bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 **Inversion Genotyping in the *Anopheles gambiae* Complex Using High-Throughput**
2 **Array and Sequencing Platforms**
3

4 R. Rebecca Love[*†], Marco Pombi[‡], Moussa W. Guelbeogo[§], Nathan R. Campbell[¶],
5 Melissa T. Stephens[†], Roch K. Dabire[††], Carlo Costantini[‡‡††], Alessandra della Torre[‡],
6 Nora J. Besansky[*†§§]
7
8 *Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556, USA
9 [†]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556,
10 USA
11 [‡]Dipartimento di Sanità Pubblica e Malattie Infettive, Istituto Pasteur Italia-Fondazione
12 Cenci-Bolognetti, Università di Roma "La Sapienza", 00185 Rome, Italy
13 [§]Centre National de Recherche et Formation sur le Paludisme (CNRFP) Ouagadougou,
14 Burkina Faso
15 [¶]GTseek LLC, Twin Falls, ID 83301, USA
16 [††]Institut de Recherche en Sciences de la Santé (IRSS)/Centre Muraz, Bobo-Dioulasso,
17 Burkina Faso
18 [‡‡]Institut de Recherche pour le Développement (IRD) UMR MIVEGEC (University of
19 Montpellier, CNRS 5290 IRD 224) Centre IRD de Montpellier, Montpellier, France
20
21
22 [§§]Corresponding author
23

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24 **Running Title:** High-throughput inversion genotyping

25

26 **Key Words:** *Anopheles gambiae*, chromosomal inversion polymorphism, polytene

27 chromosome analysis, inversion genotyping, karyotyping, malaria vector, tag SNP

28

29 **Corresponding Author:** Nora J. Besansky, Department of Biological Sciences,

30 University of Notre Dame, Notre Dame, IN 46556 USA, nbesansk@nd.edu

31

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Chromosomal inversion polymorphisms have special importance in the *Anopheles gambiae* complex of malaria vector mosquitoes, due to their role in local adaptation and range expansion. The study of inversions in natural populations is reliant on polytene chromosome analysis by expert cytogeneticists, a process that is limited by the rarity of trained specialists, low throughput, and restrictive sampling requirements. To overcome this barrier, we ascertained tag single nucleotide polymorphisms (SNPs) that are highly correlated with inversion status (inverted or standard orientation). We compared the performance of the tag SNPs using two alternative high throughput molecular genotyping approaches versus traditional cytogenetic karyotyping of the same 960 individual *An. gambiae* and *An. coluzzii* mosquitoes sampled from Burkina Faso, West Africa. We show that both molecular approaches yield comparable results, and that either one performs as well or better than cytogenetics in terms of genotyping accuracy. Given the ability of molecular genotyping approaches to be conducted at scale and at relatively low cost without restriction on mosquito sex or developmental stage, molecular genotyping via tag SNPs has the potential to revitalize research into the role of chromosomal inversions in the behavior and ongoing adaptation of *An. gambiae* and *An. coluzzii* to environmental heterogeneities.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

51 **INTRODUCTION**

52

53 A chromosomal inversion is a structural mutation that arises when a chromosome
54 segment breaks and reattaches in reverse orientation. Those that are retained as long-
55 term polymorphisms often span hundreds or thousands of genes (Wellenreuther &
56 Bernatchez 2018). Suppressed recombination in inversion heterozygotes (between
57 inverted and non-inverted orientations) preserves allelic combinations on the inverted
58 arrangement as haplotype blocks. The proposed role of inversions in adaptation to
59 environmental heterogeneities arises from the expectation that locally adapted
60 haplotype blocks can be maintained by spatially and temporally varying selection
61 (Hoffmann & Rieseberg 2008; Kirkpatrick 2010; Kirkpatrick & Barton 2006; Schaeffer
62 2008). Despite significant advances in genomic resources and technologies, a more
63 detailed understanding of the precise alleles targeted by selection and their phenotypic
64 consequences is largely lacking. One important roadblock to progress is the paucity of
65 methodologies that allow inversion genotyping at scale.

66

67 Polymorphic chromosomal inversions are abundant in all four of the major malaria-
68 transmitting mosquitoes with pan-African distributions (Ayala *et al.* 2017; Coluzzi *et al.*
69 2002), including the two sister species of the *Anopheles gambiae* complex studied here,
70 *An. gambiae sensu stricto* (hereafter *An. gambiae*) and *An. coluzzii*. It has been
71 suggested that these inversion polymorphisms promote ecological flexibility, enabling
72 the successful exploitation of heterogeneous environments across tropical Africa (Ayala
73 *et al.* 2017; Coluzzi *et al.* 2002; Costantini *et al.* 2009). Decades ago, intensive and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

74 laborious cytogenetic studies demonstrated that inversion frequencies correlate with
75 latitudinal gradients of aridity (Coluzzi *et al.* 1979), seasonal fluctuations in rainfall
76 (Rishikesh *et al.* 1985), and local microhabitat differences (Coluzzi *et al.* 1979), patterns
77 that persist stably over decades and across different geographic regions. These
78 observations suggest that inversions confer an adaptive benefit in arid environments,
79 where they achieve their highest frequencies. The epidemiological relevance of
80 inversion polymorphism for malaria transmission and control, beyond habitat expansion
81 and seasonal persistence of the mosquito disease vectors, was manifest by differences
82 in vector house resting behavior (Coluzzi *et al.* 1979; Molineaux & Gramiccia 1980).
83 The significantly greater tendency of the inversion-carrying fraction of the mosquito
84 population to rest indoors caused non-uniform exposure to indoor-based vector control,
85 reducing its efficacy.
86
87 Despite the undeniable public health importance of this phenomenon, scientific
88 understanding has barely advanced in the forty years since its initial discovery. No
89 small reason for this hiatus is the technological and logistical difficulty of inversion
90 genotyping. Polytene chromosome analysis of anopheline mosquitoes (della Torre
91 1997) is the current basis of inversion genotyping. Strongly rate-limiting, sex-specific
92 and stage-specific, it requires dissection, preparation, and microscopic analysis of
93 ovarian polytene chromosomes by expert cytogeneticists with highly specialized training
94 in the interpretation of chromosome banding patterns of the focal species.
95

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

96 DNA-based molecular assays would offer a much more rapid and widely available
97 approach to inversion genotyping. These could be applied to mosquitoes regardless of
98 sex, developmental stage, or method of preservation, and would require no more
99 training than that normally associated with any molecular entomology laboratory. More
100 than ten years ago, a rapid PCR assay was developed for genotyping of the 22-Mb 2La
101 inversion (White *et al.* 2007), one of six common inversion polymorphisms in *An.*
102 *gambiae* (all shared with *An. coluzzii* except one, 2Rj). The availability of this assay,
103 which targets 2La inversion breakpoints, simplified the search for phenotypic traits and
104 single nucleotide polymorphisms (SNPs) associated with the inverted orientation (2La).
105 Studies employing this tool in the laboratory suggested that the inverted orientation was
106 associated with increased thermal and desiccation tolerance, a thicker cuticle, higher
107 body water content, a more aggressive upregulation of heat-responsive genes such as
108 heat shock genes, and a higher energy budget relative to the alternative ($2L+^{a}$)
109 arrangement (Cassone *et al.* 2011; Cheng *et al.* 2018; Fouet *et al.* 2012; Gray *et al.*
110 2009; Reidenbach *et al.* 2014; Rocca *et al.* 2009). Importantly, this inversion
111 genotyping tool also was used to advance the study of natural populations. Inversion
112 2La was found to be associated with a reduced tendency to rest indoors and a lower
113 malaria oocyst infection prevalence, corroborating historical evidence based on
114 cytogenetic analysis (Coluzzi *et al.* 1979; Petrarca & Beier 1992; Riehle *et al.* 2017).
115 Association mapping employing population pools of alternative 2La homokaryotypes
116 revealed dozens of candidate SNPs significantly associated with desiccation tolerance
117 (Ayala *et al.* 2019).
118

