Developing and quality testing of microsatellite loci for four species of Glossina

Sophie Ravel, Modou Séré, Oliver Manangwa, Moise Kagbadouno, Mahamat Hissene Mahamat, William Shereni, Winnie A. Okeyo, Rafael Argiles-Herrero, Thierry de Meeûs

Developing and quality testing of microsatellite loci for four species of *Glossina*

Sophie Ravel[1], Modou Sere[2], Oliver Manangwa[3], Moise Kagbadouno[4], Mahamat Hissene Mahamat[5], William Shereni[6] , Winnie A. Okeyo[7], Rafael Argiles-Herrero[8], Thierry De Meeûs[1]

[1] Intertryp, IRD, Cirad, Univ Montpellier, Montpellier, France

[2] University of Dédougou, B.P. 176 Dédougou, Burkina Faso

[3] Vector and Vector Borne Disease Research Institute, P.O.Box 1026, Tanga, Tanzania

[4] Programme National de Lutte contre la THA (PNLTHA), Conakry, Guinée

[5] Institut de Recherche en Elevage pour le Développement (IRED), Ndjaména, Tchad

[6] Ministry of Lands, Agriculture, Water and Rural Resettlement, Harare, Zimbabwe

[7] Biotechnology Research Institute - KALRO: Kikuyu, Kenya

[8] Insect Pest Control Sub-Programme, Joint Food and Agriculture Organization of the United Nations/International Atomic Energy Agency Programme of Nuclear Techniques in Food and Agriculture, A-1400, Vienna, Austria.

Corresponding author: Thierry De Meeûs, Thierry.demeeus@ird.fr

Keywords: Tsetse flies; Microsatellite loci, Population genetics, Trypanosomiasis, Hetero-chromosomes.

Abstract

Microsatellite loci still represent valuable resources for the study of the population biology of non-model organisms. Discovering or adapting new suitable microsatellite markers in species of interest still represents a useful task, especially so for non-model organisms as tsetse flies (genus *Glossina*), which remain a serious threat to the health of humans and animals in sub-Saharan Africa. In this paper, we present the development of new microsatellite loci for four species of *Glossina*: two from the *Morsitans* group, *G. morsitans morsitans* (Gmm) from Zimbabwe, *G. pallidipes* (Gpalli) from Tanzania; and the other two from the *Palpalis* group, *G. fuscipes fuscipes* (Gff) from Chad, and *G. palpalis gambiensis* (Gpg) from Guinea. We found frequent short allele dominance and null alleles. Stuttering could also be found and amended when possible. Cryptic species seemed to occur frequently in all taxa but Gff. This explains why it may be difficult finding ecumenical primers, which thus need adaptation according to each taxonomic and geographic context. Amplification problems occurred more often in published old markers, and Gmm and Gpg were the most affected (stronger heterozygote deficits). Trinucleotide markers displayed selection signature in some instances (Gmm). Combining old and new loci, for Gmm, eight loci can be safely used (with correction for null alleles); and five seem particularly promising; for Gpalli, only five to three loci worked well, depending on the clade, which means that the use of loci from other species (four *morsitans* loci seemed to work well), or other new primers will need to be used; for Gff, 14 loci behaved well, but with null alleles, seven of which worked very well; and for *G. palpalis sl*, only four loci, needing null allele and stuttering corrections seem to work well, and other loci from the literature are thus needed, including X-linked markers, five of which seem to work rather well (in female only), but new markers will probably be needed. Finally, the high proportion of X-linked markers (around 30%) was explained by the non-Y DNA quantity and chromosome structure of tsetse flies studied so far.

1. Introduction

The study of the spatiotemporal genetic variation has proven a useful means to study the population biology of natural populations, especially for non-model organisms of small sizes and/or difficult to sample as parasites and their vectors (Manangwa et al., 2019). To this respect, microsatellite loci still represent valuable resources. These markers are frequent in the genome of most eukaryotes, easy to handle and highly polymorphic (Katti et al., 2001), except in some flying vertebrates like birds and bats where weight loss evolution explains their rarity (Primmer et al., 1997). Moreover, their extensive use in various organisms over a fairly long period of time have allowed the development of efficient methods to detect and cure allele mis-scoring due to the more or less frequent amplification problems met, such as null alleles, short allele dominance (SAD), allelic dropout, and stuttering (Chapuis and Estoup, 2007; Wang et al., 2012; Séré et al., 2014; Séré et al., 2017; De Meeûs, 2018; De Meeûs et al., 2019a; Manangwa et al., 2019). More modern and numerous single nucleotide polymorphisms (SNPs) may be used instead, but one can see many reasons why SNPs may be less suitable than microsatellites in population genetics studies. Given the very low polymorphism and resulting high variance expected from one SNP to the other, and given the number of loci required (at least 200) (Séré et al., 2017), it is hard to think of a reliable method that could efficiently detect loci under selection or in linkage disequilibrium, unless thousands of individuals are genotyped. Moreover, SNPs are not immune of genotyping errors (Vignal et al., 2002; Garvin et al., 2010; Ekblom and Galindo, 2011; Helyar et al., 2011; Bayerl et al., 2018). Many problems will probably emerge in a near future with their extended use, as was the case for microsatellite loci since the beginning of their utilization in the nineties (Ellegren, 2004; Richard et al., 2008) and the discovery of the first amplification problems as null alleles (Paetkau and Strobeck, 1995), stuttering (Murray et al., 1993) or SAD (Wattier et al., 1998). Consequently, discovering or adapting new suitable microsatellite markers in

species and/or populations of interest still represents a useful task, especially so for non-model organisms, for which pitfalls and drawbacks are expected with SNPs based methods.

Tsetse flies (genus *Glossina*) remain a very serious threat to the health of humans and animals in sub-Saharan Africa (Solano et al., 2010b). They indeed represent the cyclical vector of human African trypanosomiasis (HAT), also known as sleeping sickness, a lethal neglected tropical disease that was targeted for elimination by the WHO (Franco et al., 2018). Tsetse flies also transmit the animal African trypanosomiasis (AAT), also known as nagana. Nagana ravages domestic mammal livestock of sub-Saharan Africa where its cost was estimated $ 4.75 billion per year (Van den Bossche et al., 2010). Studying the population biology of these pests with microsatellite markers and population genetics tools represents an interesting step to design the best control strategies or to understand the failure of past eradication campaigns. This has been done for several species and in several countries with more or less success (for most recent reviews, see (Solano et al., 2010b; Krafsur and Maudlin, 2018; De Meeûs et al., 2019b) and references therein). Nevertheless, some difficulties occurred during the transfer of different loci from one species to another or, within the same species, from one geographic zone to the other or from one cryptic species to the other. As discussed elsewhere (Manangwa et al., 2019), this should impel the redesign of more specific primers for extant and/or for new microsatellite loci. In this paper, we present the development of new microsatellite loci for four species of *Glossina*: two from the *Morsitans* group, *G. morsitans morsitans* from Zimbabwe, and *G. pallidipes* from Tanzania; and the other two from the *Palpalis* group, *G. fuscipes fuscipes* from Chad, and *G. palpalis gambiensis* from Guinea. We discuss some basic population genetics statistics found for those and compare it with already available loci. When available, we also compare the respective performances of trinucleotide and

other kind of loci. We also provide a logical explanation for the frequency with which these markers were found on the X chromosome.

2. Material and methods

2.1. *Glossina morsitans morsitans* (Gmm)

Twenty six microsatellite markers were available from the literature for this species (Baker and Krafsur, 2001; Hyseni et al., 2011; Molecular Ecology Resources Primer Development Consortium et al., 2011) (Table 1). Twenty of these loci were evaluated as X-linked or as autosomal (Ouma et al., 2007; Hyseni et al., 2011; Molecular Ecology Resources Primer Development Consortium et al., 2011). Thus six were of unknown (untested) chromosomic location (Gmm22, Gmm127, GmcCA16C, Gmm14, Gmm15, and GmsCAG6); and five were identified as X-linked (GmmC15, GmmD03, GmmF10, GmmL17, and GmmP07) (Hyseni et al., 2011; Molecular Ecology Resources Primer Development Consortium et al., 2011); while the remaining 15 were identified as autosomal. We re-evaluated twelve of these microsatellite markers: Gmm5B, Gmm8, Gmm9B, Gmm22, Gmm127, GmsCAG6, GpCAG133, GmsCAG29B, GmmA06, GmmB20, GmmH09, and GmmK06. These loci were evaluated on an ABI 3500xl sequencer (Applied Biosytems, Waltham, Massachusetts, USA) on 23 males of *G. morsitans morsitans* from Zimbabwe, because males are expected to be homozygous at X-linked loci.

Table 1: Published primers for microsatellites in *Glossina morsitans morsitans*, their motif, first publication (Reference a), chromosome location (Chromosome) and reference for chromosome location (Reference b).

Locus name	Motif	Reference a	Chromosome	Reference b
Gmm127	dinucleotide	Baker and Krafsur 2001	Unknown	NA
Gmm14	dinucleotide	Baker and Krafsur 2001	Unknown	NA
Gmm15	dinucleotide	Baker and Krafsur 2001	Unknown	NA
Gmm22	dinucleotide	Baker and Krafsur 2001	Unknown	NA
GmcCA16C	dinucleotide	Baker and Krafsur 2001	Unknown	NA
GmsCAG6	trinucleotide	Baker and Krafsur 2001	Unknown	Ouma et al. 2007
Gmm5B	dinucleotide	Baker and Krafsur 2001	Autosomal	Ouma et al. 2007
Gmm8	dinucleotide	Baker and Krafsur 2001	Autosomal	Ouma et al. 2007
Gmm9B	dinucleotide	Baker and Krafsur 2001	Autosomal	Ouma et al. 2007
GmsCAG2	trinucleotide	Baker and Krafsur 2001	Autosomal	Ouma et al. 2007
GmsCAG29B	trinucleotide	Baker and Krafsur 2001	Autosomal	Ouma et al. 2007
GpCAG133	trinucleotide	Baker and Krafsur 2001	Autosomal	Ouma et al. 2007
GmmA06	dinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmB20	dinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmD15	dinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmH09	dinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmL03	dinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmL11	dinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmF10	dinucleotide	Hyseni et al. 2011	X-linked	Hyseni et al. 2011
GmmK22	trinucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmC15	trinucleotide	Hyseni et al. 2011	X-linked	Hyseni et al. 2011
GmmD03	trinucleotide	Hyseni et al. 2011	X-linked	Hyseni et al. 2011
GmmP07	trinucleotide	Hyseni et al. 2011	X-linked	Hyseni et al. 2011
GmmC17	tetranucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmK06	tetranucleotide	Hyseni et al. 2011	Autosomal	Hyseni et al. 2011
GmmL17	tetranucleotide	Hyseni et al. 2011	X-linked	Hyseni et al. 2011

We also designed eight di-nucleotidic microsatellite markers from the *morsitans* genome at the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/project/pathogens/Glossina/morsitans/Assemblies/) using Tandem Repeats Finder Program (Benson, 1999) (with parameters 2 7 7 80 10 150 2): Gmm7278, Gmm8710, Gmm12610, Gmm20291, Gmm26134, Gmm31657, Gmm41987 and Gmm66047. The full sequences are available in supplementary file S1. Primers, motif and product sizes are presented in Table 2. These markers were also evaluated with the same dataset as for already published loci. Conditions for extraction and amplifications were the following. Three legs (or parts of thorax when the legs were absent) were

removed from each of the flies, dried and subjected to chelex treatment in order to obtain DNA for further genotyping. Chelex treatment: 200 µl of 5% Chelex® chelating resin under stirring was added to each tube containing the dried legs. Legs were crushed using piston pellet. After incubation at 56°C for one hour, DNA was denatured at 95°C for 30 min. The tubes were then centrifuged 3 min. at 12,000 g for three min. Supernatants were used for genotyping. Ten µl of each DNA (dilution 1/5) was used in the touch-down PCR reactions (to increase the specificity of PCR) which were carried out in a thermocycler (Mastercycler, Eppendorf, Hamburg, Germany) in 20 µl final volume. Indeed, we cannot quantify the amount of DNA obtained from chelex treatment because it is not pure DNA. Based on our previous experience we know that 10 microliter of DNA previously diluted to 1/5 is sufficient to perform the PCRs reaction used in the ABI 3500XL technologies. PCR conditions for primer pairs were briefly as follows: 40 cycles and annealing temperature as described in Table 2 for the new microsatellite markers. For the old ones an annealing temperature of 48°C for GmmA06, GmmB20, GmmH09, Gmm127 and GmsCAG29B, 50°C for Gmm9B, and GpCAG133 , 54°C for Gmm22 and GmsCAG6, 55°C for Gmm5B, Gmm8 and GmmK06 was used.

