Title: Reticulate speciation and adaptive introgression in the *Anopheles gambiae* species complex.

Running Head: Speciation genomics of *Anopheles* mosquitoes

Authors: Jacob E. Crawford[1,2], Michelle M. Riehle[3], Wamdaogo M. Guelbeogo[4], Awa Gneme[4], N'Fale Sagnon[4], Kenneth D. Vernick[5], Rasmus Nielsen[2]*, Brian P. Lazzaro[1]*.
* These authors contributed equally to this work.

Affiliations:

1. Department of Entomology, Cornell University, Ithaca, NY, USA
2. Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA
3. Department of Microbiology, University of Minnesota, St. Paul, MN, USA
4. Centre National de Recherche et de Formation sur le Paludisme, 1487 Avenue de l'Oubritenga, 01 BP 2208 Ouagadougou, Burkina Faso.
5. Unit for Genetics and Genomics of Insect Vectors, Institut Pasteur, Paris, France
6. Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA

Abstract:

Anopheles gambiae, the primary vector of human malaria in sub-Saharan Africa, exists as a series of ecologically specialized subgroups that are phylogenetically nested within the *Anopheles gambiae* species complex. These species and subgroups exhibit varying degrees of reproductive isolation, sometimes recognized as distinct subspecies. We have sequenced 32 complete genomes from field-captured individuals of *Anopheles gambiae*, *Anopheles gambiae* M form (recently named *A. coluzzii*), sister species *A. arabiensis*, and the recently discovered "GOUNDRY" subgroup of *A. gambiae* that is highly susceptible to *Plasmodium*. Amidst a backdrop of strong reproductive isolation and adaptive differentiation, we find evidence for introgression of autosomal chromosomal regions among species and subgroups, some of which have facilitated adaptation. The X chromosome, however, is strongly differentiated among all species and subgroups, pointing to a disproportionately large effect of X chromosome genes in driving speciation among anophelines. Strikingly, we find that autosomal introgression has occurred from contemporary hybridization among *A. gambiae* and *A. arabiensis* despite strong divergence (~5× higher than autosomal divergence) and isolation on the X chromosome. We find a large region of the X chromosome that has swept to fixation in the GOUNDRY subgroup within the last 100 years, which may be an inversion that serves as a partial barrier to contemporary gene flow. We show that speciation with gene flow results in genomic mosaicism of divergence and introgression. Such a reticulate gene pool connecting vector species and subgroups across the speciation continuum has important implications for malaria control efforts.

Author Summary:

Subdivision of species into ecological specialized subgroups allows organisms to access a wider variety of environments and sometimes leads to the formation of species complexes. Adaptation to distinct environments tends to result in differentiation among closely related subgroups, although hybridization can facilitate sharing of globally adaptive alleles. Here, we show that differentiation and hybridization have acted in parallel in a species complex of *Anopheles* mosquitoes that vector human malaria. In particular, we show that extensive adaptive differentiation and partial reproductive

52 isolation has led to genomic differentiation among mosquito species and subgroups,
53 especially on the X chromosome. However, we also find evidence for exchange of genes
54 on the autosomes that has provided the raw material for recent rapid adaptation. For
55 example, we show that *A. arabiensis* has shared a mutation conferring insecticide
56 resistance with two subgroups of *A. gambiae* within the last 60 years, illustrating the fluid
57 nature of species boundaries among even more advanced species pairs. Our results
58 underscore the expected challenges in deploying vector-based disease control strategies
59 since many of the world's most devastating human pathogens are transmitted by
60 arthropod species complexes.
61
62

63 **Introduction:**

64 Closely related, morphologically similar, and sometimes interbreeding species
65 complexes are common in nature and challenge conceptions of species boundaries [1]. In
66 some cases, taxa diverge to exploit distinct ecological niches and thus represent incipient
67 species on a trajectory towards reproductive isolation and phenotypic differentiation [2].
68 However, an alternative, more fluid, view of species boundaries and divergence as a
69 continuum may be more appropriate in cases where the extent of reproductive isolation
70 varies across time and environmental space (reviewed in [3]). Whether permeable
71 species boundaries affect adaptive evolution among closely related taxa remains
72 contentious. On one hand, introgression may serve to homogenize diverging taxa and
73 oppose adaptive differentiation [4]. On the other hand, globally adaptive alleles may be
74 shared among subgroups, increasing the mean fitness of each [5,6]. In species complexes,
75 semi-permeable species boundaries could provide conduits for adaptive alleles to spread
76 across large distances, both geographic and environmental.

77 The *Anopheles gambiae* species complex in sub-Saharan Africa includes several
78 major vectors of malaria, which continues to place a devastating burden on local human
79 populations [7]. Prior to the 1940s, *A. gambiae* was considered a single biologically
80 variable species, but crossing studies and genetic analysis led to the naming of nine
81 morphologically similar species that vary in their geographic distribution and ecology
82 (reviewed in [8]). It is becoming increasingly appreciated from ecological distinctions
83 and recent discoveries of additional genetic substructure that, even within species,
84 *Anopheles* species frequently form partially reproductively isolated and differentiated
85 subpopulations [9–11].

86 The M and S "molecular forms" are two major subgroups within *Anopheles*
87 *gambiae sensu stricto* that are mostly reproductively isolated in the field, although they
88 are compatible in captivity [12,13]. The two forms are sympatric in West and Central
89 Africa but exploit distinct microecological niches [9,14]. The M form was recently given
90 a formal species name, *Anopheles coluzzi* [8], although in the present paper we use the
91 older terminology of 'M form' in the present work for continuity with existing literature.
92 The 'S form' retains the name *A. gambiae s.s.*, but we use 'S form' here. Internal
93 subdivisions exist even within the molecular forms [15,16] and recent evidence indicates

94 an often high level of local M-S hybridization[11], illustrating the fluid semispecies
95 boundaries within the *Anopheles gambiae* group. It is not known how many cryptic
96 subpopulations exist within *Anopheles*, or how much gene flow they share, but there is
97 evidence that subdivision may be common [11]. This is exemplified by the recently
98 discovery of the GOUNDRY subgroup in Burkina Faso, which shows considerable
99 genetic distinction from other described *A. gambiae* and is particularly permissive for
100 *Plasmodium* development [17].
101 Epidemiological modeling and vector-based malaria control strategies must
102 explicitly consider unique aspects of subgroups such as M, S, and GOUNDRY if they are
103 to effectively predict disease dynamics and responses to intervention [18]. Because many
104 of the world's most devastating human diseases are vectored by arthropods in species
105 complexes, including *Anopheles* vectors of malaria, tsetse fly vectors of African sleeping
106 sickness, and *Ixodes* tick vectors of Lyme disease [19–21], the dynamic evolution of
107 phenotypic diversity and adaptive introgression among cryptic taxa in species complexes
108 has serious implications for public health. In the specific case of malaria, the existence of
109 partially differentiated but occasionally hybridizing subspecies could complicate malaria
110 control efforts that rely on the spread of transgenes through mosquito populations or
111 conventional controls that target specific aspects of mosquito ecology or life history [22].
112 Genomic dissection of species complexes also lends insight into processes of
113 speciation. Studies have shown strong differences between sex chromosomes (X or Z)
114 and autosomes in the rates of genetic divergence and introgression and therefore
115 speciation among pairs of species in systems ranging from *Drosophila* to Hominids.
116 Genetic divergence tends to be higher on sex chromosomes relative to autosomes and
117 introgressed regions are underrepresented on the sex chromosomes [23–25].
118 Heterogeneity in levels of genetic divergence has also been shown among genomic
119 regions that vary in rates of meiotic recombination, with elevated divergence in genomic
120 regions with low recombination rates such as centromeric regions and chromosomal
121 inversions [26–28]. Analysis of genetic divergence and introgression in the *Anopheles*
122 *gambiae* species complex provides a valuable opportunity to test the generality of these
123 patterns across the speciation continuum.

