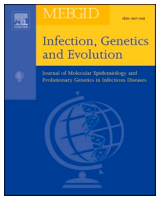
Infection, Genetics and Evolution 114 (2023) 105501

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid



Genomic evidence of sex chromosome aneuploidy and infection-associated genotypes in the tsetse fly *Glossina fuscipes,* the major vector of African trypanosomiasis in Uganda

Norah P. Saarman [a,*], Jae Hak Son [b], Hongyu Zhao [c], Luciano V. Cosme [d], Yong Kong [c], Mo Li [c], Shiyu Wang [e], Brian L. Weiss [c], Richard Echodu [f], Robert Opiro [f], Serap Aksoy [c], Adalgisa Caccone [d]

[a] *Utah State University, Logan, UT, USA*
[b] *Rutgers, The State University of New Jersey, Piscataway, NJ, USA*
[c] *Yale School of Public Health, New Haven, CT, USA*
[d] *Yale University, New Haven, CT, USA*
[e] *Emory University, Atlanta, GA, USA*
[f] *Gulu University, Gulu, Uganda*

ARTICLE INFO

Keywords:
ddRAD
Trypanosomiasis
Vector
Genome wide association
GWAS
Population genomics
Muller elements
Chromosome arms
Aneuploidy

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*). Our study improved the *Gff* genome assembly with whole genome 10× Chromium sequencing of a lab reared pupae, identified autosomal *versus* sex-chromosomal regions of the genome with ddRAD-seq data from 627 field caught *Gff*, and identified SNPs associated with trypanosome infection with genome-wide association (GWA) analysis in a subset of 351 flies. Results from 10× Chromium sequencing greatly improved *Gff* genome assembly metrics and assigned a full third of the genome to the sex chromosome. Results from ddRAD-seq suggested possible sex-chromosome aneuploidy in *Gff* and identified a single autosomal SNP to be highly associated with trypanosome infection. The top associated SNP was ~1100 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 suggest that there may be naturally occurring genetic variation in *Gff* in genomic regions in linkage disequilibrium with LCAT 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.

1. Introduction

Human and animal African trypanosomiasis (HAT and AAT, respectively) limit livestock production and represent a significant public health constraint (Muhanguzi et al., 2017; Spickler, 2018). 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 (Brun et al., 2010; Wamwiri and Changasi, 2016). 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*) (Omolo et al., 2009; Krafsur et al., 2008), which lives in lowland forests and along waterways in the western and central regions of the country (Fig. 1).

* Corresponding author.
E-mail addresses: norah.saarman@usu.edu (N.P. Saarman), jaehak.son@rutgers.edu (J.H. Son), hongyu.zhao@yale.edu (H. Zhao), luciano.cosme@yale.edu (L.V. Cosme), yong.kong@yale.edu (Y. Kong), shiyu.wang@emory.edu (S. Wang), brian.weiss@yale.edu (B.L. Weiss), robopiro@gu.ac.ug (R. Opiro), serap.aksoy@yale.edu (S. Aksoy), adalgisa.caccone@yale.edu (A. Caccone).

https://doi.org/10.1016/j.meegid.2023.105501
Received 3 May 2023; Received in revised form 7 September 2023; Accepted 11 September 2023
Available online 12 September 2023

Molecular mechanisms of protection against trypanosome infection in mammals are well described and known to involve high-density lipoprotein metabolism (Hajduk et al., 1994; Vanhamme and Pays, 2004), while the counterpart of these mechanisms in tsetse flies is an active area of research. Host-pathogen dynamics are among the strongest forces of selection (Vallender, 2004; Mears et al., 1981; Van Valen, 1973). 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 (Maudlin, 1982; Moloo et al., 1998; Krafsur and Maudlin, 2018) 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 (Gloria-Soria et al., 2016, 2018). 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 (Gloria-Soria et al., 2016).

Since the most recent GWA study in *Gff* was published (Gloria-Soria et al., 2018), 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.* 10× Chromium sequencing) with long-range contiguity. Having long-range contiguity allows assembly of longer stretches of DNA molecules especially in structurally complex genomic regions (Wallberg et al., 2019; Zhang et al., 2019; Li et al., 2019), 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 (Aksoy et al., 2005). Since approximately a third of it is on the sex chromosome (Attardo et al., 2019), this creates complications in GWA study design and interpretation (König et al., 2014) because of the different copy numbers in males *versus* females and the potential inactivation of large regions as a mechanism of dosage compensation. An additional potential complication in GWA design arises from studies since the 1980's that suggest sex-chromosome aneuploidy can occur in tsetse flies where females may be XX, XXY, or XXXY, and males may be XY, XYY, or XO (Southern, 1980), but this has not yet been investigated with population genomic data, or in *Gff*. 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 (Saarman et al., 2019). Analysis also demonstrated complex patterns of introgression that were non-uniform across the genome (Saarman et al., 2019), raising the possibility that population structure should also be accounted for non-uniformly 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, and also open the opportunity to investigate *Gff* chromosomal copy number in sex chromosome associated regions of the genome.