Table 2: Primers' description of the eight new microsatellite markers designed for *Glossina morsitans morsitans*. Range of product size is given in number of base pairs (bp) and annealing temperature (T) in °C.

Locus	Repeat motif	Range (bp)	T (° C)	Primer pair sequences (5'-3')
Gmm7278	(GA)40	232	56	F: TGTGTTAGGAAGAGCGAATAGG R: TTGCTTTTGGCATCTGTTCA
Gmm8710	(TC)38	180	55	F: ATGCATGCTTGCACTAACC R: ATTAATTATGACGGGCAAGC
Gmm12610	(AG)39	208	50	F: GGTTTTTAGTGAGCCGGTGT R: CTTATCCGAATTAAGTACAACC
Gmm20291	(TG)37	184	55	F: CAAGAGAATATTTATTTTTGCAAGC R: ACAGCGTCTACCCCAAGATG
Gmm26134	(TG)37	174	55	F: TTTAAGAAATGTTTGTTGGTTCG R: AACAGGGAACACGAGTAGGC
Gmm31657	(TA)40	242	55	F: TGCATTTAAAAACAAAGCAGTG R: GCTGCTCCTAAGTGGTCTTTTC
Gmm41987	(TA)37	201	57	F: GAAAAGCATCTGGCTCATGG R: TGTGCGAGTATGTGCATTTG
Gmm66047	(TG)39	198	57	F: TTAGGGTTCCTCCGTCCTTT R: TCATTAACAAATGGGCCAGAG

After PCR amplification, allele bands were routinely resolved on ABI 3500XL sequencer. This method allows multiplex by the use of four dyes. Allele calling was done using GENEMAPPER 4.1 software and the size standard GS600LIZ short run.

2.2. *Glossina pallidipes* (Gpalli)

Eleven loci developed for Gpalli can be found in the literature: GpA19a (dinucleotide), GpA23b (dinucleotide), GpB6b (dinucleotide), GpB20b (dinucleotide), GpC5b (trinucleotide), GpC10b (trinucleotide), GpC26b (trinucleotide), and GpD18b (trinucleotide) (Ouma et al., 2003); and GpB115 (dinucleotide), GpC101 (trinucleotide), and GpC107 (trinucleotide) (Ouma et al., 2006b). No specific screening for heterosomal loci was ever done for these markers so that we just considered that loci with strong tendencies for heterozygote deficits and generally never or rarely used in publications, to be X-linked, unless proven otherwise. Not enough DNA was left to study these loci on Tanzanian samples as for the new markers in the following. Instead, we used the data from the Nguruman escarpment in Kenya (Okeyo et al., 2017) for comparison, with loci GpC5b, GpC26b, GpC10b, GpA19a, and GpB20b, reanalyzed as described in the supplemental material of De Meeûs et al (2019b) paper (De Meeûs et al., 2019b).

For the new primers, genomic DNA was isolated from teneral males of *G. pallidipes* from the IAEA colony, using Wizard genomic DNA purification kit (Promega, Madison, WI, USA). The DNA library was prepared using the Nextera DNA sample kit (Illumina, San Diego, CA, USA). Sequencing was performed on a MiSeq Sequencer (Illumina). To search for microsatellite markers, the MIcroSAtellite identification tool (MISA) (Thiel et al., 2003) was used. For the screening of the microsatellite markers, we selected primer pairs flanking dinucleotide microsatellite motifs with a minimum of 10 repetitions.

We screened 30 primer pairs on 10 Gpalli from Tanzania. Details of primers proposed for these 30 loci are presented in the supplementary Table S1. Amplification of the expected microsatellite fragments was first observed using agarose gel electrophoresis. Sixteen primer pairs that did not yield a single, well-amplified PCR product were removed from the polymorphism survey. This allowed to retain 14 primer pairs;

PAL1, PAL2, PAL8, PAL10, PAL12, PAL13, PAL15, PAL17, PAL18, PAL19, PAL22, PAL23, PAL26, and PAL29. Primer sequences details can be found in Table 3.

Table 3: Primers' description of the 14 new microsatellite markers designed for *Glossina pallidipes*. Range of product size is given in number of base pairs (bp).

Locus	Repeat motif	Range (bp)	Primer pair sequences (5'-3')
PAL1	(TC)12	170	F: TGAACTCGACGAAGGAAG R: AAAACACCTTCCAGCATTT
PAL2	(AT)10	165	F: CCATAAAGTTCCACTTAGATCA R: CTTTTCGTTCATGCCTTC
PAL8	(AG)11	174	F: AGGCAAGTCATGCGAAA R: ACAGTCACAATGGAAGTTGG
PAL10	(CT)11	174	F: TCTATTTCATAGCGCATTTC R: GCTATTTGCCGGTTTG
PAL12	(TG)11	171	F: GTTGTGCATCGTAGTAGCC R: GCCAACAACCAAAATAACA
PAL13	(AG)14	242	F: TCCGATAGATCCTTGCGT R: CACAACGATAGGTTGCATTT
PAL15	(AT)12	129	F: CACATATTTTCGACTGGTTTG R: CGGGTCCAGCACATTAG
PAL17	(AT)11	133	F: GCCGCTCACTACGAAAC R: CCCAGTGGTGTTGTGCT
PAL18	(CT)10	148	F: CGTATTGGAAGCCCAGTT R: AGCTGAGAGTCTGGACGAA
PAL19	(CAT)10	253	F: ACATATCCATACAGATGCACAC R: ATCCTGTTTACCAACGCA
PAL22	(CT)13	211	F: TCTTCTTCATCTCCATTCAC R: CTTCACTATTCATGGCTTTT
PAL23	(AG)11	137	F: TGTGTTATACAAGGCCGA R: ATGATGGCAACGCTAAA
PAL26	(TC)16	214	F: ACGTTGTTTCAAGGGCA R: GCGAGAGGCTGAGTGAA
PAL29	(TA)10	159	F: ACACGCACCATTTCTTTG R: CGAATTTGACGTGCATAAA

These loci were then evaluated on an ABI 3500xl sequencer (Applied Biosytems, Waltham, Massachusetts, USA) for their polymorphism and presence on the X chromosome in order to select the best markers for future genotyping. To achieve these

goals, we used 22 males of *G. pallidipes* from Tanzania. Then, for comparison of basic statistics with old loci, we reanalysed the two datasets of *G. pallidipes* from Tanzania Clades A and B (Manangwa et al., 2019). DNA extraction and PCR conditions are available in that paper.

2.3. *Glossina fuscipes fuscipes* (Gff)

According to the literature, 16 dinucleotide loci were developed for this species: five by Abila et al. (Abila et al., 2008): B03, D05, B05, D101, and D12; 10 by Brown et al 2008 (Brown et al., 2008): GffB8, GffC107, GffD6, GffD109, GffA3, GffA6, GffA9, GffA112, GffB101, and GffA10; and one more of the same origin but used in (Dyer et al., 2011): GfB105 an autosomal dinucleotide locus. There was not enough DNA left to study the Ethiopian samples as below for these markers. Instead, we studied the polymorphism of five of these loci (GffA3, GffA10, GffB8, GffB101, and GffB105; GfA3, GfB8, and GfB101 were redesigned from Brown et al., 2018 to produce smaller amplicons). These were analyzed in (Dyer et al., 2011) in 10 subsamples: Kinshasa (DRC, N=43), Kisantu (DRC, N=11), Madimba (DRC, N=29); Gogara (Ethiopia, N=30), Ungoye (Kenya, N=35), Manga (Kenya, N=35), Rusinga (Kenya, N=35), Bunghazi (Uganda, N=30), Busime (Uganda, N=23), and Buvuma (Uganda, N=50).

Genomic DNA was isolated from teneral males from Gff colony from UMR Intertryp (Montpellier), using Wizard genomic DNA purification kit (Promega, Madison, WI, USA). The DNA library was prepared using the Nextera DNA sample kit (Illumina, San Diego, CA, USA). The sequencing was performed on a MiSeq Sequencer (Illumina). To search for microsatellite markers, the MIcroSAtellite identification tool (MISA) (Thiel et al., 2003) was used. For the screening of the microsatellite markers, we selected sequences with dinucleotide microsatellite motifs with a minimum of 10 repetitions and long enough flanking regions to allow for the designing of appropriate primers.

We screened 30 primer pairs on 10 Gff from Ethiopia (Supplementary Table S2). Amplification of the expected microsatellite fragments was first checked with an agarose gel electrophoresis. Eleven primer pairs that did not yield a single, well-amplified PCR product were removed from the polymorphism survey. This allowed retaining 19 primer pairs (Table 3) that were then evaluated on an ABI 3500xl sequencer (Applied Biosytems, Waltham, Massachusetts, USA) on 23 males of Gff from Chad, HAT focus of Mandoul. Three legs were removed from each fly, dried and subjected to chelex treatment as previously described (Ravel et al., 2007) in order to obtain DNA for further microsatellite genotyping. Ten µl of each DNA (dilution 1/5) was used in the touch-down PCR reactions which were carried out in a thermocycler (Mastercycler, Eppendorf, Hamburg, Germany) in 20 µl final volume. PCR conditions for primers pairs were briefly as follows: 40 cycles and annealing temperature as described in Table 4. After PCR amplification, allele bands were routinely resolved on ABI 3500XL sequencer. This method allows multiplex by the use of four dyes. Allele calling was done as described above for Gmm loci.

Table 4: Primers' description of the 19 new microsatellite markers designed for *Glossina fuscipes fuscipes*. Range of product size is given in number of base pairs (bp) and annealing temperature (T) in °C.