124 The *Anopheles* system has become a model system for studying speciation and
125 gene flow. Although substantial effort has been dedicated to understanding the status,
126 history, and genomic consequences of reproductive isolation between two members of
127 the *Anopheles gambiae* species complex, the M and S molecular forms of *A. gambiae*
128 [13,29–31], firm conclusions regarding the degree of reproductive isolation and the age
129 of the two forms remains controversial. This is due in part to the fact that these studies
130 were based on ascertained SNP-panels, low-resolution sequencing datasets, and sequence
131 datasets from small samples of laboratory mosquito colonies. In addition, they relied on
132 statistical approaches that are incapable of distinguishing between multiple confounding
133 population genetic processes. In particular, these analyses relied on measures of relative
134 divergence, which are not robust to variation in recombination rates and natural selection
135 across the genome and do not explicitly distinguish between lineage sorting of ancestral
136 polymorphism and introgression [32–35]. Relatively few studies have addressed
137 divergence among other members of the species complex and the same analytical
138 concerns apply here as well [30,36,37]. In depth analysis of full sequence data using
139 appropriate statistical approaches will greatly improve our ability to infer the
140 introgression dynamics and the evolutionary consequences.
141 In the present study, we generated whole genome sequences from 32 wild caught
142 females from the *Anopheles gambiae* species complex representing multiple points along
143 the speciation continuum within this complex. We conducted detailed population genetic
144 analysis of these data and show that 1) *A. gambiae* is more closely related to *A.*
145 *arabiensis* than to *A. merus*, and the newly discovered GOUNDRY subgroup of *A.*
146 *gambiae* diverged from the M form ~100 kya, 2) speciation is driven by the X
147 chromosome in this species complex, but introgression of autosomal regions is common
148 among species across the speciation continuum, 3) ecological specialization and local
149 adaptation may have driven many recent selective sweeps in these species, and 4)
150 introgression has facilitated adaptation in some cases.
151
152 **Results**
153 *Genome Sequencing and Population Genetic Analysis*

154 We have completely sequenced the genomes of 32 field-captured female
155 *Anopheles* mosquitoes from Burkina Faso and Guinea using the Illumina HiSeq2000
156 platform. We sequenced *A. gambiae* GOUNDRY (n=12), *A. gambiae* M form (n=10), *A.*
157 *gambiae* S form (n=1) and *Anopheles arabiensis* (n=9). Most individuals were
158 sequenced to an average read depth of 9.79x, while one individual each from
159 GOUNDRY, M form, and S form was sequenced to at least 16.44x (Table S1). We also
160 used publicly available genome sequences from *Anophleles merus* as an outgroup and a
161 previously published *An. arabiensis* genome from Tanzania for a geographically distinct
162 comparison [38]. We conducted population genetic analysis of aligned short-read data
163 using genotype likelihoods and genotype calls calculated using the probabilistic inference
164 framework ANGSD [39].
165
166 *Genetic Relatedness Among Species and Subgroups*
167 To confirm broad-scale genetic relationships among the *Anopheles* included in
168 this analysis, we calculated an unrooted neighbor-joining tree based on genome-wide
169 genetic distance (D_{xy}) at intergenic sites (Figure 1). Comparisons of genetic distance
170 averaged across the genomes in our data recovers monophyletic species and subgroups
171 with the *A. gambiae* clade more closely related to *A. arabiensis* than either is to *A. merus*,
172 consistent with recent analyses [37,40]. Specifically, the length of the branch leading to
173 the *A. merus* clade ($0.0240 \pm 7.61\text{x}10^{-4}$) is 3.64× the length of the branch leading to the *A.*
174 *arabiensis* clade ($0.0066 \pm 7.41\text{x}10^{-4}$) and their 95% bootstrap intervals do not overlap
175 indicating a closer genetic relationship between *A. arabiensis* and *A. gambiae*. In
176 contrast to the initial description of the recently discovered GOUNDRY subgroup of *A.*
177 *gambiae* as a genetic outgroup to the M and S molecular forms of *A. gambiae* [17], our
178 data indicate that GOUNDRY is actually derived from the M form ($D_{GM} = 0.0109$; 100%
179 bootstrap support) more recently than either subgroup shares a common ancestor with the
180 S form ($D_{GS} = 0.0149$; $D_{MS} = 0.0143$).
181
182 *Origins of GOUNDRY, a New Cryptic Taxon*
183 It has been hypothesized that the advent of agriculture in sub-Saharan Africa ~5-
184 10 kya played a role in driving diversification and expansion of *Anopheles* mosquitoes

185 [41]. To test whether the origin of GOUNDRY could have been associated with habitat
186 modification driven by agriculture, we fit three-epoch population historical models
187 (Figure 2; Methods) to the two-dimensional site frequency spectrum for GOUNDRY and
188 M form *A. gambiae* using *dadi* [42]. The best-fitting model (Table 1) predicts that these
189 subgroups diverged ~111,200 ya (95% CI 96,718 – 125,010), followed by a 100-fold
190 reduction in the size of both subgroups after isolation (Methods), and thus rejects any role
191 of modern agriculture in subgroup division, although it should be noted that estimates of
192 such old splits times inherently carry considerable uncertainty. Our inferred model is
193 inconsistent by an order of magnitude with agriculture as a driving force in cladogenesis
194 and more consistent with habitat fragmentation and loss due to natural causes. The
195 model supports a >500-fold population growth in M form and 19-fold growth in
196 GOUNDRY with extensive gene flow between them ~85,300 ya, consistent with a re-
197 establishment of contiguous habitat and abundant availability of bloodmeal hosts.
198 Interestingly, the model supports additional population growth in both subgroups in the
199 most recent epoch, which spans the last 10,000 years and coincides with the advent of
200 agriculture. Hybridization related to secondary contact has not led to complete
201 homogenization, however, as we conservatively identified nearly 9,000 fixed nucleotide
202 differences distributed across the genomes of the two subgroups.
203 We note that the dates reported here depend on assumptions about both the
204 physiological mutation rate as well as the number of generations per year, neither of
205 which are well known in *Anopheles* (Methods). As such, the details of these results
206 would differ somewhat if different estimates were used, but we would have to invoke
207 extreme values of these parameters in order to obtain estimates of an age of the
208 GOUNDRY-M form split that is consistent with the advent of agriculture. The resulting
209 2D spectra from the data and the best-fit model are presented in Figure 2. Residuals
210 indicating differences between the model and data are also presented. In general the fit is
211 quite good. Most of the weight in the 2D empirical spectrum is matched in the 2D
212 spectrum from the model, and the residual plot reveals that in general the model captures
213 most features of the true spectrum. Several exceptions exist, however, indicating that
214 some complexity of the true history is not captured by our model. Nonetheless, the
215 model captures the patterns in the data quite well and suggest that GOUNDRY and M