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

119 However, molecular genotyping tools that perform robustly for the common inversions
120 on the right arm of chromosome 2 (2Rj, 2Rb, 2Rc, 2Rd, 2Ru) have been lacking, with
121 the sole exception of a newly available set of polymerase chain reaction (PCR)
122 restriction fragment length polymorphism (RFLP) assays for 2Rb (Montanez-Gonzalez
123 *et al.* 2020). Two previously developed PCR genotyping assays, one that targeted the
124 breakpoints of this inversion and another that targets the breakpoints of inversion 2Rj,
125 proved unreliable or had limited geographic application in natural populations,
126 presumably due to structural variation in inversion breakpoint regions (Coulibaly *et al.*
127 2007; Lobo *et al.* 2010). The newly developed PCR-RFLP genotyping assays for 2Rb
128 (Montanez-Gonzalez *et al.* 2020), and additional genotyping assays under development
129 for 2Rc (Montanez-Gonzalez, Vallera, Calzetta, Love, Pombi, Guelbeogo, Dabire,
130 Costantini, Pichler, Petrarca, della Torre, Besansky, unpublished) exploit tag SNPs
131 inside the rearranged region whose allelic state is strongly correlated with inversion
132 orientation regardless of their position relative to the breakpoints (Love *et al.* 2019). To
133 our knowledge, no DNA-based molecular assays exist for the genotyping of the other
134 inversions in *An. gambiae* or *An. coluzzii*.
135
136 We recently described a strategy that exploited the *An. gambiae* and *An. coluzzii*
137 database of natural variation (Ag1000G; www.malariagen.net/projects/ag1000g) (Miles
138 *et al.* 2017) to identify tag SNPs predictive of inversion orientation for all six common
139 inversion polymorphisms in these species. Using these tags, we developed an
140 algorithm capable of *in silico* inversion genotyping based on SNPs called from whole
141 genome resequencing data (Love *et al.* 2019). This is a rapid and powerful approach

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

142 assuming that whole genome sequence data are already available or will be produced
143 for other reasons. However, it does not satisfy experimental designs in which genomic
144 sequence data is not otherwise required, and where its procurement would be cost-
145 prohibitive. For the requisite statistical power, studies aimed at finding significant
146 associations between inversions and behavioral or physiological phenotypes will likely
147 require thousands of specimens of known inversion genotype. Here, we develop cost-
148 effective high-throughput molecular methods of inversion genotyping to address this
149 need. Using tag SNPs ascertained in Ag1000G (Love *et al.* 2019), we compare two
150 molecular platforms that allow inversion genotyping of hundreds or thousands of
151 individual *An. coluzzii* and *An. gambiae* mosquitoes at tens or hundreds of tag SNPs
152 targeting all six inversions in a single experiment. One platform, the TaqMan
153 OpenArray (Life Technologies), referred to hereafter as OA, is a 5'-exonuclease method
154 that genotypes tag SNPs based on PCR in the presence of allele-specific probes, both
155 labeled with different reporter dyes. The other, Genotyping-in-Thousands by
156 sequencing (GT-seq), is a custom amplicon sequencing approach that genotypes tag
157 SNPs by next-generation sequencing of multiplexed PCR products (Campbell *et al.*
158 2015). Using 960 individual *An. gambiae* and *An. coluzzii* mosquitoes previously
159 karyotyped cytogenetically and up to 184 SNP markers, we show that both approaches
160 successfully predict inversion genotypes for the common polymorphic inversions in
161 these species (excluding 2Rc in *An. gambiae* and 2Rd in *An. coluzzii*). Our data
162 suggest that these methods not only offer efficiency of scale and cost, but also
163 represent substantial improvements in genotyping accuracy relative to the classical
164 cytogenetic approach.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

165 **MATERIALS AND METHODS**

166

167 **Mosquito study population**

168 Burkina Faso lies in the arid Sudan savanna belt of West Africa. In this region, *An.*
169 *gambiae* and *An. coluzzii* are highly polymorphic for chromosomal inversions (Costantini
170 *et al.* 2009). Sampling was conducted in a 35 x 65 km area located 30 km SW of the
171 capital, Ouagadougou. In total, 85 villages approximately 5 km distant from each other
172 were sampled in 2006. Mosquito collection was performed indoors in the early
173 afternoon by pyrethrum spray catch in 3-5 compounds per village. Morphological
174 identification and initial processing was performed in the field under a dissecting
175 microscope. *An. gambiae s.l*. females at the appropriate stage for polytene
176 chromosome analysis were each assigned a unique numerical code, whose value was
177 incremented by '1' with each new mosquito as the collection progressed. Ovaries of
178 each female mosquito were immediately cropped and placed in an individual 1.5 ml
179 tube containing Carnoy's fixative (1:3 glacial acetic acid:absolute ethanol), labeled with
180 its unique numerical code. The corresponding carcass was placed in an individual 1.5
181 ml tube containing a desiccant (silica gel) and a matching numeric label unique for that
182 mosquito. Ovaries were stored at -20°C, and carcasses maintained at ambient
183 temperature before further processing.

184

185 Mosquito DNA was extracted from the carcass using a CTAB method (Chen *et al.* 2010)
186 and identified to species using rDNA-based PCR assays (Favia *et al.* 1997; Scott *et al.*
187 1993). The corresponding ovaries were prepared for karyotype analysis according to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

188 standard procedures (della Torre 1997). The banding pattern was observed under a
189 phase-contrast microscope (400x) and interpreted with reference to the cytogenetic
190 map (George *et al.* 2010; Pombi *et al.* 2008). Karyotype analysis was performed on
191 >1,770 mosquitoes. The effort was divided equally between two groups without spatial
192 or temporal sampling bias, based strictly on mosquitoes with odd- versus even-valued
193 numerical codes. For this study, we selected a subset of 960 mosquitoes based on
194 their cytogenetic karyotypes, with the goal of achieving maximum possible inversion
195 genotype balance for the purpose of validating the molecular tag SNPs.
196
197 **OA and GT-seq assay design and genotyping**
198 Because the methods we used to develop and validate the set of tag SNPs for *in silico*
199 inversion genotyping (Love *et al.* 2019) were still being refined at the time the present
200 study was initiated, the initial list of candidate molecular tags differed slightly from the *in*
201 *silico* set, though overlap was extensive. One minor methodological difference in
202 ascertainment was that, for the molecular tag SNPs, genotypic concordance was not
203 based on ten bootstrap replications of a training set, but instead was based on the
204 simple percentage of mosquitoes with a matching inversion- and SNP-genotype out of
205 the total mosquito sample analyzed for that inversion in the Ag1000G variation
206 database. This was calculated separately for each of three inversion genotypes
207 [homozygous standard (*i.e.*, uninverted), heterozygous, homozygous inverted], with the
208 minimum value taken as the conservative genotypic concordance. The other
209 methodological difference was that ascertainment of tag SNPs in 2Rc was based on
210 slightly different data partitions: (i) *An. coluzzii*, and (ii) *An. gambiae* after exclusion of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

