

1 **In silico karyotyping of chromosomally polymorphic malaria mosquitoes in the**
2 ***Anopheles gambiae* complex**

3

4 R. Rebecca Love^{*}, Seth N. Redmond^{†1}, Marco Pombi[‡], Beniamino Caputo[‡], Vincenzo
5 Petrarca[‡], Alessandra della Torre[‡], The *Anopheles gambiae* 1000 Genomes
6 Consortium[§], Nora J. Besansky^{*††}

7

8 ^{*}Eck Institute for Global Health & Department of Biological Sciences, University of Notre
9 Dame, Notre Dame, IN 46556, USA

10 [†]Infectious Disease and Microbiome Program, Broad Institute, Cambridge, MA 02142,
11 USA

12 [‡]Dipartimento di Sanità Pubblica e Malattie Infettive, Istituto Pasteur Italia-Fondazione
13 Cenci-Bolognetti, Università di Roma "La Sapienza", Piazzale Aldo Moro, 5, 00185
14 Rome, Italy

15 [§]<https://www.malariagen.net/projects/ag1000g#people>

16

17 ^{††}Corresponding author

18 ¹Present address: Monash University, Institute of Vector-Borne Disease, 3800, Clayton,
19 Australia

20

21

22 **Running Title:** In silico karyotyping in malaria mosquitoes

23

24 **Key Words:** *Anopheles gambiae*, chromosomal inversion polymorphism, genomics,

25 inversion genotyping, karyotype analysis, malaria vector, tag SNP

26

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

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

29 **Abstract**

30 Chromosomal inversion polymorphisms play an important role in adaptation to
31 environmental heterogeneities. For mosquito species in the *Anopheles gambiae*
32 complex that are significant vectors of human malaria, paracentric inversion
33 polymorphisms are abundant and are associated with ecologically and epidemiologically
34 important phenotypes. Improved understanding of these traits relies on determining
35 mosquito karyotype, which currently depends upon laborious cytogenetic methods
36 whose application is limited both by the requirement for specialized expertise and for
37 properly preserved adult females at specific gonotrophic stages. To overcome this
38 limitation, we developed sets of tag SNPs inside inversions whose biallelic genotype is
39 strongly correlated with inversion genotype. We leveraged 1,347 fully sequenced *An.*
40 *gambiae* and *Anopheles coluzzii* genomes in the Ag1000G database of natural
41 variation. Beginning with principal components analysis (PCA) of population samples,
42 applied to windows of the genome containing individual chromosomal rearrangements,
43 we classified samples into three inversion genotypes, distinguishing homozygous
44 inverted and homozygous uninverted groups by inclusion of the small subset of
45 specimens in Ag1000G that are associated with cytogenetic metadata. We then
46 assessed the correlation between candidate tag SNP genotypes and PCA-based
47 inversion genotypes in our training sets, selecting those candidates with >80%
48 agreement. Our initial tests both in held-back validation samples from Ag1000G and in
49 data independent of Ag1000G suggest that when used for *in silico* inversion genotyping
50 of sequenced mosquitoes, these tags perform better than traditional cytogenetics, even

51 for specimens where only a small subset of the tag SNPs can be successfully
52 ascertained.

53 **Introduction**

54

55 A chromosomal inversion originates when a chromosome segment reverses end to end.

56 Inversions maintained in plant and animal populations as structural polymorphisms tend

57 to be large (several megabases) and contain hundreds of genes (reviewed in

58 Wellenreuther and Bernatchez 2018). Long-term balancing selection can maintain

59 these polymorphisms through millions of generations and multiple species radiations

60 (Wellenreuther and Bernatchez 2018). Because recombination is greatly reduced

61 between opposite orientations in inversion heterozygotes, inversions preserve

62 selectively advantageous combinations of alleles despite homogenizing gene flow in

63 collinear regions. Theory and mounting evidence implicate inversions in local

64 adaptation, adaptive divergence, and range expansion, though the precise molecular

65 mechanisms are rarely known (Hoffmann *et al.* 2004; Kirkpatrick and Barton 2006;

66 Hoffmann and Rieseberg 2008; Schaeffer 2008; Kirkpatrick 2010; Lowry and Willis

67 2010; Joron *et al.* 2011; Jones *et al.* 2012; Kirkpatrick and Barrett 2015; Twyford and

68 Friedman 2015; Kapun *et al.* 2016; Ayala *et al.* 2017; Fuller *et al.* 2017; Wellenreuther

69 *et al.* 2017; Wellenreuther and Bernatchez 2018). Importantly, because of occasional

70 double-crossovers and gene conversion, the suppression of gene flux is not absolute.

71 As long as inversion heterozygotes are formed in populations, any significant

72 association between an inversion and an allele within its boundaries is subject to

73 eventual erosion unless gene flux is countered by selection (Navarro *et al.* 1997;

74 Andolfatto *et al.* 2001).

75 The *Anopheles gambiae* complex is a medically important group of at least eight
76 closely related and morphologically indistinguishable mosquito sibling species from sub-
77 Saharan Africa (White *et al.* 2011; Coetzee *et al.* 2013). Three members of the complex
78 (the eponymous *Anopheles gambiae*, *Anopheles coluzzii*, and *Anopheles arabiensis*)
79 are among the most significant malaria vectors globally, responsible for a majority of the
80 435,000 malaria deaths in 2017 (World Health Organisation 2018). The ecological
81 plasticity of these three species contributes greatly to their status as major human
82 malaria vectors (Coluzzi *et al.* 2002). In contrast to the other five, these three species
83 have wide distributions across diverse biomes of tropical Africa. Not coincidentally, they
84 also segregate strikingly high numbers of paracentric inversion polymorphisms, which
85 are implicated in adaptation to seasonal and spatial environmental heterogeneities
86 related both to climatic variables and anthropogenic alterations of the landscape
87 (Coluzzi *et al.* 1979; Bryan *et al.* 1982; Coluzzi *et al.* 1985; Toure *et al.* 1998; Manoukis
88 *et al.* 2008; Costantini *et al.* 2009; Simard *et al.* 2009; Cheng *et al.* 2012; Ayala *et al.*
89 2014; Caputo *et al.* 2014; Ayala *et al.* 2017; Cheng *et al.* 2018). Some of these
90 inversions also have been associated with ecologically relevant phenotypes, including
91 desiccation and thermal tolerance (Gray *et al.* 2009; Rocca *et al.* 2009; Cassone *et al.*
92 2011; Fouet *et al.* 2012; Ayala *et al.* 2018; Cheng *et al.* 2018).

93 The sister taxa *An. gambiae* and *An. coluzzii*, the focus of the present
94 investigation, are the most closely related species in the *An. gambiae* complex, sharing
95 extensive nucleotide variation through both recent common ancestry and introgression
96 (Fontaine *et al.* 2015; Hanemaaijer *et al.* 2018), while maintaining characteristic
97 differences in ecology and behavior (Costantini *et al.* 2009; Diabate *et al.* 2009; Simard

98 *et al.* 2009; Gimonneau *et al.* 2010; Gimonneau *et al.* 2012a; Gimonneau *et al.* 2012b;
99 Dabire *et al.* 2013; Tene Fossog *et al.* 2015; Ayala *et al.* 2017). They also share four of
100 six common chromosomal inversion polymorphisms on chromosomal arm 2R (*b*, *c*, *d*, *u*)
101 and the only inversion polymorphism on chromosomal arm 2L (*a*) (Figure 1) (della Torre
102 *et al.* 2005). These inversions range in size from ~4Mb to 22Mb, and together span
103 thousands of genes and a sizeable fraction of chromosome 2: ~61% of 2R and ~38% of
104 2L polytene (euchromatic) content (Coluzzi *et al.* 2002). Inversions 2La and 2Rb are
105 found in populations throughout tropical Africa and are therefore cosmopolitan, while
106 three other inversions on 2R (*c*, *d*, and *u*) are widespread in West, very rare in Central
107 Africa, and absent from East Africa. The remaining two inversions, 2Rj and 2Rk, have
108 more restricted geographic distributions (Coluzzi *et al.* 2002; Ayala *et al.* 2017).

109 Early cytogenetic studies of *An. gambiae* and *An. coluzzii*, presumed at the time
110 to be a single heterogenous species, uncovered genetic discontinuities that led to the
111 designation of five presumed assortatively-mating ‘chromosomal forms’: FOREST,
112 SAVANNA, MOPTI, BAMAKO, and BISSAU (Coluzzi *et al.* 1985; Toure *et al.* 1998;
113 Coluzzi *et al.* 2002; della Torre *et al.* 2005). They were delineated based on stable non-
114 random associations of different sets of chromosome 2R inversions in co-occurring
115 populations, and differed in larval ecology. Subsequent DNA-based studies identified
116 fixed differences in the ribosomal DNA (rDNA), located in the pericentromeric region of
117 the X chromosome, leading to the definition of two assortatively mating M and S
118 ‘molecular forms’ of *An. gambiae* (della Torre *et al.* 2001). The molecular forms, which
119 were eventually given specific status as *An. coluzzii* (formerly M) and *An. gambiae*
120 *sensu stricto* (formerly S) (Coetzee *et al.* 2013), are incongruent with the chromosomal

121 forms. Nearly all inversion associations segregate in both species albeit at different
122 frequencies, and likely play similar roles in ecological specialization and adaptation in
123 both *An. gambiae* s.s. (hereafter, *An. gambiae*) and *An. coluzzii* (della Torre *et al.* 2005;
124 Costantini *et al.* 2009; Simard *et al.* 2009; Ayala *et al.* 2017). Hence, inversion
125 associations are indicative of environmental heterogeneities more so than intrinsic
126 reproductive boundaries.

127 Beyond a role in ecological specialization, inversions in the *An. gambiae* complex
128 are also associated with vector traits affecting malaria transmission intensity and
129 control: biting and resting behavior (Coluzzi *et al.* 1979; Riehle *et al.* 2017), seasonality
130 (Rishikesh *et al.* 1985), morphometric variation (Petrarca *et al.* 1990), and *Plasmodium*
131 infection rates (Petrarca and Beier 1992; Riehle *et al.* 2017). Although a robust
132 molecular assay is available for genotyping inversion 2La in natural populations (White
133 *et al.* 2007), 2R inversions with characterized breakpoint sequences (*j*, *b*, *c*, and *u*)
134 (Coulibaly *et al.* 2007a; Sangare 2007; Lobo *et al.* 2010) proved difficult to genotype
135 molecularly at the breakpoints (Coulibaly *et al.* 2007b; Lobo *et al.* 2010), owing to
136 extensive tracts of flanking repetitive DNA. The 2R*k* breakpoints have yet to be
137 characterized, but recent localization of the 2R*d* breakpoints in the reference genome
138 assembly using proximity-ligation sequencing (Corbett-Detig *et al.* 2019) also revealed
139 high repeat content, suggesting that repetitive DNA at inversion breakpoints will pose a
140 significant challenge both for breakpoint characterization and for molecular genotyping
141 assays targeting breakpoint regions in these species.