216 form diverged ~100 kya, have both undergone bouts of population growth and
217 experienced increased rates of hybridization in more recent evolutionary time.
218
219 *Evidence for extensive adaptive differentiation*
220 Ecological specialization typically involves natural selection on ecologically
221 relevant traits involving one or more genetic loci. To identify loci putatively involved in
222 ecological specialization and quantify the role of selection in adaptive differentiation, we
223 scanned the genomes of GOUNDRY, M form, and *A. arabiensis* for signals of recent
224 positive natural selection using SweepFinder [43]. We evaluated the global site
225 frequency spectra from intergenic sites in euchromatic regions of the autosome and X
226 chromosome separately without explicitly calling genotypes using algorithms within
227 ANGSD [39,44–46]. For M form and GOUNDRY, we established genome-wide critical
228 values by comparing empirical likelihood ratio scores to a distribution of high scores
229 obtained by applying SweepFinder to genomes simulated with coalescent simulations
230 under the demographic model inferred for these subgroups described in the section above.
231 For *A. arabiensis*, we used a 'high-score-cluster' approach to identify putative sweeps
232 where regions were considered significant if two adjacent windows were assigned
233 likelihood ratio scores from the top 0.1% genome-wide (see Methods). Consistent with
234 ongoing adaptive divergence, perhaps related to ecological specialization, we find
235 evidence for a number of recent selective sweeps that are private to each of the three
236 groups [Figure 3, Tables S2 and S3]. We find evidence of 120 recent selective sweeps in
237 the M form subgroup, compared to only 67 in GOUNDRY and 33 in the *A. arabiensis*
238 genome.
239 We were interested in determining whether GOUNDRY or M form has
240 experienced significantly more selection than the other. We scanned each genome for
241 selective windows and found that, of the 230 possible megabase windows, 36 windows
242 harbored at least one selective sweep in GOUNDRY while 58 harbored sweeps in M
243 form, indicating that M form has experienced significantly (Figure S1; $P < 0.0001$) more
244 positive selection in recent evolutionary history. This difference may reflect recent
245 exploitation of marginal ecological habitats by the M form subgroup, especially those
246 associated with human settlements [9,47].

247 We note here that the many recent selective sweeps in GOUNDRY combined
248 with the fact that GOUNDRY is partially inbred have led to unusual patterns of
249 nucleotide diversity across the genome (Figure 4). We observed strong evidence of
250 inbreeding in the GOUNDRY sample, but not in the other subgroups. Optimized
251 individual inbreeding coefficients for chromosomal arms ranged from 0 to nearly 0.7 in
252 the *A. gambiae* GOUNDRY subgroup sample, where an inbreeding coefficient of 1
253 would indicate complete lack of heterozygosity. Other than relatively high values on 3R
254 for *A. arabiensis*, inbreeding coefficients for the M form subgroup and *A. arabiensis* were
255 uniformly low. The underlying source of the high values on 3R in *A. arabiensis* is not
256 known. In GOUNDRY, the homozygous regions are uniformly distributed around the
257 genome but unique to each sampled individual, which is consistent with a high degree of
258 close inbreeding. Importantly, inbreeding affects the distribution of diversity along
259 diploid genomes, but should not obscure the signals of selective sweeps observed in
260 GOUNDRY. Specifically, selective sweeps involve the fixation of the same haplotype
261 across the population, whereas inbreeding simply brings together parental haplotypes that
262 are identical-by-descent but are not necessarily shared among breeding pairs. As such,
263 these processes have distinct effects on patterns of polymorphism that should not impact
264 the ability to detect sweeps.
265 Two of the swept loci in the M form [*TEP1* and *Resistance to dieldrin* (*Rdl*)] have
266 been identified previously using independent datasets and analytical approaches [48][13],
267 providing an important validation of our analysis pipeline. These genes are involved in
268 immunity and resistance to insecticide, respectively. We additionally find a particularly
269 intriguing novel sweep at a gene encoding a neuropeptide F (NPF; AGAP004122). In
270 *Drosophila*, NPFs have been shown to control larval behavior and feeding [49], so it is
271 tempting to speculate that adaptive differentiation of this gene may underlie the
272 previously described unique larval bottom-feeding and extended predator avoidance
273 behaviors of the M form [50]. The most notable signal of positive selection in M form
274 genomes is found near MB 15 on the X chromosome. Linkage disequilibrium among
275 SNPs at least 1,000 bp apart is exceptionally elevated across a region approximately one
276 megabase in size, suggesting that this region has been subjected to especially strong and
277 recent positive selection (Figure S2). However, it is difficult to say whether or not this

278 was a single event since our SweepFinder analysis indicates five independent peaks in the
279 log-likelihood ratio surface in this region (Table S2), even after re-analyzing clusters of
280 peaks that could represent single events falsely divided into a swarm of smaller peaks
281 (Methods). The genes nearest to the five peaks include a gene encoding a cuticular
282 protein in the RR-2 family, a gene encoding CDC-like kinase, and two genes with no
283 known function (Table S2).
284 The sweeps in the *A. arabiensis* genomes include genes encoding two proteins
285 involved in development, Eclosion hormone and Frizzled-2, and the gene encoding the
286 sulfonylurea receptor. RNAi knockdown of the sulfonylurea receptor in *Drosophila*
287 *melanogaster* significantly increases susceptibility to viral infections [51].
288 One particularly notable sweep in GOUNDRY lies near the gene encoding the
289 master gustatory and odorant receptor *Agam/Orco* (*AgOr7*) that is a required coreceptor
290 for all odorant binding proteins [52]. Expression of *AgOr7* is predominantly localized in
291 female antennae and palps and fluctuates with circadian cycles, leading to the hypothesis
292 that this gene controls temporal shifts in olfactory sensitivity [53]. Adaptation at *AgOr7*
293 could have important impacts on anthropophily, sensitivity to DEET [54], and may
294 underlie exophilic behavior in GOUNDRY [17].
295 In summary, we identified a large number of recent bouts of positive selection in
296 GOUNDRY, M form, and *A. arabiensis*, including a number of interesting loci that
297 potentially underlie important ecological and physical differences between these taxa.
298 We also showed that M form has experienced more recent positive selection than
299 GOUNDRY, which may reflect the more diverse environmental challenges that the M
300 form has experienced as it exploits novel environments and niches.
301
302 *Recent autosomal introgression*
303 Since recently diverged subgroups and species may still share polymorphism even
304 if they are completely reproductively isolated in practice, we addressed the question of
305 whether several members of the *Anopheles gambiae* species complex have introgressed
306 historically or contemporaneously by formally differentiating between introgression and
307 lineage sorting of ancestral polymorphism. To explicitly differentiate between these two
308 population genetic processes, we used the ABBA-BABA test [55,56]. This test uses the

