# The *G.m. morsitans* (Diptera: Glossinidae) genome as a source of microsatellite markers for other tsetse fly (Glossina) species

***Chaz Hyseni[1*], Jon S. Beadell[1], Zaneli Gomez Ocampo[1], Johnson O. Ouma[2], Loyce M. Okedi[3], Michael W. Gaunt[4], Adalgisa Caccone[1]***

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA

[2] Trypanosomiasis Research Centre, Kenya Agricultural Research Institute, Kukuyu, Kenya

[3] National Livestock Resources Research Institute, Tororo, Uganda

[4] London School of Tropical Medicine, London, United Kingdom

[*]E-mail: chaz.hyseni@aya.yale.edu

***Abstract***

We searched the *Glossina morsitans morsitans* genome for short sequence repeats (SSR) in order to adapt polymorphic microsatellite markers to other species of *Glossina*, *G. fuscipes fuscipes* and *G. pallidipes*, two major vectors of African trypanosomiasis. We tested 30 loci containing perfect di-, tri-, or tetranucleotide repeats. We identified seven polymorphic loci that amplified across both *G.f. fuscipes* and *G. pallidipes* samples, as well as seven additional loci that were variable in just one species. Five of these fourteen loci were homozygous in males of one or both species and are likely to be X-linked. Although the success rate of adapting SSR markers from the *G.m. morsitans* genome for use in other species was not very high, this process yielded several polymorphic markers that should be useful in future studies of tsetse ecology and evolution.

Trypanosomiasis is responsible for thousands of infected people and millions of cattle deaths with serious repercussions for public health and economic development in 37 countries in sub-Saharan Africa (http://www.fao.org/Ag/againfo/programmes/en/paat/disease.html). The tsetse species *Glossina fuscipes* and *Glossina pallidipes* are primary vectors of both human and animal forms of African trypanosomiasis (sleeping sickness and nagana, respectively). Molecular tools are increasingly being used to garner information about the population structure, reproductive behavior, and other aspects of tsetse ecology and life history (reviewed in Krafsur 2009) pertinent to the design of more effective vector control strategies. While microsatellite markers have been characterized for both *G. fuscipes fuscipes* (Abila et al. 2008, Brown et al. 2008) and *G. pallidipes* (Ouma et al. 2003, Ouma et al. 2006), many of these loci are difficult to score and exhibit high null allele frequencies. For *G.f. fuscipes*, a few additional markers have been isolated as a result of recent efforts to adapt loci from other tsetse species (Beadell et al. 2010, unpubl. data). Approximately 13 autosomal loci are now available for *G.f. fuscipes* and 8 for *G. pallidipes*. Additional autosomal markers are required for exploring fine-scale structuring of populations, for increasing the certainty of estimates of effective population size, as well as quantifying dispersal and gene flow. Vector control would also benefit from further insight into the reproductive behavior of tsetse. Development of X-linked loci would help provide such insight.

Recently, the Wellcome Trust Sanger Institute completed whole genome shotgun sequencing of *G. morsitans morsitans* (http://www.sanger.ac.uk/Projects/G_morsitans/). Here, we explore the utility of using the *G.m. morsitans* genome as a source of short sequence repeats (SSR) for use in other tsetse species by testing a panel of SSRs for cross-amplification and variability in *G.f. fuscipes* and *G. pallidipes*. We used both *MISA* (MicroSatellite identification tool, Thiel et al. 2003) and *Msatfinder* (Thurston & Field 2005) to search the *G.m. morsitans* genome. An initial screening for 1-, 2-, 3-, 4-, 5- and 6-nucleotide SSRs greater than 30bp in length produced over 6,000 results. From these we selected di-, tri- and tetranucleotide SSRs 50-200bp in size with at least 100bp of sequence flanking either side and

used a script written in the stream editor program, *sed*, to extract the up- and downstream flanking regions. These sequences were run through *Msatcommander* (Faircloth 2008), which includes a Primer3 (Rozen & Skaletsky 2000) extension. Primers were designed in *Msatcommander* by modifying the default Primer3 settings. Thirty loci were identified for cross-amplification testing.

Loci were tested on samples collected from sites in western Uganda: Kabunkanga (*G.f. fuscipes*, n=22) and Murchison Falls (*G. pallidipes*, n=23). Six of the twenty-three *G.p.* samples and eight of the twenty-two *G.f.f.* samples were male. Individual flies were sexed by inspection of genitalia in order to permit identification of X-linked loci. Since tsetse males are heterogametic, male individuals are expected to be homozygous in X-linked loci.