211 BAMAKO (Manoukis *et al.* 2008) and specimens carrying the inverted arrangement of
212 2Ru. Beginning with a ranked list of tags based on descending genotypic concordance,
213 we applied filters (described below) that narrowed the numbers of tags based on design
214 criteria unique to each platform.
215
216 *OA:* TaqMan assays were designed by the Dana-Farber/Harvard Cancer
217 Center (DF/HCC) Genotyping and Genetics for Population Sciences Core, a unit of the
218 Partners HealthCare Center for Personalized Genetic Medicine. Assays designed for
219 this platform require forward and reverse PCR primers which produce ~100 base
220 amplicons containing the tag SNP, and additionally require two allele-specific
221 fluorescently labeled 30-bp probes ('reporters') that discriminate between the reference
222 allele at the tag (VIC dye) and the alternate allele (FAM dye). Candidate SNPs were
223 filtered out if they were surrounded by runs of nucleotides and low complexity regions
224 that interfered with acceptable primer design parameters. Further filtering was
225 performed if candidate SNPs were surrounded by high frequency variants in the 25
226 bases immediately upstream or downstream of the tag SNP. High frequency was
227 defined as ≥5% in at least one inversion genotype (*i.e.*, homozygous standard,
228 heterozygous, homozygous inverted) or at least two population samples analyzed for a
229 given inversion in the Ag1000G database. [As discussed in Love *et al*. (2019), the
230 population samples analyzed varied depending upon the inversion under consideration,
231 due to inferred taxonomic or geographic population structure based on principal
232 components analysis]. Such high frequency variants were deemed likely to significantly
233 interfere with successful probe annealing. Due to the highly polymorphic nature of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

234 *An. gambiae* genome (Miles *et al.* 2017), these filters eliminated many candidate tags.
235 To ensure that we retained at least six tag SNPs per inversion for genotyping, we were
236 compelled to reduce the genotypic concordance threshold below the 0.8 level imposed
237 in Love *et al*. (2019). Even lowering the threshold to 0.7 for 2Rc tags in *An. gambiae*
238 failed to yield more than three candidates, and we declined to reduce that threshold
239 further. Minimum genotypic concordances for each inversion were 0.7 for 2Rb, 2Rc and
240 2Ru; 0.75 for 2Rd; 0.9 for 2Rj; and 0.9925 for 2La.
241
242 After filtering, we retained 54 tag SNPs in total, ranging from 6 to 11 per inversion
243 except 2Rc in *An. gambiae,* with only 3 tag SNPs. Based on these 54 tags, we selected
244 a custom 64-assay TaqMan OpenArray genotyping plate design whose 3,072 reaction
245 through-holes are divided into 48 sub-arrays, each with 64 through-holes (54 of which
246 were preloaded with a single custom assay). One such plate genotypes 48 mosquitoes
247 at 54 tags (2,592 genotypic assays).
248
249 DNA quantification of genomic DNA from 960 mosquitoes was conducted by DF/HCC
250 via picogreen-based fluorimetry; average DNA concentration was 26 ng/ul (range, 0.1-
251 58.1 ng/ul). OA requires 250 copies of a haploid genome for each individual through-
252 hole [0.0675 ng of *An. gambiae* genomic DNA, assuming a haploid genome size of 0.27
253 pg (260 Mb); (Besansky & Powell 1992)]; 64 through-holes require only ~4-5 ng DNA
254 per mosquito. DF/HCC performed the genotyping using endpoint detection of
255 fluorescent signals on the TaqMan OpenArray Genotyping System, following
256 manufacturer's specifications (Applied Biosystems, Foster City CA, USA). Conditions

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

257 for genotyping are available upon request to DF/HCC. Tag SNPs, primers and probes
258 for genotyping assays are provided in Table S1.
259
260 *GT-seq:* GTseek LLC conducted multiplex primer design and consulted on GT-seq
261 optimization. Because this multiplexed amplicon sequencing approach uses only
262 unlabeled PCR primers to produce 50-100 bp amplicons spanning a tag SNP, high
263 frequency variants neighboring the tag are not a limitation. However, the highly
264 multiplexed nature of GT-seq, allowing simultaneous amplification of up to 500 SNP loci
265 per individual for thousands of individuals, requires that the primer pool be optimized not
266 only for individual amplicons (*e.g.*, by avoiding nucleotide runs, low complexity regions,
267 and primer-dimer), but also to minimize primer interactions across loci and mis-priming
268 with other amplicons. The initial list of candidate tag SNPs ranked by concordance was
269 filtered based on the output of custom perl scripts to evaluate primer pools (Campbell *et*
270 *al.* 2015, https://github.com/GTseq). Minimum concordance values varied by inversion
271 (>0.8 for 2Rc; >0.85 for 2Rd; >0.9 for 2Rb, 2Rj, and 2Ru; >0.995 for 2La). Candidate
272 tag SNPs for 2La, which were overly abundant, were pruned by selecting every third
273 candidate from a list ordered by chromosome position.
274
275 Following Campbell *et al.* (2015), Illumina sequencing primer sites were added to locus-
276 specific forward and reverse primer sequences to create PCR1 primers, which were
277 ordered along with PCR2 primers (a set of 96 i5 and i7 indexes) from Integrated DNA
278 technologies (IDT) in 96-well plate format at a 25nmole synthesis scale and a
279 concentration of 200 µM in Tris-EDTA pH 8.0 buffer. GT-seq test libraries were

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

280 prepared and sequenced by the University of Notre Dame Genomics and Bioinformatics
281 Core Facility (GBCF) from a subset of specimens (n=192) to refine preparation
282 techniques and identify primers that produced PCR artefacts or were overrepresented.
283 Following optimization, primer pools were re-made to include only the optimized panel
284 of PCR 1 primers. Tag SNPs and PCR 1 primers for GT-seq genotyping are listed in
285 Table S2.
286
287 The final libraries prepared by the GBCF included the same 192 specimens used during
288 optimization and 765 additional specimens. They were constructed without optional
289 exo-SAP treatment following Campbell *et al.* (2015), with the following modifications to
290 PCR conditions and post library cleanup: PCR1: 95 °C – 15 min; 5 cycles [95 °C – 30s,
291 3% ramp down to 57°C–30s, 72°C–2min]; 10 cycles [95 °C–30 s, 65 °C–30 s, 72 °C–30
292 s]; 4 °C hold. PCR2: 95 °C – 15 min; 10 cycles [95°C–10s; 62°C–30s; 72°C–30s];
293 72°C–5min; 4°C hold. Following PCR2, each plate of samples was purified and
294 normalized using the Just-a-Plate 96 PCR Purification and Normalization Kit (Charm
295 Biotech) according to manufacturer's instructions. Following normalization, 10 ul of
296 each sample per 96 well plate (up to 960 ul total) was then combined into a 1.5-mL
297 Eppendorf tube, for a total of 10 tubes. From each tube, 300ul was transferred to a
298 fresh 1.5-mL Eppendorf tube for two rounds of purification using AMPure XP
299 paramagnetic beads (Beckman Coulter, Inc.) with ratios of 0.5X and 1.3X respectively.
300 Purified libraries were eluted in 35 ul 1xTE and transferred to fresh 1.5-mL tubes before
301 adding 3.5 ul buffer EB containing a 1% Tween 20 solution.
302

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

303 Each of the 10 plate libraries was quality assessed on an Agilent Bioanalyzer 2100 High
304 Sensitivity chip and quantified by qPCR using the Illumina Kapa Library Quantification
305 Kit (Roche, Cat. #KK4824). The libraries were then normalized to a concentration of 4
306 nM and pooled for sequencing. The final pooled library containing 957 *An. gambiae* and
307 *An. coluzzii* individuals was sequenced on a single lane of Illumina NextSeq 500 v2.5
308 (75 cycle) High Output flowcell using a dual indexed 75bp single-end read. Base calling
309 was done by Illumina Real Time Analysis (RTA) v2 software.
310
311 Using scripts described in the bioinformatics pipeline of Campbell *et al.* (2015) and
312 available on Github (https://github.com/GTseq), sequencing data were demultiplexed
313 into single fastq files for each individual sample. Individuals were genotyped at each
314 locus with a perl script (GTseq_Genotyper_v3.pl) that counts the occurrence of each
315 allele at a locus within individual fastq files. The ratio of allele 1 to allele 2 counts was
316 used to generate a genotype for each locus with total read counts >10, following the
317 methods and cut-offs of Campbell *et al.* (2015).
318