Locus	Repeat motif	Range (bp)	T in ° C	Primer pair sequences (5'-3')
GFF1	(TC)17	143	51	F: TTCAAGCAATCTCGCTAC R: ACTGTACCGAAGAATCAAAA
GFF3	(AG)14	186	56	F: ATGTTCAGCAAACTCAGCAA R: GGTATTCAGTGACTCCTGGG
GFF4	(AG)20	141	55	F: AGGATGCTGGCTGTGAA R: ACGCTTTCTCTCATACTTTCTCT
GFF8	(AG)18	165	54	F: TACGGCAATTTCACAAGG R: GGGAATTTCTTTCTCTTCCA
GFF10	(AG)14	184	55	F: CGGAAGCGGAAAGATAAA R: TGAGCGAAGAAGAAGATGAA
GFF11	(AC)17	149	53	F: TCATTGGTGTTTCATTGGT R: TTGCGAGTGATGGATTG
GFF12	(CT)15	127	55	F: ATATTGCGTAAAGTGGCGT R: AATTCAAGAAGTTTGTGCGA
GFF13	(TC)17	184	50	F: GTTCTCCTTTTCTATCTTTCTC R: GAATGGTCGTGTCCTCT
GFF14	(AG)16	151	54	F: CCGATAGAGGGAGAGGG R: TGATCTTGTGCTCTTGACTTT
GFF16	(AG)23	149	55	F: AGACAGAGACATCGAAGCG R: CGAATTTCAATCAGTCCCA
GFF17	(CT)18	185	54	F: TGAAATATGTACGACGACCC R: AAGAGAGTAAAGTTTCGCTTGA
GFF18	(TC)16	215	55	F: TGCTTTCAGACGATTTAGAGTT R: CGAGAACTTCATCTGCCTC
GFF20	(AC)22	169	54	F: TCCACTCAAGAAACTTGCTATT R: TTATCTGGCAACATGGGA
GFF21	(TC)16	168	53	F: TCGCACTCTCTTTCCATC R: AAATGCGTTGACTCTTGTG
GFF22	(AG)22	198	55	F: ACAGCAGCGACAACAAAA R: TTTCCATCCATCCTTCGT
GFF23	(AG)17	252	54	F: TCAAATGAAAAGCAGGGA R: TGAGCAGTCAGTAATGGTAAGA
GFF25	(AC)16	211	54	F: AAATTGCTTTTGCATCTCAC R: GAACAATAGAACGCCAACC
GFF27	(CT)24	173	50	F: ATGCGTTCAAAGGATGT R: ACAGAATAAGGCAAAGAGAA
GFF29	(GT)17	122	50	F: AAATCAAGAGGCAGTCAA R: GTTTCTACAACTTTCGGATT

2.4. *Glossina palpalis s.l.*

For this species complex, twenty loci were available from the literature: Gpg55.3, Gpg19.62, and Gpg69.22 (all dinucleotide) (Solano et al., 1997), from *G. palpalis gambiensis*; Pgp1, Pgp8, Pgp11, Pgp13, Pgp17, Pgp20, Pgp22, Pgp24, Pgp28, Pgp29, Pgp33, Pgp34, Pgp35 (all dinucleotide) (Luna et al., 2001), from *G. palpalis palpalis*; plus four loci from A. Robinson (Insect Pest Control Sub-programme, Joint Food and Agriculture Organization of the United Nations/International Atomic Energy Agency Programme of Nuclear Techniques in Food and Agriculture), that were first evaluated in different studies: B104 (dinucleotide) (Camara et al., 2006), C102 (trinucleotide) (Bouyer et al., 2007), B110 (dinucleotide) (Solano et al., 2009) (in *G. palpalis gambiensis*) and B3 (dinucleotide) (Mélachio et al., 2011) (in *G. palpalis palpalis*).

For the new loci, genomic DNA was isolated from teneral males from the *Glossina palpalis gambiensis* (Gpg) colony of UMR Intertryp (Montpellier), using Wizard genomic DNA purification kit (Promega, Madison, WI, USA). The DNA library was prepared using the Nextera DNA sample kit (Illumina, San Diego, CA, USA). Sequencing was performed on a MiSeq Sequencer (Illumina). To search for microsatellite markers, the MIcroSAtellite identification tool (MISA) (Thiel et al., 2003) was used. For the screening of the microsatellite markers, we selected primer pairs flanking dinucleotide microsatellite motifs with a minimum of 10 repetitions.

We screened 30 primer pairs of dinucleotide loci (Supplementary Table S3) on 10 Gpg from UMR Intertryp colony (Montpellier). Amplification of the expected microsatellite fragments was first observed using agarose gel electrophoresis. Fourteen primer pairs that did not yield a single, well-amplified PCR product were excluded from the polymorphism survey. This allowed retaining 16 primer pairs: pGpg1, pGpg2, pGpg4, pGpg6, pGpg7, pGpg14, pGpg15, pGpg16, pGpg20, pGpg22, pGpg24, pGpg26, pGpg27, pGpg28, pGpg29 and pGpg30 (Table 5).

Table 5: Primers' description of the 16 new microsatellite markers designed for *Glossina palpalis gambiensis*. Range of product size is given in number of base pairs (bp) and annealing temperature (T) in °C.

Locus	Repeat motif	Range (bp)	T in °C	Primer pair sequences (5'-3')
pGpg1	(AG)12	165	54	F: AAAACGAAGAGTTTACGATCAC R: AAATTCGACAACAGGATAAGAG
pGpg2	(TA)26	160	54	F: AAAATTGGATGTTGCTCTTG R: TGCATGTGCTTTCGCT
pGpg4	(TA)17	166	53	F: AAAGCTGTATCAGTCTCTGACC R: GCACCCATGTTCCTTGT
pGpg6	(TG)21	156	54	F: AAATCACATTTTCGCTTCC R: TCGAGTTTGCATTGTTGTT
pGpg7	(AG)17	211	51	F: AAATGAATAGGAAGAGGGA R: TTACTAAAGAGAGAGCATCGT
pGpg14	(TG)13	218	54	F: AAGCCGTAAAACCGGAG R: ACATCAGAACCATTAAGTATCCC
pGpg15	(GT)15	174	54	F: AAGCTGAGAGCACCGAA R: GACCACATCAAGAGCGAA
pGpg16	(AC)14	186	53	F: AAGGCAGATACACAATAAGCA R: CCAACGAATATGCAGCTAA
pGpg20	(AG)25	278	54	F: CGCATACAAGCAATGGG R: AGCAAGGAATCAGACGACA
pGpg22	(AG)18	233	54	F: CGCGTCTAATAAAATTCCC R: ACTTTCGCAGAGAGGCA
pGpg24	(GA)14	193	51	F: CGCTTGTGTCGATTTGT R: GAATAGGATGATGACGGG
pGpg26	(AG)17	193	51	F: CGGCAAATAGCCAAAA R: TGCTTGCACTTACCACTC
pGpg27	(CT)18	165	54	F: CGGCTCGGATTTATGTCT R: GTGACAGCAAAAGTCTCCAA
pGpg28	(TC)17	216	53	F: CGGTAAAGCCGCAAC R: GTCGTAATGCGAAAGAGG
pGpg29	(TC)25	239	53	F: CGGTTAGATGGCGAGTT R: TGCCCTCACCGTATTTT
pGpg30	(GA)23	261	51	F: CTATTTCGCGCTTGTTT R: ATGCAAACTTTCTCCACA

These primers were then evaluated on a 4300 DNA Analysis System (LI-COR, Lincoln, NE) using 19 males of Gpg from the HAT focus of Boffa, Guinea. The same individuals were evaluated with three of the old autosomal loci from a previous work: pGp24, B3 (all dinucleotide), and C102 (trinucleotide) (Kagbadouno et al., 2012). Three

legs were removed from each fly, dried and subjected to chelex treatment as previously described (Ravel et al., 2007) in order to obtain DNA for further microsatellite genotyping. Touch-down PCR reactions were carried out in a thermocycler (MJ Research, Cambridge, UK) in 10 μl final volume, using 1 μl of the supernatant from the extraction step. PCR conditions for primers pairs were briefly as follows: 40 cycles and annealing temperature as described in Table 4. After PCR amplification, allele bands were resolved on a 4300 DNA Analysis System (LICOR, Lincoln, NE) and named using SAGA software as previously described (Solano et al., 2009).

2.5. Population genetics data analyses

Data were formatted for Create V 1.37 (Coombs et al., 2008), which transformed the datasets into required formats according to needs.

Amplification success of loci from the literature and the new ones, or between dinucleotide and trinucleotide loci were compared with the Wilcoxon rank sum tests with the R-commander package (Fox, 2005; Fox, 2007) in R (R-Core-Team, 2020).

Polymorphism statistics were assessed with the raw number of alleles found, and Nei's unbiased estimator of genetic diversity (H_S) (Nei and Chesser, 1983). Comparisons between published and new markers and between microsatellite types (Di, tri and tetra nucleotides) were undertaken with a Wilcoxon rank sum test under R-commander package in R. Linkage disequilibrium (LD) between each pair of loci was tested with the *G*-based test over subsamples with 10000 reshuffling of genotypes. *G*-based LD test over subsamples is the most powerful combining procedure (De Meeûs et al., 2009). There are as many tests (*NT*) as locus pairs (For *L* loci, $NT=L*(L-1)/2$), and these tests are not independent from each other's. We thus used the Benjamini and Yekutieli (BY) false discovery rate (FDR) procedure to adjust *p*-values. This was done with the "p.adjust" command in *R*. Deviation from panmixia was assessed with Wright's F_{IS} (Wright, 1965),

estimated with Weir and Cockerham's unbiased estimator *f* (Weir and Cockerham, 1984). Its significance was tested with a *f*-based 10000 randomizations of alleles between individuals within subsamples. Subdivision was measured with Wright's F_{ST} (Wright, 1965), estimated with θ (Weir and Cockerham, 1984) and its significance tested with the *G*-based test (Goudet et al., 1996) and 10000 randomization of individuals among subsamples. It is indeed the most powerful procedure to combine subdivision tests across subsamples (De Meeûs et al., 2009). Standard errors of *F*-statistics (StdErrFIS and StdErrFST), and confidence intervals at 95% (95%CI) of *F*-statistics were assessed through jackknives over subsamples and 5000 bootstraps over loci. Estimations and randomization tests were handled with Fstat 2.9.4 (Goudet, 2003), updated from Fstat 1.2 (Goudet, 1995).

The behavior of the different loci was checked with the strategy presented by De Meeûs et al in several papers to detect and discriminate between Wahlund effects and allele genotyping errors due to null alleles, short allele dominance and stuttering, and cure these loci when possible (De Meeûs, 2018; De Meeûs et al., 2019a; Manangwa et al., 2019). If StdErrFIS is at least twice StdErrFST, and/or if F_{IS} and F_{ST} are positively correlated, this is in favor of amplification problems as null alleles and/or SAD. Additionally, if the number of missing data per locus explains well F_{IS} variation (correlation tests and determination coefficient computation), null alleles represent a good explanation. Most of the time, when in reasonable proportions, amplification problems should not generate too much LD (but see (De Meeûs et al., 2019a)). If the F_{IT} of a given locus is negatively correlated to the size of the alleles, and confirmed by a linear regression weighted for polymorphism (if p_i is the frequency of allele i, the corresponding weight is $p_i(1-p_i)$), the locus displays SAD. If the heterozygote deficit is due to a Wahlund effect, it must involve all loci, if a high proportions of locus pairs are in significant LD, and if a correlation between the number of time a locus is involved in a significant pair (nLDsig) and its total genetic diversity H_T (H_T/nLDsig correlation) is substantial, and this correlation is positive,

this should mean a strong Wahlund effect (admixture of subdivided subpopulations); if this correlation is negative, it should mean a very strong Wahlund effect (admixture of different species) (Manangwa et al., 2019). Correlation and regression tests were undertaken with R-Commander (Rcmdr).