DNA was extracted from fly legs using the Qiagen DNEasy Tissue Kit (Qiagen, Inc.) as per the manufacturer's protocol. We used the M13-tailed primer method (Boutin-Ganache et al. 2001) to obtain individual genotypes. Forward primers were 5'-tailed with a 15-mer M13 sequence (5'-**TCCCAGTCACGACGT**-3') and in addition to the unmodified reverse primers, reactions included a third primer with a fluorescent dye attached to the same M13 sequence. We prepared 12.5 µL reactions using 1x PCR Buffer (50mM KCl, 10mM Tris-HCl, pH 8.3) and 2 mM $MgCl_2$ (Applied Biosystems), 0.22 mM each dNTP and 3µg BSA (New England Biolabs), 5 pmol fluorescently-labeled M13 primer, 5 pmol reverse primer, 0.4 pmol M13-tailed forward primer and 0.5 units AmpliTaq Gold DNA Polymerase (Applied Biosystems). PCR amplification was done via touchdown thermal cycling: after the initial denaturation (95°C for 5 min), reactions cycled through 95°C for 30s, 60-49°C (1°C decrement/cycle) for 25s and 72°C for 30s in the first step, going through an additional 40 cycles of 95°C, 48°C and 72°C, and finally a 15-minute extension at 72°C. Annealing temperatures of 53°C to 48°C in the second step and 60-54°C to 60-49°C in the first step were tested to ensure that lower annealing temperatures did not result in stronger PCR product and lower null allele frequencies due to non-specific amplification. We tested several loci

using 5 pmol each of fluorescently-labeled forward primer and unlabeled reverse primer, which along with lower annealing temperatures enhanced PCR amplification compared to the above three-primer method. PCR amplicons were run through an Applied Biosystems 3730xl DNA Analyzer and the resulting data were analyzed using *Genemarker 1.91* (SoftGenetics).

Observed and expected heterozygosity values were calculated using *Genalex 6.2* (Peakall & Smouse 2006). We used *Genepop 4.0.10* (Rousset 2008) to test for departures from Hardy-Weinberg equilibrium (HWE) and to determine whether any linkage disequilibrium (LD) was present among loci. Loci developed in this study were also tested for linkage with autosomal loci developed for prior studies (*G.f.f.*: Beadell et al. 2010; *G.p.*: Ouma et al. 2003, Ouma et al. 2006). We estimated null allele frequencies in *Microchecker* (Oosterhout et al. 2004) using the method of Oosterhout et al. (2006) and in *FreeNA* (Chapuis & Estoup 2007) using the method of Dempster et al. (1977). Calculations in *Genepop* were performed using a burn-in of 10,000 and 1,000 batches with 10,000 iterations per batch. We performed 10,000 randomizations in *Microchecker* and *FreeNA*.

Twenty-four of the thirty loci selected from the *G.m. morsitans* genome amplified in either *G.f.f.*, *G.p.* or both species. Of these 24, only 15 loci were easily scored and polymorphic in at least one species. We only reported 14 loci here (see Table 1), because locus GmmC22, polymorphic in *G.p.*, appeared to be physically linked to a gene (GenBank accession no. AM940018.1). Although *G.p.* and *G.m.m.* share more recent ancestry than *G.f.f.* and *G.m.m.* (Dyer et al. 2008), transfer of loci from *G.m.m.* to *G.p.* was no more successful than transfer to *G.f.f.* We recovered 10 polymorphic markers for *G.p.* and 11 polymorphic markers for *G.f.f.*; only 7 loci were polymorphic in both species. In *G.f.f.*, the success rate of adapting loci from the *G.m.morsitans* genome (11 polymorphic loci out of 30 tested) was lower compared with the success rate of isolating polymorphic loci from a fuscipes-specific library (17 polymorphic loci out of 20 tested; unpubl. data).