319 **Filtering and calling multilocus inversion genotypes**

320 The procedures for filtering and calling molecular inversion genotypes were the same
321 for both OA and GT-seq platforms. Filtering steps were as follows. For each tag SNP,
322 we calculated the percentage of mosquito specimens in the sample with a genotype call
323 at that tag (the SNP call rate). If SNP call rates were <80%, the underperforming tag
324 SNPs were eliminated from further analysis. In addition, for each mosquito specimen
325 analyzed, we calculated the percentage of tag SNPs with a genotype call (the specimen

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

326 call rate). If the specimen call rate was <80%, that specimen was excluded from further
327 analysis. Note that the mosquito specimens in the sample varied according to the
328 inversion under consideration: 2La, 2Rb and 2Ru tags perform in both species, 2Rj and
329 2Rd tags are *An. gambiae*-specific, and defined subsets of 2Rc tags (referred to in this
330 work as 2Rc_col and 2Rc_gam) apply respectively to *An. coluzzii* or *An. gambiae*
331 individuals.
332
333 To calculate the multilocus inversion genotype for each specimen, we converted the raw
334 genotype data for individual tag SNPs to the count of alternate alleles (if necessary),
335 where '0' is a homozygote for the reference allele, '1' is a heterozygote carrying one
336 reference allele, and '2' is a homozygote for the alternate allele. Next, we averaged the
337 number of alternate alleles present across all tag SNPs in a given inversion, and binned
338 this average to produce a predicted inversion genotype (0-0.67, 0; 0.68-1.33, 1; 1.34-2,
339 2). Multilocus molecular genotypes were then compared to each other and with
340 cytogenetically determined inversion genotypes.
341
342 **Code and data availability**
343 Supplemental files are available at FigShare. Table S1 contains OA tag SNP ID
344 numbers, locations, reference and alternate alleles, forward and reverse primer
345 sequences, and probe sequences. Table S2 contains GT-seq tag SNP ID numbers,
346 locations, reference and alternate alleles, and forward and reverse primer sequences.
347 Table S3 contains specimen ID numbers and inversion genotypes (cytogenetic, OA,
348 and GT-seq) for each individual mosquito. Code used to generate the data can be

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

349 found on Github (https://github.com/GTseq and
350 https://github.com/rrlove/molec_karyo_notebooks).
351

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

352 # RESULTS AND DISCUSSION

353

354 ## OA

355 Custom OA plates were used to genotype 960 individual *An. gambiae* and *An. coluzzii*
356 mosquitoes at 54 tag SNP loci. The SNP call rate for one of the 54 loci fell below the
357 80% threshold (77.7%) and was eliminated from the panel. After filtering, the SNP call
358 rate averaged 99.4% for the remaining 53 tags (range, 98.3%-100%). Three specimens
359 were dropped from analysis due to the belated determination that their cytogenetic
360 genotypes were ambiguous. Four additional mosquito specimens were dropped from
361 further OA analysis due to unacceptably low specimen call rates (ranging from 17.5% to
362 52.6%). The remaining 953 specimens had an average specimen call rate of 99.3%
363 (range, 87.7%-100%). The final number of OA tags per each inversion, and their
364 approximate genomic position within the inversion, are shown in Table 1 and Figure 1.

365

366 ## GT-seq

367 Sequencing from one NextSeq lane included the pooled GT-seq library of 957 *An.*
368 *gambiae* and *An. coluzzii,* as well as another GT-seq pooled library of 235 *An. funestus*
369 mosquitoes pertaining to an independent experiment to be described elsewhere. This
370 produced ~359M total reads, of which ~236M could be assigned to the 957 *An.*
371 *gambiae* and *An. coluzzii* specimens based on their barcode sequences. Read counts
372 from each of the ten *An. gambiae-An. coluzzii* sample plates ranged from 17.5M to
373 27.6M reads per plate and read counts per individual mosquito averaged 246,925 (SD
374 85,614). The tag SNP call rate was below the 80% threshold for three tags, which were

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

375 subsequently dropped from the genotyping panel. For the remaining 131 tags, SNP call
376 rates averaged 98.2% (range, 80.25% to 100%). As was the case for OA analysis,
377 three specimens were dropped due to ambiguous cytogenetic genotypes. Of the
378 remaining 954, two were considered to have failed because they had specimen call
379 rates below the 80% threshold, and were thus dropped. Those 952 within acceptable
380 limits had average specimen call rates of 98.2% (range, 93.1% to 100%). The final
381 number of GT-seq tags per each inversion, and their approximate genomic position
382 within the inversion, are shown in Table 1 and Figure 1.
383

384 **Concordance**

385 The inversion genotypes inferred for each specimen by the three methods
386 (cytogenetics, OA, and GT-seq) are provided in Table S3. We compared these
387 genotypes to assess their concordance. Due to our filtering rules, not every specimen
388 had genotype calls by both molecular methods. We focused our assessment on the
389 subset of specimens that were successfully genotyped with all three methods (435 for
390 *An. gambiae*, and 513 for *An. coluzzii*). As summarized in Figure 2 and Table 2, over
391 90% of the relevant mosquito samples had concordant genotypes for all three methods
392 with the notable exception of *An. gambiae* genotyped for 2Rc, where three-way
393 concordance fell to ~81%. We discuss the special case of 2Rc in *An. gambiae* more
394 fully below; here, we concentrate on the five other inversions and 2Rc in *An. coluzzii*.
395

396 A strikingly high number of specimens had multilocus molecular genotypes inferred from
397 both OA and GT-seq that agreed, but were jointly discordant with cytogenetics. Except

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

398 for 2Rj with negligible discordance (and correspondingly low levels of polymorphism in
399 our sample), the cytogenetic versus multilocus molecular discordance affected from 14
400 to 73 mosquito specimens per inversion, representing 3% to 8% of the mosquito
401 samples (mean, 5%). Although cytogenetic karyotyping may be considered the gold
402 standard for inversion genotyping, two important considerations lend considerable
403 confidence to molecular genotypes, particularly when both molecular approaches
404 concur. First, while none of the tag SNPs are deterministic (*i.e.*, none is perfectly and
405 invariably correlated with inversion orientation), OA and GT-seq infer genotypes based
406 on multiple predictive tags scored per inversion, thus providing weight of numbers.
407 Second, the final set of tags used for OA and GT-seq are almost completely non-
408 overlapping, an outcome produced by distinct filters imposed on the initial list of
409 candidate tags during assay development (see Methods; Figure 3). Accordingly,
410 agreement between both molecular methods is even stronger evidence in favor of the
411 inferred molecular genotype than that provided by one or the other molecular method by
412 itself. Table 2 shows that the two molecular methods agree at least 95% of the time (an
413 average of 98%), except in the case of 2Rc in *An. gambiae*.
414
415 We hypothesized that some genotypic discordances, specifically those in which the two
416 molecular methods agree but conflict with cytogenetics, are caused by cytogenetic
417 errors rather than systematic biases in the molecular approaches. This is difficult to
418 demonstrate conclusively, because the specimens used to compare the three
419 genotyping methods have not been subjected to whole genome sequencing.
420 Furthermore, it was not possible to double-check the cytogenetic karyotypes of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