309 *D* statistic to compare the distribution of alleles on the four taxon tree ((H1,H2),H3),O),
310 where H1 and H2 are sister taxa and H3 and O are the outgroups. Under the null
311 hypothesis of a perfect tree structure and no gene flow, the number of derived mutations
312 that are shared only between the genomes of H2 and H3 (ABBA) is expected to equal the
313 number of those that are shared only between H1 and H3 (BABA). *D* is then calculated
314 as the standardized difference between the numbers of ABBAs and the number of
315 BABAs with an expectation that *D* is zero under the null hypothesis [55]. Significant
316 excess sharing of derived alleles between H3 and either H1 or H2 will result in a non-
317 zero *D* and provides evidence of introgression. In our case, such tests are not appropriate
318 because the alternative to no introgression is not only asymmetric introgression between
319 an outgroup and one of the ingroups, but also potentially approximately equal amounts of
320 introgression between the outgroup and each of the two ingroups. For that reason we use
321 a modified test in which we instead consider the length distribution of fragments of
322 shared ancestry. Since haplotypes are broken down by recombination over successive
323 generations, the length distribution of shared haplotypes among populations is
324 informative regarding the time since the most recent introgression event [57,58]. After *t*
325 generations, the mean length of a shared haplotype is approximately $(r(1-m)(t-$
326 $1))^{-1}$, where *r* is the recombination rate and *m* is the proportion of introgressed
327 individuals in the population [58]. Since *t* will be smaller for introgressed haplotypes
328 than for shared ancestral haplotypes, the mean length of introgressed haplotypes will be
329 longer resulting in longer clusters of excess ABBAs or BABAs. We used patterns of
330 linkage disequilibrium to approximate this expectation and deviations from it by
331 calculating variance in the *D* statistic among ~250 Kb blocks, Var[D_{BLOCK}], to detect
332 physical clusters of correlated genomic segments consistent with an excess of long shared
333 haplotypes, and compared it to its null distribution generated by permuting fragments of
334 size ~2.5 Kb in order to manually break up possible correlations (see Methods). If the
335 observed value of Var[D_{BLOCK}] is significantly larger than that predicted from permuted
336 fragments of size ~2.5 Kb, which is >10x the average length of the decay of linkage
337 disequilibrium, we conclude that the genomes harbor introgressed haplotypes.
338 We applied this test using several four-taxon trees and find evidence for recent
339 introgression among species and subgroups of the *A. gambiae* species complex. We

340 calculated the *D* statistic using a four-taxon tree with M form and GOUNDRY as the
341 sister taxa (H1 and H2, respectively), *A. merus* as the outgroup (O), and either S form or
342 *A. arabiensis* as H3. After comparing the Var[D_{BLOCK}] in the empirical data to
343 Var[D_{BLOCK}] from 10^4 randomly permutated genomes, we find significant evidence for
344 introgression between the M form-GOUNDRY clade and S form [$P < 0.0001$, Figure 5].
345 Using M form and GOUNDRY as the ingroups again, but *A. arabiensis* as the test taxon
346 (H3), we also find that the Var[D_{BLOCK}] is significantly larger than all 10^4 permuted
347 genomes [$P < 0.0001$, Figure 5]. In fact, Var[D_{BLOCK}] is larger by a factor of 3.7 in the S
348 form test (0.0122) and a factor of 6.7 *A. arabiensis* test (0.0268) than the largest value of
349 the 10^4 permuted genomes (Figure 5). These results indicate that the observation of
350 shared polymorphism between *A. gambiae* and *A. arabiensis* and between *A. gambiae* M
351 and S forms can be attributed in part to introgression.
352
353 *Contemporary introgression between* A. gambiae *and* A. arabiensis
354 We can also ask whether introgression has occurred via contemporary
355 hybridization by comparing signals of introgression with sympatric and allopatric
356 populations. If introgression has occurred recently, we expect stronger affinity among
357 sympatric relative to allopatric populations [33,59,60]. We tested whether introgression
358 between *A. gambiae* M form and *A. arabiensis* has been recent with sympatric *A.*
359 *arabiensis* versus allopatric *A. arabiensis* from Tanzania. In this case, we used the
360 standard ABBA-BABA test since we are explicitly testing a simple '1-taxon' model
361 (Figure 5). We tested for introgression using a four-taxon tree of high-coverage
362 individuals with the two *A. arabiensis* individuals as the ingroups (H1 and H2) with M
363 form as the outgroup (H3) and find a significant excess of shared derived mutations
364 between M form *A. gambiae* and sympatric *A. arabiensis* relative to allopatric *A.*
365 *arabiensis* from Tanzania ($D = -0.0542$, Block Jackknife Z-score = -13.1533, $P =$
366 1.63×10^{-39}; Table 2). Similarly, we used GOUNDRY as H3 and find evidence for
367 significant excess introgression between sympatric *A. arabiensis* and GOUNDRY ($D = -$
368 0.0441, Block Jackknife Z-score = -11.7559, $P = 6.58\times10^{-32}$; Table 2). In line with recent
369 evidence of contemporary hybridization between *A. gambiae* and *A. arabiensis* in
370 Uganda [61], our results provide strong evidence that introgression continues to occur via

371 contemporary hybridization in Burkina Faso as well that has the potential to impact the
372 evolution of both ecologically and epidemiologically relevant traits.
373
374 *Introgressed genes*
375 Introgressed haplotypes may contain genes with potentially important phenotypic
376 effects and variation along the genome in concentration of introgressed haplotypes can be
377 indicative of barriers to introgression, so we partitioned the signal of introgression to
378 identify introgressed chromosomal segments. To do so, we established genome-wide
379 significance thresholds by comparing each empirical D_{BLOCK} value to the distribution of
380 the most extreme D_{BLOCK} value from each of the 10^4 permuted genomes (Methods).
381 Since each D_{BLOCK} is polarized, we can identify windows that show significant
382 introgression with each of the sister taxa (i.e. M form and GOUNDRY) and H3 (Figure
383 6). After conservatively correcting for multiple testing genome-wide, we find significant
384 evidence of introgression between M form and both S form and *A. arabiensis*, and the
385 proportions of the genome that are represented by significant windows are 1.1% and
386 3.6% with S form and *A. arabiensis,* respectively. These introgressed chromosomal
387 blocks include 97 annotated protein coding sequences introgressed between the M and S
388 forms and 543 introgressed between M form and *A. arabiensis* (Table S4). Moreover, we
389 find strong evidence for introgression of the pericentromeric region on chromosome 2L
390 between the M and S forms, which contrasts starkly with previous suggestions that this
391 region may be a barrier to introgression and important for speciation between these taxa
392 [13,29,30]. A recent study based on SNP-genotype data identified the same signal of
393 introgression in this genomic region and attributed the high frequency of the introgressed
394 haplotype to the sharing of an adaptive insecticide resistance allele at the *kdr* locus [62].
395 One window that shows an exceptionally strong introgression signal between M
396 form and *A. arabiensis* harbors the GABA receptor gene, also known as the *resistance to*
397 *dielrin* locus because of the role of this receptor in conferring resistance to the insecticide
398 dieldrin and related insecticides [63]. Our finding that the *Rdl* locus has been
399 introgressed between *A. gambiae* and *A. arabiensis* is contradictory to previous reports
400 that these species have acquired resistance at this locus through independent but
401 convergent mutations [64].