Allele numbers varied from 2 to 8 for *G.f.f.* and 2 to 16 for *G.p*. Observed heterozygosities ranged from 0.09 to 0.86 for *G.f.f.* and 0.13 to 0.87 for *G.p*. After applying sequential Bonferroni correction (Holm 1979), the only significant LD among loci characterized in this study was observed between GmmL17 and GmmP07 in the *G.p*. samples. When taking into account previously developed loci, significant linkage was, again, only observed in *G.p.*, between GmmL11 and locus B115 of Ouma et al (2006). The only locus to exhibit significant deviation from HWE after sequential Bonferroni correction was locus GmmL17 in *G.f.f*.

Based on the observation that all male samples are homozygous, locus GmmL17 is likely X-linked. Using the criterion of male homozygosity, we identified five loci with potential X-linkage: GmmC15, GmmD03, GmmF10, GmmL17 and GmmP07. While *G.p.* males were homozygous for locus GmmC15, however, there was one heterozygous male in *G.f.f.* X-linkage of loci GmmF10, GmmL17 and GmmP07 is corroborated by low p-values for HWE tests and high (above 0.1) null allele frequencies, and loci GmmL17 and GmmP07 were, at least in *G.p.*, also significantly linked to each other. Fluorescence in situ hybridization (FISH) has been used to map microsatellite markers (Stratikopoulos et al. 2008). FISH will be employed in future studies to confirm location in the X chromosome of the loci we suspect to be X-linked.

The putative X-linked loci we identified will be valuable for remating studies, where SSRs are used to estimate the number of unique male contributors to female sperm pools (Bonizzoni et al. 2002). The other markers reported here should provide additional power for resolving fine-scale spatial structuring of tsetse populations and determining the temporal stability of these populations. These autosomal markers, in addition to the 13 that are already developed for *G.f.f.* and 8 for *G.p.*, will also allow more accurate estimation of effective population size, of gene flow and migration rates, and will aid in

characterizing other facets of tsetse ecology and life history relevant to the design and improvement of vector control strategies.

***Acknowledgement***

Funding was provided by the NIH (R01AI068932) and WHO-TDR (80132) grants.

***References***

Abila PP, Slotman MA, Parmakelis A*, et al.* (2008) High levels of genetic differentiation between Ugandan Glossina fuscipes fuscipes populations separated by Lake Kyoga. *PLoS Neglected Tropical Diseases*, **2**, e242.

Beadell JS, Hyseni C, Abila PP*, et al.* (2010) Phylogeography and population structure of Glossina fuscipes fuscipes in Uganda: implications for control of tsetse. *PLoS Neglected Tropical Diseases*, **4**, e636.

Bonizzoni M, Katsoyannos BI, Marguerie R*, et al.* (2002) Microsatellite analysis reveals remating by wild Mediterranean fruit fly females, Ceratitis capitata. *Molecular Ecology*, **11**, 1915-1921.

Boutin-Ganache I, Raposo M, Raymond M, Deschepper CF (2001) M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. *Biotechniques*, **31**, 24-27.

Brown JE, Komatsu KJ, Abila PP*, et al.* (2008) Polymorphic microsatellite markers for the tsetse fly Glossina fuscipes fuscipes (Diptera: Glossinidae), a vector of human African trypanosomiasis. *Molecular Ecology Resources*, **8**, 1506-1508.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621-631.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, **39**, 1-38.

153 Dyer NA, Lawton SP, Ravel S*, et al.* (2008) Molecular phylogenetics of tsetse flies (Diptera: Glossinidae)
154 based on mitochondrial (COI, 16S, ND2) and nuclear ribosomal DNA sequences, with an
155 emphasis on the palpalis group. *Molecular Phylogenetics and Evolution*, **49**, 227-239.
156 Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-
157 specific primer design. *Molecular Ecology Resources*, **8**, 92-94.
158 Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*,
159 **6**, 65-70.
160 Krafsur ES (2009) Tsetse flies: genetics, evolution, and role as vectors. *Infection, Genetics and Evolution*,
161 **9**, 124-141.
162 Ouma JO, Cummings MA, Jones KC, Krafsur ES (2003) Characterization of microsatellite markers in the
163 tsetse fly, Glossina pallidipes (Diptera: Glossinidae). *Molecular Ecology Notes*, **3**, 450-453.
164 Ouma JO, Marquez JG, Krafsur ES (2006) New polymorphic microsatellites in Glossina pallidipes (Diptera:
165 Glossinidae) and their cross-amplification in other tsetse fly taxa. *Biochemical Genetics*, **44**, 471-
166 477.
167 Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for
168 teaching and research. *Molecular Ecology Notes*, **6**, 288-295.
169 Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows
170 and Linux. *Molecular Ecology Resources*, **8**, 103-106.
171 Rozen S, Skaletsky H (2000) Primer3 on the www for general users and for biologist programmers. In:
172 *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds Krawetz S, Misener S),
173 pp. 365-386. Humana Press, Totowa, New Jersey.
174 Stratikopoulos EE, Augustinos AA, Petalas YG*, et al.* (2008) An integrated genetic and cytogenetic map
175 for the Mediterranean fruit fly, Ceratitis capitata, based on microsatellite and morphological
176 markers. *Genetica*, **133**, 147-157.