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,

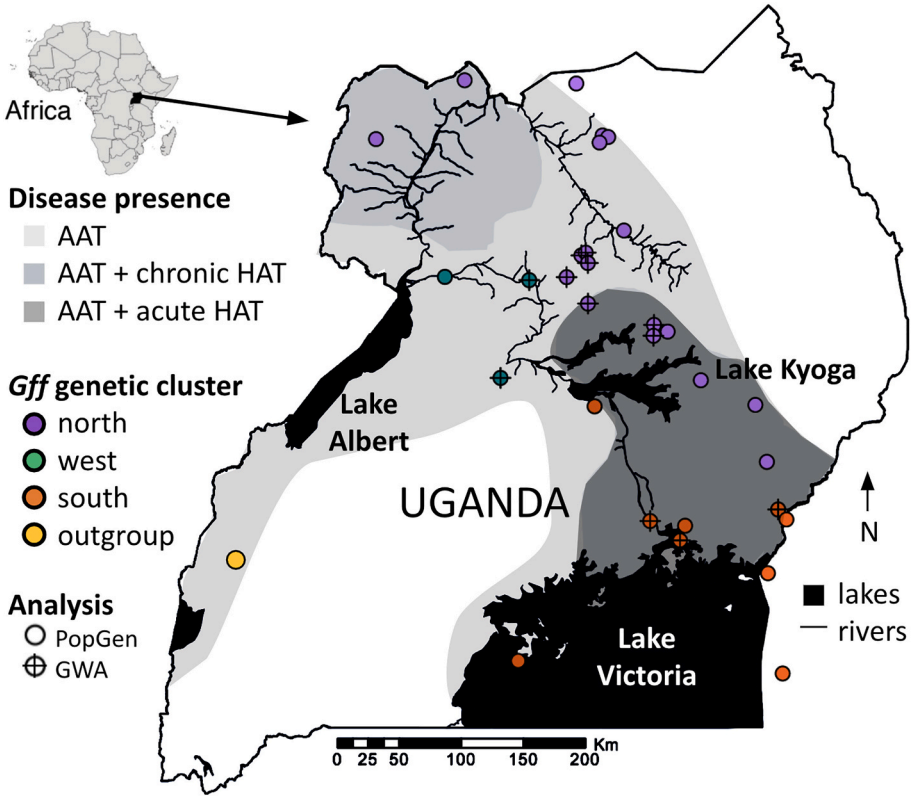


Fig. 1. Map of the study area and sampling sites. The inset in upper left indicates the location of Uganda in Africa, disease presence is indicated in different shades of grey (AAT only = light grey, AAT and chronic HAT = medium grey, AAT and acute HAT = dark grey), *Glossina fuscipe fuscipes (Gff)* 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). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 (Peterson et al., 2012) 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 41,970 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.

2. Material and methods

2.1. Whole-genome sequencing sampling and library prep

To improve the genome assembly of *Gff* we completed whole genome sequencing with 10× Chromium GenCode Technology (10× 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 10× Genomics® Sample Preparation Demonstrated Protocol DNA Extraction from Single Insects (10X Genomics, 2018), and was suspended in 1× TE Buffer. Genomic DNA was sent to the Yale Center for Genomic Analysis for quality control including a pulse field gel, library preparation, and 10× Genomics Genome Sequencing.

2.2. 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 (Wood and Salzberg, 2014). 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 10× Genomics (Weisenfeld et al., 2017) with –maxreads = 220,000,000. The annotations for the assembly were lifted over from the existing *Gff* assembly available from NCBI under the name "Glossina_fuscipes-3.0.2" (Glossina Genomes Consortium, 2014) and Vectorbase under the name "GfusI1" (Giraldo-Calderón et al., 2015; Attardo and Aksoy, 2020) using the UCSC liftOver suite of programs (Hinrichs, 2006).

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* (Attardo et al., 2019). 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 (Weller and Foster, 1993; Vicoso and Bachtrog, 2015). Scaffolds were assigned to MEs following (Attardo et al., 2019) based on results from an OrthoDB (Zdobnov et al., 2017) search for 1:1 orthologs with *Drosophila melanogaster* genes available from (Attardo et al., 2019). 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 (Attardo et al., 2019).

2.3. 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 4.2). Both 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 (Echodu et al., 2013; Saarman et al., 2019). 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.

2.4. 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 (Gloria-Soria et al., 2016) using a modified version of the Peterson ddRAD protocol (Peterson et al., 2012) with restriction enzymes *Nla*III and *Mlu*CI. 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.

