	13679764	93649
South	10	3	119	39	9	9	1.66%	13223360	86746
Outgroup	1	0	15	0	NA	NA	NA	11342751	51055
Overall	33	12	627	351	174	177	3.64%	14092162	80256

We estimated linkage disequilibrium (LD) blocks and rate of decay (measured as the rate of r^2 decrease per kb) using PLINK v. 1.90-beta4.4 [29–31]. Median LD block length ranged from ~ 2 kb in the north in autosomal ME E to ~ 170 kb in the south in sex chromosome element F (Supplementary Table S3 online; Supplementary Fig. S1 online). Mean LD decay rate ranged from 0.0003 per kb in the north in autosomal element D to 0.1573 per kb in the west in autosomal element B (Supplementary Table S4 online; Supplementary Fig. S2 online). When averaged across the genome, the median LD block length was ~ 8 kb for the north, ~ 19 kb for the west, and ~ 89 kb for the south (Supplementary Table S3 online), and the mean LD decay rate was ~ 0.0252 per kb for the north, 0.0247 per kb for the west, and 0.0021 per kb for the south (Supplementary Table S4 online). Together, these results indicated different LD patterns in the three genetic clusters and MEs, with relatively short blocks and fast decay in the north and in autosomal elements, and relatively long blocks and slow decay in the south and sex chromosome associated scaffolds (Supplementary Fig. S1 online; Supplementary Fig. S2 online).

Male *versus* female relative read depth of coverage and heterozygosity were calculated to test the hypothesis of male haploidy in sex chromosome associated scaffolds (MEs A, D, and F). In species with heteromorphic sex chromosomes with male heterogamety (XY), males are haploid XY (small and degenerated Y) and females are diploid XX. When male and female reads are mapped to scaffolds, relative read coverage of males *versus* females can be used to identify sex-linked sequences. Therefore, we expect that relative coverage ($\log_2$ male/female coverage) is close to 0 (equal between males and females) in the autosome-linked scaffolds (MEs B, C, and E), but is close to -1 (a half ratio for males *versus* females) in X-linked scaffolds. Male/female coverage is close to 0 in the autosomes while the male/female read-depth is significantly decreased in the X-linked sequences (Supplementary Fig. S3 online), supporting lower ploidy in males in X-linked scaffolds. It is also expected that relative heterozygosity ($\log_2$ male/female heterozygosity) is similar between males and females in autosomes and is reduced in X-linked sequences due to haploid males and diploid females for the X. We show that the level of heterozygosity between males and females are similar in the autosomes and reduced in the X-linked sequences (Supplementary Fig. S4 online), further supporting lower ploidy in males than females in X-linked scaffolds. Taken together, results from comparison of male/female coverage and heterozygosity validate that MEs A, D, and F are on the chromosomes, while MEs B, D, and E are autosomal [19].

PCA identified different population structures in the autosomal *versus* sex chromosome associated SNPs (Fig. 2, Supplementary Fig. S5 online). For the autosomal SNPs, PC 1 explains ~ 39.5% of the variation, and PC 2 explains ~ 13.9% of the variation, and identifies 4 distinct groups, including the outgroup and the three genetic clusters in the north, west, and south of the country (Fig. 2). For the sex chromosome associated SNPs, PC 1 explains ~ 31.27% of the variation, and PC 2 explains ~ 19.3% of the variation, and in contrast to the pattern found in autosomal regions, the three genetic clusters in the north, west and south of the country do not cluster into distinct groups according to geographic origin. Instead, there is more variation within than between the north, west, and south genetic clusters, with at least three distinct clusters within any one genetic cluster (Fig. 2). These differences in the autosomal *versus* sex chromosome associated PCA results hold true for analysis completed for each ME separately (Supplementary Fig. S5 online). Thus, taken together, PCA results indicate genomic structure with unresolved origins in the sex chromosome associated MEs.

ddRAD-seq genome-wide association (GWA)

GWA analysis was conducted on autosomal and sex chromosome associated SNPs separately to allow us to follow best practices of considering only females in the sex chromosome analysis. For the autosomal MEs, GWA was conducted using 50,960 SNPs with a subset of 351 flies originating from 12 sampling sites to create a balanced study design (174 infected and 177 control; Table 2; Supplementary Table S2 online). Analysis was performed with the R package “statgenGWAS” v. 1.0.7 [32] with ME specific kinship matrices in a mixed model [33, 34]. Genetic variance was 0.2747, residual variance was 1.2470e-05, and the genomic control correction was applied with inflation-factor 0.89. The Q-Q plot revealed a close match of observed and expected p-values, indicating a low rate of false-positives and lending support for the reliability of the results. SNP 546904:87:- had significant association of genotype with infection status after Bonferroni correction for multiple testing (p-value = 1.80e-11; Fig. 3), and was in LD with two other SNPs that did not meet the genome-wide significance threshold (1e-6): SNP 1443046:51:+ (p-value = 1.96e-6) and SNP 1550026:39:- (p-value = 4.83e-6; Supplementary Table S5 online).

The top SNP 546904:87:- was at position 11393113 on scaffold_2 (NW_023998416.1). The minor allele (T) for SNP 546904:87:- was associated with low infection rate (Fig. 3), and was found in homozygous form in only 16 flies (none of them infected) from site KAF in the west genetic cluster. The major allele (A) for SNP 546904:87:- was found in homozygous form in 308 flies (172 were infected) and was geographically widespread. The heterozygous genotype (AT) was found in 27 flies (2 were infected) from KAF and APU (west and north genetic clusters). The SNP is within a ~ 2 Mb LD block that spans positions 9,938,197 to 12,058,501 on scaffold_2 and contains 136 liftOver gene annotations (Supplementary Table S6 online). The closest gene annotation to trypanosome associated SNP 546904:87:- was GFUI026799 *lecithin cholesterol acyltransferase* (LCAT), and was 1167 bp downstream of the SNP.

We also conducted a female-only GWA analysis with the sex chromosome associated SNPs to remove the risk of mis-scoring haploid male genotypes as diploid homozygous genotypes (i.e. best practices). Female-only GWA analysis of the sex-chromosome association MEs was completed using 19,423 SNPs from 209 flies (120 infected and 89 control; Supplementary Table S2 online). Using females alone removed the risk of mis-scoring haploid males, but did not remove the risk of mis-scoring females with sex-chromosome aneuploidy, a phenomenon known to occur in *Glossina spp.* [35]. The Q-Q plot revealed lower than expected p-values,

indicating excessive false-positives (Supplementary Note online; Supplementary Table S7 online) possibly due to the unresolved genome architecture of the sex chromosomes in *Gff* (see population genomics results). Given that these results indicate unreliable identification of associated SNPs with this approach, we limit our interpretation of results to the autosomal elements.