Thiel T, Michalek W, Varshney RK, 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.

Thurston M, Field D (2005) MSATFINDER: Detection and characterization of microsatellites.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535-538.

van Oosterhout C, Weetman D, Hutchinson WF (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, **6**, 255-256.

**Table 1.**

| Locus | Primer Sequence (5'-3') | G.m.m genome: Sequence ID | G.m.m genome: (Motif)Units [Fragment Size] | Species | Allele Size Range | No. of Flies (% Amplified) | No. of Alleles | $H_O$ | $H_E$ | H-W(P) | X-linked | Null (O et al) | Null (D et al) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***GmmA06*** | F: ACTTCCATGTTATGTTCGTTGC<br>R: TGCCTTAGTTGAGAAACTCTGC | Tfly_23-t487a06.q1k | (AC)31 [209] | ***G.f.f*** | 146-174 | 22 (100%) | 7 | 0.727 | 0.699 | 0.728 | | -0.035 | 0.000 |
| | | | | ***G.p*** | 154-166 | 23 (100%) | 3 | 0.130 | 0.127 | 1.000 | | -0.066 | 0.000 |
| ***GmmB20*** | F: AAATGCATGTCTAACTGTCCG<br>R: AGCAAAAGGCAACTAAAGTGATG | Tfly_23-t572b20.q1k | (GT)33 [222] | ***G.f.f*** | 171-181 | 22 (100%) | 3 | 0.500 | 0.635 | 0.419 | | 0.100 | 0.068 |
| | | | | ***G.p*** | 221-281 | 23 (100%) | 16 | 0.870 | 0.895 | 0.344 | | 0.004 | 0.000 |
| ***GmmC15*** | F: ACTGCATCTGCCTCTGTCG<br>R: TGAACGAGAAAATGTGAATGGTAAG | Tfly_23-t601c15.p1k | (ATT)20 [208] | ***G.f.f*** | 193-211 | 22 (100%) | 4 | 0.409 | 0.506 | 0.178 | | 0.080 | 0.068 |
| | | | | ***G.p*** | 190-211 | 23 (100%) | 5 | 0.609 | 0.723 | 0.086 | - | 0.063 | 0.066 |
| ***GmmC17*** | F: TGCGCTTTGAACGGAACG<br>R: CTATGCCGCCTGGCTTATC | Tfly_23-t506c17.p1k | (ATGT)14 [224] | ***G.f.f*** | 184-188 | 22 (100%) | 2 | 0.091 | 0.089 | 1.000 | | -0.047 | 0.000 |
| | | | | ***G.p*** | 190-202 | 23 (100%) | 2 | 0.304 | 0.322 | 1.000 | | 0.016 | 0.011 |
| ***GmmD03*** | F: TGCACTTCACCGATTGCAC<br>R: GTTGAAAGTTGGTTGTGACGC | Tfly_23-t605d03.q1k | (AAT)20 [189] | ***G.f.f*** | 151-160 | 22 (100%) | 3 | 0.136 | 0.132 | 1.000 | - | -0.070 | 0.000 |
| | | | | ***G.p*** | No amplification | | | | | | | | |
| ***GmmD15*** | F: GCATCACACTTTGCTTGCG<br>R: CGTTGGAAACTAGACATCTCACG | Tfly_23-t535d15.q1k | (GT)59 [227] | ***G.f.f*** | 149-159 | 22 (100%) | 5 | 0.545 | 0.606 | 0.409 | | 0.017 | 0.056 |
| | | | | ***G.p*** | No amplification | | | | | | | | |
| ***GmmF10*** | F: TGCCTTTCGATAGAGAAACCATC<br>R: ACCTGGACACTTATACCGCTC | Tfly_23-t606f10.q1k | (AC)42 [248] | ***G.f.f*** | No amplification | | | | | | | | |
| | | | | ***G.p*** | 186-200 | 23 (100%) | 6 | 0.565 | 0.773 | 0.049 | - | 0.120 | 0.117 |