142 Failure to account for the presence of inversions is a barrier to a more
143 comprehensive understanding of epidemiologically relevant mosquito behavior and

144 physiology. Inversion-blind analysis of population data can mislead population genetic
145 inference, and create spurious associations in genome-wide association studies (Seich
146 Al Basatena *et al.* 2013; Houle and Marquez 2015). Powerful genomic resources exist
147 for *An. gambiae*, including a high-quality reference genome assembly (Holt *et al.* 2002)
148 and a database of genomic variation (Ag1000G) based on deep genome re-sequencing
149 of thousands of mosquitoes from natural populations across Africa (Miles *et al.* 2017).
150 Unfortunately, inversion genotypes are not automatically revealed by genome re-
151 sequencing, as reads are mapped to their position in the reference genome assembly,
152 not their position in the re-sequenced mosquito genome. Despite advancing genome
153 technology, the only method currently available to determine the *An. gambiae* karyotype
154 is a method perfected half a century ago (Coluzzi 1968) involving cytological analysis of
155 ovarian nurse cell polytene chromosomes (Coluzzi *et al.* 2002; Pombi *et al.* 2008). At
156 best, such cytological analysis is severely rate-limiting because it is laborious and
157 requires highly specialized training. At worst, it is prohibitive because it requires proper
158 preservation of chromosomes harvested only from ovaries of adult females at a specific
159 gonotrophic stage; suitable polytenization is absent at other gonotrophic stages as well
160 as in males (della Torre 1997). While salivary glands of late fourth instar larvae also
161 contain chromosomes with an adequate degree of polytenization, and the banding
162 patterns of salivary and ovarian chromosomes are homologous in principle, most bands
163 are difficult to homologize due to a different pattern of chromosome ‘puffing’ (della Torre
164 1997), rendering this alternative impractical. To overcome these impediments, our goal
165 is to develop broadly accessible computational and molecular methods of genotyping
166 chromosomal inversions in individual specimens of *An. gambiae* and *An. coluzzii*.

167 Here, we exploit the Ag1000G database and leverage the subset of cytologically
168 karyotyped specimens within that database to develop a computational approach for
169 karyotyping applicable to whole genome sequence data. We identify multiple tag single
170 nucleotide polymorphisms (SNPs) significantly associated with inversions across
171 geography that collectively predict with high confidence the genotypes of six common
172 polymorphic inversions on chromosome 2 (*a, j, b, c, d, u*) in individually sequenced
173 genomes of *An. coluzzii* and *An. gambiae*. We then apply this approach to data
174 generated independently of Ag1000G to show that our approach has wider utility, even
175 for specimens where only a small subset of the tag SNPs can be successfully
176 ascertained.
177

178

179 **Methods**

180

181 *Mosquito genotype data*

182 Variant call data used for the discovery of inversion tag SNPs were accessed from
183 Ag1000G (Miles *et al.* 2017) and Vector Observatory (VOBS; Table S1), projects of the
184 Malaria Genomic Epidemiology Network (MalariaGEN; <https://www.malariagen.net/>)
185 that provide catalogs of genomic sequence variation based on individual wild-collected
186 *An. gambiae* and *An. coluzzii* mosquitoes sampled from multiple African countries and
187 the Mayotte archipelago. With the exception of four atypical samples (see next section),
188 we verified species identifications as reported in Ag1000G and VOBS using principal
189 component analysis (PCA) of biallelic SNPs on the X chromosome. We excluded any
190 specimens with more than 50,000 missing genotypes on chromosomal arm 2R (N=9),
191 and any specimens subjected to whole genome amplification (WGA) prior to genomic
192 sequencing (N=44), as PCA revealed strong biases associated with WGA. After
193 filtering, we retained variant call data from 1,347 mosquitoes (Table S2).

194

195 *Karyotype imputation by local PCA*

196 Cytological karyotype information derived from phase contrast microscopy of ovarian
197 polytene chromosomes (della Torre 1997) was available only for a relatively small
198 subset of specimens (N=373) in Ag1000G/VOBS (hereafter, Ag1000G for brevity).
199 Thus, as a first step toward discovering SNPs putatively predictive of inversion status
200 (tag SNPs), we imputed karyotypes computationally at each of six focal inversions

201 (Figure 1), using local PCA (where ‘local’ refers to windows of the genome
202 corresponding to chromosomal rearrangements). Ma and Amos (2012) showed that
203 applying PCA to SNP genotypes in a window of the genome containing an inversion
204 polymorphic in population genomic data (an approach that we call ‘PCA karyotyping’)
205 produces a pattern of three equidistant clusters (stripes) in a plot of the first two
206 principal components, assuming adequate numbers of each of three possible inversion
207 genotypes: inverted and uninverted (standard) homokaryotypes, and heterokaryotypes.
208 The two flanking stripes represent alternative homokaryotypes, and the middle stripe
209 represents the inversion heterokaryotype, a 1:1 “admixture” between the two
210 homokaryotype classes (Ma and Amos 2012).

211 To apply this approach, we combined specimens from both species (*An.*
212 *gambiae* and *An. coluzzii*) and different geographic localities into a single
213 metapopulation sample of 1,347 mosquitoes (Tables S1, S2). We identified a set of
214 biallelic SNPs within inversion boundaries (Table S3) with potentially informative levels
215 of polymorphism [minor allele count ≥ 3 and minimum alternate allele frequency (MAF)
216 ≥ 0.15 for all inversions except 2Rd, for which the MAF threshold was reduced to 0.03].
217 As 2Rd overlaps 2Ru in the genome (Figure 1), we limited consideration to only those
218 SNPs found outside (proximal to) 2Ru for PCA karyotyping of 2Rd (Table S3). Next, we
219 converted mosquito genotypes at these SNPs to a count of the number of alternate
220 alleles (‘0’ if both matched the reference allele, ‘1’ or ‘2’ if one or both matched the
221 alternate allele, respectively). Using the scikit-allel Python package v1.1.9 (Miles and
222 Harding 2017), we then applied PCA to the resulting matrix of alternate allele counts,
223 and represented the output as a scatter plot of the first two principal components for

224 each mosquito in the population sample. The correct genotype corresponding to the
225 two homokaryotype stripes was determined based on the inclusion in a given stripe of
226 mosquitoes with cytologically determined karyotype. Based on this classification,
227 mosquitoes without cytologically determined karyotypes were assigned a PCA
228 karyotype.

229 The distinction between stripes was not always sharp; the stripes could be
230 diffuse and oblique rather than tightly clustered. In extreme cases, stripes were not
231 initially discernable. Through an iterative process of 'leave one population sample out'
232 followed by PCA, we determined that absence of a clear three-stripe pattern was
233 attributable to some or all of the same four atypical source populations, in particular,
234 those from Kenya, Mayotte, The Gambia, and Guinea Bissau. The Kenyan sample has
235 been found to display signs of extreme inbreeding (Miles *et al.* 2017), and Mayotte is an
236 island whose mosquito population is plausibly subject both to inbreeding and a degree
237 of isolation from mainland samples. The Gambia and Guinea Bissau are localities with
238 unusually high degrees of hybridization and introgression between *An. gambiae* and *An.*
239 *coluzzii* (Caputo *et al.* 2008; Oliveira *et al.* 2008; Caputo *et al.* 2011; Marsden *et al.*
240 2011; Weetman *et al.* 2012; Nwakanma *et al.* 2013). Where necessary, we removed
241 these population samples, as well as two *An. gambiae*-*An. coluzzii* hybrid specimens
242 from Burkina Faso and Guinea Conakry, and repeated the PCA. In addition, successful
243 PCA karyotyping of 2Rd and 2Rj required the removal of all *An. coluzzii* specimens
244 owing to taxonomic structuring of variation. Accordingly, PCA karyotyping was
245 successful on all (2La) or subsets (all 2R inversions) of the 1,347 specimens (Table S4).
246

247 *Discovery of SNPs predictive of inversion orientation*

248 The PEST reference genome assembly for *An. gambiae* (AgamP4; Giraldo-Calderon *et*
249 *al.* 2015) was derived from a colony whose karyotype was homozygous standard with
250 respect to all common chromosomal inversions in this species. We therefore had the
251 general expectation that an individual SNP might be a good predictor of chromosomal
252 inversion orientation if the reference allele is strongly associated with the standard
253 arrangement and the alternate allele is strongly associated with the inverted
254 arrangement within and across population samples. As shown in Figure 2 in overview,
255 we assessed SNP genotype-inversion genotype concordance for each inversion in
256 individual mosquitoes, limiting our assessment to potentially more informative, higher
257 frequency biallelic SNPs inside inversion boundaries (*i.e.*, those with $MAF \geq 5\%$). We
258 converted both the SNP genotype and the corresponding mosquito's PCA-based
259 inversion genotype to single numbers, representing the count of alternate alleles (0, 1,
260 or 2) in the case of SNP genotype, and the count of inverted chromosomes (0, 1, or 2)
261 in the case of inversion genotype. Successfully performing tags are expected to have a
262 SNP genotype that correlates strongly with the PCA-based inversion genotype.

263 More formally, we sought to identify candidate tag SNPs using the procedure
264 illustrated in Figure 3 (applied separately for each inversion). Specimens assigned a
265 PCA-based karyotype for a focal inversion were divided into a training sample used for
266 tag SNP discovery (75%) and a validation sample that was held in reserve until a later
267 time (25%), using the `model_selection` module of the scikit-learn Python package
268 (v0.19.2) (Pedregosa *et al.* 2011). We ensured that both partitions were balanced with
269 regard to inversion genotypes but randomized in all other respects. For robust

270 identification of candidate tag SNPs within the training sample, we masked all SNP
271 genotypes inside the inversion boundaries with a genotype quality (GQ) below 20.

272 Next, we created ten bootstrap replicates of the training sample (Figure 3). Each
273 of the ten replicates consisted of sub-samples of 75% of the full training sample, chosen
274 at random with respect to all variables except inversion genotype balance. For each
275 bootstrap replicate at each interrogated SNP (biallelic, $MAF \geq 5\%$), we calculated the
276 SNP genotype-inversion genotype concordance for each mosquito in the sample, as
277 described above (Figure 2). Genotypic concordance at each SNP interrogated in a
278 given bootstrap replicate was expressed as the percentage of mosquitoes for which the
279 number of alternate SNP alleles matched the number of inverted chromosomes.

280 Because an imbalance among inversion genotypes could lead to false-positive tag
281 SNPs, we calculated concordance separately for the three inversion genotypes in each
282 of the ten bootstrap replicates. We then averaged the concordance scores across the
283 ten replicates, by inversion genotype. To generate a single, conservative tag SNP
284 concordance statistic, we used the minimum of the three mean values. Note that
285 because the mosquito composition differed among bootstrap replicates, some SNPs
286 were not evaluated in all ten, if they did not pass our filters in one or more iterations.
287 Finally, to eliminate SNP positions with high levels of missing genotypes, we also
288 calculated for each inversion genotype in each bootstrap replicate the percentage of
289 mosquitoes with SNP genotype calls at the candidate tag (the 'call rate'), and averaged
290 across the ten replicates.