2.5. 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 (Rochette et al., 2019) using the –retain-headers flag. FastQC v. 0.11.6 (Andrews, 2010) was used on these processed reads to identify any read quality problems or overrepresented sequences. CutAdapt v. 1.18 (Martin, 2011) 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 (Li and Durbin, 2009) with default settings was used to align processed reads to the new 10× assembly and were sorted with Samtools v. 0.1.19 (Li et al., 2009). BAM file outputs were then filtered to remove reads assembled with >3 mismatches to the genome reference with BamTools v. 2.5.1 (filter -tag 'NM:<4') (Barnett et al., 2011), and analyzed for SNPs with two separate programs, STACKS2 (Rochette et al., 2019) and BamTools v. 2.5.1 MPILEUP (Barnett et al., 2011). STACKS2 and MPILEUP were chosen because of their proven track record of providing repeatable identification of SNPs in single copy genes in non-model organisms (Rochette et al., 2019; Barnett et al., 2011). Overlap between these two software was retained for final analyses to improve repeatability of SNP identification for the loci used in this study since there is evidence that different software identify different sets of SNPs (Mielczarek and Szyda, 2016; Baes et al., 2014; Hwang et al., 2015). SNPs from each software were independently filtered to remove SNPs within areas flagged by RepeatMasker v. 4.0.7 (Smit et al., 2013–2015) with species set as *Drosophila*, and for genotyping missingness with an iterative strategy with PLINK v. 1.90-beta4.4 (Purcell et al., 2007a, 2007b), 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 (O'Leary et al., 2018).

Overlap between STACKS2 and MPILEUP was determined with VCFtools 0.1.16 (Danecek et al., 2011) 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 >10 and <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).

2.6. ddRAD-seq population genomics

LD was characterized using PLINK (Gaunt et al., 2007; Taliun et al., 2014; Purcell et al., 2007a, 2007b). 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" (Fox et al., 2019) 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" (Wickam, 2016). 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 (Wigginton et al., 2005; Graffelman and Moreno, 2013). PCA was performed for each ME and for all scaffolds assigned to autosomes *versus* sex chromosomes using PLINK.

2.7. ddRAD-seq genome-wide association (GWA)

GWA analysis was performed with the R package "statgenGWAS" v. 1.0.7 (Van Rossum et al., 2021) 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 (Rincent et al., 2014; VanRaden, 2008). 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 41,970 SNPS and 351 individuals (142 males, 209 females), while the sex-chromosome dataset contained 19,423 SNPS and 209 individuals (all females).

3. Results

3.1. Whole-genome sequencing, assembly and annotation

The DNA extraction from a *Gff* pupae whole body used for 10× Chromium GenCode Technology (10× 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 (Wood and Salzberg, 2014) and were removed. *De novo* assembly with the 10× Genomics supernova software package (Weisenfeld et al., 2017) yielded a total assembly of 390.09 Mb, with 7986 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) (Glossina Genomes Consortium, 2014).

We transferred 18,538 gene annotations from the existing assembly onto the new assembly using the UCSC liftOver suite of programs (Hinrichs, 2006). Of these genes, 6345 had 1:1 *Drosophila melanogaster* orthologs identified using OrthoDB (Zdobnov et al., 2017). Assignment to MEs (six chromosomal elements of Diptera species known as MEs A-F) were made following the method outlined (Attardo et al., 2019), 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).

3.2. 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

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×	56×
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	20,138	18,538
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%