| ***GmmH09*** | F: ACCCAGATAACCTATATTGCTCG<br>R: CGTTCAGGCAGATACGAAAATTG | Tfly_23-t508h09.p1k | (AT)27 [191] | ***G.f.f*** | 148-182 | 22 (100%) | 7 | 0.864 | 0.829 | 0.121 | | -0.042 | 0.000 |
| | | | | ***G.p*** | Monomorphic | | | | | | | | |
| ***GmmK06*** | F: TAACGTGCATGTGCGTGTG<br>R: CCATCAATACGAGCAGACCG | Tfly_23-t504k06.q1k | (ATGT)12 [154] | ***G.f.f*** | Monomorphic | | | | | | | | |
| | | | | ***G.p*** | 123-131 | 23 (100%) | 3 | 0.435 | 0.414 | 1.000 | | -0.056 | 0.000 |
| ***GmmK22*** | F: ACGCTTACGTTTCCGTTACAC<br>R: AAGCTAACCGAACCAGCAC | Tfly_23-t513k22.p1k | (GTT)19 [228] | ***G.f.f*** | Monomorphic | | | | | | | | |
| | | | | ***G.p*** | 192-198 | 23 (100%) | 3 | 0.478 | 0.569 | 0.288 | | 0.069 | 0.038 |
| ***GmmL03*** | F: ACAGTCCAATTTTCGCCCG<br>R: GGCCAACAATGTCATAAACCG | Tfly_23-t545l03.q1k | (AT)28 [210] | ***G.f.f*** | 181-187 | 21 (95.5%) | 3 | 0.381 | 0.528 | 0.290 | | 0.129 | 0.079 |
| | | | | ***G.p*** | Monomorphic | | | | | | | | |
| ***GmmL11*** | F: CCACCACTAACAACGACAGC<br>R: TGGCTGGTTACAAGATTGCAC | Tfly_23-t516l11.p1k | (AT)27 [249] | ***G.f.f*** | 216-246 | 22 (100%) | 8 | 0.682 | 0.736 | 0.194 | | 0.020 | 0.032 |
| | | | | ***G.p*** | 250-252 | 23 (100%) | 2 | 0.696 | 0.464 | 0.020 | | -0.448 | 0.000 |
| ***GmmL17*** | F: CGTACATGCAAGGCAGAGC<br>R: TCAACTGAAACCGAAAGAGC | Tfly_23-t608l17.q1k | (GGTT)10 [273] | ***G.f.f*** | 283-299 | 22 (100%) | 4 | 0.227 | 0.496 | 0.001* | + | 0.233 | 0.191 |
| | | | | ***G.p*** | 277-285 | 23 (100%) | 2 | 0.217 | 0.322 | 0.165 | + | 0.129 | 0.092 |
| ***GmmP07*** | F: ACTGACATATTGAGTTGAAAGGGG<br>R: TCTTCCGTTAAATACAGAGTGCAG | Tfly_23-t516p07.q1k | (ATT)18 [234] | ***G.f.f*** | 213-234 | 22 (100%) | 4 | 0.227 | 0.357 | 0.038 | + | 0.137 | 0.121 |
| | | | | ***G.p*** | 231-246 | 22 (100%) | 4 | 0.348 | 0.563 | 0.010 | + | 0.168 | 0.142 |

Allele size ranges represent size ranges of amplified DNA minus the 15-mer M13 sequence. The asterisk (*) indicates significance after sequential Bonferroni correction. The plus sign (+) indicates a putative X-linked locus, based on the absence of heterozygous males from both species. The minus sign (-) indicates a potentially X-linked locus based on only one species. Heterozygosity calculated using *Genalex 6.2*: $H_O$ - observed heterozygosity and $H_E$ - expected heterozygosity. P-values for the Hardy-Weinberg (H-W(P)) equilibrium test computed with *Genepop 4.0.10.* Null allele frequencies estimated in *Microchecker* based on Oosterhout et al. (2006) (Null (O et al)) and *FreeNA* based on Dempster et al. (1977) (Null (D et al)).