Discussion

Results from this study provide an improved genome assembly for *Gff*, reliably assigned scaffolds to genomic regions associated with autosomes *versus* sex-chromosomes, and identified a single SNP on autosomal ME B strongly associated with trypanosome infection in field-collected *Gff.* The genome assembly was improved from an N50 scaffold size of 0.6 Mb to 9.52 Mb and increased the number of scaffolds > 10 Mb from zero to six (Table 1). Population genomics analysis of LD blocks, male *versus* female coverage (Supplementary Fig. S3 online) and heterozygosity (Supplementary Fig. S4 online), and PCA (Fig. 2) confirmed broadly accurate assignment of 55.12% of the new assembly to autosomal and 34.66% to sex chromosome associated regions (Table 1). Results from the autosomal GWA analysis indicated a low false-positivity rate and identified a single SNP with highly significant association with infection.

Improved genome assembly

The genome assembly was improved from an N50 scaffold size of 0.6 Mb to 9.52 Mb and increased the number of scaffolds > 10 Mb from zero to six (Table 1). Our ortholog search against *D. melanogaster* MEs (chromosome arms) assigned 169 scaffolds and 350 Mb (89.74%) to MEs (Table 1) with 55.12% assigned to autosomal elements, and 34.66% assigned to sex-chromosome associated elements (Supplementary Table S1 online). This suggests a large proportion of the assembled genome is associated with the sex-chromosomes in *Gff.* A major limitation of our analysis is reliance on the assumption of syntenic conservation between tsetse scaffolds and Drosophila chromosomal structures. Any deviation from complete conservation could result in mis-assigned scaffolds (scaffolds assigned to sex chromosomes when actually autosomal, or *visa versa*) that could be difficult to detect if correctly assigned scaffolds made up most of the sequence in each ME and overwhelmed the signal. Nonetheless, the differences in population genomics results from these two categories (scaffolds assigned as autosomal *versus* sex chromosome) provide support for general accuracy in these assignments (see directly below).

Population genomics confirms chromosome identification and hints at possible aneuploidy

Population genomics results support our identification of sex chromosomes *versus* autosomes using an ortholog search against *D. melanogaster* MEs (chromosome arms). Evidence includes striking differences between sex chromosome and autosome patterns of LD, sequencing statistics, and genetic structure. Additionally, relative male *versus* female coverage and heterozygosity statistics did not meet expectations of a classic XY sex-determination system, raising the possibility of sex chromosome aneuploidy in *Gff*, a phenomenon known to occur in tsetse flies [35, 36]. Nonetheless, results provide evidence of diploidy in the regions assigned to autosomes (Table 1; Supplementary Table S1 online), and therefore, supports the use of autosomal SNPs in our subsequent GWA analysis.

Striking differences in LD – We found striking differences in LD patterns across chromosome elements and genetic clusters, with relatively short blocks and fast decay in autosomal elements (especially in the north), and relatively long blocks and slow decay in sex chromosome associated elements (especially in the south; Supplementary Fig. S1 online; Supplementary Fig. S2 online). There are multiple possible explanations for this. One possibility is that there are inversions on the sex chromosome, which would result in reduced recombination around the inversions and relatively long blocks and slow decay. Another possibility is that miss-scored genotypes caused by sex chromosome anueploidy creates a false signal of LD in sex chromosomes. In this scenario, additional copies of chromosomal elements would cause genotyping error and false associations among SNPs. The observed within-species variation in LD among genetic clusters regardless of the genomic region (generally fast decay in the north, slow decay in the south; Supplementary Fig. S1 online; Supplementary Fig. S2 online) could be caused by differences in sampling among the genetic clusters: Number of sampling sites ranges from three to 20 (Table 1). LD estimates are known to be sensitive to sample size [37], thus variation in the number of sampling sites could influence LD estimates directly. Additionally, the size of the geographic range of each genetic cluster ranges widely from a minimum area of ~ 10,000 km^2 to a maximum area of ~ 50,000 km^2 (Fig. 1). LD estimates are known to be sensitive to overall genetic diversity present [38], which in turn is expected to be positively correlated with geographic range of sampling. Thus, variation in geographic range could also influence LD estimates.

Sequencing statistics suggest deviation from the classic XY sex-determination system – Male *versus* female coverage (Supplementary Fig. S3 online) and heterozygosity (Supplementary Fig. S4 online) supports reliable assignment of 55.12% of the new assembly to autosomal and 34.66% to sex chromosome associated regions (Table 1). Equal male *vs* female coverage and heterozygosity in autosomal elements and lower relative coverage (-0.5) and heterozygosity (range from − 1 to -2) in males in sex chromosome associated elements (Supplementary Fig. S3 online, Supplementary Fig. S4 online) provides support for reliable assignment and lower ploidy in males relative to females in sex chromosomes. The theoretical relative coverage with strict haploidy in males is -1 [39–41]. Thus, the observed relative coverage of -0.5 (Supplementary Fig. S3 online) suggests that *Gff* does not follow the classic XY sex-determination system. Expectations of heterozygosity are less well defined because it depends on genetic diversity, which alters both female heterozygosity and expected relative heterozygosity [39–41]. Thus, the extreme difference between males and females observed here (Supplementary Fig. S4 online) further supports that *Gff* does not follow the classic XY sex-determination system.

Differences in population structure – PCA (Fig. 2) identified genetic structure in autosomes that closely matches geography and previous studies. We observed three well defined genetic clusters that align with sampling sites from north of Lake Kyoga, west of the Victoria Nile, and south of Lake Kyoga (north, west, south, respectively). These findings align with the overall population structure found with multiple clustering analysis using both microsatellite markers and genomic ddRAD SNPs [21]. A very different pattern of genetic structure is obtained from the sex chromosomes, with at least three distinct clusters within each geographically based genetic cluster. Although this is somewhat surprising given the large number of analyses from genomic DNA loci that have indicated three distinct genetic clusters, when revisiting the [21] results, it is apparent that there is a faint pattern of intra-cluster variation that likely corresponds to the SNPs from sex chromosomes, which were not identified or filtered in that previous study. It is difficult to determine the causal forces that are responsible for the striking difference in the population structure observed in the sex chromosomes *versus* the autosomes, but results point

to several possibilities. One possible explanation is the presence of chromosome inversions or low-recombining regions on the sex chromosomes. Inversions and low-recombining regions are known to differentiate populations more strongly than genomic regions outside of these regions [42], and to create genomically localized heterogeneity (i.e., contrasting patterns among genomic regions) in population structure [43]. Another possible explanation is sex-chromosome aneuploidy, a phenomenon known to occur in tsetse flies [35]. Sex-chromosome aneuploidy occurs in tsetse flies where females may be XX, XXY, or XXXY, and males may be XY, XYY, or XO [35, 36, 44]. When present, aneuploidy would cause extra copies of the genes on these chromosome arms in both males and females. In this scenario, signals of intra-cluster genetic structure would have originated from the variation between the 1–3 copies of the X chromosome present in any one individual.