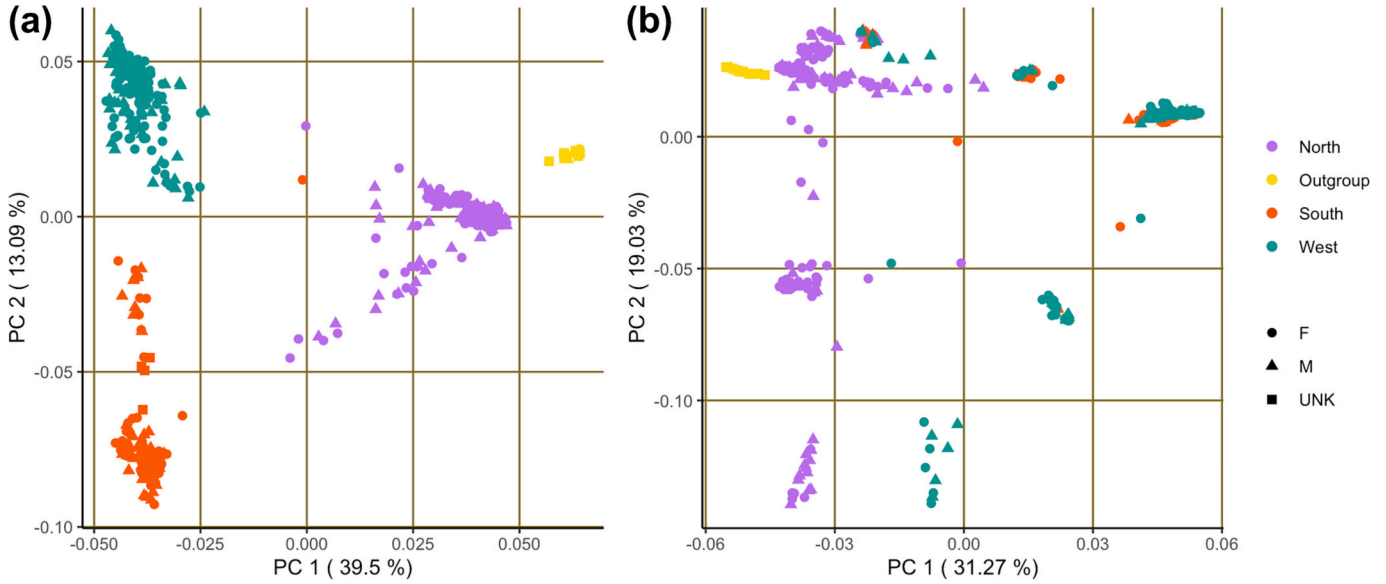


Fig. 2. Principal components analysis of genomic variation of all individuals in the population genomics dataset for (a) autosome associated regions of the genome (MEs B, C, E), and (b) sex chromosome associated regions of the genome (MEs A, D, F). Shape indicates sex (female = circle, male = triangle, unknown = square), and colour indicates the common genetic cluster of the sample's place of origin (north = blue, west = purple, south = orange, outgroup = yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

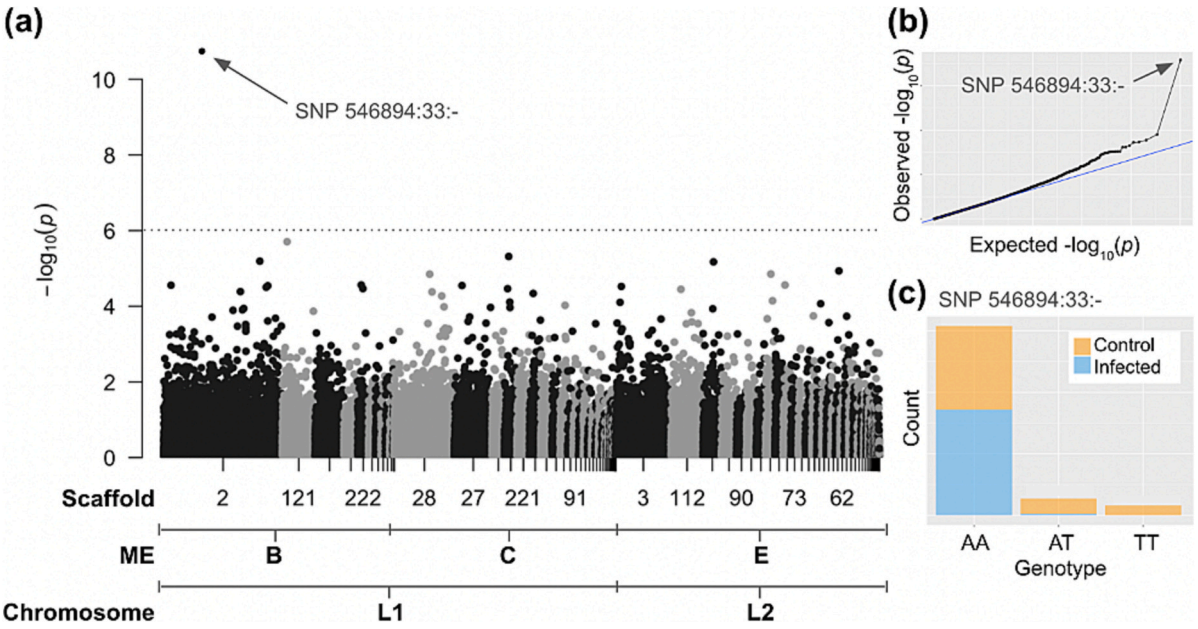


Fig. 3. Results from GWA for autosome associated regions showing the (a) Manhattan plot of observed -$\log_{10}(p)$ for GWA for each SNP arranged by scaffold number and Muller element assignment (ME), with arrow pointing to p-value of top SNP "546,894:33:-", significance threshold shown with dotted horizontal line, (b) Q-Q plot of observed *versus* expected -log10(p) indicating top SNP 546894:33:- has a highly significant signal of association with infection status, and (c) count of each genotype for top SNP 546894:33:- colored by infection status (yellow = control, blue = infected). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distant phylogenetic relationship (Beadell et al., 2010).

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 (Purcell et al., 2007a, 2007b; Gaunt et al., 2007; Taliun et al., 2014). 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; Fig. A.1 Appendix A). 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; Fig. A.2 Appendix A). 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 (Fig. A.1 Appendix A) and slow decay in the south and sex chromosome associated scaffolds (Fig. A.2 Appendix A).