402 We used the same approach to identify genomic blocks shared between
403 GOUNDRY and two H3 taxa, S form and *A. arabiensis* (Table 2). We find that windows
404 representing 3.2% of the GOUNDRY genome and harboring 369 protein-coding
405 sequences share a significant excess of derived mutations with the S form. In addition,
406 we find that 3.5% of the GOUNDRY genome harboring 499 protein coding sequences
407 shares a significant excess of derived mutations with *A. arabiensis*. In line with results
408 above showing evidence of recent introgression among *A. gambiae* and sympatric *A.*
409 *arabiensis* relative to allopatric *A. arabiensis*, we find that the windows with significant
410 evidence of introgression with sympatric (Burkina Faso) *A. arabiensis* cover slightly
411 more of the genome than windows with significant evidence of introgression with
412 allopatric (Tanzania) *A. arabiensis* (Table 2). Although only relatively small percentages
413 of these individual genomes show significant evidence of recent introgression, hundreds
414 of protein-coding genes are involved in each case such that homogenization of these
415 genomic regions may have large effects on phenotypic evolution, exhibited most
416 prominently by the sharing of presumably functional mutations at two insecticide loci in
417 the M form comparisons.
418 While identifying introgressed regions provides insight into the homogenizing
419 effects of hybridization, identifying genomic regions depauperate of recent introgression
420 can provide evidence for genomic barriers to introgression. Despite evidence for
421 considerable introgression on the autosomes, we find no evidence of recent introgression
422 of X chromosome sequence among any subgroups or species, which suggests a
423 disproportionately large role for the X in speciation among these taxa.
424
425 *Introgressed chromosomal inversions*
426 Partitioning the ABBA-BABA test into genomic blocks can also reveal possible
427 evidence for introgressive sharing of chromosomal inversions (Figure 6). *A. arabiensis* is
428 fixed for the inverted 2La^a arrangement that spans ~22 MB of chromosome 2L and is
429 also nearly fixed in the subgroups of *A. gambiae* (i.e. the M and S forms) in most
430 Savanna regions of West Africa [65]. However, both forms of the inversion (2La^a and
431 2La^+) are segregating in GOUNDRY [17] and we analyzed both GOUNDRY forms here.
432 Consistent with previous studies [30,65–68], we find strong evidence that the inverted

433 form has introgressed between *A. gambiae* and *A. arabiensis.* The 2L*a* region exhibits
434 extremely strong excess sharing of derived mutations when M form (2La^a) is H1,
435 GOUNDRY (2La^+) is H2, and *A. arabiensis* is H3 (2La^a). This result reflects excess
436 BABAs, or alleles shared between the 2La^a haplotypes of M form and *A. arabiensis*. In a
437 second comparison similar to the previous except that GOUNDRY is represented by its
438 2La^a haplotype, the strong and extended reduction in *D* across the region is not observed.
439 The strong reduction in *D* when the two inversion forms are contrasted provides evidence
440 consistent with the introgression of the 2La^a form of this inversion between *A. gambiae*
441 and *A. arabiensis* before the split of M form and GOUNDRY. We also find a similar
442 extended reduction in *D* on 2R coinciding with the known location of the 2R*b*
443 chromosomal inversion (Figure 6). This reduction is also consistent with introgression of
444 one form of this inversion between M form and *A. arabiensis,* although we don't know
445 the inversion karyotypes for the individuals analyzed here, so it is more difficult to
446 interpret this signal. In principle, these results are consistent with introgression, but
447 chromosomal inversions are known to suppress recombination to some extent [69–71]
448 such that ancestral haplotypes may be maintained over time violating some of the
449 assumptions our modified ABBA-BABA test. However, this confounding scenario can
450 be ruled out in the case of 2L*a* since we find a strong reduction in genetic divergence
451 (D_{xy}) between the 2La^a haplotypes of *A. arabiensis* and both M form and GOUNDRY
452 relative to surrounding genomic regions (Figure 7), which is consistent with introgression
453 but not ancestral haplotype maintenance. As a result, this region harbors genomic patterns
454 of divergence and allele sharing that deviate from the genome at large in a comparison-
455 specific fashion. Interestingly, we also find local deviations from this pattern that
456 provide evidence for recent introgression events that are independent from the inversion
457 introgression event (Figure 6). The combination of ABBA-BABA tests and genetic
458 divergence among the chromosomal forms of these taxa provide strong evidence that 2L*a*
459 has introgressed and suggestive evidence that 2R*b* may have crossed species boundaries.
460

461 *Introgression facilitates adaptation*

462 Natural selection is expected to remove most introgressed genetic material due to
463 ecological misfit or Dobzhanski-Muller Incompatibilities [72,73], especially between

464 more distant species, but some introgressed alleles may be selectively favored in the
465 recipient subgroup. As just described above, we find clear evidence for introgression of
466 the large 2L*a* inversion between *A. gambiae* and *A. arabiensis*, which is in line with
467 several previous studies [30,65,67,68]. A number of studies have associated this
468 inversion with various putatively adaptive phenotypes and shown that the frequencies of
469 2L*a* karyotypes correlate strongly with environmental factors such as aridity in West and
470 Central Africa [65,74]. Taken together, these observations indicate that the 2La^a
471 haplotype has crossed species boundaries and may have facilitated ecological and
472 physiological adaptation.
473 To identify other putatively adaptive introgression events, we examined the
474 intersection between our scans for selection and for introgression (both described above).
475 Since we cannot determine the direction of introgression robustly with the framework
476 used here, we identified selective sweeps in either the donor or recipient species and
477 found 17 selective events that may be due to selection on alleles that were originally
478 introgressed from another taxa. We find a large region on chromosome 2L (~21.5-23.5
479 MB) that is enriched for introgression of derived mutations between GOUNDRY (2La^a)
480 and *A. arabiensis* that harbors a recent sweep in *A. arabiensis* at the gene AGAP005855,
481 which has no known function.
482 The most striking example of these putative adaptive introgression events
483 involves allele sharing at the GABA receptor on 2L known to be involved in insecticide
484 resistance in both *A. gambiae* and *A. arabiensis* ([64]; Figure 8). Our scan for selection
485 points to a selective sweep in the M form subgroup at the boundary of the coding
486 sequence for this gene ($P_{gen} < 0.005$, $P_{loc} < 1.26\text{x}10^{-06}$). However, we also find evidence,
487 albeit below the significance threshold at the genome level, for selection in both
488 GOUNDRY and *A. arabiensis* (Figure 8). There is an independent signal of selection in
489 GOUNDRY ($P_{gen} > 0.05$, $P_{loc} < 2.67\text{x}10^{-04}$) upstream of the M form signal. We do not
490 find LLR values above background levels that would be consistent with recent selection
491 in *A. arabiensis*, but we find that LD is slightly elevated and nucleotide diversity reduced
492 relative to neighboring regions, consistent with a historical sweep at this locus (Figure 8).
493 It is important to note that the model implemented in SweepFinder is specifically aimed
494 at detecting very recent and complete sweeps [43], so it is not surprising that we might