Genome-wide association

GWA analysis indicated a low false-positive rate in autosomal regions and identified a single SNP 546904:87:- with highly significant association with infection status in *Gff* (p-value = 1.80E-11; Fig. 3). This result implies that there is natural genetic variation in *Gff* that can provide protection against trypanosome infection. The SNP 546904:87:- allele associated with zero infection (T; Fig. 3) is in low frequency (16 homozygous form AT, 27 heterozygous form TT), and only exists in homozygous form TT in a single region of Uganda (west genetic cluster) that has high trypanosome infection rates (11.9%; Table 2). This supports the concept of spatially patchy and temporally variable selection imposed by trypanosomiasis such that selection is strongest where infection rates are highest, but that the overall strength of selection experienced in the wild is insufficient to cause fixation of this protective allele in *Gff*, even on a local scale.

Although there is low likelihood that the associated SNP 546904:87:- detected in our analysis is the functional polymorphism itself, SNP 546904:87:- is part of a ~ 2 Mb LD block containing 136 genes, suggesting that there could be multiple and interacting linked functional mutations causing association with infection status. The closest annotation is GFUI026799 *lecithin cholesterol acyltransferase* (LCAT), a gene involved in cholesterol metabolism [13, 14, 19, 45–47]. The orientation of the trypanosome associated SNP 546904:87:- is just 1167 bp upstream of the LCAT gene generates the hypothesis that LCAT is involved in *Gff*'s trypanosome infection response, either through linked DNA sequence polymorphism within the gene or through an expression level response mediated by polymorphism within LCAT regulatory components.

LCAT is a particularly interesting candidate gene because of its role in forming the cholesterol esters found in apolipoprotein L-I (ApoL-I), known to be the lytic component of high-density lipoprotein (HDL; also known as the trypanosome lytic factor) in humans [48, 49]. In humans, LCAT transfers a fatty acid from the sn-2 position of lecithin (phosphatidylcholine) to cholesterol, forming the trypanosome lytic core of HDL, ApoL-I [49]. It is unlikely that this precise human mechanism of defense against trypanosomes operates in tsetse flies because of the vast differences in physiology, and because trypanosomes appear to be fully resistant to lysis by ApoL-I once they have entered the midgut of the tsetse fly and have transformed into their procyclic form [49]. However, because trypanosomes remain in their bloodstream form and susceptible to lysis by HDL until they reach tsetse's midgut [50], the possibility remains that a similar mechanism of HDL mediated lysis could occur early during infection establishment within the fly. Furthermore, other trypanolytic factors operate in the tsetse's midgut [51], opening the possibility that HDL, and thus LCAT, may be involved in lysis through an altogether different mechanism once the trypanosomes reach the midgut. Finally, HDL plays a broad role in disruption of host metabolism during an infection with trypanosomes [52], opening the possibility that LCAT is affected by or

involved in the disruption of lipid metabolism in tsetse flies following trypanosome infection. This involvement in lipid metabolism may underlie the association between the linked SNP 546904:87:- and trypanosome infection status in *Gff*. Additionally, disruption of tsetse lipid metabolism could directly impact the ability of procyclic trypanosomes to sustain an infection in their fly vector, as parasites at this stage of their developmental cycle use environmental fatty acids to maintain their metabolic homeostasis [53].

Undoubtedly, more research is needed to test this hypothesis and establish definitive association between trypanosome infection in *Gff* and sequence polymorphism and/or show an expression level response in the candidate gene LCAT. Indeed, the Aksoy lab at the Yale School of Public health has several experiments underway investigating differential expression in infected *versus* uninfected *Gff* that can provide additional evidence as to which genes are associated with reduced trypanosome infection rates and the possible contribution of LCAT. These and other functional studies are needed to confirm and unravel the potential mechanism of the association between genotype at SNP 546904:87:- and *Gff* infection status.

Conclusion

This study highlights the importance of identifying and accounting for genome architecture and population structure in genome wide association studies. Our use of OrthoDB to assign scaffolds to MEs proved effective in identifying and excluding sex chromosomes from our analysis, allowing us to account for element-specific patterns of LD. Together, these additions to the study design significantly increased our power to detect GWA candidate SNPs beyond what was previously possible.

Our GWA analysis identified one autosomal SNP 546904:87:- that was highly associated with *Gff* infection status (Fig. 3). The allele associated with low infection rate in *Gff* was in low frequency in the wild, and was found in heterozygous form in only one locality in the west of Uganda. Results indicate natural variation in wild populations of *Gff* that may provide some protection against trypanosome infection. The SNP identified as highly associated with infection status was proximal to a promising gene candidate, LCAT, that we hypothesize has DNA sequence polymorphism and/or an expression level response to trypanosome infection in *Gff.* These findings can inform our understanding of the mechanisms of the tsetse's natural defenses against trypanosome infection, identify gene pathways involved in defenses, are hypothesis-forming for future studies, and ultimately can be made use of in future adaptive vector control programs.

Methods

Whole-genome sequencing sampling and library prep

To improve the genome assembly of *Gff* we completed whole genome sequencing with 10X Chromium GenCode Technology (10x Genomics, USA) from a single *Gff* pupa originating from the Yale School of Public Health insectary's *Gff* colony (IAEA, Vienna, Austria). High molecular weight genomic DNA was extracted from the whole body of the pupa following the 10x Genomics® Sample Preparation Demonstrated Protocol DNA Extraction from Single Insects [54], and was suspended in 1X TE Buffer. Genomic DNA was sent to the Yale Center for Genomic Analysis for quality control including a pulse field gel, library preparation, and 10X Genomics Genome Sequencing.

Genome assembly and annotation

The raw sequences were scanned for bacteria, viral, archaea, and cloning vector sequences using Kraken 2 with confidence threshold as 0.0 [23]. Any raw sequencing reads that are classified were excluded for de novo assembly of the fly genome. The *de novo* assembly of the *Gff* genome was completed with the supernova software package from 10X Genomics [24] with –maxreads = 220000000. The annotations for the assembly were lifted over from the existing *Gff* assembly available from NCBI under the name "Glossina_fuscipes-3.0.2" [25] and Vectorbase under the name "GfusI1" [55, 56] using the UCSC liftOver suite of programs [26].