291 The procedure just described returned from 99 to 349 candidate tag SNPs for
292 five inversions, but only two for 2Rc (Table 1). We therefore adopted a modified

293 approach to control for suspected population structure. One possible source of
294 structure was the haplotype configuration of 2Rc with respect to the flanking inversions
295 (2Rb and 2Ru) (Figure 1). The inverted orientation of 2Rc is in almost perfect linkage
296 disequilibrium with the inverted orientation of either 2Rb (as haplotype '2Rbc') or 2Ru
297 (as haplotype '2Rcu'). In a ~50-year cytogenetic database compiled from samples
298 collected in many parts of sub-Saharan Africa (described in Pombi *et al.* 2008), only four
299 specimens were ever recorded as carrying the inverted orientation of 2Rc
300 unaccompanied by either 2Rb or 2Ru (V. Petrarca, unpublished data). A second
301 source, not mutually exclusive, was population structure between *An. coluzzii*, *An.*
302 *gambiae*, and the BAMAKO chromosomal form that is subsumed taxonomically within
303 *An. gambiae* but is at least partially reproductively isolated and genetically differentiated
304 (Manoukis *et al.* 2008; Love *et al.* 2016). Although 2Rc occurs in all three taxa, there is
305 a strong karyotype imbalance among them in natural populations and in Ag1000G. For
306 example, of 70 *An. coluzzii* with 2Rc in Ag1000G, at least 49 (70%) carried the 2Rbc
307 haplotype (haplotypes of the other specimens could not be inferred unambiguously).
308 Similarly, of 64 non-BAMAKO *An. gambiae* with 2Rc, 62 (97%) carried the 2Rbc
309 haplotype. On the other hand, all 45 BAMAKO, by definition, carried 2Rcu. We initially
310 partitioned our sample by species, but the inclusion of BAMAKO in the *An. gambiae*
311 partition resulted in very few candidate tags concordant with inversion genotype (N=17).
312 Ultimately, we retained two data partitions (*An. coluzzii* and non-BAMAKO *An.*
313 *gambiae*), eliminating a third BAMAKO partition due to the fixation of 2Rc in this taxon
314 (Coluzzi *et al.* 1985). From the non-BAMAKO *An. gambiae* partition (hereafter, *An.*
315 *gambiae* for brevity), we omitted two of only three specimens carrying 2Ru (AZ0267-C

316 from Mali and AV0043 C from Guinea), guided by PCA. As described above, both data
317 partitions were split into training (75%) and validation (25%) sets, and ten bootstrap
318 replicates of each training set were analyzed.

319 Ultimately, the candidate tag SNPs chosen (Table 1) met the following three
320 criteria: they were (i) analyzed in at least eight of the ten bootstrap replicates; (ii) called
321 at a rate greater than 90% within each karyotype class; and (iii) concordant with
322 karyotype more than 80% of the time within each karyotype class (99.5% for 2La).
323 Their approximate physical position relative to the span of each inversion is illustrated in
324 Figure S1.

325

326 *Validation of candidate tag SNPs in Ag1000G*

327 We interrogated the candidate tag SNPs in the validation samples from Ag1000G that
328 had been held aside during the discovery phase (Figure 3). For each mosquito in the
329 validation set, we masked genotypes inside the focal inversion with GQ scores less than
330 20. Next, among the retained SNPs, we identified those corresponding to candidate
331 tags and converted their diploid genotypes to a count of the number of alternate alleles.
332 Finally, the number of alternate alleles at each tag SNP was summed across tags and
333 averaged to provide an overall computational karyotype score. We compared this mean
334 score to the PCA-based karyotype.

335

336 *Testing tag SNPs in data independent of the Ag1000G pipeline*

337 We also explored the efficacy of our tag SNPs for computational karyotyping in wild-
338 caught mosquitoes subject to whole genome sequencing and variant calling by

339 individual investigators, for which corresponding cytological karyotypes had been
340 determined through phase microscopy (Figure 3). We used specimens originating from
341 southern Mali, 8 *An. gambiae* BAMAKO chromosomal form (Fontaine *et al.* 2015; Love
342 *et al.* 2016) and 17 *An. coluzzii* (Main *et al.* 2015), whose variant calls and cytogenetic
343 metadata are publicly accessible (Table S5). These data include specimens sequenced
344 to much lower coverage than the standard adhered to by Ag1000G. We followed the
345 same procedure described for the Ag1000G validation set to computationally karyotype
346 these specimens, and compared their computational and cytologically determined
347 karyotypes.

348

349 *Genetic distance trees to assess inversion history*

350 We compared patterns of relatedness near the breakpoints of all six inversions using
351 unrooted neighbor-joining (NJ) trees. For each inversion, we used biallelic SNPs with a
352 MAF of 0.01 found within 5 kb upstream and downstream of the distal and proximal
353 breakpoints (15 kb for 2Rd). Total numbers of SNPs for each inversion were: 2La, 596;
354 2Rj, 909; 2Rb, 428; 2Rc, 2141; 2Rd, 955; 2Ru, 1110. Using the python package
355 *anhima*, we converted the number of alternate alleles at these SNPs into a Euclidean
356 distance matrix, and then constructed neighbor-joining trees using all 1,347 specimens.
357 To assess support for the nodes of the 2Rc tree, we used the transfer bootstrap
358 estimate (TBE; Lemoine *et al.* 2018), a statistic that measures the number of taxa that
359 must be transferred to make a given branch of a reference tree match the closest
360 equivalent branch in a bootstrap tree. To calculate this statistic, we imported the matrix
361 of alternate allele counts into R (v. 3.5.1, “Feather Spray”; R Core Team 2018) and used

362 the `dist()` function of base R to construct the Euclidean distance matrix. We then used
363 the `nj()` function in the `ape` package (v. 5.2) to construct the neighbor joining tree, and
364 the `boot.phylo()` function to generate 1,000 bootstrap trees. We used these trees as
365 input to `booster` (Lemoine *et al.* 2018), which calculates the TBE for each node.

366

367 *Code and data availability*

368 All genomic sequence data and variant call files used in this study are located in open
369 data repositories as specified in Tables S1 and S2. The *An. gambiae* AgamP4
370 reference assembly is available through VectorBase (<https://www.vectorbase.org/>). All
371 custom code necessary to reproduce this analysis can be found at

372 https://github.com/rrlove/comp_karyo_notebooks and <https://github.com/rrlove/ingenos>.

373 The complete set of tag SNPs, together with a custom script for computational
374 karyotyping, which calculates the mean inversion genotype across the relevant tag
375 SNPs, can be found at <https://github.com/rrlove/compkaryo>.

376

377 **Results**

378

379 After filtering, we retained the genotype data from 1,347 individually sequenced *An.*
380 *coluzzii* and *An. gambiae* mosquitoes from the Ag1000G repository of natural genomic
381 sequence variation, representing population samples from 13 West, Central, and East
382 African countries and the island of Mayotte (Tables S1, S2).

383

384 *Patterns of genetic variation at inversion breakpoints*

385 To gain insight into the relative roles of inversion history, taxonomic status, and
386 geographic location in structuring genetic variation for each inversion, we reconstructed
387 neighbor-joining trees based on SNPs in the immediate vicinity of the breakpoints
388 (Figure 4). The resulting dendrograms, color-coded by inversion genotype, taxon and
389 African country, indicate little clustering on the basis of geographic location; outlier
390 population samples are those with a history of inbreeding or hybridization (see
391 Methods). On the other hand, with the notable exception of 2La, taxonomic status is an
392 important factor structuring inversion variation between *An. gambiae* and *An. coluzzii*.
393 Moreover, BAMAKO specimens appear to comprise a differentiated outlier clade within
394 the larger *An. gambiae* cluster. It is interesting to note that for inversion 2Rc, taxonomic
395 status appears to be a more decisive factor than inversion genotype. All three 2Rc
396 inversion genotypes cluster within their respective species (supported by bootstrap at
397 90%, or 98% if dendrograms are constructed after removing outlier samples from The
398 Gambia, Guinea-Bissau and Kenya; not shown). Further investigation is required to

399 determine whether this pattern results from a monophyletic inversion that subsequently
400 differentiated between taxa, or from independent inversion events in the two taxa.

401

402 *Inversion karyotype imputation by PCA*

403 Only 373 of the 1,347 mosquitoes were associated with metadata that included
404 cytologically determined inversion karyotypes. As discovery of candidate tag SNPs
405 requires provisional inversion genotype assignments, we applied local PCA to assign
406 genotypes for individual inversions on chromosome 2, following Ma and Amos (2012).
407 A recognized limitation to this population-level approach, beyond the fact that it cannot
408 be applied to individual mosquitoes, is that its success depends upon the presence of
409 all three inversion genotypes in the sample under study. For this reason, and with the
410 goal of finding the most flexible solution to inversion genotyping across geography and
411 taxa, we began with PCA based on the entire set of 1,347 mosquitoes, under the
412 simplifying assumption that the expected ‘three-stripe’ signal on a PCA plot would not
413 be overwhelmed by geographic or population structure. Only in the case of 2La could
414 genotype assignments be confidently inferred from the combination of all 1,347
415 specimens. For inversions on 2R, from one to four admixed (*An. gambiae*-*An. coluzzii*)
416 or geographic outlier populations (highly inbred or island samples) had to be removed
417 from analysis before a three-genotype pattern could be discerned on the PCA plot
418 (Tables S2, S4; see Methods). Additionally, for 2Rd and 2Rj, *An. gambiae*-*An. coluzzii*
419 population structure dominated the PCA. Taken together with the fact that 2Rj has yet
420 to be found in *An. coluzzii* (Coluzzi *et al.* 2002; della Torre *et al.* 2005), we removed all
421 341 *An. coluzzii* specimens (Tables S2, S4) prior to PCA karyotyping of 2Rd and 2Rj in

422 *An. gambiae*. Ultimately, PCA karyotypes were imputed for 780-1,347 mosquitoes,
423 depending upon the inversion (Table S4).

424

425 *Tag SNP ascertainment and validation in Ag1000G*

426 Dividing the Ag1000G samples into training (75%) and validation (25%) sets for each
427 inversion, and working within the training sets using a bootstrapping procedure, we
428 screened for candidate tag SNPs in the five 2R inversions and 2La (see Methods for
429 details). Candidate tag SNPs were those whose genotypes were concordant with the
430 corresponding PCA genotypes, when averaged across ten bootstrap replicates, for
431 more than 80% of the specimens that were scored (99.5% for 2La). The number of
432 candidate tags ranged from 99 (2Rj) to 349 (2Rb) excluding 2Rc, in which only two
433 candidates were found due to population structure between *An. gambiae* and *An.*
434 *coluzzii* (Figure 4; Table 1). Partitioning the 2Rc sample by taxon (and omitting
435 BAMAKO; see Methods) resulted in 49 and 57 tags for *An. gambiae* and *An. coluzzii*,
436 respectively (Table 1). Notably, there was no overlap between the two sets of tags.

437 To assess the performance of these candidate tags, we used them to predict
438 karyotypes in the held-out validation sets of Ag1000G specimens. For each inversion
439 and specimen, we calculated a computational karyotype score representing the average
440 genotype inferred across all candidate tag SNPs ascertained (see Methods).

441 Histograms of resulting computational karyotype scores generally showed tight
442 clustering around the three theoretical genotypic optima (0, 1, 2), reflecting close
443 agreement among specimens (Figure 5). For each specimen in a validation set, we
444 then compared the computational karyotype score to its PCA karyotype, and tallied the

445 number of disagreements (Table 2). All except one specimen had matching PCA and
446 computational karyotype scores. This exception, one of 254 (0.4%) assignments for
447 2Rc in *An. gambiae*, involved a specimen carrying 2Ru (AZ0267-C) already noted as an
448 outlier (see Methods).

449

450 *Performance of tag SNPs in resequencing data independent of Ag1000G*

451 Previous studies re-sequenced *An. gambiae* or *An. coluzzii* mosquitoes from Mali
452 whose karyotypes had been determined from the polytene chromosome banding
453 pattern (Main *et al.* 2015; Love *et al.* 2016). Although sample size is limited, these data
454 allow a direct comparison of cytogenetic and *in silico* karyotyping under less ideal
455 conditions—lower sequencing depth, with variant calls made independently of the
456 Ag1000G pipeline. For each specimen and inversion, we calculated computational
457 karyotype scores (averaged across all tag SNPs that could be ascertained in a
458 specimen) (Tables S5, S6). Histograms of these scores by inversion, similar to those
459 based on Ag1000G validation sets, reveal clustering of scores around the three
460 genotypic optima provided that taxon-specific tags (2Rc-*coluzzii* and 2Rc-*gambiae*) are
461 applied to the conspecific taxon, and heterospecific applications (including use of 2Rc-
462 *gambiae* tags to genotype BAMAKO) are avoided (Figure 6, Figure S2).