495 see remnants of a putative historical sweep that is not picked up by SweepFinder.
496 Together, these results suggest that a resistance allele arose first and adaptively fixed in *A.*
497 *arabiensis*, and adaptive fixations occurred more recently in M form and GOUNDRY.
498 We find strong evidence that haplotypes involving the *Rdl* locus have been
499 introgressed among *Anopheles.* It is necessary to consider the genomic background when
500 interpreting data at the *Rdl* locus since if falls within the 2L*a* chromosomal inversion. On
501 one hand genetic divergence is relatively low across the inverted 2La^a region between *A.*
502 *gambiae* and *A. arabiensis*, since this inversion has introgressed between these species.
503 Conversely, there is exceptionally high genetic divergence between GOUNDRY 2La^+
504 chromosomes and the M form subgroup (fixed for 2La^a in this sample set) and *A.*
505 *arabiensis* (fixed for 2La^a as a species) chromosomes (Figure 7). Close inspection of
506 ABBA-BABA tests and patterns of pairwise genetic divergence at the *Rdl* region
507 suggests that both GOUNDRY and M form have independently received introgressed
508 material in this region from *A. arabiensis* in a *2*L*a*-inversion karyotype dependent fashion.
509 Specifically, the ABBA-BABA test involving the tree ((M form 2La^a, GOUNDRY 2La^a),
510 *A. arabiensis* 2La^a), *A. merus* 2La^a) indicates evidence for significant allele sharing
511 between *A. arabiensis* and M form at this locus (Figures 6 and 8). Together with the fact
512 that genetic divergence between these two taxa is low at *Rdl* relative to immediately
513 surrounding regions (Figure 8), these results provide strong evidence for introgression
514 between M form and *A. arabiensis* at the *Rdl* locus. However, we also see evidence for
515 an independent introgression event between *A. arabiensis* and the population of 2La^+
516 chromosomes in GOUNDRY. In the ABBA-BABA test based on the tree ((M form
517 2La^a , GOUNDRY 2La^+), *A. arabiensis* 2La^a), *A. merus* 2La^a), the region inside the *2*L*a*
518 inversion is almost entirely negative and nearly -1, reflecting the strong excess of allele
519 sharing between the introgressed 2La^a chromosomes from *A. arabiensis* and M form.
520 However, amidst this pattern, we see a large region of positive *D* statistics at the *Rdl*
521 locus indicating excess derived mutation sharing between *A. arabiensis* and GOUNDRY
522 2La^+ chromosomes (Figures 6 and 8). Consistent with this result, we also see a large and
523 extended reduction in genetic divergence between GOUNDRY 2La^+ and both M form
524 and *A. arabiensis*, indicating recent introgression between GOUNDRY and one of these
525 groups. However, genetic divergence is exceptionally low between GOUDNRY and *A.*

526 *arabiensis* across the entire putatively introgressed region, while divergence is reduced
527 between M form and GOUNDRY only where genetic divergence is also low between M
528 form and *A. arabiensis* (Figure 8). These observations provide evidence that an
529 introgression event occurred first from *A. arabiensis* into M form followed by a second
530 independent introgression event from *A. arabiensis* into GOUNDRY.
531 In summary, the evidence for positive selection at this locus in these groups in
532 combination with the signals of independent introgression events lead us to hypothesize
533 that a mutation conferring insecticide resistance arose and adaptively fixed first in *A.*
534 *arabiensis* and then was shared successively with the M form and GOUNDRY subgroups
535 where it underwent positive selection in these groups as well. Since insecticide pressure
536 on mosquito populations in West Africa, and thus selective pressure on this variant,
537 began in the 1950s [75], we infer that introgression of this adaptive allele among these
538 taxa has occurred within the last 60 years.
539
540 *Xag-Xbcd inversion complex is barrier to introgression*
541 We have shown above that some genomic regions have recently introgressed
542 between the genomes of members of the *Anopheles gambiae* species complex, which
543 raises the question of whether certain genomic regions contain barriers to such
544 introgression. We hypothesized that long term differences in introgression along the
545 genome will be reflected in patterns of genetic divergence, since genetic divergence will
546 be partially determined in part by selection for and against introgressed material.
547 Empirical and theoretical work has led to the hypothesis that hybrid sterility factors or
548 locally adapted genetic variants in genomic regions with restricted recombination in
549 hybrids play an especially important role in formation and maintenance of species
550 boundaries in the face of gene flow because the hitchhiking process involved in selection
551 against maladaptive introgressed variants will affect larger genomic swaths in lowly
552 recombining regions [71,76–79]. In particular, chromosomal inversions and centromeric
553 regions are thought to play a role in species formation when hybridization is ongoing,
554 although currently available empirical evidence supporting a role of centromeric regions
555 is confounded by the effects of natural selection on linked sites [32,33,35]. Empirical
556 and theoretical evidence also supports the hypothesis that the X chromosome may play an

557 important role in maintaining species boundaries ('large-X effect'; [23,25,80,81]). To
558 test for genomic barriers among taxa in the *Anopheles gambiae* species complex, we
559 asked whether sequence divergence (D_{xy}) between taxa at various distances along the
560 speciation continuum differed among genomic regions as expected under a model of
561 differential rates of introgression. However, D_{xy} is an estimator of $2\mu t + 4N\mu$, where μ is
562 the mutation rate, t is the number of generations since the species split, and N is the
563 effective population size of the ancestral population, but $4N\mu$ can vary among genomic
564 regions for reasons unrelated to differential introgression (i.e. variable mutation rate or
565 effects of natural selection on linked sites; Figure 4). Therefore, we jointly analyzed
566 nucleotide diversity (π) as a proxy for $4N\mu$ among genomic regions to avoid confounding
567 intra-population effects with differential introgression (see Methods).
568 Hybrid male sterility maps to the X chromosome in *A. gambiae-A. arabiensis*
569 crosses and two large X-linked chromosomal inversion complexes [X*ag* and X*bcd*]
570 suppress recombination on the X in hybrids of these species, so we hypothesized that less
571 frequent introgression may lead to exceptionally high sequence divergence on the X
572 relative to other genomic regions. In support of this hypothesis, we found that nucleotide
573 diversity on the X is significantly lower than on the autosomes (Mann-Whitney (hereafter
574 M-W) $P < 2.2\times10^{-16}$), but that genetic divergence is significantly higher on the X (M-W
575 $P < 2.2\times10^{-16}$; Figure 9). Moreover, genetic divergence is significantly higher in the
576 region harboring inversions (M-W $P < 2.2\times10^{-16}$) than in the surrounding chromosome.
577 Nucleotide diversity is also higher in this region relative to the centromere-proximal
578 region (M-W $P < 2.2\times10^{-16}$), where nucleotide diversity is especially low, presumably
579 due to the effects of linked selection on neutral genetic variation. While we cannot
580 formally rule out an elevated mutation rate inside the inverted region, there is no reason
581 to expect the inverted region to be more mutable. This excess genetic divergence on the
582 X is consistent with our results in the ABBA-BABA analysis above showing that the X
583 chromosome lacks evidence of recent introgression among these taxa. In general, these
584 observations are in line with previous analyses of genetic differentiation among *A.*
585 *arabiensis* and *A. gambiae*, as well as with laboratory backcrossing experiments
586 [30,37,82], indicating that introgression is particularly inhibited on the X chromosome,
587 and that the X chromosome plays a disproportionately large role in driving speciation.

588 In the comparison between M form and *A. arabiensis,* we generally find no
589 evidence for elevated divergence across the autosome, indicating a long history of
590 introgression and few genomic barriers. However, inspection of the genomic distribution
591 of nucleotide diversity and divergence (Figures 4 and 7, respectively) reveals that the
592 difference between nucleotide divergence and inter-species diversity is especially high in
593 the pericentromeric region of chromosome 3. Nucleotide diversity is reduced in this
594 region, presumably resulting in part from the effects of intra-population positive and
595 negative selection on linked sites. However, the relative increase in D_{xy}, which is not
596 affected by positive selection, suggests that migrant alleles have been selected against in
597 this region such that this region harbors both fewer shared polymorphisms as well as
598 more private mutations relative to the nearby freely recombining regions, for example.
599 This raises the question of what genetic factors might be involved in local adaptation or
600 genetic incompatibilities in this region. We find evidence for adaptive divergence at
601 multiple loci in this region, including one that involves the gene encoding *Frizzled-2*, a
602 member of the *Wnt* receptor signaling pathway involved in development [83] that could
603 underlie developmental divergence among these taxa.
604 Overall, we show that rates of introgression among *A. gambiae* and *A. arabiensis*
605 have differed dramatically among genomic regions resulting in exceptionally high
606 divergence on the X chromosome, intermediate divergence in the pericentromeric region
607 of 3L, and divergence levels similar to subgroup comparisons within *A. gambiae* across
608 most of the autosome. These data are consistent with the X chromosome playing a
609 primary and disproportionately large role in speciation among these taxa and a secondary
610 role for centromeric regions.
611
612 *Novel X-linked chromosomal inversion in GOUNDRY is a barrier to introgression*
613 The most striking sweep from the selection scan of GOUNDRY genomes above
614 covers 1.67 Mb on the X chromosome and results in nearly complete absence of
615 polymorphism across this region (Figure 4), despite read depths in the region comparable
616 to neighboring genomic regions (Figure S3). The remarkably large size of the region
617 devoid of diversity would imply exceptionally strong positive selection under standard
618 rates of meiotic recombination. For comparison, previously identified strong sweeps