421 specimens with discordances in the majority of cases, because neither slides nor
422 ovaries were available. However, there is strong evidence consistent with cytogenetic
423 error. During sampling in the field, specimens were assigned numerical identifiers that
424 incremented by "1" throughout the process; cytogenetic karyotyping was later split
425 between two institutions on the basis of even- or odd-numbered identifiers (see
426 Methods). This procedure virtually eliminates the possibility of biases owing to temporal
427 or spatial heterogeneities during the course of the mosquito sampling. Because even-
428 and odd-numbered specimens should be random subsamples from the same
429 populations, we would expect no difference in discordance rates between them. This
430 was not what we found. Of the specimens assayed by both molecular methods in this
431 study, 280 were odd-numbered and 677 even-numbered. Focusing on the inversions
432 with the largest numbers of specimens whose cytogenetic and joint molecular
433 genotypes disagreed (2La, 73; 2Rb, 41; 2Ru, 56; Table 2), the combined 170 such
434 discordances occurred disproportionately in even-numbered specimens: 169 of the 170.
435 Analyses of 2x2 contingency tables demonstrated highly significant departures from the
436 null hypothesis (by Chi-square and Fisher exact probability tests), consistent with the
437 notion that cytogenetic error disproportionately affecting the even-numbered specimens
438 is responsible for these genotypic discrepancies (~5%). If this is the case, then based
439 on the fact that both molecular approaches agree >95% of the time, we suggest that the
440 true error rate for either molecular approach is <5%, probably closer to ~2%.
441
442 It is important to recognize that although the tag SNPs assayed by the two molecular
443 approaches are largely non-overlapping for technical reasons, the assumptions

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

444 underlying the ascertainment of the initial set of candidate tags were the same. The
445 implication is that if those assumptions are violated in natural populations, both
446 approaches may agree on the wrong genotype. The tags were ascertained in the
447 Ag1000G variation database, whose content was heavily biased toward *An. gambiae* at
448 the time of their discovery (Love *et al.* 2019). Available samples of *An. coluzzii* were
449 more limited in numbers and geographic representation, although Burkina Faso was
450 one of two countries represented for this species. In addition to the issue of sampling
451 limitations is the issue of unsuspected (cryptic) population structure that could affect the
452 performance of these tags. Population structure could arise from several non-mutually
453 exclusive scenarios: (i) a lack or reduction of connectivity between natural populations;
454 (ii) local heterogeneity in selection pressures acting on targets inside the inversion;
455 and/or (iii) violation of the assumption that the focal inversion arose uniquely (*i.e.*, has a
456 monophyletic origin). We suspect that at least one of these scenarios applies to 2Rc
457 tags ascertained in *An. gambiae*, probably explaining their lower rate of apparent
458 success in genotyping (based on lower concordance values across the board; Table 2).
459 Previous work has shown that applying candidate tags to a taxon in which they are not
460 valid has the effect of downwardly biasing the average number of inferred alternate
461 alleles (Love *et al.* 2019). Consistent with this, if 2Rd tags in the present study were
462 inappropriately applied to *An. coluzzii*, the vast majority (38 of 39) of specimens
463 genotyped cytogenetically as heterozygotes would be molecularly genotyped as
464 standard homozygotes (Table S3). Based on this, we expect the systematic
465 underestimation of the number of alternate alleles to produce a distinctive pattern where
466 ‘true’ standard homozygotes are correctly identified, but heterozygotes and inverted

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

467 homozygotes would be incorrectly genotyped molecularly as standard homozygotes.
468 Table 3 shows the distribution of discordant genotypes between cytogenetics and joint
469 molecular methods when broken down by genotypic class: standard homozygotes,
470 heterozygotes, and inverted homozygotes. While we have no objective measure of
471 which specimens are 'true' heterozgyotes and 'true' inverted homozygotes, it is
472 noteworthy that the discordances for all inversions other than 2Rc in *An. gambiae* either
473 skew toward molecular genotypes of '1' or '2', or they are roughly equally distributed
474 between '1' or '2' and '0'. The pattern for 2Rc in *An. gambiae* is distinctive, in that the
475 skew is strongly toward molecular genotypes of '0', which is consistent with tags that
476 may not be appropriately suited for the *An. gambiae* population in which they are
477 applied. Further study is both required and merited, to understand the cause(s) of
478 population structure between the populations used to develop the 2Rc tags for in *An.*
479 *gambiae* in Ag1000G and those used to test the tags in the present study. Interestingly,
480 2Rc contains cytochrome P450 genes implicated in insecticide resistance in *An.*
481 *gambiae* and *An. coluzzii* (Love *et al.* 2016; Main *et al.* 2015), and more broadly,
482 Coluzzi and colleagues (2002) observed that the region spanned by 2Rc is involved in
483 many rearrangements that differentiate members of the *An. gambiae* species complex,
484 leading these authors to propose that this region may have ecological relevance with
485 respect to larval breeding site adaptations.
486
487 Although we find good concordance between both molecular genotyping approaches,
488 GT-seq more often agreed with cytogenetics than did OA (Table 2). This is not
489 surprising, given the hybridization-based nature of OA and the extremely high levels of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

490 nucleotide diversity found in *An. coluzzii* and *An. gambiae* (Miles *et al.* 2017). In
491 addition, highly concordant candidate tag SNPs that also had low polymorphism in the
492 ~50 bp immediately surrounding the tag, as required by OA, were sufficiently rare that
493 we were compelled to lower the concordance threshold to find enough candidates
494 suitable for assay design, and consequently the total number of OA tags per inversion is
495 smaller than for GT-seq (Table 1). These factors likely compound to lower the
496 performance of OA compared to GT-seq. Furthermore, as detailed by Campbell *et al.*
497 (2015), genotyping costs are lower for GT-seq compared with OA. Nevertheless, if the
498 number of tag SNPs to be genotyped is low (50-100) and the number of samples high
499 (10^2 to 10^3), OA remains a cost effective option and is still widely used (Campbell *et al.*
500 2015).
501

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

502 **CONCLUSIONS**

503

504 Chromosomal inversions have been viewed as instruments of ecotypic differentiation in
505 anopheline mosquitoes (Coluzzi 1982). Insights into their adaptive significance as
506 balanced polymorphisms and their possible role in behavioral variation, optimal habitat
507 choice, and the speciation process (Coluzzi 1982) were gained from extensive polytene
508 chromosome analyses largely conducted in the pre-genomic era (Coluzzi *et al.* 2002;
509 Coluzzi *et al.* 1979; Manoukis *et al.* 2008; Toure *et al.* 1998). Now, with access to
510 reference genome assemblies and powerful functional genomics tools, the potential
511 exists to probe molecular mechanisms and deepen our understanding. Yet, a major
512 limitation to progress in this area has been the strict requirement for polytene
513 chromosome analysis, which not only limits samples but also demands rare cytogenetic
514 expertise whose throughput is low. Here we demonstrate that tag SNPs highly
515 correlated with inversion status can be used for joint molecular genotyping of common
516 inversions in *An. gambiae* and *An. coluzzii* across the genome (*i.e.*, for karyotyping).
517 Molecular genotyping methods, both OA and GT-seq, can be performed at scale and
518 the results are comparable or superior to traditional cytogenetic karyotyping. These
519 tools invite a renewal of investigations into the role of chromosomal inversions in
520 anopheline behavior and environmental adaptation.

521

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

522 **ACKNOWLEDGEMENTS**

523

524 We thank the technicians of IRD/IRSS/Centre Muraz of Burkina Faso for their
525 fundamental contribution to the field work, as well as the inhabitants of the sampled
526 villages for their kind collaboration. We thank M. Kern and R. Montañez-Gonzalez
527 (University of Notre Dame) for assistance with DNA extraction. The Notre Dame Center
528 for Research Computing provided technical support. We thank Dana-Farber/Harvard
529 Cancer Center in Boston, MA, for the use of the Genotyping and Genetics for
530 Population Sciences Core, which provided SNP analysis using the Taqman OpenArray
531 Genotyping System. We acknowledge the Notre Dame Genomics and Bioinformatics
532 Core Facility for assistance with GT-seq project development, library preparation, and
533 sequencing. This work was supported by the National Institutes of Health (R01
534 AI125360 awarded to NJB). During this work, NJB was supported by Target Malaria,
535 which receives core funding from the Bill & Melinda Gates Foundation and from the
536 Open Philanthropy Project Fund, an advised fund of Silicon Valley Community
537 Foundation. Dana-Farber/Harvard Cancer Center is supported in part by an NCI Cancer
538 Center Support Grant # NIH 5 P30 CA06516.