Scaffolds were assigned to MEs (six chromosomal elements of Diptera species known as MEs A-F) to allow us to distinguish scaffolds within autosomal (elements B, C, E) *versus* sex-chromosome associated regions (elements A, D, F) based on comparative genomic analysis in *Gff* [19]. These assignments assumed conservation of gene content in chromosome arms (MEs) across fly species, a reasonable assumption given the ubiquity of this pattern in studies to date [57, 58]. Scaffolds were assigned to MEs following [19] based on results from an OrthoDB [27] search for 1:1 orthologs with *Drosophila melanogaster* genes available from [19]. First, each gene with a 1:1 ortholog was assigned to a ME based on the *D. melanogaster* chromosome map, and then scaffolds were assigned to a ME if the majority (> 50%) of genes with 1:1 orthologs were assigned to a single ME. This allowed us to assign scaffolds to autosomal *versus* sex-chromosome associated regions [19].

ddRAD-seq sampling

ddRAD-seq data was used to score single nucleotide polymorphisms (SNPs) for two analyses: (i) population genomics analysis used data from 627 flies from geographically diverse origins to confirm assignment of SNPs on autosomal *versus* sex chromosome associated genomic regions and to investigate population structure. (ii) Genome-wide association (GWA) analysis used data from 351 flies with a balanced study design (174:179 infected/control) to identify SNPs associated with trypanosome infection.

Population genomics analysis was completed with 627 flies originating from 33 geographically diverse sampling sites that spanned the three major genetic clusters of *Gff* found in Uganda, plus an outgroup chosen because of its more distant relationships in mitochondrial DNA phylogenetic analysis (Fig. 1; Supplementary Table S2 online). Comparisons of read coverage, heterozygosity and principal components analysis among sexes and geographic origin allowed us to confirm the scaffolds within autosomal *versus* sex-chromosome associated regions of the genome, and to identify general patterns of genomic structure in the SNPs scored for this study (see section of population genomics below). Both of these components of knowledge greatly aided in appropriate study design for the GWA.

GWA analysis was completed with a subset of 351 flies originating from 12 sampling sites to create a balanced study design with a total of 174 infected and 177 uninfected flies. Care was also taken to balance infection status among the sexes and genetic clusters from the north, west, and south of the country (Table 2). Of the infected flies, 48/66/6 were females from the north/west/south, respectively, and 28/23/3 were males from the north/west/south, respectively. Of the uninfected flies, 40/42/7 were females from the north/west/south, respectively, and 39/47/2 were males from the north/west/south, respectively (Supplementary Table S1 online).

All tsetse flies used in this study were collected using biconical Challier-Laveissiere traps set out in groups of 10–15 traps within a radius of 2 km, a field protocol that reliably traps unrelated individuals [21, 59]. Flies were

collected between January of 2014 and December of 2018 (Supplementary Table S1 online), sexed, dissected to determine midgut infection status microscopically, and preserved in 95% ethanol in screw cap vials. Specimens were stored at 4˚C for a maximum of 4 years before DNA extraction.

ddRAD-seq library preparation

DNA for the ddRAD protocol was extracted from the heads, thorax, wings, and legs of *Gff* using DNAeasy blood and tissue extraction kits (Qiagen, Valencia, CA), with a preliminary step added for tissue pulverization using the Qiagen Bead-beater system. We then quantified DNA extractions with a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA), and proceeded with only individuals having higher than 500 ng total yield of genomic DNA.

ddRAD sequencing libraries were prepared following [14] using a modified version of the Peterson ddRAD protocol [22] with restriction enzymes NlaIII and MluCI. We created eight ddRAD sequencing libraries of 32 pooled individuals each, which were sent to the Yale Center for Genome Analysis for 75 bp paired-read sequencing with the Illumina HiSeq 2500 platform under the “high-output” mode, and were sequenced in lanes shared with randomly sheared libraries.

ddRAD-seq bioinformatics

The ddRAD library raw sequence reads were de-multiplexed, quality filtered, and filtered for unambiguous barcodes using the “process_radtags” script from the STACKS2 v. 2.59 software [60] using the –retain-headers flag. FastQC v. 0.11.6 [61] was used on these processed reads to identify any read quality problems or overrepresented sequences. CutAdapt v. 1.18 [62] trimmed 5 bp of low quality data from the 3’ end of each read, and removed remaining overrepresented sequences (i.e. remaining Illumina TruSeq adaptors, ClassII mariner transposons, and mtDNA).

BWA mem [63] with default settings was used to align processed reads to the new 10X assembly and were sorted with Samtools v. 0.1.19 [64]. BAM file outputs were then filtered to remove reads assembled with more than 3 mismatches to the genome reference with BamTools v. 2.5.1 (filter -tag 'NM:<4') [65], and analyzed for SNPs with two separate programs, STACKS2 [60] and BamTools v. 2.5.1 MPILEUP [65]. Only overlap between these two software was retained for final analyses to reduce bias in SNP calls because there is strong evidence that different software identify different sets of SNPs [66–68]. SNPs from each software were independently filtered to remove SNPs within areas flagged by RepeatMasker v. 4.0.7 [69] with species set as *Drosophila*, and for genotyping missingness with an iterative strategy with PLINK v. 1.90-beta4.4 [70], alternating between per locus and per individual filters (–geno and –mind) applied at the following levels < 70%, < 65%, < 60%, < 55%, and < 50%, which is a strategy that has been shown to outperform hard cutoff filters [71].

Overlap between STACKS2 and MPILEUP was determined with VCFtools 0.1.16 [72] using the STACKS2 file as the main input, filtering to retain only positions that also existed in the MPILEUP file. Once combined, the final VCF file was filtered with VCFtools to retain only biallelic SNPs with minimum minor allele count of 3, mean depth greater than 10 and less than 500, and minimum genotyping rate of 50% per locus. The final population genomics dataset contained 627 individuals and 96,965 SNPS (63,652 autosomal, 29,069 sex-chromosome associated). This dataset was split into six files corresponding to the MEs identified using VCFtools (flag –bed).

ddRAD-seq population genomics

LD was characterized using PLINK [30, 31, 70]. We estimated pairwise r^2 (using the options –r2 –ld-window-r2 0), and LD block size (options –blocks –blocks-max-kb 200) for all polymorphic sites for each Mueller element and major genetic cluster (north, west, south) individually. We estimated the LD decay curves using r^2 estimates from PLINK with a fitting algorithm from the package "ngsLD" [73] with 100 bootstraps, and we tested different bin sizes until we obtained the smallest confidence intervals possible (final bin size chosen was 25 bp). We generated LD plots in R using the built-in functions and the R package "ggplot2" [74]. Read depth of coverage statistics were estimated using VCFtools (flag "–geno-depth") to confirm identification of sex chromosomes. Heterozygosity was estimated and tests for Hardy-Weinberg equilibrium were performed with PLINK [75, 76]. PCA was performed for each ME and for all scaffolds assigned to autosomes *versus* sex chromosomes using PLINK.

ddRAD-seq genome-wide association (GWA)

GWA analysis was performed with the R package "statgenGWAS" v. 1.0.7 [32] built under R v. 3.6.2 with ME specific kinship matrix in a mixed model, a method described to give a considerable improvement in power [33, 34]. First, SNP files from the appropriate MEs were combined to create two to separate datasets corresponding to the autosomal (MEs B, C, E) and the sex-chromosome (MEs A, D, F) associated regions of the genome. The autosomal dataset contained 50,960 SNPS and 351 individuals (142 males, 209 females), while the sex-chromosome dataset contained 19,423 SNPS and 209 individuals (all females).