2.6. Comparisons between species and markers

Controlling the possible effect of species, we undertook comparisons between old and new loci and between trinucleotide and non-trinucleotide loci with multiple regressions under Rcmdr. The response variables tested were: the proportion of missing genotypes, local genetic diversity H_S, heterozygote deficit F_{IS}, and the degree of subdivision F_{ST}. Explanatory variables were the context of the study (species and/or sampling zone), the status of the locus (old or new), the motif of the locus (trinucleotide or not) (because trinucleotide loci might display selection signatures (Berté et al., 2019)), and interaction between the context and other variables. The significance of each explanatory variable was tested with *F* tests (command "anova" of R) between the complete model (with all variables) and the model without the variable to be tested. The series of *p*-values obtained was finally FDR adjusted with the Benjamini and Hochberg (BH) procedure (Benjamini and Hochberg, 1995) with R ("p.adjust").

3. Results

All usable data were combined into the Supplementary Table S4.

3.1. *Glossina morsitans morsitans*

Among the 12 loci tested from the literature, one did not amplify properly (Gmm9B) and one was located on the X chromosome (Gmm127) (all males homozygous at that locus for the three alleles found). This was surprising since this locus displayed a negative

F_{IT} when it was first published (Baker and Krafsur, 2001). This is why, for other loci that we did not tested ourselves, we did not rely on F_{IT} computed in (Baker and Krafsur, 2001) to consider if a locus was likely X-linked or not. The three loci Gmm22, GmcCA16C, Gmm14 were never studied elsewhere for the identification of their chromosomal locations and were thus removed from the equation for the computation of the proportion of heterosomal loci. As a result, from a total number of 23 usable old loci, six were determined as X-linked. For the eight newly designed loci, two were located on the X chromosome (Gmm8710 and Gmm12610) and two others did not amplify properly: Gmm7278 (only five genotypes out of 23 individuals); and Gmm31657 (no amplification for three individuals and uninterpretable profiles for the others). Thus, out of the six usable new loci, two were X-linked.

Across all loci (old and new ones), eight among 29 usable loci occurred in the X chromosome (i.e.28%).

Finally, 14 primer pairs were retained to be used for the genotyping of 23 males of Gmm from Zimbabwe: Gmm5B, Gmm8, Gmm22, GmsCAG6, GpCAG133, GmsCAG29B, GmmA06, GmmB20, GmmH09, and GmmK06 (old markers); and Gmm20291, Gmm66047 Gmm26134 and Gmm41987 (new markers). There were three trinucleotide markers: GmsCAG6, GpCAG133, and GmsCAG29B; and one tetranucleotide marker: GmmK06.

For old markers, there was no missing genotypes, the number of alleles was 9 on average, with H_S=0.678 and F_{IS}=0.281. Genetic diversity parameters appeared smaller for trinucleotides with 4 alleles and H_S=0.52, as compared to the other kind of loci with 11 alleles, and H_S=0.7456, differences which were significant for the number of alleles, but marginally not for H_S (respective *p*-values: 0.0206 and 0.0667). An important proportion of locus pairs were in significant LD (20%). No allele frequency reached 0.6, hence all loci could be kept for the BY procedure (De Meeûs, 2014), after which no locus pair stayed

significant (all *p*-values>0.4549). Nevertheless, given the small size of the sample, these tests may have been individually weakly powerful. We thus still chose to undertake the H_T/nLDsig correlation test. The correlation was negative (ρ=-0.4767) but not significant (two sided *p*-value=1636). There was an important, highly significant and highly variable heterozygote deficit with F_{IS}=0.281 in 95%CI=[0.196, 0.384] (*p*-value<0.0001). This was poorly explained (if any) by the motif of the locus (*p*-value=0.1095), even if trinucleotides tended to display higher F_{IS}=0.5 than the others (di and tetra-nucleotides) (F_{IS}=0.3). The standard error of F_{IS} was only 1.6 times the standard error of F_{ST}, meaning that amplification problems (null alleles or SAD) represented a poor explanation. Accordingly, we observed no missing genotypes and weak SAD signatures. For Gmm5b, SAD was not significant for the correlation test (ρ=-0.3724, *p*-value=0.1297), but marginally significant with the weighted regression test (*p*-value=0.0492). Additionally, though SAD was not significant for GmmH09, a detailed check showed that the heterozygote deficit at this locus was the result of the two main alleles (with cumulated frequency of p=0.761) of this locus, which happened also to be the smallest ones. A one sided Wilcoxon test comparing the F_{IS} of these two alleles of size≤169, with the others of size ≥178 provided a marginally not significant test (*p*-value=0.0667). This locus can seriously be suspected to suffer from SAD. A strong correlation was found between F_{IS} and F_{ST} (ρ=0.7356, *p*-value=0.0077). In fact, variation of F_{ST} across loci was mainly explained by the motif (*p*-value=0.00333), with trinucleotides displaying much bigger F_{ST}=0.22 as compared to other marker kinds (F_{ST}=0.03). Null alleles may be present for the dinucleotide locus GmmH09, which presented very high F_{IS}=0.457 and F_{ST}=0.208, the remaining F_{IS} probably came from a Wahlund effect and marginal SAD for Gmm5b and may be GmmH09, and F_{ST} variation probably resulted from some kind of selection signature affecting trinucleotide loci.

For new loci, missing data varied from 0 to 7 (3 on average), which was significantly more than for old markers (*p*-value=0.0043). Number of alleles (here 8) and H_S (here

0.6295) did not significantly differ from old loci (*p*-values>0.9). Two locus pairs (33%) appeared in significant LD (none stayed significant with BY). When combining old and new markers, we found 22 locus pairs in significant LD (24%). No loci displayed any allele with an average allele frequency above or equal to 0.9, hence all could be kept for the BY procedure, after which no locus pair stayed in significant LD (*p*-values>0.4449). Given the low power of individual tests, due to small sizes of samples, and because of the increasing severity of BY correction with the number of tests handled (here 91 tests), it was worth exploring further what caused such an important proportion of significant LD tests again. The number of times a locus was involved in a significant LD pair (nLDsig) was negatively correlated with its total genetic diversity (H_T) (Spearman's ρ=-0.5683, *p*-value=0.034). According to Manangwa et al criterion (Manangwa et al., 2019), this could be symptomatic of a Wahlund effect produced by the co-occurrence of highly divergent lineages (i.e. cryptic species) in the different subsamples.

Interestingly, trinucleotide loci were also more often implicated in a significant LD pair together (100% of the time) than the other kind of locus pairs (22% of the time) and, when tested with a logistic regression with R-Commander, the difference was highly significant (Chi square test, *p*-value=0.0029).

There was an important heterozygote deficit (F_{IS}=0.447 in 95%CI=[0.235, 0.431], *p*-value<0.0001), which was not significantly different from old markers (*p*-value=0.1035). This heterozygote deficit can largely be explained by the Wahlund effect detected above (but see below). With only four loci, jackknives or bootstraps over loci could not be undertaken. We found evidence of SAD for two loci Gmm20291 and Gmm26134: For Gmm20291, there was a strong signal (ρ=-0.7986, *p*-value=0.0009), confirmed by the weighted regression (*p*-value=0.00712); Gmm26134 also displayed a significant signal with the correlation method (ρ=-0.8281, *p*-value=0.0209) and the weighted regression (*p*-value=0.0195). Because loci with SAD or under selection are not useful in the missing

data/FIS correlation/regression tests, we combined old and new loci without SAD or selection. We found a positive correlation between the number of missing genotypes (putative null homozygotes) and F_{IS} (ρ=0.5394), though not significant (p-value=0.1058), with R^2=0.74. In addition to selection on trinucleotides, SAD and Wahlund effect, F_{IS} variation across GmmK06, GmmA06, GmmB20, Gmm8, Gmm22, Gmm66047, and Gmm41987 could be explained by null alleles.

All population genetics statistics are compiled in the supplementary Table S5.

3.2. *Glossina pallidipes*

While combining the F_{IS} observed in different papers (Ouma et al., 2003; Ouma et al., 2006a; Ouma et al., 2011; Okeyo et al., 2018; De Meeûs et al., 2019b), we found that seven loci (among 11) were obviously autosomal: GpA19a, GpA23b, GpB20b, GpC5b, GpC10b, GpC26b, and GpB115. Nevertheless, with the 22 Tanzanian males studied in the present paper, GpB115 appeared as duplicated and could not be used for proportion computations. From an unpublished data set kindly provided by Winnie Okeyo on 29 males and 31 females from Ruma (Kenya), we could confirm that GpB6b was not X-linked. We considered that the three other loci (GpD18b, GpC101 and GpC107) from the literature were potentially X-linked (i.e. 3/10).

Among the 14 new loci, three (PAL2, PAL10 and PAL29) did not amplify properly and three (PAL8, PAL17 and PAL18) were located on the X chromosome. This means that three interpretable new loci out of 11 were X-linked.

Globally, the proportion of heterosomal microsatellite loci was then 6/21=29%.

For the old loci, only 12 missing genotypes were found over all loci and the difference between di- and tri-nucleotidic loci was not significant (p-value>0.5). The average number of alleles was 7 (9 for dinucleotides and 6 for trinucleotides), and H_S=0.59 (0.65 for dinucleotides and 0.56 for trinucleotides). The differences between di- and tri-

nucleotidic loci were never significant (*p*-values>0.1). No locus displayed any allele with frequency above 0.7. Two locus pairs appeared in significant LD (20% of all possible pairs, *p*-values<0.03). None stayed significant after BY adjustment (*p*-value>0.39). Nonetheless, the correlation between H_T and the number of time a locus is involved in a significant LD pair was positive and significant (ρ=0.9487, *p*-value=0.0138). According to Manangwa et al's criterion (Manangwa et al., 2019), subsamples may sometimes contain flies from different subpopulations. There was a weak and non-significant heterozygote deficit: F_{IS}=0.029 in 95%CI=[-0.006, 0.055] (two sided *p*-value=0.1412). The standard error of F_{IS} was four times the one of F_{ST}, suggesting amplification problems (null alleles, SAD) at some loci. Correlations between F_{IS} and F_{ST} or between F_{IS} and number of missing genotypes were positive but not significant (*p*-values>0.3) and no SAD test appeared significant (*p*-values>0.3). Though the heterozygote deficit was weak and not significant, it may be explained by rare null alleles. Variation of F_{ST} across loci did not reveal any particular outlier behavior.