Male *versus* female relative read depth of coverage and

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 and sample sizes (N), number infected, number uninfected (control), regional infection rate, and average number of reads and ddRAD-tags retained.

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%	14,874,020	69,857
West	3	1	201	190	89	89	11.91%	13,679,764	93,649
South	10	3	119	39	9	9	1.66%	13,223,360	86,746
Outgroup	1	0	15	0	NA	NA	NA	11,342,751	51,055
Overall	33	12	627	351	174	177	3.64%	14,092,162	80,256

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 (Fig. A.3 Appendix A), 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 (Fig. A.4 Appendix A), 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 (Attardo et al., 2019).

PCA identified different population structures in the autosomal *versus* sex chromosome associated SNPs (Fig. 2, Fig. A.5 Appendix A). 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 (Fig. A.5 Appendix A). Thus, taken together, PCA results indicate genomic structure with unresolved origins in the sex chromosome associated MEs.

3.3. 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 41,970 SNPs that made all data filter cuttoffs 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 (Van Rossum et al., 2021) with ME specific kinship matrices in a mixed model (Rincent et al., 2014; VanRaden, 2008). 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 546894:33:- 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; Supplementary Table S5 online) and were likely linked because of association with the phenotype rather than because of physical linkage with SNP 546894:33:-, as they were further from SNP 546894:33:- than the average LD decay rate (~2500 bp for Muller element B; Fig. A.2 Appendix A).

The top SNP 546894:33:- was at position 11,393,113 on scaffold_2 (NW_023998416.1). The minor allele (T) was associated with low (zero) infection (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) 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 546894:33:- was GFUI026799 *lecithin cholesterol acyltransferase* (LCAT), and was 1167 bp downstream of the SNP.

We repeated GWA in a subset of the data in a subsample of individuals from the populations where the minor allele (T) associated with low infection rates was found, namely samples from the west, which was represented in the final analysis by 36,121 scored SNPs from 177 flies (88 infected and 89 control; Supplementary Table S2 online). Results from this subset of the data corroborated the statistical association of infection status and the TT genotype at SNP 546894:33:- found in the overall dataset (Appendix B).

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 17,354 SNPs from 209 flies (120 infected and 89 control; Supplementary Table S2 online). Using females only 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.* (Gooding and Krafsur, 2005). Results indicated excessive false-positives (Appendix C), with lower-than-expected *p*-values (Fig. C.1 Appendix C, 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 discussion to the GWA of autosomal elements.

4. 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 (Fig. A.3 Appendix A) and heterozygosity (Fig. A.4 Appendix A), 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.

4.1. 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 Section 4.2).

4.2. 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 (Gooding and Krafsur, 2005; Maudlin, 1979). 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.

4.2.1. 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; Fig. A.1 Appendix A; Fig. A.2 Appendix A). 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; Fig. A.1 Appendix A; Fig. A.2 Appendix A) 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 (Ardlie et al., 2002), 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 (Li and Merila, 2011), which in turn is expected to be positively correlated with geographic range of sampling. Thus, variation in geographic range could also influence LD estimates.

4.2.2. Sequencing statistics suggest deviation from the classic XY sex-determination system

Male *versus* female coverage (Fig. A.3 Appendix A) and heterozygosity (Fig. A.4 Appendix A) 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 (Fig. A.3 Appendix A, Fig. A.4 Appendix A) 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 (Palmer et al., 2019; Hansen et al., 2022; Vicoso, 2019). Thus, the observed relative coverage of −0.5 (Fig. A.3 Appendix A) 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 (Palmer et al., 2019; Hansen et al., 2022; Vicoso, 2019). Thus, the extreme difference between males and females observed here (Fig. A.4 Appendix A) further supports that *Gff* does not follow the classic XY sex-determination system.

4.2.3. 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 (Saarman et al., 2019). 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 Saarman et al. (2019) ddRAD-seq 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 (Li and Ralph, 2019), and to create genomically localized heterogeneity (*i.e.*, contrasting patterns among genomic regions) in population structure (Mérot et al., 2021). Another possible explanation is sex-chromosome aneuploidy, a phenomenon known to occur in tsetse flies (Gooding and Krafsur, 2005). Sex-chromosome aneuploidy occurs in tsetse flies where females may be XX, XXY, or XXXY, and males may be XY, XYY, or XO (Gooding and Krafsur, 2005; Mérot et al., 2021; Southern, 1980). 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.

4.3. Genome-wide association

GWA analysis indicated a low false-positive rate in autosomal regions and identified a single SNP 546894:33:- with highly significant association with infection status in *Gff* (*p*-value = 1.80E-11; Fig. 3). Average LD decay for this Muller element is ~2500 bp, making detection of physically linked SNPs with a similar association with the phenotype of interest of very low probability. Thus, the "helicopter" pattern detected is not interpreted as evidence of a spurious association. The SNP 546894:33:- 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). Results imply there is natural genetic variation in *Gff* that can provide protection against trypanosome infection, and supports the concept of spatially patchy and temporally variable selection imposed by trypanosomiasis. Results are consistent with positive selection acting to increase the frequency of the minor allele at SNP 546894:33:- most strongly where infection rates are highest. However, the strength of positive selection experienced at this SNP is apparently insufficient to cause fixation of this protective allele in *Gff,* even on a local scale.