Declarations

Acknowledgements

We acknowledge financial support from the Fogarty International Center (FIC) at the National Institutes of Health's (NIH's) Global Infectious Diseases Training Grant (award number D43TW007391), and from the Foundation for the NIH's Research Project Grant Program (award numbers AI068932 and 5T32AI007404-24). We acknowledge Alfonse Okello, Calvin Owora and Constant Khizza, and the rest of the Gulu University field team for help with sample collection, and Andrea Gloria-Soria, Augustine W. Dunn, and Carol Mariani for the smooth transfer of technical information from previous related projects in the Caccone lab.

Authors' contributions

Conceptualization: BLW, SA, AC. Formal analysis: NPS, JHS, LVC, YK. Project Administration and Funding Acquisition: SA, AC. Investigation: NPS, JHS, LVC, YK, ML, SW, BLW, RE, RO. Methodology: NPS, HZ, ML, SW, AC. Resources: HZ, YK, RE, SA, AC. Supervision: HZ, SA, RE, AC. Visualization: NPS, JHS, LVC. Writing – original draft: NPS, AC. Writing – review & editing: NPS, JHS, HZ, BLW, RE, RO, SA, AC. All authors read and approved the final version of the manuscript.

Data availability Statement

All data generated or analyzed during this study are included within this published article, associated supplementary information files, or publicly accessible online databases. The 10X Chromium genomic sequences generated and analyzed during the current study are available in the NCBI *via* BioProject accession number PRJNA596165 (www.ncbi.nlm.nih.gov/bioproject/596165), which includes links to associated

BioSample, Sequence Read Archive (SRA), and gene annotation accessions. ddRADseq DNA sequence files for this study have been deposited in the Sequence Read Archive (SRA) at NCBI under BioProject accession number PRJNA498097 (www.ncbi.nlm.nih.gov/bioproject/PRJNA498097). Resulting variant data for this study have been deposited in the European Variation Archive (EVA) at EMBL-EBI [77] under accession number PRJEB53725 (www.ebi.ac.uk/eva/?eva-study=PRJEB53725).

Additional information

Ethics approval and consent to participate is not applicable, as there were no human subjects involved in this study. There are no competing interests to report.

References

1. Muhanguzi, D. *et al.* African animal trypanosomiasis as a constraint to livestock health and production in Karamoja region: a detailed qualitative and quantitative assessment. *BMC Veterinary Research* **13**, 355 (2017).
2. Spickler, A. R. *African Animal Trypanosomiasis*. https://www.cfsph.iastate.edu/Factsheets/pdfs/trypanosomiasis_african.pdf (2018).
3. Brun, R., Blum, J., Chappuis, F. & Burri, C. Human African trypanosomiasis. *The Lancet* **375**, 148–159 (2010).
4. Wamwiri, F. N. & Changasi, R. E. Tsetse Flies (*Glossina*) as Vectors of Human African Trypanosomiasis: A Review. *BioMed Research International* **2016**, 1–8 (2016).
5. Omolo, M. O. *et al.* Prospects for Developing Odour Baits To Control Glossina fuscipes spp., the Major Vector of Human African Trypanosomiasis. *PLoS Neglected Tropical Diseases* **3**, e435 (2009).
6. Krafsur, E. S., Marquez, J. G. & Ouma, J. O. Structure of some East African *Glossina fuscipes fuscipes* populations. *Medical and Veterinary Entomology* **22**, 222–227 (2008).
7. Vallender, E. J. Positive selection on the human genome. *Human Molecular Genetics* **13**, R245–R254 (2004).
8. Mears, J. G. *et al.* Sickle gene. Its origin and diffusion from West Africa. *Journal of Clinical Investigation* **68**, 606–610 (1981).
9. Van Valen, L. A new evolutionary law. *Evolutionary Theory* **1**, 1–30 (1973).
10. Maudlin, I. Inheritance of susceptibility to *Trypanosoma congolense* infection in *Glossina morsitans*. *Annals of Tropical Medicine & Parasitology* **76**, 225–227 (1982).
11. Moloo, S. K., Kabata, J. M., Waweru, F. & Gooding, R. H. Selection of susceptible and refractory lines of *Glossina morsitans centralis* for *Trypanosoma congolense* infection and their susceptibility to different pathogenic *Trypanosoma* species. *Medical and Veterinary Entomology* **12**, 391–398 (1998).
12. Krafsur, E. S. & Maudlin, I. Tsetse fly evolution, genetics and the trypanosomiases - A review. *Infection, Genetics and Evolution* **64**, 185–206 (2018).
13. Gloria-Soria, A. *et al.* Patterns of Genome-Wide Variation in *Glossina fuscipes fuscipes* Tsetse Flies from Uganda. *G3 Genes|Genomes|Genetics* **6**, 1573–1584 (2016).