463 In the BAMAKO sample of Love *et al.* (2016) where mean sequencing depth
464 ranged from 9-10x, there was concordance in karyotype assignments between
465 cytogenetic and computational methods for five inversions including 2La, even though
466 only 10-12 2La tags were ascertained (Table S5, S6). However, as expected for

467 BAMAko, the *An. gambiae* 2Rc tags failed. Due to the extreme geospatial restriction of
468 BAMAko, this specific problem is limited in scope.

469 In the *An. coluzzii* sample of Main *et al.* (2015), mean sequencing coverage
470 varied widely (4-66x; Tables S5, S6). The impact of very low sequencing coverage on
471 the success of computational karyotyping is illustrated by specimens 04SEL021 and
472 04SEL02 (4x and 5x, respectively). For 04SEL021, there is no apparent disagreement
473 between the cytogenetic and mean computational genotype scores for any of the six
474 inversions. Nevertheless, for those inversions classified as heterozygotes both
475 cytogenetically and computationally (2La, 2Rb, 2Rc), the proportion of tags whose
476 genotype matches the mean computational score drops drastically to ~30% (Table S5),
477 likely because true heterozygous sites are often scored as homozygous either for the
478 reference or alternate allele (0 or 2) due to low sequencing coverage. (Indeed, using
479 chromosome 3L, we confirmed the expected drop in the rate of heterozygosity with
480 decreasing coverage in these 17 specimens; data not shown). Low coverage alone is
481 less likely to bias computational scores toward zero or two. For 04SEL02 (5x
482 coverage), where cytogenetic versus computational discrepancies occur at 2Rb and
483 2Ru, the computational karyotype is supported respectively by 81% of 208 tags and
484 >94% of 57 tags, favoring the computational genotype by weight of evidence. The
485 remaining six specimens with discordant inversion genotypes were sequenced to at
486 least 10x coverage. In these cases, when the computational genotype score signaled
487 '1' in contradiction to a homokaryotypic cytogenetic genotype (02SEL85, 02SEL006,
488 02SEL009, 01Osel134), the proportion of tags supporting the computational genotype
489 ranged from 65% to >92%. For other types of genotypic disagreements between

490 methods, the computational score was supported by >80% of tags scored. Overall,
491 these results suggest that computational karyotyping using tag SNPs can be successful
492 in data derived independently of Ag1000G (Tables S5, S6), though care should be
493 taken when this approach is applied to very low coverage samples.

494

495 *Performance of tag SNPs against cytogenetically karyotyped Ag1000G specimens*

496 We compared the cytogenetic karyotype assignments for 373 specimens in Ag1000G to
497 their corresponding computational karyotype assignments (Table 3). Conflicts were few
498 overall, and for every inversion, all but one conflict (involving specimen AZ0267-C, the
499 exceptional *An. gambiae* carrier of the 2Ru inversion) could be attributed to errors in the
500 cytogenetically assigned scores, as genotypes imputed from both PCA and tag SNPs
501 contradict the cytogenetic assignment. Visual reference back to the PCA plots clearly
502 confirmed that for specimens whose cytogenetic and tag SNP assignments differed and
503 for whom PCA karyotypes could be determined, their locations on the plot strongly
504 agreed with the tag SNP genotype (Figure S3). Considering that we ascertained tens or
505 hundreds of tags per specimen, and that the proportion of tags whose SNP genotype
506 matched the computational score was greater than 83% in all except the unusual
507 specimen AZ0267-C (Table 3), the computational scores more confidently predict the
508 true inversion genotype than traditional cytogenetics for these five inversions. The most
509 dramatic example is with respect to inversion 2Ru, where we noted an unusually large
510 number of erroneous cytogenetic genotypes of '1' (N=18/29) conflicting with both PCA
511 and computational assignments of '0'. It is not immediately clear what could lead to
512 such an elevated rate of cytogenetic error (which otherwise is ~4%), but it is possible

513 that the 2Ru heterozygous loop was mistaken either for a loop created by a rare
514 inversion (sensu Pombi *et al.* 2008) in the same chromosomal region, or for a 2Rd loop
515 in samples from regions where 2Ru is rare (as supported by the fact all 11 cytogenetic
516 errors in *An. gambiae* were found in samples from the same small region in Cameroon,
517 six of which were scored computationally as '1' for 2Rd).

518 Our results also highlight the pitfalls of using taxon-specific tags to genotype
519 other taxa, or populations with high levels of admixture between taxa (Table S7). As
520 expected, we find elevated numbers of cytogenetic-computational disagreements when
521 (i) 2Rc-*gambiae* tags are applied to BAMAKO (60% of the 45 specimens), (ii) 2Rd-
522 *gambiae* tags are used to genotype *An. coluzzii*, and (iii) 2Rd-*gambiae* tags are applied
523 to admixed *An. gambiae*-*An. coluzzii* populations such as those from Guinea Bissau.
524 These disagreements involve specimens carrying inverted arrangements according to
525 cytogenetic analysis which are not tagged as inverted computationally, due to the lack
526 of correlation between tags and the inverted orientation in the heterospecific genetic
527 background.

528 **Discussion**

529 Analysis of the Ag1000G database allowed us to develop the first standardized
530 computational karyotyping of the six main polymorphic chromosomal inversions in the
531 major malaria vectors *An. coluzzii* and *An. gambiae*, despite the fact that only a small
532 subset of specimens in the database had cytogenetic karyotype assignments (Figure 7).
533 Direct comparison of computational karyotype scores with the cytogenetic assignment
534 for the same specimen in Ag1000G suggests that computational karyotyping
535 outperforms traditional cytogenetics in terms of accuracy, given that assignments are
536 based on tens or hundreds of individual tags. Preliminary testing on specimens
537 sequenced and computationally processed by individual laboratories outside of
538 Ag1000G standards suggests that our tag SNPs have the potential to perform well,
539 even on specimens sequenced to much lower depth. Our approach not only has a
540 lower error rate compared to classical cytogenetics, but also is more widely applicable
541 (regardless of mosquito gender, physiological status, or method of preservation), more
542 widely accessible to those without specialized expertise, higher throughput, and
543 therefore, ultimately cheaper to implement at scale. This method can now be used to
544 predict inversion genotypes in previously sequenced data sets for which ecological and
545 behavioral data may already be available. Even more important, easy large-scale
546 adoption of this approach allows for new and properly powered association studies to
547 be performed on ecologically and epidemiologically relevant mosquito phenotypes,
548 studies that that would have been prohibitively ambitious when relying on cytogenetic
549 karyotyping. In addition, this method can now facilitate sequencing experiments for
550 which inversion karyotype is relevant at scales. Expanding the possibilities further,

551 molecular assays based on these results that will allow inversion genotyping without
552 whole genome sequencing are under development.

553 However, some important limitations exist. Computational karyotyping is strictly
554 dependent upon tag SNPs that are strongly correlated with inversion status, a
555 contingency that depends upon representative sampling. Although Ag1000G is
556 populated by samples derived from multiple countries in West, Central and East Africa,
557 *An. coluzzii* is underrepresented, as is southern Africa (Miles *et al.* 2017). Even more
558 importantly, with the exception of the cosmopolitan inversions 2La and 2Rb, the inverted
559 orientation of other rearrangements (2Rj, 2Rc, 2Rd, and 2Ru) is underrepresented in
560 the Ag1000G data that was available at the time of our analysis. It is clear that
561 population structure is an especially important factor in the application and further
562 development of tags for 2Rc and 2Rd. The current taxon-specific tags should not be
563 used to genotype heterospecific specimens (including BAMAKO) or samples from areas
564 where high rates of interspecific hybridization are known. The presence of strong
565 population structure means that correlations between tags and the inverted orientation
566 characteristic of the target taxon cannot be assumed in a different taxon. The absence
567 of correlation should downwardly bias the computational score, resulting in false
568 negatives when genotyping true inverted homozygotes and heterozygotes. Finally, our
569 inversion breakpoint dendrograms raise the possibility that at least one cytologically-
570 recognized inversion, 2Rc, may have arisen repeatedly at the molecular level, a result
571 that requires further investigation beyond the scope of this study. With the exception of
572 2Rc, 2Rd, and 2Rj, for which we developed taxon-specific tags, our approach implicitly
573 assumed that inversions shared by *An. gambiae* and *An. coluzzii* are monophyletic, and

574 may yield unexpected results if this assumption is violated. Accordingly, these tools
575 should be applied with caution, and there is ample room for improvement as more data
576 become available. Fortunately, our standardized approach makes it easy to
577 accommodate improvements. The success of our method thus far suggests that the
578 general approach may be suitable for studying inversions more broadly, in additional
579 malaria vectors as well as other systems where inversions are implicated in local
580 adaptation.

581 Nearly twenty years ago, Coluzzi and colleagues predicted that the then-newly-
582 assembled *An. gambiae* reference genome would facilitate our analyses of polymorphic
583 chromosomal inversions in the *An. gambiae* complex (Coluzzi *et al.* 2002). Our work
584 continues the realization of that prediction by providing, for the first time, cross-continent
585 diagnostics for multiple inversions. These computational diagnostics, and the molecular
586 diagnostics that they leverage, take us one step closer to understanding the contribution
587 of chromosomal inversions to the deadly facility of *An. gambiae* and *An. coluzzii* for
588 vectoring malaria.

589

590 **Acknowledgements**

591 We thank the Notre Dame Center for Research Computing for technical support, and C.
592 Liu, C. Sweet, and J. Young for helpful discussions. We thank M. Kern and R.
593 Montanez-Gonzalez for assistance with DNA extraction. This work was supported by
594 the National Institutes of Health (R01 AI125360 awarded to NJB). During this work,
595 NJB was supported by Target Malaria, which receives core funding from the Bill &
596 Melinda Gates Foundation and from the Open Philanthropy Project Fund, an advised
597 fund of Silicon Valley Community Foundation.

598

599

600

601 **Figure Legends**

602

603 **Figure 1.** Diagrammatic representation of the common polymorphic inversions (labeled
604 brackets) on chromosome 2 in *An. gambiae*. Polytene chromosome map modified from
605 Figure 1 and poster in Coluzzi *et al.* (2002). CT, centromere.

606

607 **Figure 2.** Assessment of correspondence between SNP and inversion genotype in each
608 mosquito. For each chromosomal rearrangement and mosquito, biallelic SNP
609 genotypes inside rearrangement boundaries were converted to a number representing
610 the count of alternative alleles (relative to the AgamP4 reference). We applied PCA to
611 the resulting matrix to assign each individual mosquito an inversion genotype. The
612 expectation is that the PCA-based genotype, expressed as the number of inverted
613 chromosomes at the focal rearrangement, should match the number of alternative
614 alleles at SNPs predictive of inversion status (tag SNPs).

615

616 **Figure 3.** Overview of experimental design. For each inversion, the appropriate
617 Ag1000G sample of mosquitoes that had been successfully karyotyped by PCA was
618 partitioned into a training set (75%) and a validation set (25%). Ten bootstrap replicates
619 of the training set were created from a random sample of 75% of the full training set.
620 For each bootstrap replicate and each mosquito, higher frequency biallelic SNPs within
621 inversion breakpoints were interrogated for genotypic concordance with the PCA-based
622 genotype. Results were summarized across the ten replicates to create a set of
623 candidate tag SNPs with concordance rates exceeding 80%. Candidate tags were used

624 to genotype the held-out validation set, and the computational karyotype score
625 computed across tags was compared to the PCA-based karyotype. Candidate tags
626 were also used to interrogate mosquitoes sequenced independently of Ag1000G, and
627 the computational karyotype score was compared to the associated cytogenetically
628 determined karyotype.