Although there is low likelihood that the associated SNP 546894:33:- detected in our analysis is the functional polymorphism itself, SNP 546894:33:- 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 (Gloria-Soria et al., *2016, 2018;* Attardo et al., 2019; The VEuPathDB Project Team, 2022; Team TVeP, 2022; ELIXIR, 2022). The orientation of the trypanosome associated SNP 546894:33:- at 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 the molecular pathway that initiates trypanosome lysis and protection in mammals. LCAT is involved 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 (Hajduk et al., 1994; Vanhamme and Pays, 2004). 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 (Vanhamme and Pays, 2004). 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 (Vanhamme and Pays, 2004). However, because trypanosomes remain in their bloodstream-form and are therefore susceptible to lysis by HDL until they reach tsetse's midgut (Matthews, 2005), it remains possible 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 (Weiss and Aksoy, 2011) 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 (Miao and Ndao, 2014), potentially indicating 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 546894:33:- 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 (Ray et al., 2018).

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. Further functional and comparative studies are needed to confirm and unravel the potential mechanism of the association between genotype at SNP 546894:33:- and *Gff* infection status. For example, a double-stranded-RNA-mediated gene interference (RNAi) experiment designed to knockdown expression of LCAT and test for infection susceptibility could confirm or refute a functional role of LCAT in protecting tsetse flies from trypanosome infection. In the same vein of research, 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.

5. 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. Population genomics results in sex chromosome associated SNPs of differences in male *versus* female read depth, heterozyosity, as well as complex population structure suggest aneuploidy, a finding that calls for further research in sex chromosome evolution in *Gff*.

Whole genome resequencing and population genomics additions to the study design significantly increased our power to detect GWA candidate SNPs beyond what was previously possible. GWA analysis identified one autosomal SNP 546894:33:- 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 homozygous 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.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2023.105501.

Additional information

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

CRediT authorship contribution statement

Norah P. Saarman: Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Jae Hak Son:** Formal analysis, Investigation, Visualization, Writing – review & editing. **Hongyu Zhao:** Methodology, Resources, Supervision, Writing – review & editing. **Luciano V. Cosme:** Formal analysis, Investigation, Visualization. **Yong Kong:** Formal analysis, Investigation, Resources. **Mo Li:** Investigation, Methodology. **Shiyu Wang:** Investigation, Methodology. **Brian L. Weiss:** Conceptualization, Investigation, Writing – review & editing. **Richard Echodu:** Investigation, Resources, Supervision, Writing – review & editing. **Robert Opiro:** Investigation, Writing – review & editing. **Serap Aksoy:** Conceptualization, Project

administration, Funding acquisition, Resources, Supervision, Writing – review & editing. **Adalgisa Caccone:** Conceptualization, Project administration, Funding acquisition, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

None.

Data availability

Sequence data from this article have been deposited with the GenBank NCBI data libraries under Project No. PRJNA596165 (www.ncbi.nlm.nih.gov/bioproject/596165) and PRJNA498097 (www.ncbi.nlm.nih.gov/bioproject/PRJNA498097), and variant data have been deposited in the EMBL-EBI data libraries (Cezard et al., 2022) under Accession No. PRJEB53725 (www.ebi.ac.uk/eva/?eva-study=PRJEB53725).

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.

References

10x Genomics, 2018. 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.

Aksoy, S., et al., 2005. A case for a Glossina genome project. Trends Parasitol. 21, 107–111.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ardlie, K.G., Kruglyak, L., Seielstad, M., 2002. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3, 299–309.

Attardo, G.M., Aksoy, S., 2020. VectorBase Release 57 GfusI1.8 (GCA_000671735.1) Glossina fuscipes IAEA Genome Sequence and Annotation. https://vectorbase.org/vectorbase/app/record/dataset/TMPTX_gfusIAEA.

Attardo, G.M., et al., 2019. Comparative genomic analysis of six Glossina genomes, vectors of African trypanosomes. Genome Biol. 20, 187.

Baes, C.F., et al., 2014. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. BMC Genomics 15, 948.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P., Marth, G.T., 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27, 1691–1692.

Beadell, J.S., et al., 2010. Phylogeography and population structure of Glossina fuscipes fuscipes in Uganda: implications for control of tsetse. PLoS Negl. Trop. Dis. 4, e636.

Brun, R., Blum, J., Chappuis, F., Burri, C., 2010. Human African trypanosomiasis. Lancet 375, 148–159.