14. Gloria-Soria, A. *et al.* Uncovering Genomic Regions Associated with *Trypanosoma* Infections in Wild Populations of the Tsetse Fly *Glossina fuscipes*. *G3 Genes|Genomes|Genetics* **8**, 887–897 (2018).
15. Wallberg, A. *et al.* A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. *BMC Genomics* **20**, 275 (2019).
16. Zhang, L., Zhou, X., Weng, Z. & Sidow, A. Assessment of human diploid genome assembly with 10x linked-reads data. *GigaScience* **8**, (2019).
17. Li, Q. *et al.* A chromosome-scale genome assembly of cucumber (*Cucumis sativus* L.). *GigaScience* **8**, (2019).
18. Aksoy, S. *et al.* A case for a Glossina genome project. *Trends in Parasitology* **21**, 107–111 (2005).
19. Attardo, G. M. *et al.* Comparative genomic analysis of six Glossina genomes, vectors of African trypanosomes. *Genome Biology* **20**, 187 (2019).
20. König, I. R., Loley, C., Erdmann, J. & Ziegler, A. How to include chromosome X in your genome-wide association study. *Genetic Epidemiology* **38**, 97–103 (2014).
21. Saarman, N. P. *et al.* The population genomics of multiple tsetse fly (*Glossina fuscipes fuscipes*) admixture zones in Uganda. *Molecular Ecology* **28**, 66–85 (2019).
22. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model Species. *PLoS ONE* **7**, e37135 (2012).
23. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46 (2014).
24. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Research* **27**, 757–767 (2017).
25. Glossina Genomes Consortium. NCBI Glossina_fuscipes-3.0.2 Genome Assembly Report. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_000671735.1 (2014).
26. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34**, D590–D598 (2006).
27. Zdobnov, E. M. *et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research* **45**, D744–D749 (2017).
28. Beadell, J. S. *et al.* Phylogeography and Population Structure of Glossina fuscipes fuscipes in Uganda: Implications for Control of Tsetse. *PLoS Neglected Tropical Diseases* **4**, e636 (2010).
29. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
30. Gaunt, T. R., Rodríguez, S. & Day, I. N. Cubic exact solutions for the estimation of pairwise haplotype frequencies: Implications for linkage disequilibrium analyses and a web tool “CubeX.” *BMC Bioinformatics* **8**, 428 (2007).
31. Taliun, D., Gamper, J. & Pattaro, C. Efficient haplotype block recognition of very long and dense genetic sequences. *BMC Bioinformatics* **15**, 10 (2014).
32. Van Rossum, B.-J. *et al.* statgenGWAS: Genome Wide Association Studies. R package version 1.0.7. https://CRAN.R-project.org/package (2021).

33. Rincent, R. *et al.* Recovering Power in association mapping panels with variable levels of linkage disequilibrium. *Genetics* **197**, 375–387 (2014).
34. VanRaden, P. M. Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414–4423 (2008).
35. Gooding, R. H. & Krafsur, E. S. Tsetse genetics: Contributions to biology, systematics, and control of tsetse flies. *Annual Review of Entomology* **50**, 101–123 (2005).
36. Maudlin, I. Chromosome polymorphism and sex determination in a wild population of tsetse. *Nature* **277**, 300–301 (1979).
37. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3**, 299–309 (2002).
38. Li, M.-H. & Merila, J. Population differences in levels of linkage disequilibrium in the wild. *Molecular Ecology* **20**, 2916–2928 (2011).
39. Palmer, D. H., Rogers, T. F., Dean, R. & Wright, A. E. How to identify sex chromosomes and their turnover. *Molecular Ecology* **28**, 4709–4724 (2019).
40. Hansen, C. C. R., Westfall, K. M. & Pálsson, S. Evaluation of four methods to identify the homozygotic sex chromosome in small populations. *BMC Genomics* **23**, 160 (2022).
41. Vicoso, B. Molecular and evolutionary dynamics of animal sex-chromosome turnover. *Nature Ecology & Evolution* **3**, 1632–1641 (2019).
42. Li, H. & Ralph, P. Local PCA shows how the effect of population structure differs along the genome. *Genetics* **211**, 289–304 (2019).
43. Mérot, C. *et al.* Locally adaptive inversions modulate genetic variation at different geographic scales in a seaweed fly. *Molecular Biology and Evolution* **38**, 3953–3971 (2021).
44. Southern, D. Chromosome diversity in tsetse flies. In *Insect Cytogenetics* (eds. Blackman, R., Hewitt, G. & Ashburner, M.) **10**, 225–243 (Oxford: Blackwell Science, 1980).
45. The VEuPathDB Project Team. OrthMCL DB Release 6.10 21 group record OG6_101376. https://orthomcl.org/orthomcl/app/record/group/OG6_101376#Sequences (2022).
46. Team TVeP. Vectorbase GFUI026799 *lecithin-cholesterol acyltransferase* gene record https://vectorbase.org/vectorbase/app/record/gene/GFUI026799#MetabolicPathways (2022).
47. ELIXIR. PFAM Lecithin cholesterol acyltransferase gene family report. http://pfam.xfam.org/family/PF02450 (2022).
48. Hajduk, S. L. S., Hager, K. M. K. & Esko, J. D. J. Human high density lipoprotein killing of African trypanosomes. *Annual Review of Microbiology* **48**, (1994).
49. Vanhamme, L. & Pays, E. The trypanosome lytic factor of human serum and the molecular basis of sleeping sickness. *International Journal for Parasitology* **34**, 887–898 (2004).
50. Matthews, K. R. The developmental cell biology of *Trypanosoma brucei*. *Journal of Cell Science* **118**, 283–290 (2005).
51. Weiss, B. & Aksoy, S. Microbiome influences on insect host vector competence. *Trends in Parasitology* **27**, 514–522 (2011).

52. Miao, Q. & Ndao, M. *Trypanosoma cruzi* infection and host lipid metabolism. *Mediators of Inflammation* **2014**, 1–10 (2014).
53. Ray, S. S., Wilkinson, C. L. & Paul, K. S. Regulation of Trypanosoma brucei Acetyl Coenzyme A Carboxylase by Environmental Lipids. *mSphere* **3**, (2018).
54. x Genomics. *10x Genomics: Sample Preparation Demonstrated Protocol: DNA Extraction from Single Insects*. https://support.10xgenomics.com/de-novo-assembly/sample-prep/doc/demonstrated-protocol-dna-extraction-from-single-insects (2018).
55. Giraldo-Calderón, G. I. *et al.* VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research* **43**, D707–D713 (2015).
56. Attardo, G. M. & Aksoy, S. VectorBase Release 57 GfusI1.8 (GCA_000671735.1) *Glossina fuscipes* IAEA Genome Sequence and Annotation. https://vectorbase.org/vectorbase/app/record/dataset/TMPTX_gfusIAEA (2020).
57. Weller, G. L. & Foster, G. G. Genetic maps of the sheep blowfly *Lucilia cuprina*: linkage-group correlations with other dipteran genera. *Genome* **36**, 495–506 (1993).
58. Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in Diptera. *PLOS Biology* **13**, e1002078 (2015).
59. Echodu, R. *et al.* Genetically distinct *Glossina fuscipes fuscipes* populations in the Lake Kyoga region of Uganda and its relevance for human African trypanosomiasis. *BioMed Research International* **2013**, (2013).
60. Rochette, N. C., Rivera-Colón, A. G. & Catchen, J. M. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology* **28**, 4737–4754 (2019).
61. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
62. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).
64. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
66. Mielczarek, M. & Szyda, J. Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics* **57**, 71–79 (2016).
67. Baes, C. F. *et al.* Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics* **15**, 948 (2014).
68. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports* **5**, 17875 (2015).
69. Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. http://www.repeatmasker.org (2013-2015).
70. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81:559–75.

71. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Molecular Ecology. 2018;27:3193–206.
72. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
73. Fox EA, Wright AE, Fumagalli M, Vieira FG. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics. 2019;35:3855–6.
74. Wickam H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
75. Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg Equilibrium. The American Journal of Human Genetics. 2005;76:887–93.
76. Graffelman J, Moreno V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. Statistical Applications in Genetics and Molecular Biology. 2013;12.
77. Cezard T, Cunningham F, Hunt SE, Koylass B, Kumar N, Saunders G, Shen A, Silva AF, Tsukanov K, Venkataraman S, Flicek P, Parkinson H, Keane TM. The European Variation Archive: a FAIR resource of genomic variation for all species. Nucleic Acids Research. 2021; gkab960.

Figures

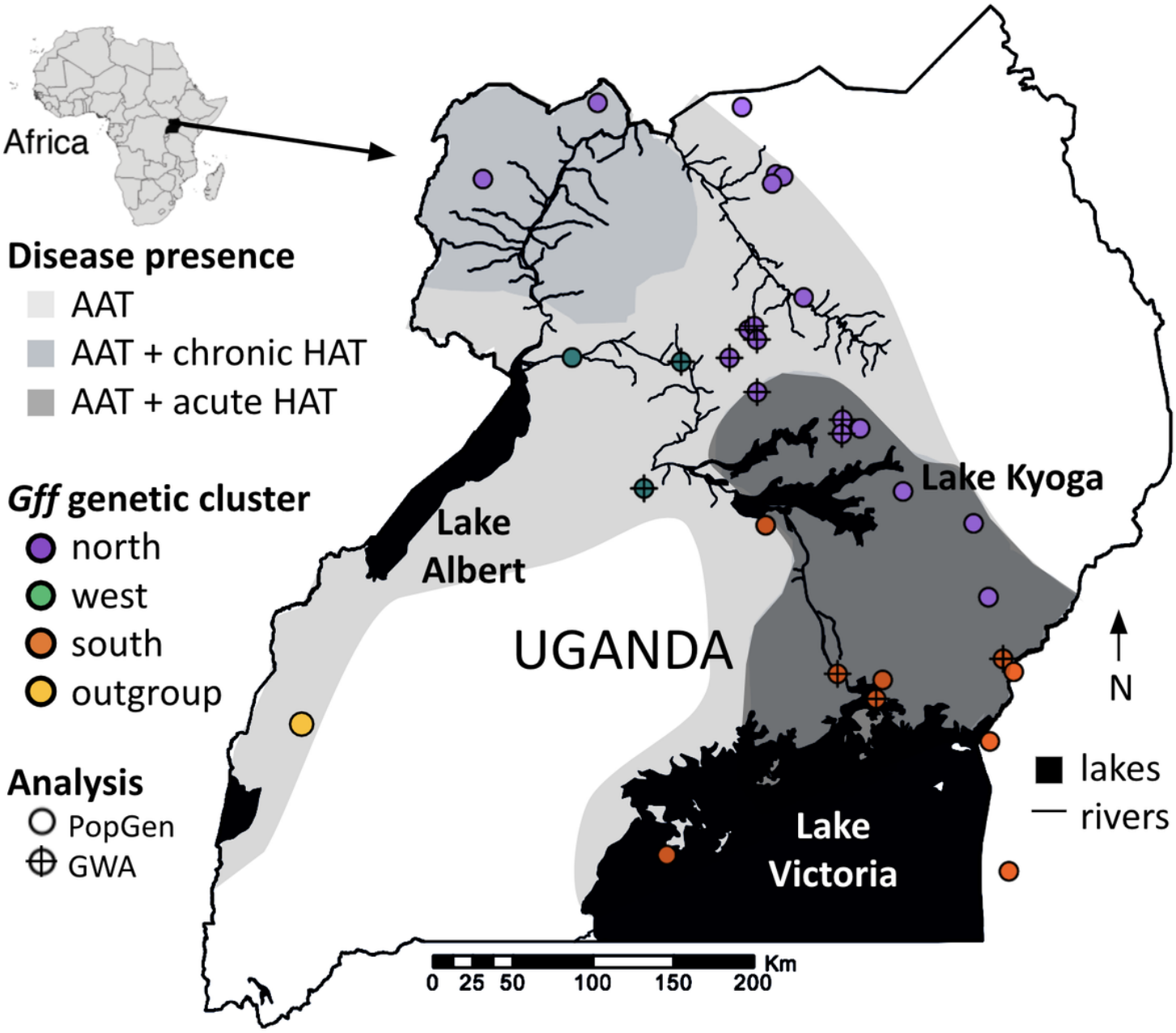

Figure 1

Map of the study area and sampling sites. The inset in upper left indicates the location of Uganda in Africa, disease presence is indicated in different shades of grey (AAT only = light grey, AAT and chronic HAT = medium grey, AAT and acute HAT = dark grey), *Glossina fuscipe fuscipes (Gff)* sampling points are indicated with symbols colored by genetic cluster (north = blue, west = purple, south = orange), and the analysis performed is indicated by the symbol (population genomics (PopGen) analysis = circle, genome wide association (GWA) analysis = circle with cross-hairs).