629
630 **Figure 4.** Neighbor-joining dendrograms reconstructed from 1,347 *An. gambiae* and *An.*
631 *coluzzii* mosquitoes from Ag1000G, using biallelic SNPs within 5 kb of inversion
632 breakpoints (15 kb for 2Rd) having a minimum minor allele frequency of 0.01. Columns
633 represent the same inversion dendrogram, alternately color-coded by inversion
634 genotype as determined from PCA (first row), taxon (second row), or geographic source
635 (third row). Some specimens that could not be karyotyped by PCA for inversions 2Rc,
636 2Rd, 2Rj, and 2Ru had cytogenetically determined karyotypes, which were used in
637 place of PCA for color-coding the inversion genotype. 'None' refers to mosquitoes that
638 were not assigned an inversion genotype either by PCA or cytogenetically; 'Other' refers
639 to mosquitoes that were not identified taxonomically.

640
641 **Figure 5.** Histograms of computational karyotyping scores calculated by interrogating
642 tag SNPs in *An. gambiae* and *An. coluzzii* mosquitoes from the Ag1000G validation
643 sets. Note that these mean scores cluster around 0, 1, and 2.

644
645 **Figure 6.** Histograms of computational karyotyping scores calculated by interrogating
646 tag SNPs in *An. gambiae* and *An. coluzzii* mosquitoes re-sequenced independently of

647 the Ag1000G pipeline, often at lower sequencing depth. Scores cluster near 0, 1, and 2
648 with little dispersion except when taxon-specific tag SNPs are applied to a different
649 taxon (indicated by an asterisk).

650

651 **Figure 7.** Map of the study area with the frequency of *An. gambiae* and *An. coluzzii*
652 inversion genotypes inferred for up to six polymorphic chromosome 2 inversions
653 summarized by country (and the island of Mayotte; see Table S2). Blue connecting
654 lines point to *An. coluzzii* samples, while white connecting lines point to *An. gambiae*
655 and hybrid/outlier populations. Image: Visible Earth, NASA. Produced with cartopy
656 v.0.17.1.

657

658

659 Literature Cited

- 660 Andolfatto, P., F. Depaulis, and A. Navarro, 2001 Inversion polymorphisms and nucleotide
661 variability in *Drosophila*. *Genet. Res.* 77: 1-8.
- 662 Ayala, D., P. Acevedo, M. Pombi, I. Dia, D. Boccolini *et al.*, 2017 Chromosome inversions and
663 ecological plasticity in the main African malaria mosquitoes. *Evolution* 71: 686-701.
- 664 Ayala, D., A. Ullastres, and J. Gonzalez, 2014 Adaptation through chromosomal inversions in
665 *Anopheles*. *Front Genet* 5: 129.
- 666 Ayala, D., S. Zhang, M. Chateau, C. Fouet, I. Morlais *et al.*, 2018 Association mapping
667 desiccation resistance within chromosomal inversions in the African malaria vector
668 *Anopheles gambiae*. *Mol. Ecol.*
- 669 Bryan, J. H., M. A. Di Deco, V. Petrarca, and M. Coluzzi, 1982 Inversion polymorphism and
670 incipient speciation in *Anopheles gambiae s. str.* in The Gambia, West Africa. *Genetica*
671 59: 167-176.
- 672 Caputo, B., D. Nwakanma, F. P. Caputo, M. Jawara, E. C. Oriero *et al.*, 2014 Prominent intra-
673 specific genetic divergence within *Anopheles gambiae* sibling species triggered by habitat
674 discontinuities across a riverine landscape. *Mol. Ecol.* 23: 4574-4589.
- 675 Caputo, B., D. Nwakanma, M. Jawara, M. Adiamoh, I. Dia *et al.*, 2008 *Anopheles gambiae*
676 complex along The Gambia river, with particular reference to the molecular forms of *An.*
677 *gambiae s.s.* *Malar. J.* 7: 182.
- 678 Caputo, B., F. Santolamazza, J. L. Vicente, D. C. Nwakanma, M. Jawara *et al.*, 2011 The "far-
679 west" of *Anopheles gambiae* molecular forms. *PLoS One* 6: e16415.
- 680 Cassone, B. J., M. J. Molloy, C. Cheng, J. C. Tan, M. W. Hahn *et al.*, 2011 Divergent
681 transcriptional response to thermal stress by *Anopheles gambiae* larvae carrying
682 alternative arrangements of inversion 2La. *Mol. Ecol.* 20: 2567-2580.
- 683 Cheng, C., J. C. Tan, M. W. Hahn, and N. J. Besansky, 2018 A systems genetic analysis of
684 inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proc. Natl. Acad.*
685 *Sci. U. S. A.* 115: E7005-E7014.
- 686 Cheng, C., B. J. White, C. Kamdem, K. Mockaitis, C. Costantini *et al.*, 2012 Ecological
687 genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing
688 approach. *Genetics* 190: 1417-1432.
- 689 Coetsee, M., R. H. Hunt, R. Wilkerson, A. della Torre, M. B. Coulibaly *et al.*, 2013 *Anopheles*
690 *coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex.
691 *Zootaxa* 3619: 246-274.
- 692 Coluzzi, M., 1968 Cromosomi politenici delle cellule nutrici ovariche nel complesso gambiae del
693 genere *Anopheles*. *Parassitologia* 10: 179-183.
- 694 Coluzzi, M., V. Petrarca, and M. A. DiDeco, 1985 Chromosomal inversion intergradation and
695 incipient speciation in *Anopheles gambiae*. *Bollettino di Zoologia* 52: 45-63.
- 696 Coluzzi, M., A. Sabatini, A. della Torre, M. A. Di Deco, and V. Petrarca, 2002 A polytene
697 chromosome analysis of the *Anopheles gambiae* species complex. *Science* 298: 1415-
698 1418.
- 699 Coluzzi, M., A. Sabatini, V. Petrarca, and M. A. Di Deco, 1979 Chromosomal differentiation and
700 adaptation to human environments in the *Anopheles gambiae* complex. *Trans. R. Soc.*
701 *Trop. Med. Hyg.* 73: 483-497.
- 702 Corbett-Detig, R., I. Said, M. Calzetta, M. Genetti, J. McBroome *et al.*, 2019 Fine-mapping
703 complex inversion breakpoints and investigating somatic pairing in the *Anopheles*

- 704 *gambiae* species complex using proximity-ligation sequencing. *BioRxiv* doi:
705 <https://doi.org/10.1101/662114>.
- 706 Costantini, C., D. Ayala, W. M. Guelbeogo, M. Pombi, C. Y. Some *et al.*, 2009 Living at the
707 edge: biogeographic patterns of habitat segregation conform to speciation by niche
708 expansion in *Anopheles gambiae*. *BMC Ecol.* 9: 16.
- 709 Coulibaly, M. B., N. F. Lobo, M. C. Fitzpatrick, M. Kern, O. Grushko *et al.*, 2007a Segmental
710 duplication implicated in the genesis of inversion 2Rj of *Anopheles gambiae*. *PLoS One*
711 2: e849.
- 712 Coulibaly, M. B., M. Pombi, B. Caputo, D. Nwakanma, M. Jawara *et al.*, 2007b PCR-based
713 karyotyping of *Anopheles gambiae* inversion 2Rj identifies the BAMAKO chromosomal
714 form. *Malar. J.* 6: 133.
- 715 Dabire, K. R., S. Sawadodgo, A. Diabate, K. H. Toe, P. Kengne *et al.*, 2013 Assortative mating
716 in mixed swarms of the mosquito *Anopheles gambiae* s.s. M and S molecular forms, in
717 Burkina Faso, West Africa. *Med. Vet. Entomol.* 27: 298-312.
- 718 della Torre, A., 1997 Polytene chromosome preparation from anopheline mosquitoes, pp. 329-
719 336. in *Molecular Biology of Disease Vectors: A Methods Manual*, edited by J.M.
720 Crampton, C.B. Beard and C. Louis. Chapman & Hall, London.
- 721 della Torre, A., C. Fanello, M. Akogbeto, J. Dossou-yovo, G. Favia *et al.*, 2001 Molecular
722 evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. *Insect*
723 *Mol. Biol.* 10: 9-18.
- 724 della Torre, A., Z. Tu, and V. Petrarca, 2005 On the distribution and genetic differentiation of
725 *Anopheles gambiae* s.s. molecular forms. *Insect Biochem. Mol. Biol.* 35: 755-769.
- 726 Diabate, A., A. Dao, A. S. Yaro, A. Adamou, R. Gonzalez *et al.*, 2009 Spatial swarm segregation
727 and reproductive isolation between the molecular forms of *Anopheles gambiae*. *Proc.*
728 *Biol. Sci.* 276: 4215-4222.
- 729 Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey *et al.*, 2015 Extensive
730 introgression in a malaria vector species complex revealed by phylogenomics. *Science*
731 347: 1258524.
- 732 Fouet, C., E. Gray, N. J. Besansky, and C. Costantini, 2012 Adaptation to aridity in the malaria
733 mosquito *Anopheles gambiae*: chromosomal inversion polymorphism and body size
734 influence resistance to desiccation. *PLoS One* 7: e34841.
- 735 Fuller, Z. L., G. D. Haynes, S. Richards, and S. W. Schaeffer, 2017 Genomics of natural
736 populations: Evolutionary forces that establish and maintain gene arrangements in
737 *Drosophila pseudoobscura*. *Mol. Ecol.* 26: 6539-6562.
- 738 Gimonneau, G., J. Bouyer, S. Morand, N. J. Besansky, A. Diabate *et al.*, 2010 A behavioral
739 mechanism underlying ecological divergence in the malaria mosquito *Anopheles*
740 *gambiae*. *Behav. Ecol.* 21: 1087-1092.
- 741 Gimonneau, G., M. Pombi, M. Choisy, S. Morand, R. K. Dabire *et al.*, 2012a Larval habitat
742 segregation between the molecular forms of the mosquito, *Anopheles gambiae* in a rice
743 field area of Burkina Faso, West Africa. *Med. Vet. Entomol.* 26: 9-17.
- 744 Gimonneau, G., M. Pombi, R. K. Dabire, A. Diabate, S. Morand *et al.*, 2012b Behavioural
745 responses of *Anopheles gambiae* sensu stricto M and S molecular form larvae to an
746 aquatic predator in Burkina Faso. *Parasit Vectors* 5: 65.
- 747 Giraldo-Calderon, G. I., S. J. Emrich, R. M. MacCallum, G. Maslen, E. Dialynas *et al.*, 2015
748 VectorBase: an updated bioinformatics resource for invertebrate vectors and other
749 organisms related with human diseases. *Nucleic Acids Res.* 43: D707-713.