Cezard, T., Cunningham, F., Hunt, S.E., Koylass, B., Kumar, N., Saunders, G., Shen, A., Silva, A.F., Tsukanov, K., Venkataraman, S., Flicek, P., Parkinson, H., Keane, T.M., 2022. The European Variation Archive: a FAIR resource of genomic variation for all species. Nucleic Acids Res 50, D1216–D1220.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., et al., 2011. The variant call format and VCFtools. Bioinformatics. 27, 2156–2158.

Echodu, R., et al., 2013. Genetically distinct Glossina fuscipes fuscipes populations in the Lake Kyoga region of Uganda and its relevance for human African trypanosomiasis. Biomed. Res. Int. 2013.

ELIXIR, . PFAM Lecithin cholesterol acyltransferase gene family report. http://pfam.xfam.org/family/PF02450.

Fox, E.A., Wright, A.E., Fumagalli, M., Vieira, F.G., 2019. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics. 35, 3855–3856.

Gaunt, T.R., Rodríguez, S., Day, I.N., 2007. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool "CubeX.". BMC Bioinformatics 8, 428.

Giraldo-Calderón, G.I., et al., 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. 43, D707–D713.

Gloria-Soria, A., et al., 2016. Patterns of genome-wide variation in Glossina fuscipes fuscipes tsetse flies from Uganda. G3 Genes Genomes Genet. 6, 1573–1584.

Gloria-Soria, A., et al., 2018. Uncovering genomic regions associated with Trypanosoma infections in wild populations of the tsetse fly Glossina fuscipes. G3 Genes Genomes Genet. 8, 887–897.

Glossina Genomes Consortium, 2014. NCBI Glossina_fuscipes-3.0.2 Genome Assembly Report. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_000671735.1.

Gooding, R.H., Krafsur, E.S., 2005. Tsetse genetics: contributions to biology, systematics, and control of tsetse flies. Annu. Rev. Entomol. 50, 101–123.

Graffelman, J., Moreno, V., 2013. The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat. Appl. Genet. Mol. Biol. 12.

Hajduk, S.L.S., Hager, K.M.K., Esko, J.D.J., 1994. Human high density lipoprotein killing of African trypanosomes. Annu. Rev. Microbiol. 48.

Hansen, C.C.R., Westfall, K.M., Pálsson, S., 2022. Evaluation of four methods to identify the homozygotic sex chromosome in small populations. BMC Genomics 23, 160.

Hinrichs, A.S., 2006. The UCSC genome browser database: update 2006. Nucleic Acids Res. 34, D590–D598.

Hwang, S., Kim, E., Lee, I., Marcotte, E.M., 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci. Rep. 5, 17875.

König, I.R., Loley, C., Erdmann, J., Ziegler, A., 2014. How to include chromosome X in your genome_wide association study. Genet. Epidemiol. 38, 97–103.

Krafsur, E.S., Maudlin, I., 2018. Tsetse fly evolution, genetics and the trypanosomiases - a review. Infect. Genet. Evol. 64, 185–206.

Krafsur, E.S., Marquez, J.G., Ouma, J.O., 2008. Structure of some East African Glossina fuscipes fuscipes populations. Med. Vet. Entomol. 22, 222–227.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25, 1754–1760.

Li, M.-H., Merila, J., 2011. Population differences in levels of linkage disequilibrium in the wild. Mol. Ecol. 20, 2916–2928.

Li, H., Ralph, P., 2019. Local PCA shows how the effect of population structure differs along the genome. Genetics 211, 289–304.

Li, H., et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, Q., et al., 2019. A chromosome-scale genome assembly of cucumber (*Cucumis sativus* L.). GigaScience 8.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17 (10).

Matthews, K.R., 2005. The developmental cell biology of Trypanosoma brucei. J. Cell Sci. 118, 283–290.

Maudlin, I., 1979. Chromosome polymorphism and sex determination in a wild population of tsetse. Nature 277, 300–301.

Maudlin, I., 1982. Inheritance of susceptibility to Trypanosoma congolense infection in Glossina morsitans. Ann. Trop. Med. Parasitol. 76, 225–227.

Mears, J.G., et al., 1981. Sickle gene. Its origin and diffusion from West Africa. J. Clin. Investig. 68, 606–610.

Mérot, C., et al., 2021. Locally adaptive inversions modulate genetic variation at different geographic scales in a seaweed fly. Mol. Biol. Evol. 38, 3953–3971.

Miao, Q., Ndao, M., 2014. Trypanosoma cruzi infection and host lipid metabolism. Mediat. Inflamm. 2014, 1–10.

Mielczarek, M., Szyda, J., 2016. Review of alignment and SNP calling algorithms for next-generation sequencing data. J. Appl. Genet. 57, 71–79.