539

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

540
FIGURE LEGENDS

541

542 **Figure 1.** Locations of tag SNPs assayed for each inversion for OA and GT-seq. The
543 bottom panel labeled 2Rdu shows inversions 2Ru and 2Rd on the same panel, as 2Ru
544 is wholly encompassed by 2Rd. Vertical lines indicate SNPs common to both methods.

545

546 **Figure 2.** Concordance heat map of genotypes imputed by cytogenetics (CYT), OA,
547 and GT-seq. Each row is an individual mosquito, and each column compares inversion
548 genotypes derived from three genotyping approaches for a given inversion (a, 2La; j,
549 2Rj; b, 2Rb; c, 2Rc; d, 2Rd; u, 2Ru). Rows are grouped by species; 2Rj and 2Rd tags
550 are not applicable in *An. coluzzii*. Green represents 3-way genotypic concordance;
551 yellow, concordance between OA and GT-seq; purple, concordance between CYT and
552 GT-seq; black, concordance between CYT and OA; gray is missing data; red is 3-way
553 discordance.

554

555 **Figure 3**. Venn diagrams showing degree of overlap between tag SNPs developed for
556 *in silico* inversion genotyping by Love *et al*. (2019) and those developed in this study for
557 OA and GT-seq.

558

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

559 **LITERATURE CITED**

560

561 Ayala D, Acevedo P, Pombi M, *et al.* (2017) Chromosome inversions and ecological plasticity in
562 the main African malaria mosquitoes. *Evolution*, **71**, 686-701.
563 Ayala D, Zhang S, Chateau M, *et al.* (2019) Association mapping desiccation resistance within
564 chromosomal inversions in the African malaria vector Anopheles gambiae. *Molecular*
565 *Ecology*, **28**, 1333-1342.
566 Besansky NJ, Powell JR (1992) Reassociation kinetics of Anopheles gambiae (Diptera:
567 Culicidae) DNA. *Journal of Medical Entomology*, **29**, 125-128.
568 Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-
569 seq): A cost effective SNP genotyping method based on custom amplicon sequencing.
570 *Molecular Ecology Resources*, **15**, 855-867.
571 Cassone BJ, Molloy MJ, Cheng C, *et al.* (2011) Divergent transcriptional response to thermal
572 stress by *Anopheles gambiae* larvae carrying alternative arrangements of inversion 2La.
573 *Molecular Ecology*, **20**, 2567-2580.
574 Chen H, Rangasamy M, Tan SY, Wang H, Siegfried BD (2010) Evaluation of five methods for
575 total DNA extraction from western corn rootworm beetles. *PloS One*, **5**, e11963.
576 Cheng C, Tan JC, Hahn MW, Besansky NJ (2018) A systems genetic analysis of inversion
577 polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proceedings of the National*
578 *Academy of Sciences of the United States of America*, **115**, E7005-E7014.
579 Coluzzi M (1982) Spatial distribution of chromosomal inversions and speciation in anopheline
580 mosquitoes. In: *Mechanisms of Speciation* (ed. Barigozzi C), pp. 143-153. Alan R. Liss,
581 Inc., New York.
582 Coluzzi M, Sabatini A, Della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome
583 analysis of the *Anopheles gambiae* species complex. *Science*, **298**, 1415-1418.
584 Coluzzi M, Sabatini A, Petrarca V, Di Deco MA (1979) Chromosomal differentiation and
585 adaptation to human environments in the *Anopheles gambiae* complex. *Transactions of*
586 *the Royal Society of Tropical Medicine and Hygiene*, **73**, 483-497.
587 Costantini C, Ayala D, Guelbeogo WM, *et al.* (2009) Living at the edge: biogeographic patterns
588 of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*.
589 *BMC Ecology*, **9**, 16.
590 Coulibaly MB, Pombi M, Caputo B, *et al.* (2007) PCR-based karyotyping of *Anopheles gambiae*
591 inversion 2Rj identifies the BAMAKO chromosomal form. *Malaria Journal*, **6**, 133.
592 Della Torre A (1997) Polytene chromosome preparation from anopheline mosquitoes. In:
593 *Molecular Biology of Disease Vectors: A Methods Manual* (eds. Crampton JM, Beard
594 CB, Louis C), pp. 329-336. Chapman & Hall, London.
595 Favia G, Della Torre A, Bagayoko M, *et al.* (1997) Molecular identification of sympatric
596 chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive
597 isolation. *Insect Molecular Biology*, **6**, 377-383.
598 Fouet C, Gray E, Besansky NJ, Costantini C (2012) Adaptation to aridity in the malaria mosquito
599 *Anopheles gambiae*: chromosomal inversion polymorphism and body size influence
600 resistance to desiccation. *PloS One*, **7**, e34841.
601 George P, Sharakhova MV, Sharakhov IV (2010) High-resolution cytogenetic map for the
602 African malaria vector Anopheles gambiae. *Insect Molecular Biology*, **19**, 675-682.
603 Gray EM, Rocca KA, Costantini C, Besansky NJ (2009) Inversion 2La is associated with
604 enhanced desiccation resistance in *Anopheles gambiae*. *Malaria Journal*, **8**, 215.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