- 750 Gray, E. M., K. A. Rocca, C. Costantini, and N. J. Besansky, 2009 Inversion 2La is associated
751 with enhanced desiccation resistance in *Anopheles gambiae*. *Malar. J.* 8: 215.
- 752 Hanemaaijer, M. J., T. C. Collier, A. Chang, C. C. Shott, P. D. Houston *et al.*, 2018 The fate of
753 genes that cross species boundaries after a major hybridization event in a natural
754 mosquito population. *Mol. Ecol.* 27: 4978-4990.
- 755 Hoffmann, A. A., and L. H. Rieseberg, 2008 Revisiting the impact of inversions in evolution:
756 from population genetic markers to drivers of adaptive shifts and speciation? *Annual*
757 *Review of Ecology Evolution and Systematics* 39: 21-42.
- 758 Hoffmann, A. A., C. M. Sgro, and A. R. Weeks, 2004 Chromosomal inversion polymorphisms
759 and adaptation. *Trends Ecol. Evol.* 19: 482-488.
- 760 Holt, R. A., G. M. Subramanian, A. Halpern, G. G. Sutton, R. Charlab *et al.*, 2002 The genome
761 sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129-149.
- 762 Houle, D., and E. J. Marquez, 2015 Linkage disequilibrium and inversion-typing of the
763 *Drosophila melanogaster* genome reference panel. *G3 (Bethesda)* 5: 1695-1701.
- 764 Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli *et al.*, 2012 The genomic basis
765 of adaptive evolution in threespine sticklebacks. *Nature* 484: 55-61.
- 766 Joron, M., L. Frezal, R. T. Jones, N. L. Chamberlain, S. F. Lee *et al.*, 2011 Chromosomal
767 rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*
768 477: 203-206.
- 769 Kapun, M., D. K. Fabian, J. Goudet, and T. Flatt, 2016 Genomic evidence for adaptive inversion
770 clines in *Drosophila melanogaster*. *Mol. Biol. Evol.* 33: 1317-1336.
- 771 Kirkpatrick, M., 2010 How and why chromosome inversions evolve. *PLoS Biol.* 8: e1000501.
- 772 Kirkpatrick, M., and B. Barrett, 2015 Chromosome inversions, adaptive cassettes and the
773 evolution of species' ranges. *Mol. Ecol.* 24: 2046-2055.
- 774 Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and speciation.
775 *Genetics* 173: 419-434.
- 776 Lemoine, F., J. B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Davila Felipe *et al.*, 2018
777 Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* 556: 452-
778 456.
- 779 Lobo, N. F., D. M. Sangare, A. A. Regier, K. R. Reidenbach, D. A. Bretz *et al.*, 2010 Breakpoint
780 structure of the *Anopheles gambiae* 2Rb chromosomal inversion. *Malar. J.* 9: 293.
- 781 Love, R. R., A. M. Steele, M. B. Coulibaly, S. F. Traore, S. J. Emrich *et al.*, 2016 Chromosomal
782 inversions and ecotypic differentiation in *Anopheles gambiae*: the perspective from
783 whole-genome sequencing. *Mol. Ecol.* 25: 5889-5906.
- 784 Lowry, D. B., and J. H. Willis, 2010 A widespread chromosomal inversion polymorphism
785 contributes to a major life-history transition, local adaptation, and reproductive isolation.
786 *PLoS Biol.* 8: e1000500.
- 787 Ma, J., and C. I. Amos, 2012 Investigation of inversion polymorphisms in the human genome
788 using principal components analysis. *PLoS One* 7: e40224.
- 789 Main, B. J., Y. Lee, T. C. Collier, L. C. Norris, K. Brisco *et al.*, 2015 Complex genome evolution
790 in *Anopheles coluzzii* associated with increased insecticide usage in Mali. *Mol. Ecol.* 24:
791 5145-5157.
- 792 Manoukis, N. C., J. R. Powell, M. B. Touré, A. Sacko, F. E. Edillo *et al.*, 2008 A test of the
793 chromosomal theory of ecotypic speciation in *Anopheles gambiae*. *Proceedings of the*
794 *National Academy of Science U S A* 105: 2940-2945.

- 795 Marsden, C., Y. Lee, C. Neimen, M. Sanford, J. Dinis *et al.*, 2011 Asymmetric introgression
796 between the M and S molecular forms of the malaria vector, *Anopheles gambiae*,
797 maintains divergence despite extensive hybridisation. *Mol. Ecol.* 20: 4983-4994.
- 798 Miles, A., and N. J. Harding, 2017 scikit-allel: A Python package for exploring and analysing
799 genetic variation data. <http://github.com/cggh/scikit-allel>. 10.5281/zenodo.597309.
- 800 Miles, A., N. J. Harding, G. Bottà, C. S. Clarkson, T. Antão *et al.*, 2017 Genetic diversity of the
801 African malaria vector *Anopheles gambiae*. *Nature* 552: 96-100.
- 802 Navarro, A., E. Betran, A. Barbadilla, and A. Ruiz, 1997 Recombination and gene flux caused by
803 gene conversion and crossing over in inversion heterokaryotypes. *Genetics* 146: 695-709.
- 804 Nwakanma, D. C., D. E. Neafsey, M. Jawara, M. Adiamoh, E. Lund *et al.*, 2013 Breakdown in
805 the process of incipient speciation in *Anopheles gambiae*. *Genetics* 193: 1221-1231.
- 806 Oliveira, E., P. Salgueiro, K. Palsson, J. L. Vicente, A. P. Arez *et al.*, 2008 High levels of
807 hybridization between molecular forms of *Anopheles gambiae* from Guinea Bissau. *J.*
808 *Med. Entomol.* 45: 1057-1063.
- 809 Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, 2011 Scikit-learn:
810 Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825-2830.
- 811 Petrarca, V., and J. C. Beier, 1992 Intraspecific chromosomal polymorphism in the *Anopheles*
812 *gambiae* complex as a factor affecting malaria transmission in the Kisumu area of Kenya.
813 *Am. J. Trop. Med. Hyg.* 46: 229-237.
- 814 Petrarca, V., G. Sabatinelli, M. A. Di Deco, and M. Papakay, 1990 The *Anopheles gambiae*
815 complex in the Federal Islamic Republic of Comoros (Indian Ocean): some cytogenetic
816 and biometric data. *Parassitologia* 32: 371-380.
- 817 Pombi, M., B. Caputo, F. Simard, M. A. Di Deco, M. Coluzzi *et al.*, 2008 Chromosomal
818 plasticity and evolutionary potential in the malaria vector *Anopheles gambiae sensu*
819 *stricto*: insights from three decades of rare paracentric inversions. *BMC Evol. Biol.* 8:
820 309.
- 821 R Core Team, 2018 R: A language and environment for statistical computing, R Foundation for
822 Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- 823 Riehle, M. M., T. Bukhari, A. Gneme, W. M. Guelbeogo, B. Coulibaly *et al.*, 2017 The
824 *Anopheles gambiae* 2La chromosome inversion is associated with susceptibility to
825 *Plasmodium falciparum* in Africa. *Elife* 6.
- 826 Rishikesh, N., M. A. Di Deco, V. Petrarca, and M. Coluzzi, 1985 Seasonal variations in indoor
827 resting *Anopheles gambiae* and *Anopheles arabiensis* in Kaduna, Nigeria. *Acta Trop.* 42:
828 165-170.
- 829 Rocca, K. A., E. M. Gray, C. Costantini, and N. J. Besansky, 2009 2La chromosomal inversion
830 enhances thermal tolerance of *Anopheles gambiae* larvae. *Malar. J.* 8: 147.
- 831 Sangare, D. M., 2007 Breakpoint analysis of the *Anopheles gambiae s.s.* chromosome 2Rb, 2Rc,
832 and 2Ru inversions in *PhD Thesis, Graduate Program in Biological Sciences, University*
833 *of Notre Dame*. University of Notre Dame, Notre Dame, IN.
- 834 Schaeffer, S. W., 2008 Selection in heterogeneous environments maintains the gene arrangement
835 polymorphism of *Drosophila pseudoobscura*. *Evolution* 62: 3082-3099.
- 836 Seich Al Basatena, N. K., C. J. Hoggart, L. J. Coin, and P. F. O'Reilly, 2013 The effect of
837 genomic inversions on estimation of population genetic parameters from SNP data.
838 *Genetics* 193: 243-253.

- 839 Simard, F., D. Ayala, G. C. Kamdem, J. Etouna, K. Ose *et al.*, 2009 Ecological niche partitioning
840 between the M and S molecular forms of *Anopheles gambiae* in Cameroon: the
841 ecological side of speciation. *BMC Ecol.* 9: 17.
- 842 Tene Fossog, B., D. Ayala, P. Acevedo, P. Kengne, I. Ngomo Abeso Mebuy *et al.*, 2015 Habitat
843 segregation and ecological character displacement in cryptic African malaria mosquitoes.
844 *Evol Appl* 8: 326-345.
- 845 Toure, Y. T., V. Petrarca, S. F. Traore, A. Coulibaly, H. M. Maiga *et al.*, 1998 The distribution
846 and inversion polymorphism of chromosomally recognized taxa of the *Anopheles*
847 *gambiae* complex in Mali, West Africa. *Parassitologia* 40: 477-511.
- 848 Twyford, A. D., and J. Friedman, 2015 Adaptive divergence in the monkey flower *Mimulus*
849 *guttatus* is maintained by a chromosomal inversion. *Evolution* 69: 1476-1486.
- 850 Weetman, D., C. S. Wilding, K. Steen, J. Pinto, and M. J. Donnelly, 2012 Gene flow-dependent
851 genomic divergence between *Anopheles gambiae* M and S forms. *Mol. Biol. Evol.* 29:
852 279-291.
- 853 Wellenreuther, M., and L. Bernatchez, 2018 Eco-Evolutionary Genomics of Chromosomal
854 Inversions. *Trends Ecol. Evol.* 33: 427-440.
- 855 Wellenreuther, M., H. Rosenquist, P. Jaksons, and K. W. Larson, 2017 Local adaptation along an
856 environmental cline in a species with an inversion polymorphism. *J. Evol. Biol.* 30: 1068-
857 1077.
- 858 White, B. J., F. H. Collins, and N. J. Besansky, 2011 Evolution of *Anopheles gambiae* in relation
859 to humans and malaria. *Annual Review of Ecology Evolution and Systematics* 42: 111-
860 132.
- 861 White, B. J., F. Santolamazza, L. Kamau, M. Pombi, O. Grushko *et al.*, 2007 Molecular
862 karyotyping of the 2La inversion in *Anopheles gambiae*. *Am. J. Trop. Med. Hyg.* 76: 334-
863 339.
- 864 World Health Organisation, 2018 *World Malaria Report: 2018*,
865 <https://www.who.int/malaria/publications/world-malaria-report-2018/report/en/>.
866
- 867