Moloo, S.K., Kabata, J.M., Waweru, F., Gooding, R.H., 1998. Selection of susceptible and refractory lines of Glossina morsitans centralis for Trypanosoma congolense infection and their susceptibility to different pathogenic Trypanosoma species. Med. Vet. Entomol. 12, 391–398.

Muhanguzi, D., et al., 2017. African animal trypanosomiasis as a constraint to livestock health and production in Karamoja region: a detailed qualitative and quantitative assessment. BMC Vet. Res. 13, 355.

O'Leary, S.J., Puritz, J.B., Willis, S.C., Hollenbeck, C.M., Portnoy, D.S., 2018. These aren't the loci you'e looking for: principles of effective SNP filtering for molecular ecologists. Mol. Ecol. 27, 3193–3206.

Omolo, M.O., et al., 2009. Prospects for developing odour baits to control Glossina fuscipes spp., the major vector of human African Trypanosomiasis. PLoS Negl. Trop. Dis. 3 e435.

Palmer, D.H., Rogers, T.F., Dean, R., Wright, A.E., 2019. How to identify sex chromosomes and their turnover. Mol. Ecol. 28, 4709–4724.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al., 2007a. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Purcell, S., et al., 2007b. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Ray, S.S., Wilkinson, C.L., Paul, K.S., 2018. Regulation of Trypanosoma brucei acetyl coenzyme A carboxylase by environmental lipids. mSphere 3.

Rincent, R., et al., 2014. Recovering power in association mapping panels with variable levels of linkage disequilibrium. Genetics 197, 375–387.

Rochette, N.C., Rivera-Colón, A.G., Catchen, J.M., 2019. Stacks 2: analytical methods for paired_end sequencing improve RADseq_based population genomics. Mol. Ecol. 28, 4737–4754.

Saarman, N.P., et al., 2019. The population genomics of multiple tsetse fly (Glossina fuscipes fuscipes) admixture zones in Uganda. Mol. Ecol. 28, 66–85.

Smit, A.F.A., Hubley, R., Green, P., 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Southern, D., 1980. Chromosome diversity in tsetse flies. In: Blackman, R., Hewitt, G., Ashburner, M. (Eds.), Insect Cytogenetics, 10. Blackwell Science, Oxford, pp. 225–243.
Spickler, A.R., 2018. African Animal Trypanosomiasis. https://www.cfsph.iastate.edu/Factsheets/pdfs/trypanosomiasis_african.pdf.
Taliun, D., Gamper, J., Pattaro, C., 2014. Efficient haplotype block recognition of very long and dense genetic sequences. BMC Bioinformatics 15, 10.
Team TVeP, . Vectorbase GFUI026799 lecithin-cholesterol acyltransferase gene record. https://vectorbase.org/vectorbase/app/record/gene/GFUI026799#MetabolicPathways.
The VEuPathDB Project Team, . OrthMCL DB Release 6.10 21 group record OG6_101376. https://orthomcl.org/orthomcl/app/record/group/OG6_101376#Sequences.
Vallender, E.J., 2004. Positive selection on the human genome. Hum. Mol. Genet. 13, R245–R254.
Van Rossum, B.-J., et al., 2021. statgenGWAS: Genome Wide Association Studies. R package version 1.0.7. https://CRAN.R-project.org/package.
Van Valen, L., 1973. A new evolutionary law. Evol. Theor. 1, 1–30.
Vanhamme, L., Pays, E., 2004. The trypanosome lytic factor of human serum and the molecular basis of sleeping sickness. Int. J. Parasitol. 34, 887–898.
VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423.
Vicoso, B., 2019. Molecular and evolutionary dynamics of animal sex-chromosome turnover. Nat. Ecol. Evol. 3, 1632–1641.
Vicoso, B., Bachtrog, D., 2015. Numerous transitions of sex chromosomes in Diptera. PLoS Biol. 13, e1002078.
Wallberg, A., et al., 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics 20, 275.
Wamwiri, F.N., Changasi, R.E., 2016. Tsetse flies (Glossina) as vectors of human African trypanosomiasis: a review. Biomed. Res. Int. 2016, 1–8.
Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B., 2017. Direct determination of diploid genome sequences. Genome Res. 27, 757–767.
Weiss, B., Aksoy, S., 2011. Microbiome influences on insect host vector competence. Trends Parasitol. 27, 514–522.
Weller, G.L., Foster, G.G., 1993. Genetic maps of the sheep blowfly Lucilia cuprina_: linkage-group correlations with other dipteran genera. Genome 36, 495–506.
Wickam, H., 2016. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag, New York. Available from. https://ggplot2.tidyverse.org.
Wigginton, J.E., Cutler, D.J., Abecasis, G.R., 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. 76, 887–893.
Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46.
Zdobnov, E.M., et al., 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 45, D744–D749.
Zhang, L., Zhou, X., Weng, Z., Sidow, A., 2019. Assessment of human diploid genome assembly with 10x linked-reads data. GigaScience 8.