For the new loci, the eight loci retained (PAL1, PAL12, PAL13, PAL15, PAL19, PAL22, PAL23, PAL26) were used to analyze the genotypes of *G. pallidipes* from Tanzania Clades A and B separately (see (Manangwa et al., 2019)). For the 173 individuals of Clade A, the number of missing genotypes was highly variable, with 25% missing genotypes on average, varying from 0% (loci PAL12 and PAL 19), to 1% for PAL15, 2% for PAL26, 33% for PAL1, 45% for PAL13 and more than 60% for PAL22 and PAL23. The average number of alleles varied from 3 to 13 and averaged 8; H_S=0.54, varied from 0.02 to 0.7; and F_{IS}=0.18, varying from -0.014 to 0.38. One locus pair was in significant LD (*p*-value=0.0017), which did not stay significant after BY correction (*p*-value=0.1301). Removing PAL15 (displaying one allele with *p*>0.9), did not change these observations. The average heterozygote deficit observed was highly significant (two sided *p*-value<0.0002) and highly variable across loci. The standard error of F_{IS} was seven times

the one of F_{ST}. The correlation between F_{IS} and F_{ST} was positive (ρ=0.3234) but not significant (p-value=0.2173). It was also positive but significant between the number of missing genotypes and F_{IS} (ρ=0.6347, p-value=0.0454), and missing data explained more than 25% of F_{IS} variation (R^2=0.2528). Loci PAL1 and PAL13 displayed significant SAD, with p-value=0.0332 and p-value<0.0001 respectively for the correlation tests; and p-value=0.0012 and p-value<0.0009 respectively for the regression test, with R^2=0.7508 and R^2=0.725 respectively. When these loci were removed, together with PAL22 and PAL23, which displayed prohibitive proportions of missing data, there was still a significant global F_{IS}=0.148 (two sided p-value<0.0002), but missing data explained more than 82% of F_{IS} variation across the four remaining loci (ρ=0.6325, p-value=0.1838). The two loci, PAL1 and PAL13, presented the highest F_{ST}=0.017 and F_{ST}=0.04 respectively, the last of which appeared significantly above the average over loci: F_{ST}= 0.007 in 95%CI=[0.006, 0.021]. For Clade B, sample comprised 17 individuals, and locus PAL23 had to be removed (no amplification for any of the 17 individuals in that clade). The proportion of missing data varied between 0% (PAL15 and PAL19), 25% for PAL1, 41% for PAL12 and more than 50 % for PAL13, PAL22 and PAL26. There was between 2 and 6 alleles per locus (4 in average), and genetic diversity averaged H_S=0.59 and varied between 0.4 and 0.83, while a small heterozygote deficit was observed with F_{IS}=0.037, which varied between -0.25 to 0.227. No LD test displayed a significant p-value (all p-values>0.137). The heterozygote deficit was not significant (two-sided p-value=0.2844), but with a substantial variation across loci. The standard error of F_{IS} was 2.5 times the one of F_{ST} (amplification problems). The correlation between F_{IS} and F_{ST} was negative and it was positive between the number of missing genotypes and F_{IS}, but not significant (ρ=0.1441, p-value=0.3789). Null alleles may explain weakly the variation of F_{IS} across loci. SAD was marginally significant for PAL12 with the correlation test (ρ=-0.8208, p=value=0.04429), not confirmed by the regression test(p-value=0.0816, R^2=0.6896); for PAL 19 (ρ=-1, p=value=0.0417),

confirmed by the regression method (p-value=0.0009, R^2=0.9981); as for PAL22 (ρ=-1, p=value=0.0417; R^2=0.9105, p-value=0.0458). When we excluded the loci suspected of SAD, together with PAL13 (too weakly polymorphic), F_{IS}=-0.112 and its variation across the three remaining loci, given the small number of points, is largely explained by the number of missing data (R^2=0.8251).

Comparisons between data from old and new loci were almost impossible. Indeed, it is probable that flies studied in Kenya (Okeyo et al., 2017) do not belong to any of the two clades found in Tanzania (Manangwa et al., 2019), and Tanzanian sample could not be studied with old loci (not enough DNA left). Only two loci were common to both studies, GmmC17 and GpCAG133 (initially developed for *G. morsitans*, see above). If alleles were labelled the same in both studies, then only allele 197 at locus GpCAG133 was shared by both studies but with very different allele frequencies, and all other alleles were private for both loci. Even if some mismatch occurred, allele frequency distributions would anyway imply strong allele frequency differences between the three clades: Clade A and Clade B from Tanzania, and a probable third one in the Kenyan sample studied.

All population genetics statistics are compiled in the supplementary Table S5.

3.3. *Glossina fuscipes fuscipes*

According to Beadell et al (Beadell et al., 2010); eight of the 16 loci from the literature were X-linked (50%): B03, D12, C107, GffD6, GffD109, GffA6, GffA9, and GffA112.

Among the 19 new loci we examined, six did not amplify properly (GFF1, GFF13, GFF20, GFF22, GFF25 and GFF29) and four were located on the X chromosome (GFF10, GFF11, GFF14, GFF17) (4/13 i.e. 31%).

This also means that among the 29 loci designed and usable for *G. fuscipes* to date (GFF1, GFF13, GFF20, GFF22, GFF25 and GFF29 excluded), 12 loci were found on the X chromosome, hence 41%.

Finally, nine loci were retained for the analyses of 23 Chadian males: GFF3, GFF4, GFF8, GFF12, GFF16, GFF18, GFF21, GFF23 and GFF27.

For Dyer et al.'s markers, there was a substantial amount of missing genotypes (19% on average), with two loci, GffA10 and GffB8 displaying no less than 50% and 33% of missing genotypes respectively, the three other loci displaying between 3% and 5% of missing data. Only one pair of locus was in significant LD (*p*-value=0.0207). No locus displayed any allele with an average allele frequency above or equal to 0.9, hence all could be kept for the BY procedure, after which no locus pair stayed significant (*p*-value>0.6). Genetic diversity was H_S=0.698 (0.66 to 0.74), with, on average, 29 alleles per locus (15 to 42). There was an important and highly significant heterozygote deficit: F_{IS}=0.354 in 95%CI=[0.123, 0.591] (*p*-value<0.0001). The standard error of F_{IS} was nine times the one of F_{ST}. There was a positive correlation between F_{IS} and F_{ST} (ρ=0.6), but it was not significant (*p*-value=0.175). However, F_{IS} was strongly correlated to the number of missing data (ρ=1, *p*-value=0.0083). Missing data (putative null homozygotes) explained as much as 76% of F_{IS} variation across loci. One locus, GfB105, presented a highly significant SAD (ρ=-0.5107, *p*-value<0.0004), even with the regression method (*p*-value=0.0003, R^2=0.2896). Without this locus, missing data (null homozygotes) explained no less than 93% of F_{IS} variation. This leaves around 7% to be explained by GfB105, where around 20% of F_{IS} was explained by SAD, and the remaining, with no less than 17 missing genotypes, being probably explained by null alleles as well.

Regarding the new markers, we subdivided the Chadian sample in the Bodo village into two zones: North and South. There were eight, one and four missing genotypes for GFF8, GFF12 and GFF27 respectively, i.e. around 6%. This is apparently much less than

for the Dyer et al.'s markers, though the difference was marginally not significant (*p*-value=0.052). Only two locus pairs appeared in significant LD, which did not stay significant after "BY" correction (*p*-values>0.6) (no allele reached the threshold frequency of 0.9 at any locus). Genetic diversity appeared reasonably high, with H_s=0.717, varying from 0.5 to 0.85. There was on average six alleles per locus, varying from two to eight. Compared to Dyer et al.'s markers, the difference was highly significant for the number of alleles (*p*-value=0.0032), and not significantly different for H_S (*p*-value=0.2977). There was a moderate and not significant heterozygote deficit, with F_{IS}=0.071 in 95%CI=[-0.045, 0.184] (*p*-value=0.093) and the difference with Dyer et al.'s markers was marginally not significant with the Wilcoxon rank sum test(*p*-value=0.06). Here, because some loci displayed positive and other negative values, we computed two sided *p*-values. The resulting global value was *p*-value=0.093. Accordingly, null alleles explained rather little of the observed F_{IS}, except perhaps for GFF12 and GFF27 that presented the highest values and accordingly one and four missing genotypes respectively. These missing genotypes could then have corresponded to homozygous individual for the null alleles. On the contrary, GFF8 presented eight missing genotypes but a negative F_{IS}, which means that many unspecific allelic dropouts occurred, that affected any kind of genotypes (heterozygous or homozygous). No SAD was found either (*p*-values>0.12), even with the regression method (*p*-values>0.33). Finally, the standard error of F_{IS} was four times the one for F_{ST}, confirming the probable role of null alleles.

For both old and new loci, F_{ST} variation across loci was moderate and did not show any evidence of signatures of selection of any kind.

All population genetics statistics are compiled in the supplementary Table S5.

3.4. *Glossina palpalis sl*

Among the 20 loci found in the literature, seven happened to be located on the X chromosome (35%): Gpg55.3 and Gpg19.62 (Solano et al., 1999); Pgp20 (Luna et al., 2001); Pgp11 (Camara et al., 2006), B104 and Pgp13 (Bouyer et al., 2007); and B110 (Solano et al., 2009).

Among the 16 new markers, six did not amplify correctly: pGpg7, pGpg22, pGpg26, pGpg28, pGpg29 and pGpg30; and four appeared X-linked (40% of interpretable loci): pGpg2, pGpg4, pGpg14 and pGpg24. This left us with six usable loci: pGpg1, pGpg6, pGpg15, pGpg16, pGpg20, pGpg27.

This means that globally, among 30 usable loci, 11 were X-linked, hence 37%.

With old markers, number of alleles varied from 3 for B3, 5 for C102 to 7 forPgp24, averaging 5; H_S=0.72, varying from 0.468 for B3 to 0.829 for C102 and 0.863 forPgp24; there was an important heterozygote deficit with F_{IS}=0.386, varying from 0.199 for C102, 0.476 for Pgp24and 0.545 for B3. There was around 30% of missing data for the three loci. No locus pair appeared in significant LD (*p*-values>0.057). The heterozygote deficit was highly significant (*p*-value<0.0001) and mainly explained by SAD in Pgp24 (ρ=-0.8929, *p*-value=0.0062; R^2=0.6697, *p*-value=0.0244), and in C102 (ρ=-1, *p*-value=0.0083; R^2=0.8909, *p*-value=0.0158), and by a probable substantial frequency of null alleles (p_n=0.5657, estimated as the square root of the frequency of missing genotypes) for B3. For B3, we assumed a single reproductive panmictic unit (population subdivision is indeed not significant, *p*-value=0.5205), and if we estimate the real frequency of visible alleles as $[x_i+(\Sigma x_{ij})/2]/(1+2p_n)$, where x_i is the frequency of phenotype [*i*] (homozygous + putative heterozygous with the null allele), Σx_{ij} is the frequency of heterozygous states of allele *i*, and p_n is the frequency of null alleles estimated as $p_n=\sqrt{f(\varnothing)}$ (square root of missing genotype frequency). We then compared observed with expected frequencies of the different phenotypes under the panmictic hypothesis with three visible alleles (161, 183 and 191) and a null allele of frequency p_n. We then computed the corresponding χ^2, and

obtained seven phenotypic classes ([161], [183], [191], 161/183, 161/191,183/191 and Ø). The resulting chi-square was χ^2=7.4. With 7-4=3 degrees of freedom (seven classes minus null allele frequency, allele frequencies of 161 and 183, and total sample size) this value was marginally not significant (*p*-value=0.0611). Given the size of the sample, null allele represents a satisfactory explanation for the observed F_{IS} at locus B3, but with a frequency as big as p_n=0.57, which may appear prohibitive (Séré et al., 2017).