619 associated with insecticide resistance span approximately 40 Kb and 100 Kb in freely
620 recombining genomic regions of *Drosophila melanogaster* and *D. simulans,* respectively
621 [84,85]. The swept region in GOUNDRY is marked by especially sharp edges relative to
622 the other footprints of selection in our data (Figures 3 and 4), implying that
623 recombination has been suppressed at the boundaries this region. Collectively, these
624 observations suggest that the swept region may contain a small chromosomal inversion,
625 which we have named X*h* in keeping with inversion naming conventions in the
626 *Anopheles* system. Notably, this pattern is virtually identical to the pattern of diversity in
627 a confirmed X-linked inversion discovered in African populations of *D. melanogaster*
628 [86]. The region in GOUNDRY includes 92 predicted protein coding sequences (Table
629 S3), including the *white* gene, two members of the gene family encoding the TWDL
630 cuticular protein family (*TWDL8* and *TWDL9*), and five genes annotated with immune
631 function (*CLIPC4, CLIPC5, CLIPC6, CLIPC10, PGRPS1*). The lack of diversity in the
632 region implies that the presumed X*h* inversion has a single recent origin and was quickly
633 swept to fixation in GOUNDRY. We estimated the age of the haplotype (see Methods)
634 inside the sweep region to be 78 years with a standard deviation of 9.15 by assuming that
635 all mutations in the region are new and estimating the mean time since the most recent
636 common ancestor of the sampled haplotypes, representing the time of the sweep. Such
637 extraordinarily recent adaptation in an otherwise old subgroup is consistent with the
638 selection pressures related to 19th and 20th century human activity such as insecticide
639 pressure or widespread habitat modification.
640 We hypothesized that this putative X-linked chromosomal inversion in
641 GOUNDRY may serve as a barrier to introgression with M form, which we showed in
642 the demographic inference above has occurred at least in autosomal regions. We used the
643 same logic as for the comparison with *A. arabiensis* to compare divergence among
644 genomic regions and find that genetic divergence among M form and GOUNDRY
645 generally scales with (M form) nucleotide diversity across the genome (Figure 9),
646 consistent with broadly neutral divergence, suggesting that the rate of introgression
647 among these subgroups is qualitatively similar genome-wide. Although the distributions
648 of genetic divergence in 10 Kb windows are significantly different among genomic
649 regions (M-W $P < 2.2\times10^{-16}$), the rank order of genetic divergence among genomic

650 regions is the same as for nucleotide diversity in these regions. This suggests that
651 differences in genetic divergence among regions can in large part be explained by
652 differences in nucleotide diversity. However, the difference between the X chromosome
653 at large and the putative X*h* inversion is slightly larger for the divergence measure than
654 for nucleotide diversity, suggesting that divergence in the putative X inversion may be
655 present that is too small to be detectable with this particular test of distributions.
656 Since we expect that the putative X*h* inversion is likely too young for measurable
657 differences in the distributions of D_{xy} to have accumulated, we tested for excess
658 divergence in the X*h* inversion using several additional more sensitive approaches.
659 When divergence along the X chromosome is explicitly scaled by the mutation rate
660 inferred from levels of polymorphism in the M form sample (D_a), the putatively adaptive
661 X*h* inversion between GOUNDRY and M form is proportionally much more divergent
662 than is the remainder of the X chromosome ($\overline{D}_a$ (X*h*) = 0.0022, $\overline{D}_a$ (non-X*h*) = 0.0013;
663 M-W $P < 4.89\times10^{-08}$; Figure S4). Although relative measures of divergence, such as D_a,
664 are known [32,33] to be confounded by reductions in nucleotide related to natural
665 selection on linked sites, for example, we believe that this analysis is robust to these
666 concerns because the comparison is among only X-linked windows and the region of
667 interest is in a region of the chromosome that is highly diverse in subgroups where it has
668 not been swept (Figure 4). We also directly tested whether the X*h* region is significantly
669 more diverged than similarly sized windows on the X by defining a new statistic, ΔD_{xy}, to
670 compare the average D_{xy} among 10 kb windows within the putative inverted region (n =
671 166) with the average D_{xy} among windows in every other 1.67 MB window on the X that
672 does not overlap with the inverted region (n = 1,499; Methods). A positive ΔD_{xy}
673 indicates that the average value inside the inversion is greater than the average value in
674 the window outside the inversion. A negative value indicates that the average value is
675 higher in the window outside the inversion. We calculated this statistic between M and
676 GOUNDRY (see Methods) and found that ΔD_{xy} was significantly positive ($P < 10^{-5}$) for
677 every window in the comparisons between M and GOUNDRY (Figure S5). Interestingly,
678 the windows immediately adjacent to the inversion show the smallest ΔD_{xy} values (Figure
679 S5), suggesting that the introgression-limiting effects of the inversion have spread beyond
680 the breakpoints. Both of these tests indicate that sequence divergence between M and

681 GOUNDRY is greater inside the putative inversion relative to the X as a whole, which is
682 not likely to reflect accumulation of many new private mutations inside the inversion, but
683 rather a greater proportion of shared polymorphisms outside the inversion consistent with
684 higher rates of introgression outside the inversion.

685 The effects of the putative X*h* inversion on the GOUNDRY X chromosome as a
686 barrier to gene flow may extend beyond the boundaries of the inverted region, resulting
687 in very recent accumulation of divergence on the X chromosome that is confounded in
688 our analysis with inherent evolutionary differences between the autosomes and the X
689 chromosome. One approach to estimate differences in divergence is to compare
690 divergence between a focal pair of subgroups (GOUNDRY and M form) to divergence
691 between one of the focal groups and an outgroup (GOUNDRY and S form) in order to
692 scale by differences in mutation rate and other population genetic parameters among
693 regions. This approach estimates what is known as Relative Node Depth (RND =
694 D_{GM}/D_{GS}, where subscripts G, M, and S indicate GOUNDRY, M form, and S form
695 respectively), and a higher RND indicates greater divergence between the focal groups
696 [87]. We find that RND is 0.7797 on the autosomes and 0.8058 on the X, indicating
697 lower relative genetic divergence between GOUNDRY and M form on the autosomes
698 than on the X. To explicitly test whether such a pattern could be obtained under a pure
699 split model with no gene flow, we obtained expected values of Relative Node Depth
700 (RND) by assuming a phylogeny where M form and GOUNDRY form a clade with S
701 form as the outgroup and using coalescent theory under this model to calculate expected
702 RND (Methods).