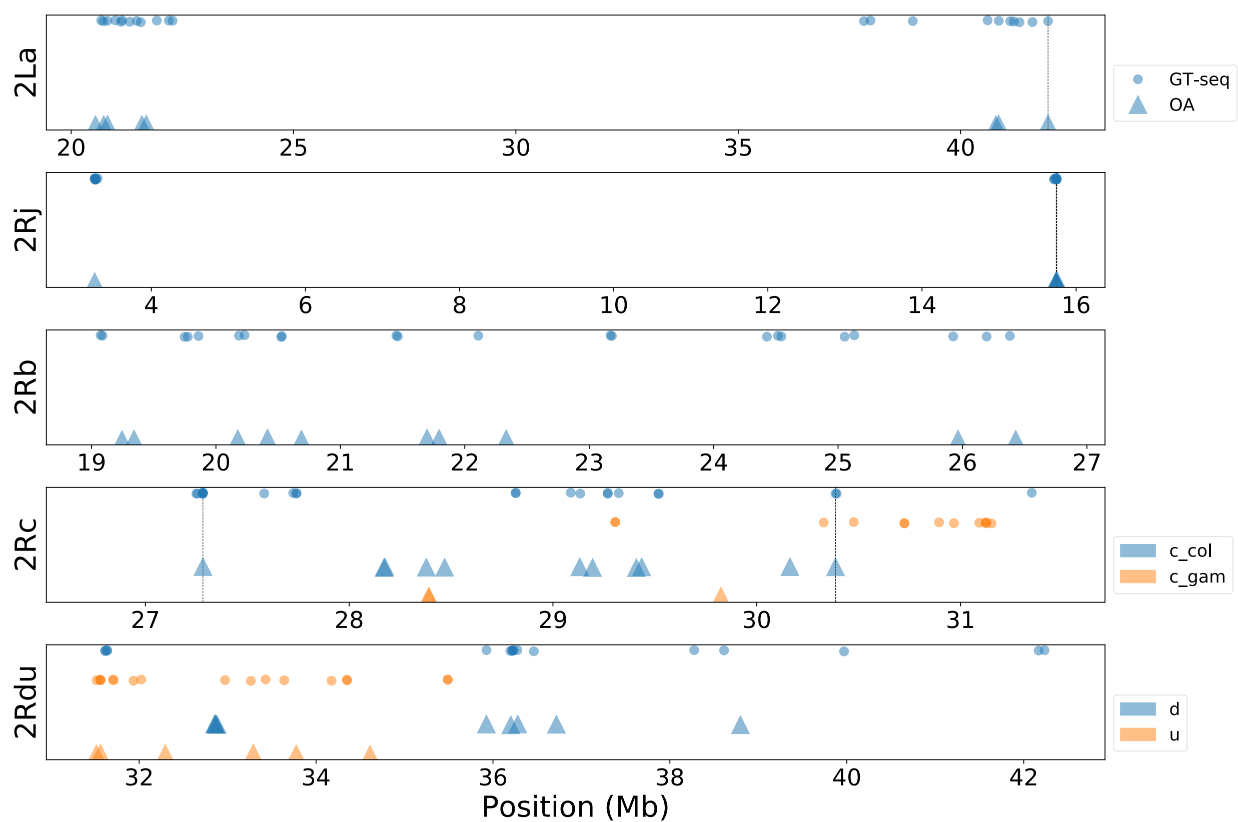
605 Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: from
606 population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of*
607 *Ecology Evolution and Systematics*, **39**, 21-42.
608 Kirkpatrick M (2010) How and why chromosome inversions evolve. *PLoS Biology*, **8**, e1000501.
609 Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation.
610 *Genetics*, **173**, 419-434.
611 Lobo NF, Sangare DM, Regier AA*, et al.* (2010) Breakpoint structure of the Anopheles gambiae
612 2Rb chromosomal inversion. *Malaria Journal*, **9**, 293.
613 Love RR, Redmond SN, Pombi M*, et al.* (2019) In silico karyotyping of chromosomally
614 polymorphic malaria mosquitoes in the *Anopheles gambiae* complex. *G3 (Bethesda)*, **9**,
615 3249-3262.
616 Love RR, Steele AM, Coulibaly MB*, et al.* (2016) Chromosomal inversions and ecotypic
617 differentiation in Anopheles gambiae: the perspective from whole-genome sequencing.
618 *Molecular Ecology*, **25**, 5889-5906.
619 Main BJ, Lee Y, Collier TC*, et al.* (2015) Complex genome evolution in Anopheles coluzzii
620 associated with increased insecticide usage in Mali. *Molecular Ecology*, **24**, 5145-5157.
621 Manoukis NC, Powell JR, Touré MB*, et al.* (2008) A test of the chromosomal theory of ecotypic
622 speciation in *Anopheles gambiae*. *Proceedings of the National Academy of Science U S*
623 *A*, **105**, 2940-2945.
624 Miles A, Harding NJ, Bottà G*, et al.* (2017) Genetic diversity of the African malaria vector
625 *Anopheles gambiae*. *Nature*, **552**, 96-100.
626 Molineaux L, Gramiccia G (1980) *The Garki Project. Research on the Epidemiology and*
627 *Control of Malaria in the Sudan Savanna of West Africa* World Health Organization,
628 Geneva, Switzerland.
629 Montanez-Gonzalez R, Pichler V, Calzetta M*, et al.* (2020) Highly specific PCR-RFLP assays
630 for karyotyping the widespread 2Rb inversion in malaria vectors of the Anopheles
631 gambiae complex. *Parasit Vectors*, **13**, 16.
632 Petrarca V, Beier JC (1992) Intraspecific chromosomal polymorphism in the *Anopheles gambiae*
633 complex as a factor affecting malaria transmission in the Kisumu area of Kenya.
634 *American Journal of Tropical Medicine and Hygiene*, **46**, 229-237.
635 Pombi M, Caputo B, Simard F*, et al.* (2008) Chromosomal plasticity and evolutionary potential
636 in the malaria vector *Anopheles gambiae sensu stricto*: insights from three decades of
637 rare paracentric inversions. *BMC Evolutionary Biology*, **8**, 309.
638 Reidenbach KR, Cheng C, Liu F*, et al.* (2014) Cuticular differences associated with aridity
639 acclimation in African malaria vectors carrying alternative arrangements of inversion
640 2La. *Parasit Vectors*, **7**, 176.
641 Riehle MM, Bukhari T, Gneme A*, et al.* (2017) The *Anopheles gambiae* 2La chromosome
642 inversion is associated with susceptibility to *Plasmodium falciparum* in Africa. *Elife*, **6**.
643 Rishikesh N, Di Deco MA, Petrarca V, Coluzzi M (1985) Seasonal variations in indoor resting
644 *Anopheles gambiae* and *Anopheles arabiensis* in Kaduna, Nigeria. *Acta Tropica*, **42**, 165-
645 170.
646 Rocca KA, Gray EM, Costantini C, Besansky NJ (2009) 2La chromosomal inversion enhances
647 thermal tolerance of *Anopheles gambiae* larvae. *Malaria Journal*, **8**, 147.
648 Schaeffer SW (2008) Selection in heterogeneous environments maintains the gene arrangement
649 polymorphism of *Drosophila pseudoobscura*. *Evolution*, **62**, 3082-3099.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

650 Scott JA, Brogdon WG, Collins FH (1993) Identification of single specimens of the *Anopheles*
651 *gambiae* complex by the polymerase chain reaction. *American Journal of Tropical*
652 *Medicine and Hygiene*, **49**, 520-529.
653 Toure YT, Petrarca V, Traore SF*, et al.* (1998) The distribution and inversion polymorphism of
654 chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa.
655 *Parassitologia*, **40**, 477-511.
656 Wellenreuther M, Bernatchez L (2018) Eco-Evolutionary Genomics of Chromosomal Inversions.
657 *Trends in Ecology & Evolution*, **33**, 427-440.
658 White BJ, Santolamazza F, Kamau L*, et al.* (2007) Molecular karyotyping of the 2La inversion
659 in *Anopheles gambiae*. *American Journal of Tropical Medicine and Hygiene*, **76**, 334-
660 339.
661
662