For the six new loci, there was 11 alleles per locus on average, H_S=0.816, F_{IS}=0.356, and number of missing data reached 8 on average (33%), varying from 0 (pGpg20 and pGpg27), 20% (pGpg15), 48% (pGpg16), to more than 60% (pGpg1 and pGpg6). Despite a marginally significant difference for the number of alleles (*p*-value=0.0489), none of the comparisons between old and new markers appeared significant (all other *p*-values>0.38). No locus pair appeared in significant LD (*p*-value>0.07). The heterozygote deficit was highly significant (*p*-value<0.0001) and the standard error of F_{IS} was three times the one of F_{ST} (amplification problems). F_{ST} was very small (-0.0148, *p*-value=0.9999), even when loci pGpg1 and pGpg6 (60% missing genotypes) were removed (F_{ST}=0.022, *p*-value=0.687). The correlation between F_{IS} and F_{ST} was negative, even when pGpg1 and pGpg6 were removed. Correlation between missing data and F_{IS} was positive but not significant (ρ=0.4058, *p*-value=0.2123), and negative without pGpg1 and pGpg6. Null alleles thus explained poorly the F_{IS} and its variation across loci. There was a marginally not significant SAD at locus pGpg20 (ρ=-0.3827, *p*-value=0.0585, R^2=0.1464, *p*-value=0.093). For pGpg27, the signal was even weaker (ρ=-0.3578, *p*-value=0.0792, R^2=0.0408, *p*-value=0.4368). Other amplification problems are at stake. Subsample are too small for Micro-Checker (Van Oosterhout et al., 2004) stuttering test. We nevertheless used the stuttering correction strategy of De Meeûs et al (De Meeûs et al., 2019a) on all loci, to check for a possible effect. For pGpg1 alleles 163-165 were pooled with 161, and 171 with 169; for pGpg6, allele 150 was pooled with

148, 154 with 152, and 158 and 160 with 156; for pGpg15, alleles 174 and 176 were pooled with 172; for pGpg16, alleles 176 and 178 were pooled with 174, and 184-190 with 182; for pGpg20, alleles 256 and 258 were polled with 254, and 266-268 with 264; and for pGpg27, alleles 149-161 were pooled with 147, and 167-181 with 165. We reanalysed this amended dataset with Fstat. This changed almost nothing for pGpg1, pGpg6 and pGpg20. However, for pGpg15, pGpg16 and pGpg27, this produced much lower F_{IS} values (from 0.15 to 0.053, from 0.139 to -0.161, and from 0.205 to -0.092, respectively). Thus, stuttering mainly (if not entirely) explains F_{IS} for pGpg15, pGpg16 and pGpg27, stuttering and null alleles explain it for pGpg 15, and the 12 missing genotypes at pGpg16 are probably due to not specific allelic dropouts that affected randomly any kind of genotype.

All population genetics statistics are compiled in the supplementary Table S5.

3.5. Comparisons between species and markers over all datasets

Regarding amplification success, we checked missing data proportions and F_{IS} differences between loci. No significant effect could be found for the proportion of missing data (all *p*-values>0.3). Heterozygote deficits (amplification problems in heterozygous individuals) appeared higher for old markers (*p*-value=0.0158). The effect of species was significant too (*p*-value=0.0017): the highest was for Gmm, followed by Gpg, Gpalli Clade A from Tanzania, Gff, Gpalli Clade B from Tanzania and Kenyan clade of this species displayed the smallest F_{IS}. Nevertheless, the interaction between locus status and species was significant (*p*-value=0.0158) (old markers displayed a higher F_{IS} only for the members of the *Palpalis* group).

Regarding genetic diversity (H_S), no effect of species or status of the marker displayed a significant influence (all *p*-values>0.4).

Finally, regarding subdivision measures (F_{ST}), the context was highly significant, as expected because the geographic range was very different across the different studies

(very small for Gpg and very large for the old markers of Gff from Dyer et al (*p*-value=0.0004). The effect of the microsatellite motif and its interaction with the context of the study were marginally not significant (*p*-value=0.0774 in both cases), suggesting that trinucleotides did not display a selection signature in all contexts.

All population genetics statistics are compiled in the supplementary Table S5.

4. Discussion

For Gmm, the proportion of X-linked markers was the smallest (28%). There was, at least in the Zimbabwean samples of the present study, different probable cryptic species coexisting within the same sites. This will thus require special attention for the studies of this species, in particular in the Zimbabwean area. There was also evidence of null alleles at some loci, requiring the use of adapted techniques to compute genetic distances (Chapuis and Estoup, 2007). Short allele dominance was very strong for loci Gmm20291 and Gmm26134. These two loci will require a cautious interpretation of the smallest peaks in apparent homozygous profile, or will need to be disregarded. For Gmm5B, SAD was less obvious and needs to be confirmed on bigger data sets. Because of small subsample sizes, stuttering could not be tested. Nevertheless, it seemed that most, if not all, heterozygote deficits here were explained by a Wahlund effect and null alleles. This will need to be explored on bigger samples, using the same strategy as described in Manangwa et al paper (Manangwa et al., 2019) to separate the different clades that coexist in Zimbabwe (at least). Finally, eight loci can be safely used (with correction for null alleles): GmmK06, GmmA06, GmmB20, Gmm8, Gmm22, GmmH09, Gmm66047, and Gmm41987; and five seem particularly promising (lowest F_{IS}): GmmA06, GmmB20, Gmm8, Gmm22, and Gmm66047.

For Gpalli, the proportion of X-linked markers was 29%. The apparent complexity of hidden biodiversities of this taxon, and heterogeneous use of different marker kinds, made

any comparison of the respective performances of new and older markers difficult. A Kruskal-Wallis test undertaken with Rcmdr for F_{IS} and the proportion of missing genotypes did not allow unveiling a significant difference between old markers from Kenya, new markers in Clade A or B from Tanzania. Old markers from the literature did not specifically identify X-linked markers for that species. Nevertheless, for new markers X-linked markers reached a classical proportion of 27%. Depending on the clade, some new loci displayed prohibitive proportions of amplification failures or short allele dominance, which will require particular caution in future use. Given the very small F_{IS}, and the fact that effective population sizes are known to be substantial in the Nguruman zone of Kenya (see supplementary material in (De Meeûs et al., 2019b)), the particular behaviour of the five old loci that we used regarding LD can be attributed to a very weak Wahlund effect and/or an artifactual consequence of the combined effects of substantial genetic diversities of these loci, importance subsample sizes, and small number of loci, which increased the power of the LD tests to an unusually high level. The subdivision of this taxon into cryptic species is problematic for advising general good markers. For the Kenyan clade, GpC5b, GpC26b, GpC10b, GpA19a, and GpB20b worked well but with some loci displaying null alleles, the best loci being GpC5b and GpC26b; for the Tanzanian clade A, PAL12, PAL15, PAL19, PAL26 worked well, and PAL12 and PAL15 displayed the best behaviour; finally for the Tanzanian clade B, PAL1, PAL15, PAL26 worked very well. Other loci from other species must be tried and tested or other new loci designed for each of these clades. Nevertheless, the prediction is that new cryptic species will occur in further studies from other places, with unpredictable performances of available loci. According to previous studies (De Meeûs et al., 2019b; Manangwa et al., 2019), Gmc17, GmK22, GmL11 and GmA06, may represent additional useful resources for *G. pallidipes*.

For Gff, the proportion of X-linked markers was 41%. New markers presented slightly better markers with less amplification problems. In particular, if GfB105, had to be

used, particular attention should be taken to small (almost invisible) peaks for alleles of larger sizes at apparent homozygous individuals. But given the strength of the signature we detected, we would rather advise not using such a locus. Loci GfA10 and GffB8, with 50% and 33% missing data respectively, that would mean very important null allele frequencies (around 70 and 57%), unless new and more efficient primers were used, should also be avoided. With such important null allele frequencies, it is known that extant correction procedures cannot avoid prohibitive variance of estimates of genetic distances due to the disturbing effects of null alleles (Séré et al., 2017). For this species, 14 loci behaved rather well, though more or less affected by null alleles: D05, B05, D101; GffA3, GffB101, GFF3, GFF4, GFF8, GFF12, GFF16, GFF18, GFF21, GFF23, and GFF27. According to their low F_{IS}, best loci probably would be: GffA3, GffB101, GFF4, GFF8, GFF16, GFF18, and GFF21.

For *G. palpalis sl*, the proportion of X-linked loci was 37%. Recent markers did not significantly behave better that old ones since most displayed important amplification problems (null alleles, SAD, allelic dropouts and stuttering). Among the old loci tested, only B3 appeared directly usable with appropriate adjustments for null alleles. Nevertheless, with a frequency as big as 0.57 for null alleles, corrections may not work very well (Séré et al., 2017). For Pgp24 and C102, strong SAD will require a very cautious interpretation of these loci if to be used. For new markers, two loci (pGpg1 and pGpg6) appeared very difficult to use with actual primers, given the prohibitive proportion of amplification failures observed. One locus, pGpg20, with very important F_{IS}=0.3, may suffer from SAD, but also of important allelic dropouts in heterozygous individuals as testified by the absence of missing genotypes at this locus. This left us with only four loci: B3, pGpg15, pGpg16, and pGpg27, the best ones being pGpg15, pGpg16, and pGpg27. Nevertheless, all would require stuttering correction, and correction for null alleles for pGpg15. This does not leave a lot of usable loci, and other loci from the literature will be needed, as some X-linked ones

that work rather well on female subsamples only, as Pgp24, C102, XPgp13, and XB104 (Berté et al., 2019); or XGpg55.3, and XPgp11 (Solano et al., 2010a). Nevertheless, the taxon *G. palpalis* is known to be subdivided into several entities (cryptic species or ecotypes) (Dyer et al., 2009; De Meeûs et al., 2015). This may explain the difficulties to transpose loci from one context to another and suggests that more loci need to be developed for that taxon. For the closely related *G. tachinoides*, which was not studied here, and for which no specific loci were ever developed, a swift glance at the available literature allowed to highlight a few loci: C102, Pgp29, Pgp28, and XPgp13 (for females) (Kone et al., 2011; Adam et al., 2014). This taxon also would need new and specific loci to be developed.

Markers performances need to be improved, especially for Gpalli and Gpg. Trinucleotide loci displayed some kind of selection signature (recent or present) in some instances. As far as demographic inferences are concerned, it will be wiser carefully using those in future studies, or simply avoid it. Nevertheless, these markers may help in identifying genomic zones that are indeed under selection, in particular for the resistance to control devices, as was suggested recently for GpCAG133 in *G. palpalis palpalis* (Berté et al., 2019).

The high proportions of X-linked markers in tsetse flies were not a surprise when one has a glance to the particular chromosomic structure of tsetse flies. We computed the relative proportion of the X chromosome length as regard to the total chromosome lengths, excluding the Y chromosome, where obviously no microsatellite loci can be met, using several papers: (Southern and Pell, 1973; Amos and Dover, 1981) for Gmm; (Southern and Pell, 1973; Amos and Dover, 1981; Willhoeft, 1997) for Gpalli; (Southern and Pell, 1973; Willhoeft, 1997) for *G. fuscipes* and *G. palpalis*. For this, we used the average number of supernumerary chromosomes (called B or S depending on the author) found in Gmm (four S chromosomes) (Southern and Pell, 1973), and the fact that one S

chromosome was found in the Zimbabwean colony of Gpalli in (Willhoeft, 1997). We used the proportions of the total length represented by X as the explanatory variable of a multiple regression weighted with the number of studies where a given species was available, against the proportion of X-linked microsatellite loci that we found in the present paper as the response variable. The regression was not significant (*p*-value=0. 2716). Nevertheless, the correlation was significant (Spearmans ρ=0.9487, *p*-value=0.0257, one sided test). It means that, except on the Y where they are absent, microsatellite markers are more or less evenly distributed across the L1, L2 and the supernumerary (B or S) chromosomes that are met in the *Morsitans* group. The proportion of X-linked loci was particularly more important in the *Palpalis* group (with no B chromosome), as compared to Gpalli (one B chromosome). This may be because B chromosome can host more microsatellite loci than the other chromosome types (or that there are more than one B chromosome in the Gpalli we worked with). If we assume that the relationship with the number of B chromosomes is linear, and since L1, L2 and X chromosomes do not vary across species, with the exception of the *Fusca* group (Willhoeft, 1997), and since B chromosomes are absent from the *Palpalis* group and are on average 10 in *G. austeni* (Southern and Pell, 1973), we can predict that the proportion of microsatellite on the X chromosome would be 39% for *G. tachinoides* (average between what we found on *G. palpalis* and *G. fuscipes*); and 26 % on average for the *Austeni* group (with an average of 10 S chromosomes (Southern and Pell, 1973)). If the Gpalli of our data had more than a single B chromosome, this prediction would represent an upper bound. For the *Fusca* group, the uncertainty of chromosomal structure found in *G. brevipalpis* did not allow any computation (Willhoeft, 1997). Interestingly, we could not find a single study on the population genetics of any member of the *Fusca* group with microsatellite loci, although the complete genome of *G. brevipalpis* is already available, a probable consequence of

the low veterinary and nil medical interests of this group of tsetse flies in most regions (Attardo et al., 2019).