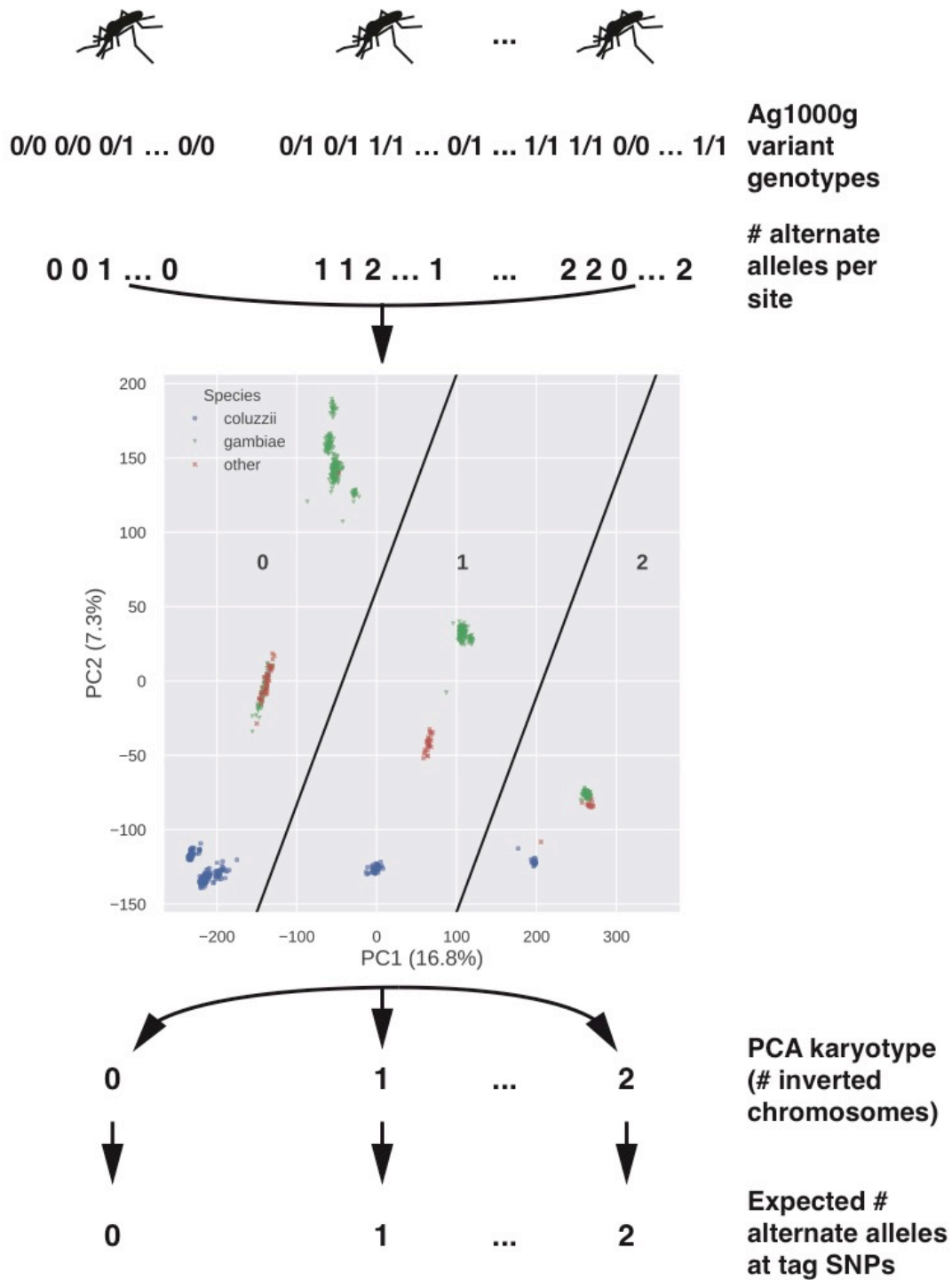
868 Figure 1.

869



871

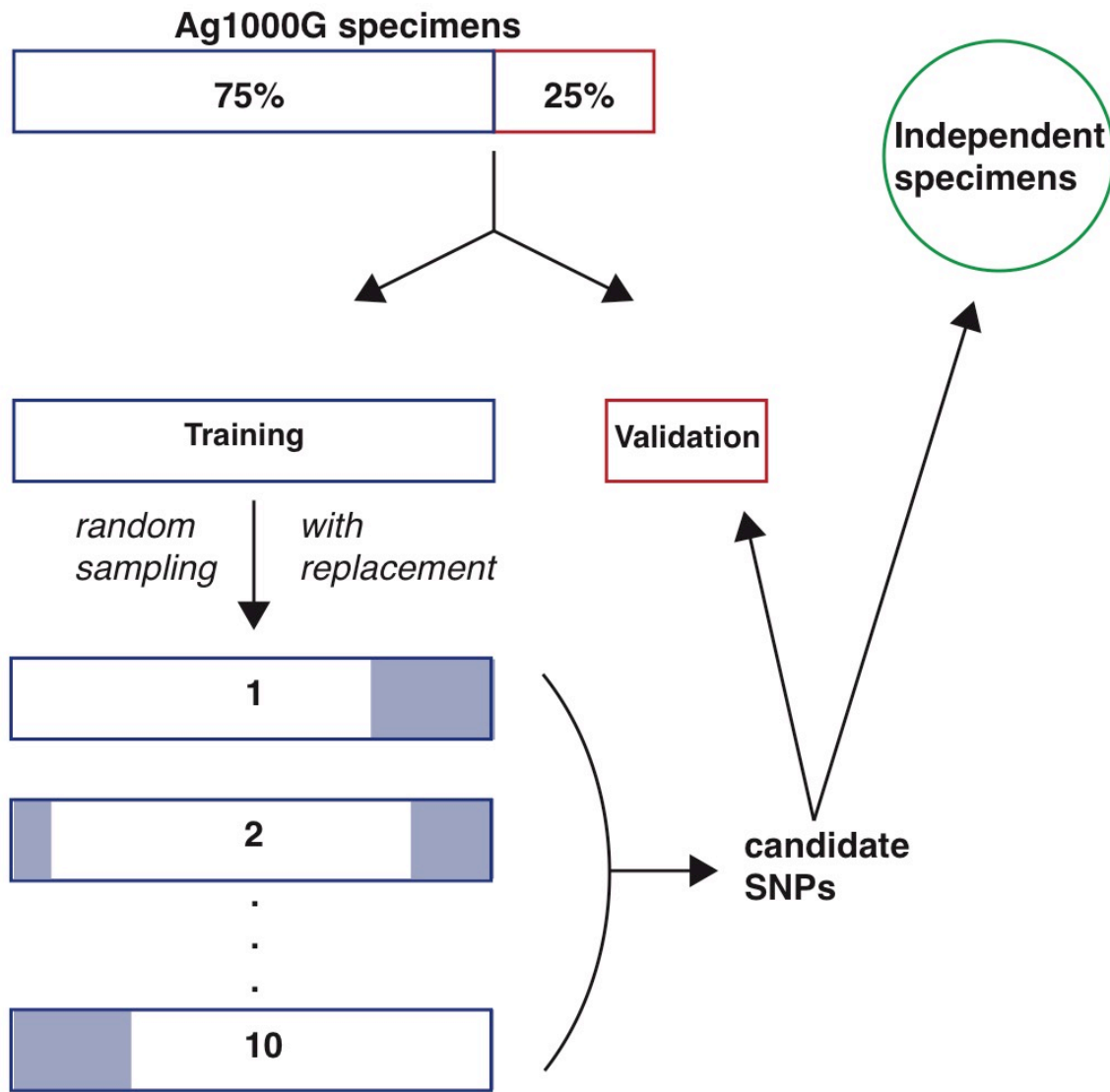
872 Figure 2.



873

874

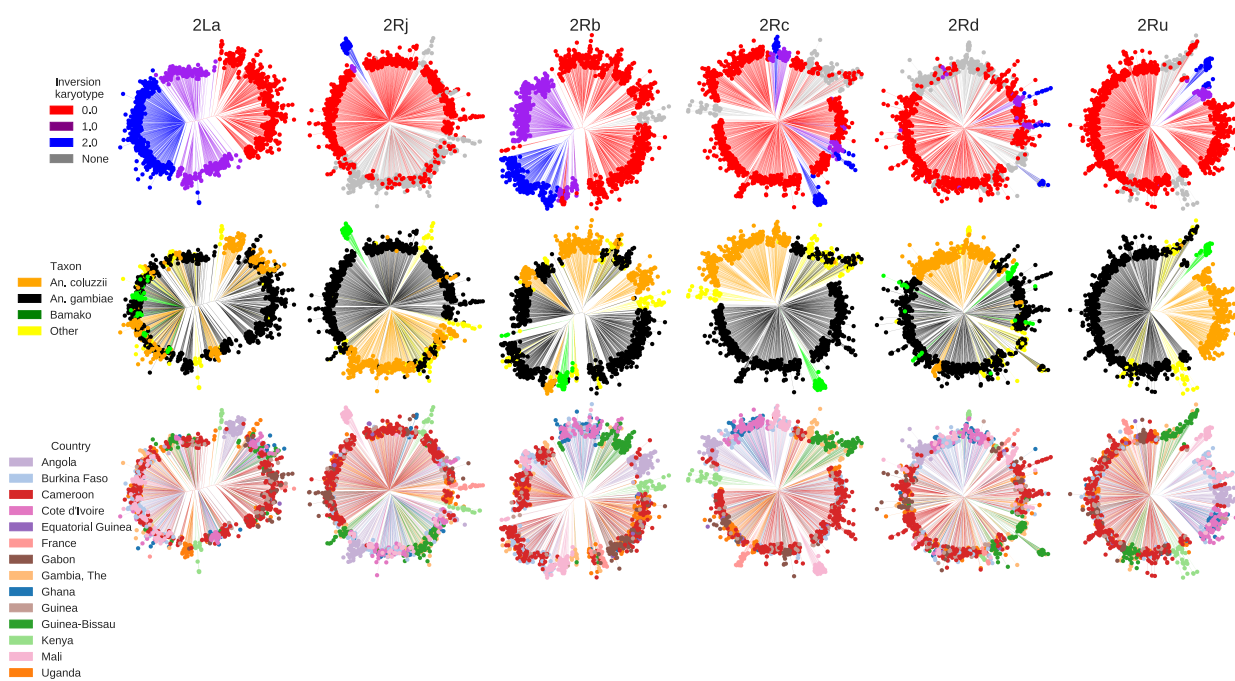
875 Figure 3.