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

663 **Figure 1.**



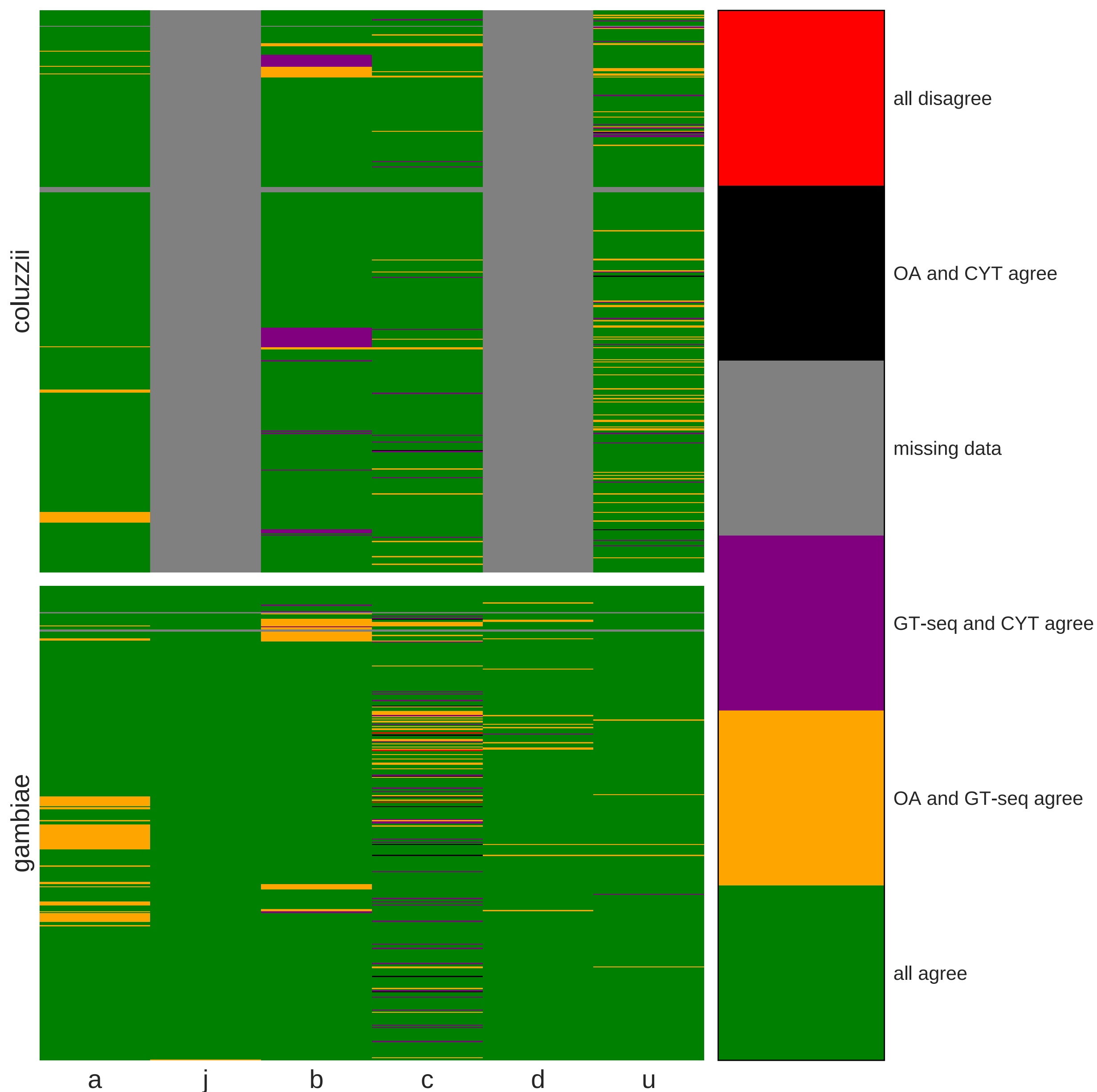
664

665

666

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

667 **Figure 2.**



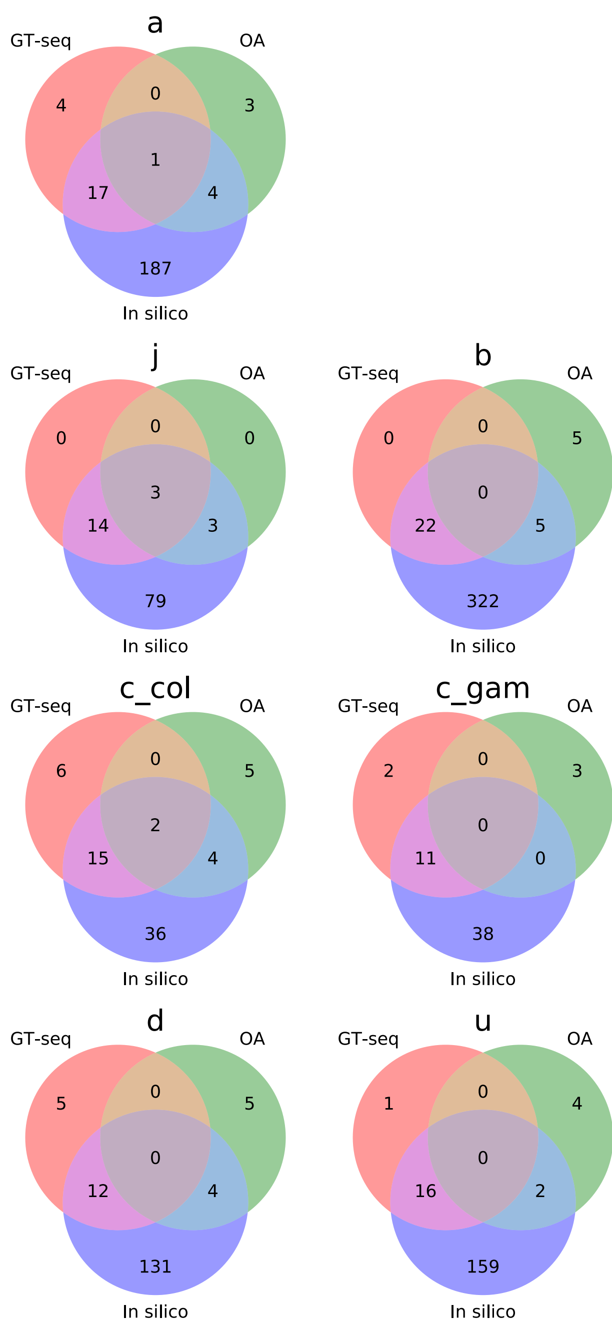
668

669

670

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

671 Figure 3.



672

673

674

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

675 Table 1. The number of tag SNPs by inversion and molecular method.

676

Inversion	GT-seq tag SNPs	OA tag SNPs
2La	22	8
2Rj	17	6
2Rb	22	10
2Rc_col	23	11
2Rc_gam	13	3
2Ru	17	6
2Rd	17	9
Total	**131**	**53**

677 Table 2. Concordance of genotypes imputed by cytogenetics (CYT), OA, and GT-seq for each inversion.

	2La, N (%)	2Rj, N (%)	2Rb, N (%)	2Rc-col, N (%)	2Rc-gam, N (%)	2Rd, N (%)	2Ru, N (%)
Concordance:							
Three-way	875 (92.30)	434 (99.77)	864 (91.14)	483 (94.15)	352 (80.92)	420 (96.55)	870 (91.77)
Discordance:							
CYT vs (GT-seq + OA)	73 (7.70)	1 (0.23)	41 (4.32)	18 (3.51)	39 (8.97)	14 (3.22)	56 (5.91)
All other	0 (0)	0 (0)	43 (4.54)	12 (2.34)	44 (10.11)	1 (0.23)	22 (2.32)
(CYT + GT-seq) vs OA	0 (0)	0 (0)	43 (4.54)	11 (2.14)	31 (7.13)	1 (0.23)	19 (1.05)
(CYT + OA) vs GT-seq	0 (0)	0 (0)	0 (0)	1 (0.19)	10 (2.30)	0 (0)	3 (0.32)
Three-way	0 (0)	0 (0)	0 (0)	0 (0)	3 (0.69)	0 (0)	0 (0)
Specimens assayed by all 3 methods	**948**	**435**	**948**	**513**	**435**	**435**	**948**

678

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.25.114793; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

679 Table 3. Concordance between CYT and both molecular methods by inversion
680 genotype (*0*, *1*, *2*) for specimens for which both molecular methods agree. Shown are
681 the numbers of specimens scored for each pairwise comparison.

		GT-seq + OA		
	CYT	*0*	*1*	*2*
2La	*0*	**1**	0	0
	1	0	**80**	2
	2	1	70	**794**
2Rb	*0*	**124**	1	7
	1	1	**491**	9
	2	7	16	**249**
2Rd	*0*	**389**	7	0
	1	7	**31**	0
	2	0	0	**0**
2Ru	*0*	**778**	39	10
	1	7	**84**	0
	2	0	0	**8**
2Rc_col	*0*	**153**	7	0
	1	10	**235**	0
	2	0	1	**95**
2Rc_gam	*0*	**309**	1	0
	1	34	**43**	0
	2	4	0	**0**

682 CYT, cytogenetics; OA, Open Array
683
684