In the present work, we did not systematically tested new primer pairs across all available samples of different taxa. This would indeed had increased the bench work and data analyses beyond what we planned to invest on that matter, with very little proportions of expected positive results, given the medium success within each taxon. Nevertheless, all new primers described in the present paper may be worth trying on other samples from different tsetse fly species in future studies, one by one, because it may provide a few but useful supplementary markers to such studies.

Aknowledgement

This study was financed by the International Atomic Energy Agency (IAEA), Austria, and by the Institut de Recherche pour le Développement (IRD), France. TdM would like to thank Jean-Mathieu Bart for his communication skills; Drs Johnson Ouma and Elliot Krafsur; and particularly Norah Saarman, for very useful discussions.

References

Abila, P.P., Slotman, M.A., Parmakelis, A., Dion, K.B., Robinson, A.S., Muwanika, V.B., Enyaru, J.C.K., Lokedi, L.M., Aksoy, S., Caccone, A., 2008. High levels of genetic differentiation between Ugandan *Glossina fuscipes fuscipes* populations separated by Lake Kyoga. PLoS Negl. Trop. Dis. 2, e242.

Adam, Y., Bouyer, J., Dayo, G.K., Mahama, C.I., Vreysen, M.J.B., Cecchi, G., Abd-Alla, A.M.M., Solano, P., Ravel, S., De Meeûs, T., 2014. Genetic comparison of *Glossina tachinoides* populations in three river basins of the upper west region of Ghana and implications for tsetse control. Infect. Genet. Evol. 28, 588–595.

Amos, A., Dover, G., 1981. The distribution of repetitive DNAs between regular and supernumerary chromosomes in species of *Glossina* (Tsetse): a two-step process in the origin of supernumeraries. Chromosoma 81, 673-690.

Attardo, G.M., Abd-Alla, A.M.M., Acosta-Serrano, A., Allen, J.E., Bateta, R., Benoit, J.B., Bourtzis, K., Caers, J., Caljon, G., Christensen, M.B., Farrow, D.W., Friedrich, M., Hua-Van, A., Jennings, E.C., Larkin, D.M., Lawson, D., Lehane, M.J., Lenis, V.P., Lowy-Gallego, E., Macharia, R.W., Malacrida, A.R., Marco, H.G., Masiga, D., Maslen, G.L., Matetovici, I., Meisel, R.P., Meki, I., Michalkova, V., Miller, W.J., Minx, P., Mireji, P.O., Ometto, L., Parker, A.G., Rio, R., Rose, C., Rosendale, A.J., Rota-Stabelli, O., Savini, G., Schoofs, L., Scolari, F., Swain, M.T., Takáč, P., Tomlinson, C., Tsiamis, G., Van Den Abbeele, J., Vigneron, A., Wang, J., Warren, W.C., Waterhouse, R.M., Weirauch, M.T., Weiss, B.L., Wilson, R.K., Zhao, X., Aksoy, S., 2019. Comparative genomic analysis of six *Glossina* genomes, vectors of African trypanosomes. Genome Biology 20, 187.

Baker, M.D., Krafsur, E.S., 2001. Identification and properties of microsatellite markers in tsetse flies *Glossina morsitans* sensu lato (Diptera: Glossinidae). Mol. Ecol. Notes 1, 234-236.

Bayerl, H., Kraus, R.H.S., Nowak, C., Foerster, D.W., Fickel, J., Kuehn, R., 2018. Fast and cost-effective single nucleotide polymorphism (SNP) detection in the absence of a reference genome using semideep next-generation Random Amplicon Sequencing (RAMseq). Mol. Ecol. Res. 18, 107-117.

Beadell, J.S., Hyseni, C., Abila, P.P., Azabo, R., Enyaru, J.C., Ouma, J.O., Mohammed, Y.O., Okedi, L.M., Aksoy, S., Caccone, A., 2010. Phylogeography and population structure of *Glossina fuscipes fuscipes* in Uganda: implications for control of tsetse. PLoS Negl. Trop. Dis. 4, e636.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573-580.

Berté, D., De Meeus, T., Kaba, D., Séré, M., Djohan, V., Courtin, F., N'Djetchi, K.M., Koffi, M., Jamonneau, V., Ta, B.T.D., Solano, P., N'Goran, E.K., Ravel, S., 2019. Population genetics of *Glossina palpalis palpalis* in sleeping sickness foci of Côte d'Ivoire before and after vector control. Infect. Genet. Evol. 75.

Bouyer, J., Ravel, S., Dujardin, J.P., De Meeûs, T., Vial, L., Thévenon, S., Guerrini, L., Sidibe, I., Solano, P., 2007. Population structuring of *Glossina palpalis gambiensis* (Diptera: Glossinidae) according to landscape fragmentation in the Mouhoun river, Burkina Faso. J. Med. Entomol. 44, 788-795.

Brown, J.E., Komatsu, K.J., Abila, P.P., Robinson, A.S., Okedi, L.M., Dyer, N., Donnelly, M.J., Slotman, M.A., Caccone, A., 2008. Polymorphic microsatellite markers for the tsetse fly *Glossina fuscipes fuscipes* (Diptera: Glossinidae), a vector of human African trypanosomiasis. Mol. Ecol. Res. 8, 1506-1508.

Camara, M., Caro-Riano, H., Ravel, S., Dujardin, J.P., Hervouet, J.P., De Meeûs, T., Kagbadouno, M.S., Bouyer, J., Solano, P., 2006. Genetic and morphometric evidence for

population isolation of *Glossina palpalis gambiensis* (Diptera : Glossinidae) on the Loos islands, Guinea. J. Med. Entomol. 43, 853-860.

Chapuis, M.P., Estoup, A., 2007. Microsatellite null alleles and estimation of population differentiation. Mol. Biol. Evol. 24, 621-631.

Coombs, J.A., Letcher, B.H., Nislow, K.H., 2008. CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. Mol. Ecol. Res. 8, 578–580.

De Meeûs, T., 2014. Statistical decision from *k* test series with particular focus on population genetics tools: a DIY notice. Infect. Genet. Evol. 22, 91-93.

De Meeûs, T., 2018. Revisiting F_{IS}, F_{ST}, Wahlund effects, and Null alleles. J. Hered. 109, 446-456.

De Meeûs, T., Bouyer, J., Ravel, S., Solano, P., 2015. Ecotype evolution in *Glossina palpalis* subspecies, major vectors of sleeping sickness. PLoS Negl. Trop. Dis. 9, e0003497.

De Meeûs, T., Chan, C.T., Ludwig, J.M., Tsao, J.I., Patel, J., Bhagatwala, J., Beati, L., 2019a. Deceptive combined effects of short allele dominance and stuttering: an example with *Ixodes scapularis*, the main vector of Lyme disease in the U.S.A. bioRxiv 622373, ver. 4 peer-reviewed and recommended by Peer Community In Evolutionary Biology, doi: https://doi.org/10.1101/622373.

De Meeûs, T., Guégan, J.F., Teriokhin, A.T., 2009. MultiTest V.1.2, a program to binomially combine independent tests and performance comparison with other related methods on proportional data. BMC Bioinformatics 10, 443.

De Meeûs, T., Ravel, S., Solano, P., Bouyer, J., 2019b. Negative density dependent dispersal in tsetse flies: a risk for control campaigns? Trends Parasitol. 35, 615-621.

Dyer, N.A., Furtado, A., Cano, J., Ferreira, F., Afonso, M.O., Ndong-Mabale, N., Ndong-Asumu, P., Centeno-Lima, S., Benito, A., Weetman, D., Donnelly, M.J., Pinto, J., 2009.

Evidence for a discrete evolutionary lineage within Equatorial Guinea suggests that the tsetse fly *Glossina palpalis palpalis* exists as a species complex. Mol. Ecol. 18, 3268-3282.

Dyer, N.A., Ravel, S., Choi, K.S., Darby, A.C., Causse, S., Kapitano, B., Hall, M.J., Steen, K., Lutumba, P., Madinga, J., Torr, S.J., Okedi, L.M., Lehane, M.J., Donnelly, M.J., 2011. Cryptic diversity within the major trypanosomiasis vector *Glossina fuscipes* revealed by molecular markers. PLoS Negl. Trop. Dis. 5, e1266.

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107, 1–15.

Ellegren, H., 2004. Microsatellites: Simple sequences with complex evolution. Nat Rev Genet 5, 435-445.

Fox, J., 2005. The R commander: a basic statistics graphical user interface to R. J. Stat. Software 14, 1–42.

Fox, J., 2007. Extending the R commander by "plug in" packages. R News 7, 46–52.

Franco, J.R., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., Simarro, P.P., Zhao, W., Argaw, D., 2018. Monitoring the elimination of human African trypanosomiasis: Update to 2016. PLoS Negl. Trop. Dis. 12, e0006890.

Garvin, M.R., Saitoh, K., Gharrett, A.J., 2010. Application of single nucleotide polymorphisms to non-model species: a technical review. Mol. Ecol. Res. 10, 915-934.

Goudet, J., 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. J. Hered. 86, 485-486.

Goudet, J., 2003. Fstat (ver. 2.9.4), a program to estimate and test population genetics parameters. Available at http://www.t-de-meeus.fr/Programs/Fstat294.zip, Updated from Goudet (1995).

Goudet, J., Raymond, M., De Meeûs, T., Rousset, F., 1996. Testing differentiation in diploid populations. Genetics 144, 1933-1940.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R., Nielsen, E.E., 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol. Ecol. Res. 11, 123-136.

Hyseni, C., Beadell, J.S., Gomez-Ocampo, Z., Ouma, J.O., Okedi , L.M., Gaunt, M.W., Caccone, A., 2011. The *G.m. morsitans* (Diptera: Glossinidae) genome as a source of microsatellite markers for other tsetse fly (*Glossina*) species. Unpublished manuscript available at https://www.researchgate.net/profile/Chaz_Hyseni/publication/259291686_The_Gm_morsitans_Diptera_Glossinidae_genome_as_a_source_of_microsatellite_markers_for_other_tsetse_fly_Glossina_species/links/02e7e52ac8fe0324e5000000/The-Gm-morsitans-Diptera-Glossinidae-genome-as-a-source-of-microsatellite-markers-for-other-tsetse-fly-Glossina-species.pdf?origin=publication_detail; database formerly available in Molecular Ecology Resouces Database that no longer exists.

Kagbadouno, M.S., Camara, M., Rouamba, J., Rayaisse, J.B., Traoré, I.S., Camara, O., Onikoyamou, M.F., Courtin, F., Ravel, S., De Meeûs, T., Bucheton, B., Jamonneau, V., Solano, P., 2012. Epidemiology of sleeping sickness in boffa (Guinea): where are the trypanosomes? PLoS Negl. Trop. Dis. 6, e1949.

Katti, M.V., Ranjekar, P.K., Gupta, V.S., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18, 1161-1167.

Kone, N., Bouyer, J., Ravel, S., Vreysen, M.J.B., Domagni, K.T., Causse, S., Solano, P., De Meeûs, T., 2011. Contrasting population structures of two vectors of African trypanosomoses in Burkina Faso: consequences for control. PLoS Negl. Trop. Dis. 5, e1217.