876

877 Figure 4.

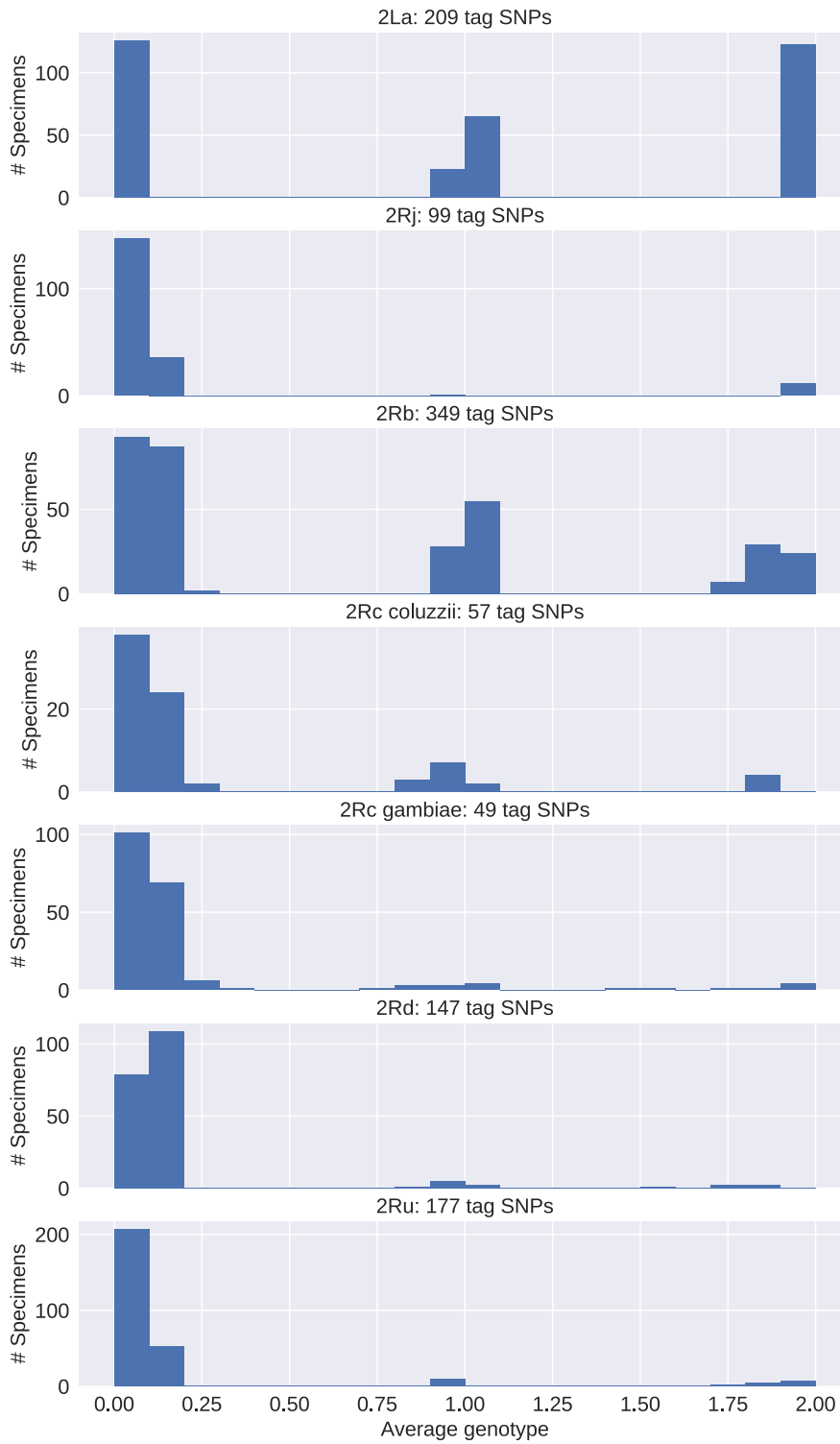
878



879

880

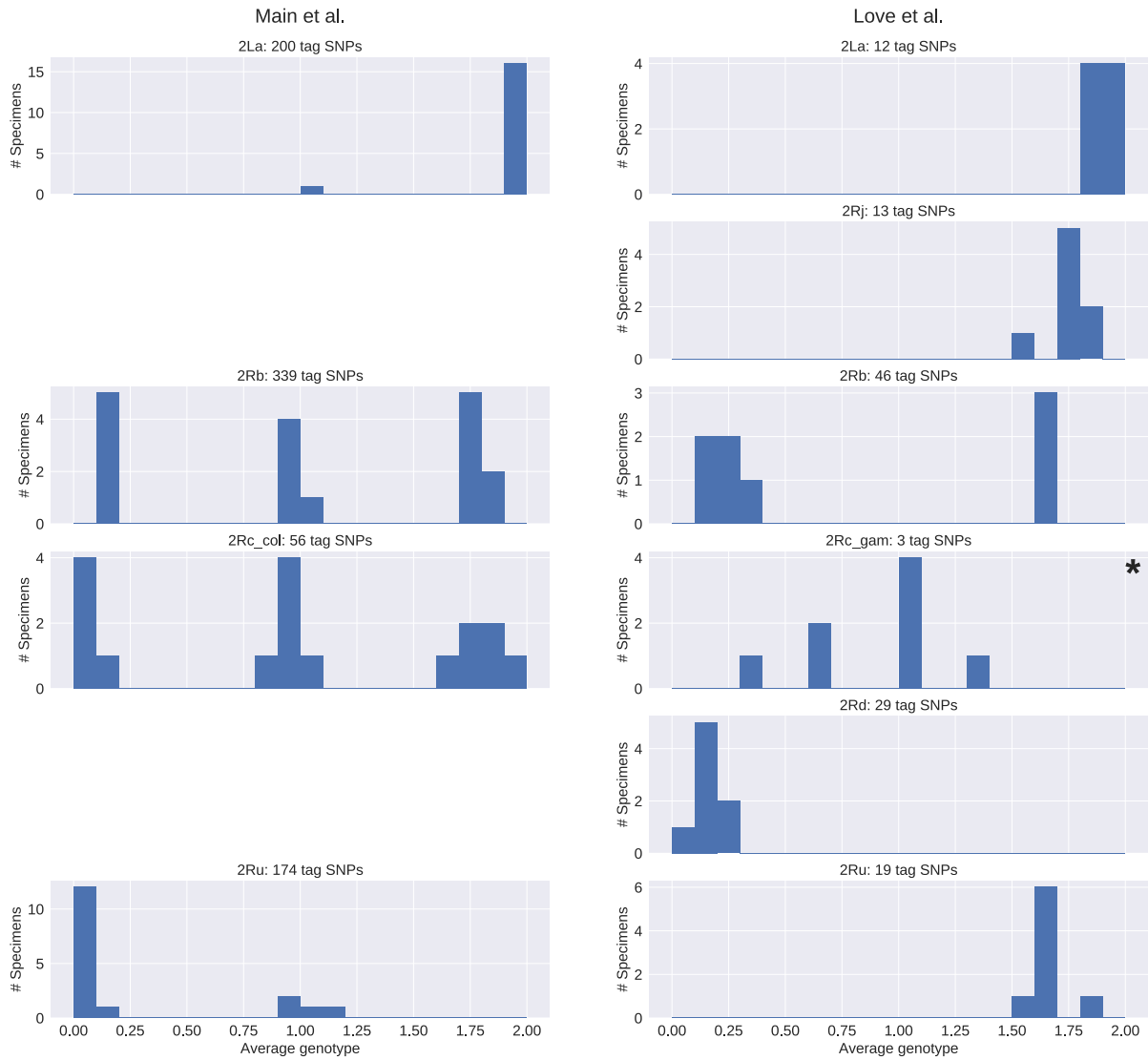
881 Figure 5.



882

883

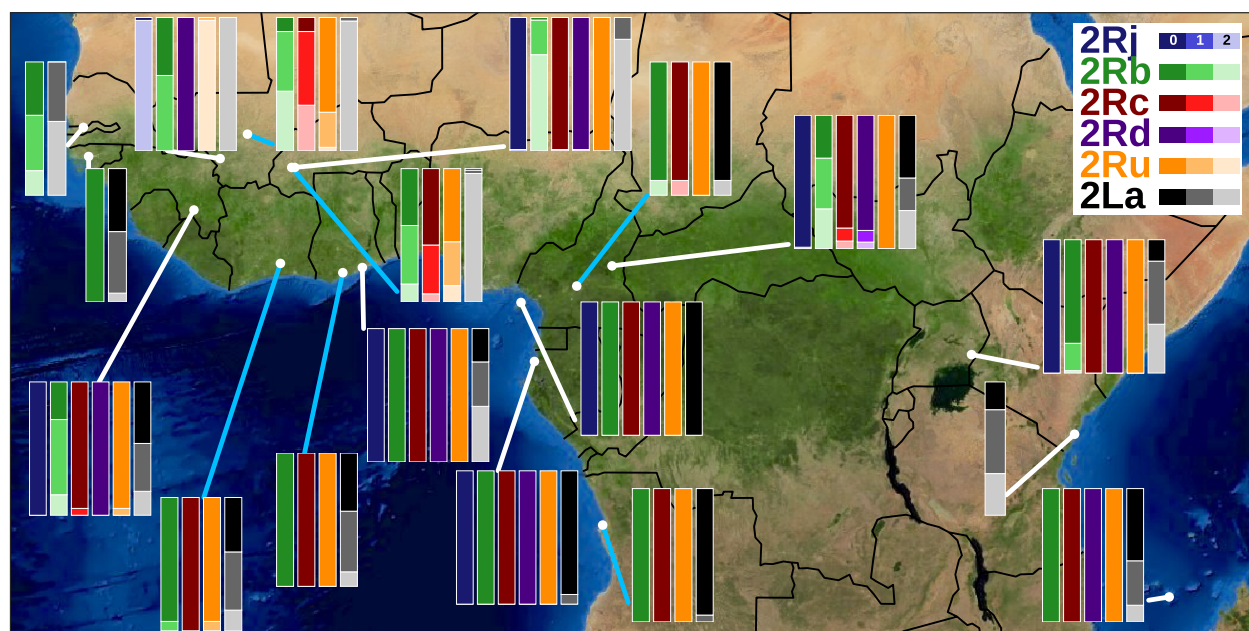
884 Figure 6.



885

886

887 Figure 7.



888

889

890 **Table 1.** Candidate tag SNPs predictive of inversion genotype in Ag1000G data

Inversion	Concordance Threshold	No. Tags
2La	>0.995	209
2Rj-gambiae	>0.8	99
2Rb	>0.8	349
2Ru	>0.8	177
2Rd-gambiae	>0.8	147
2Rc	>0.8	2
2Rc-coluzzii	>0.8	57
2Rc-gambiae	>0.8	49

891

892

893 **Table 2.** Mismatches between PCA and computational karyotypes in the Ag1000G validation sets

Inversion	Total specimens	No. tags scored	Matching karyotypes		Mismatched karyotypes	
			No. specimens	% tags supporting score	No. Specimens	% tags supporting score
2La	337	168-203	337	93.6-100	0	--
2Rj- <i>gambiae</i>	195	94-99	195	83.8-100	0	--
2Rb	325	304-349	325	77.5-97.7	0	--
2Rc- <i>coluzzii</i>	80	55-57	80	78.9-100	0	--
2Rc- <i>gambiae</i>	196	45-49	195	59.6-100	1	67.3
2Rd- <i>gambiae</i>	201	128-147	201	55.2 ¹ -95.9	0	--
2Ru	286	124-177	286	76.6-100	0	--

894 ¹Next highest value is 70.1%.

895

896 **Table 3.** Discrepancies between cytogenetic and computational karyotypes in Ag1000G mosquitoes analyzed.

Inversion Tags	Partition	CYT	Specimens (N)	Specimens with discrepancies		
				Mismatch CYT-TAG (%)	Match TAG-PCA (%)	No. tag SNPs scored (% matching TAG)
2La		0	117	5 (4.3)	5 (100)	200-203 (99.5-100)
		1	68	5 (7.4)	5 (100)	193-203 (100)
		2	160	2 (1.3)	2 (100)	201-203 (99.5-100)
2Rj-gambiae	<i>gambiae</i>	0	236	0 (0)	--	–
		1	4	0 (0)	--	–
		2	45	0 (0)	--	–
2Rb		0	127	2 (1.6)	2 (100)	348 (85.3-87.6)
		1	124	4 (3.2)	4 (100)	346-349 (86.8-93.7)
		2	121	6 (5.0)	6 (100)	331-348 (88.8-91.6)
2Rc-gambiae	<i>gambiae</i> ¹	0	184	7 (3.8)	7 (100)	48-49 (83.7-98.0)
		1	32	3 (9.4)	2 (66.7)	47-49 (42.9 ² -91.5)
		2	24	2 (8.3)	2 (100)	48-49 (90.0-91.8)
2Rc-coluzzi	<i>coluzzii</i>	0	13	1 (7.7)	1 (100)	56 (87.5)
		1	25	0 (0)	--	–
		2	16	0 (0)	--	–
2Rd-gambiae	<i>gambiae</i>	0	234	9 (3.8)	9 (100)	143-147 (84.9-96.6)
		1	28	4 (14.3)	4 (100)	146-147 (88.4-93.9)
		2	22	3 (13.6)	3 (100)	146-147 (89.1-91.8)
2Ru	<i>col+gam</i>	0	263	1 (0.38)	1 (100)	176 (85.2)
		1	29	18 (62.1)	18 (100)	170-177 (88.7-99.4)
		2	47	1 (2.1)	1 (100)	176 (97.2)

897 CYT, cytogenetic genotype; TAG, computational genotype; PCA, genotype inferred by PCA.

898 ¹*An. gambiae* excluding BAMAko

899 ²This value corresponds to one of three non-BAMAko *An. gambiae* carriers of the 2Ru inversion, AZ0267-C. The next highest value
900 is 85.7.

901

902

903