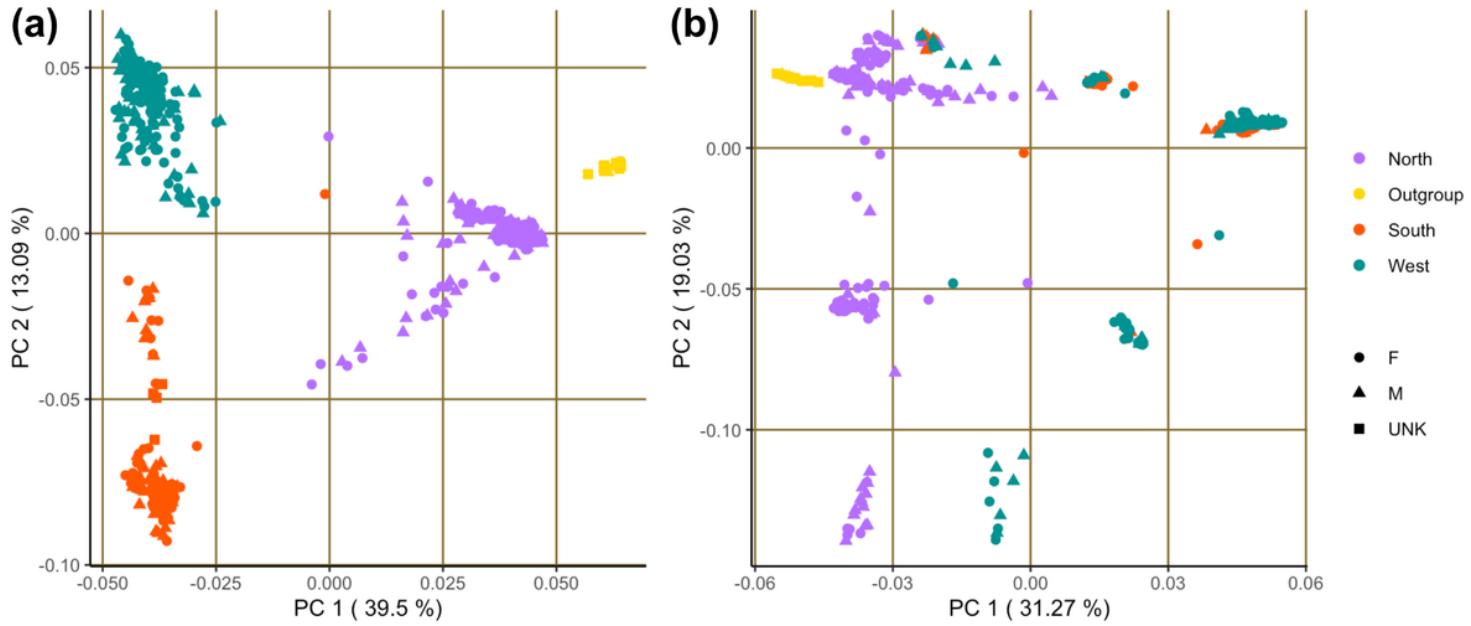

Figure 2

Principal components analysis of genomic variation of all individuals in the population genomics dataset for (a) autosome associated regions of the genome (MEs B, C, E), and (b) sex chromosome associated regions of the genome (MEs A, D, F). Shape indicates sex (female = circle, male = triangle, unknown = square), and color indicates the common genetic cluster of the sample's place of origin (north = blue, west = purple, south = orange, outgroup = yellow).

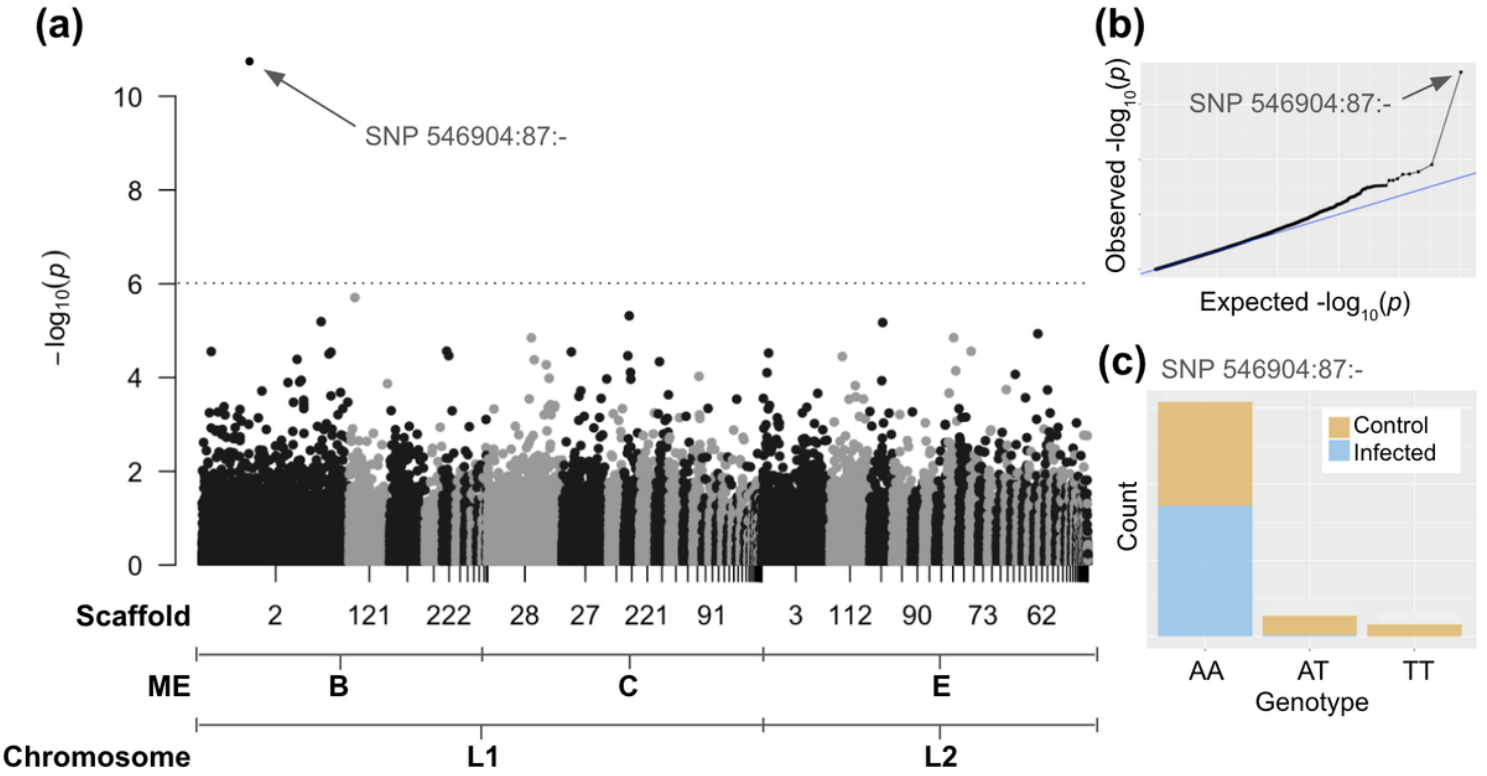

Figure 3

Results from GWA for autosome associated regions showing the **(a)** Manhattan plot of observed -log10(*p*) for GWA for each SNP arranged by scaffold number and Muller element assignment (ME), with arrow pointing to p-value of top SNP “546904:87:-”, significance threshold shown with dotted horizontal line, **(b)** Q-Q plot of observed versus expected -log10(p) indicating top SNP 546904:87:- has a highly significant signal of association with infection status, and **(c)** count of each genotype for top SNP 546904:87:- colored by infection status (yellow = control, blue = infected).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppFiguresS1S5.docx
- SuppNote.docx
- SuppTableS1.xlsx
- SuppTableS2.xlsx
- SuppTableS3.xlsx
- SuppTableS4.xlsx
- SuppTableS5.xlsx
- SuppTableS6.xlsx
- SuppTableS7.xlsx