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

Luna, C., Bonizzoni, M.B., Cheng, Q., Aksoy, S., Zheng, L., 2001. Microsatellite polymorphism in the tsetse fies (Diptera: Glossinidae). J. Med. Entomol. 38, 376–381.

Manangwa, O., De Meeûs, T., Grébaut, P., Segard, A., Byamungu, M., Ravel, S., 2019. Detecting Wahlund effects together with amplification problems : cryptic species, null alleles and short allele dominance in Glossina pallidipes populations from Tanzania. Mol. Ecol. Res. 19, 757-772.

Mélachio, T., Tito, Tanekou, Simo, G., Ravel, S., De Meeûs, T., Causse, S., Solano, P., Lutumba, P., Asonganyi, T., Njiokou, F., 2011. Population genetics of *Glossina palpalis palpalis* from central African sleeping sickness foci. Parasites and Vectors 4, 140.

Molecular Ecology Resources Primer Development Consortium, Agata, K., Alasaad, S., Almeida-Val, V.M.F., Alvarez-Dios, J.A., Barbisan, F., Beadell, J.S., Beltran, J.F., Benitez, M., Bino, G., Bleay, C., Bloor, P., Bohlmann, J., Booth, W., Boscari, E., Caccone, A., Campos, T., Carvalho, B.M., Climaco, G.T., Clobert, J., Congiu, L., Cowger, C., Dias, G., Doadrio, I., Farias, I.P., Ferrand, N., Freitas, P.D., Fusco, G., Galetti, P.M., Gallardo-Escarate, C., Gaunt, M.W., Ocampo, Z.G., Goncalves, H., Gonzalez, E.G., Haye, P., Honnay, O., Hyseni, C., Jacquemyn, H., Jowers, M.J., Kakezawa, A., Kawaguchi, E., Keeling, C.I., Kwan, Y.S., La Spina, M., Lee, W.O., Lesniewska, M., Li, Y., Liu, H.X., Liu, X.L., Lopes, S., Martinez, P., Meeus, S., Murray, B.W., Nunes, A.G., Okedi, L.M., Ouma, J.O., Pardo, B.G., Parks, R., Paula-Silva, M.N., Pedraza-Lara, C., Perera, O.P., Pino-Querido, A., Richard, M., Rossini, B.C., Samarasekera, N.G., Sanchez, A., Sanchez, J.A., Santos, C.H.D., Shinohara, W., Soriguer, R.C., Sousa, A.C.B., Sousa, C.F.D., Stevens, V.M., Tejedo, M., Valenzuela-Bustamante, M., Van de Vliet, M.S., Vandepitte, K., Vera, M., Wandeler, P., Wang, W.M., Won, Y.J., Yamashiro, A., Yamashiro, T., Zhu, C.C., 2011. Permanent genetic resources added to molecular ecology resources database 1 December 2010-31 January 2011. Mol. Ecol. Res. 11, 586-589.

Murray, V., Monchawin, C., England, P.R., 1993. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. Nucleic Acids Res. 21, 2395-2398.

Nei, M., Chesser, R.K., 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47, 253-259.

Okeyo, W.A., Saarman, N.P., Bateta, R., Dion, K., Mengual, M., Mireji, P.O., Ouma, C., Okoth, S., Murilla, G., Aksoy, S., Caccone, A., 2018. Genetic Differentiation of *Glossina pallidipes* Tsetse Flies in Southern Kenya. Am. J. Trop. Med. Hyg. 99, 945-953.

Okeyo, W.A., Saarman, N.P., Mengual, M., Dion, K., Bateta, R., Mireji, P.O., Okoth, S., Ouma, J.O., Ouma, C., Ochieng, J., Murilla, G., Aksoy, S., Caccone, A., 2017. Temporal genetic differentiation in *Glossina pallidipes* tsetse fly populations in Kenya. Parasit. Vect. 10, 471.

Ouma, J.O., Beadell, J.S., Hysen, i.C., Okedi, L.M., Krafsur, E.S., Aksoy, S., Caccone, A., 2011. Genetic diversity and population structure of *Glossina pallidipes* in Uganda and western Kenya. Parasites and Vectors 4, 122.

Ouma, J.O., Cummings, M.A., Jones, K.C., Krafsur, E.S., 2003. Characterization of microsatellite markers in the tsetse fly, *Glossina pallidipes* (Diptera: Glossinidae). Mol. Ecol. Res. 3, 450-453.

Ouma, J.O., Marquez, J.G., Krafsur, E.S., 2006a. Microgeographical breeding structure of the tsetse fly, *Glossina pallidipes* in south-western Kenya. Med. Vet. Entomol. 20, 138-149.

Ouma, J.O., Marquez, J.G., Krafsur, E.S., 2006b. New polymorphic microsatellites in *Glossina pallidipes* (Diptera: Glossinidae) and their cross-amplification in other tsetse fly taxa. Biochem Genet 44, 471-477.

Ouma, J.O., Marquez, J.G., Krafsur, E.S., 2007. Patterns of genetic diversity and differentiation in the tsetse fly *Glossina morsitans morsitans* Westwood populations in East and southern Africa. Genetica 130, 139-151.

Paetkau, D., Strobeck, C., 1995. The molecular basis and evolutionary history of a microsatellite null allele in bears. Mol Ecol 4, 519-520.

Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Møller, A.P., Ellegren, H., 1997. Low Frequency of Microsatellites in the Avian Genome. Genome Res. 7, 471-482.

R-Core-Team, 2020. R: A Language and Environment for Statistical Computing, Version 3.6.3 (2020-02-29) Ed. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Ravel, S., De Meeûs, T., Dujardin, J.P., Zeze, D.G., Gooding, R.H., Dusfour, I., Sane, B., Cuny, G., Solano, P., 2007. The tsetse fly *Glossina palpalis palpalis* is composed of several genetically differentiated small populations in the sleeping sickness focus of Bonon, Côte d'Ivoire. Infect. Genet. Evol. 7, 116-125.

Richard, G.-F., Kerrest, A., Dujon, B., 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. 72, 686–727.

Séré, M., Kabore, J., Jamonneau, V., Belem, A.M.G., Ayala, F.J., De Meeûs, T., 2014. Null allele, allelic dropouts or rare sex detection in clonal organisms : simulations and application to real data sets of pathogenic microbes. Parasites and Vectors 7, art. 331.

Séré, M., Thévenon, S., Belem, A.M.G., De Meeûs, T., 2017. Comparison of different genetic distances to test isolation by distance between populations. Heredity 119, 55-63.

Solano, P., De la Rocque, S., Cuisance, D., Geoffroy, B., De Meeûs, T., Cuny, G., Duvallet, G., 1999. Intraspecific variability in natural populations of *Glossina palpalis gambiensis* from West Africa, revealed by genetic and morphometric analyses. Med. Vet. Entomol. 13, 401-407.

Solano, P., Duvallet, G., Dumas, V., Cuisance, D., Cuny, G., 1997. Microsatellite markers for genetic population studies in *Glossina palpalis* (Diptera: Glossinidae). Acta Trop. 65, 175-180.

Solano, P., Kaba, D., Ravel, S., Dyer, N.A., Sall, B., Vreysen, M.J., Seck, M.T., Darbyshir, H., Gardes, L., Donnelly, M.J., De Meeus, T., Bouyer, J., 2010a. Population genetics as a tool to select tsetse control strategies: suppression or eradication of Glossina palpalis gambiensis in the Niayes of Senegal. PLoS Negl. Trop. Dis. 4, e692.

Solano, P., Ravel, S., Bouyer, J., Camara, M., Kagbadouno, M.S., Dyer, N., Gardes, L., Herault, D., Donnelly, M.J., De Meeus, T., 2009. The population structure of Glossina palpalis gambiensis from island and continental locations in Coastal Guinea. PLoS Negl. Trop. Dis. 3, e392.

Solano, P., Ravel, S., De Meeûs, T., 2010b. How can tsetse population genetics contribute to African trypanosomiasis control? Trends Parasitol. 26, 255-263.

Southern, D.I., Pell, P.E., 1973. Chromosome relationships and meiotic mechanisms of certain *morsitans* group tsetse flies and their hybrids. Chromosoma 44, 319-334.

Thiel, T., Michalek, W., Varshney, R.K., Graner, A., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics 106, 411-422.

Van den Bossche, P., Séphane, d.L.R., Hendrickx, G., Bouyer, J., 2010. A changing environment and the epidemiology of tsetse-transmitted livestock trypanosomiasis. Trends Parasitol. 26, 236-243.

Van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M., Shipley, P., 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4, 535-538.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34, 275-305.

Wang, C., Schroeder, K.B., RosenbergO, N.A., 2012. A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. Genetics 192, 651-669.

Wattier, R., Engel, C.R., Saumitou-Laprade, P., Valero, M., 1998. Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). Mol. Ecol. 7, 1569-1573.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 38, 1358-1370.

Willhoeft, U., 1997. Fluorescence in situ hybridization of ribosomal DNA to mitotic chromosomes of tsetse flies (Diptera: Glossinidae: *Glossina*). Chromosome Res. 5, 262-267.

Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to system of mating. Evolution 19, 395-420.

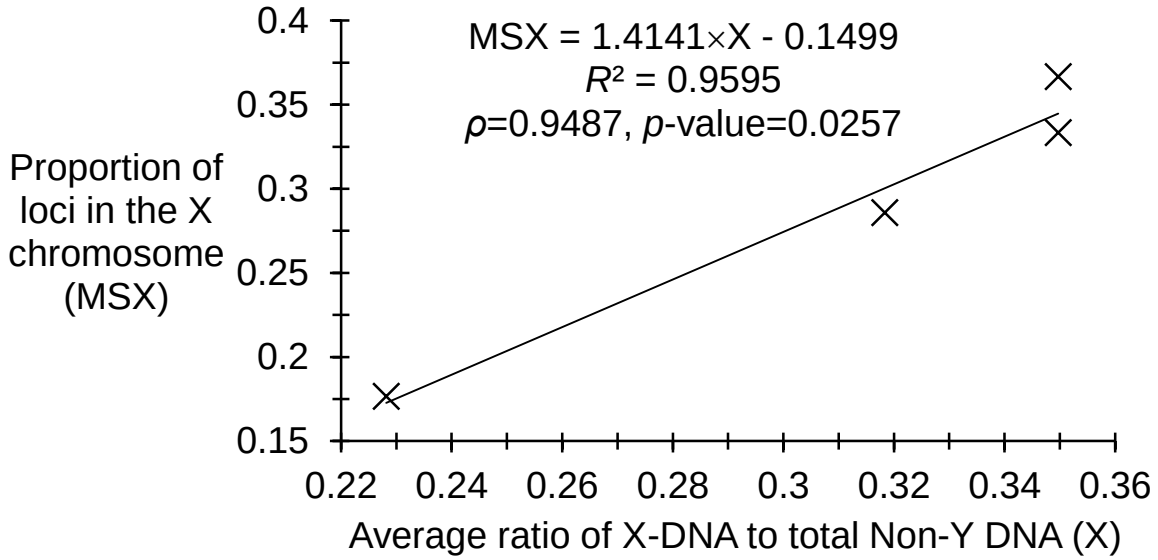

Illustration of the correlation that links the ratio of the length of DNA in the X-chromosome to the average total non-Y chromosomes length and the proportion of X-linked microsatellite markers found. The test is the Spearman's rank correlation test and the regression parameters are presented for the sake of illustration