703 Our analytical results support the hypothesis that D_{GM} is downwardly biased
704 relative to D_{GS} on the autosomes as a result of higher rates of gene flow on the autosomes
705 relative to the X. We find that under some parameter combinations (Figure 10), RND
706 decreases with increasing effective M-GOUNDRY effective population size, which could
707 result in a smaller RND value on the autosomes since the autosomes should have an
708 effective size at least as big as the X. However, most parameter combinations suggest
709 that this pattern is unexpected (i.e. most regions of the curves predict that RND should
710 increase with increasing effective population size), and the estimate for the ancestral
711 effective size of M-GOUNDRY we obtained in a separate demographic analysis above

712 suggests that these subgroups exist in a parameter space where the RND function is
713 consistently increasing with increasing effective sizes. Taken together with the
714 demographic inference, the above results suggest that, after having initially diverged
715 ecologically approximately 100,000 years ago, genomic barriers to introgression have
716 only been established within the last 100 years, presumably owing to the accumulation
717 and extended effects of locally adapted loci or genetic incompatibility factors within large
718 swept X*h* region on the GOUNDRY X chromosome
719 Collectively, the data show that nucleotide-diversity corrected divergence is
720 higher inside the putative inverted region, the inverted region as a chromosomal segment
721 is the most diverged of all segments of the same size on the X chromosome, and the X
722 chromosome as a whole is more diverged among GOUNDRY and M form relative to the
723 autosomes. The most parsimonious explanation for these patterns is that, although very
724 few new mutations have accumulated inside of X*h* since its origin less than 100 years ago,
725 ongoing introgression among M form and GOUNDRY has led to a greater density of
726 shared polymorphism and therefore lower sequence divergence in non-inverted regions of
727 the X chromosome relative to the inversion, especially distal to the inversion breakpoints.
728 These results lead us to conclude that while cladogenesis of GOUNDRY and M form
729 ~100 kya by other means established some degree of temporally fluctuating reproductive
730 isolation, the recently derived putative X*h* inversion now serves as a genomic barrier to
731 introgression, and the effects of selection against migrant haplotypes have begun to
732 extend to linked sites outside the inversion breakpoints.
733
734 *Collinear X chromosomes are barriers to introgression*
735 The X chromosomes of the *A. gambiae* M and S forms are collinear with that of *A.*
736 *merus*, so recombination is not expected to be suppressed in contemporary or historical
737 hybrids among these groups. To test whether sequence divergence is exceptionally high
738 on the X relative to the autosomes in these comparisons despite the lack of inverted
739 regions, we compared genetic divergence and nucleotide diversity among genomic
740 regions as described above. We find that genetic divergence between M and S scales
741 with nucleotide diversity across the genomic regional classes (Figure 9), consistent with
742 few long term barriers to introgression. However, inspection of chromosomal

743 distributions of divergence reveals a slight elevation in genetic divergence in a
744 chromosomal region where nucleotide diversity is low relative to the rest of the
745 chromosome (~16-20 MB; Figures 4 and 7) and recombination is known to be
746 proportionally rare in M-S hybrids [88]. In contrast, genetic divergence between *A.*
747 *gambiae* (M form) and *A. merus* does not scale with nucleotide diversity (Figure 9).
748 Nucleotide diversity is lower on the X than on the autosomes (M-W $P < 2.2\times10^{-16}$), but
749 genetic divergence is significantly higher on the X (M-W $P < 2.2\times10^{-16}$). Moreover, we
750 find that pericentromeric and telomeric regions are significantly more diverged than the
751 freely recombining autosome (M-W $P < 2.2\times10^{-16}$), but nucleotide diversity is
752 significantly lower in these regions (M-W $P < 2.2\times10^{-16}$). Although we could not test for
753 recent introgression among *A. gambiae* and *A. merus* using the ABBA-BABA tests above,
754 this pattern of excess divergence in some genomic regions is in contrast to what we
755 would expect under a divergence in allopatry model and is more consistent with historical
756 introgression in autosomal regions among *A. gambiae* and *A. merus*, likely in the
757 evolutionary period following the initial species split. Overall, these findings are
758 consistent with a large role for the X in driving speciation even among subgroups with
759 collinear X chromosomes and pericentromeric regions diverging secondarily.
760

761 **Discussion**

762 We present one of the first analyses of complete genome sequences from multiple
763 species and subgroups from the *Anopheles gambiae* species complex, including the newly
764 discovered cryptic GOUNDRY subgroup of *A. gambiae*, and demonstrate the power of
765 multi-subgroup analysis to comprehensively dissect the mosaic pattern of divergence,
766 adaptation, and introgression across the genome at base-pair resolution. Using average
767 genetic distance across whole genomes, we show clear monophyly of each subgroup and
768 species and confirm that *A. gambiae* is more closely related to *A. arabiensis* than to *A.*
769 *merus.* We also show that, in contrast to initial findings [17], the newly discovered
770 GOUNDRY subgroup of *A. gambiae* falls genetically within the molecular forms of *A.*
771 *gambiae* and is not an outgroup. Our demographic analysis suggests that GOUNDRY
772 has existed for approximately 100,000 years and represents a recent example of the
773 frequent speciation dynamics in *Anopheles* that we think is common.

774 We also show that the genomes of these species and subgroups have been
775 subjected to extensive natural selection in recent evolutionary history, likely related to
776 local adaptation to distinct ecological niches, and we provide the first base-pair resolution
777 maps of recent selective events. We annotated each putative selective event with an
778 associated gene and provide many actionable candidates for follow-up experiments.
779 Interestingly, the selection maps are almost entirely distinct, reflecting ongoing adaptive
780 divergence among these taxa. Despite extensive overlap in their geographic distributions,
781 it is likely that these taxa differ in important ecological and physiological phenotypes
782 [8,14]. For example, recent evidence shows that *A. gambiae* M form and *A. arabiensis*
783 use distinct strategies for surviving the dry season in West Africa and re-establishing
784 local populations after the first rain [89]. In contrast, some ecological pressures are
785 shared among these species as evidenced by the apparent selective benefit of sharing the
786 2L*a* chromosomal inversion. Moreover, the presence of insecticides like dieldrin has
787 made insecticide resistance alleles globally beneficial among the species studied here.
788 Interestingly, we show that the M form has experienced significantly more positive
789 natural selection than GOUNDRY. Since we know so little about GOUNDRY, it is
790 difficult to speculate on the fundamental differences underlying this discrepancy. One
791 possibility though is that the M form subgroup, having expanded into a presumably much
792 larger and ecologically varied geographic distribution, continues to face new adaptive
793 challenges, while GOUNDRY may exist in a relatively small, specialized and
794 ecologically homogeneous environment with fewer novel selective pressures. Substantial
795 field and functional studies focused on the ecology, mating behavior, and physiology of
796 these subgroups would help to explain the selective pressures underlying their divergence.
797 We show that the physical locations of putatively introgressed genomic segments
798 coincide with the location of selective sweeps in one or both of the exchanging taxa and
799 suggest that some of these selective events were driven by advantageous alleles shared
800 through introgression. Adaptive introgression has been documented in other systems as
801 well as *Anopheles* [6,62], and appears to be an important evolutionary channel for
802 functional genetic variants that can facilitate phenotypic adaptation. In addition to the
803 compelling case of introgression of a *Rdl* insecticide resistance allele (Figure 8), we
804 identify 17 other candidate sweeps that may have involved an introgressed beneficial

805 allele. We note that identifying selective sweeps and identifying introgressed haplotypes
806 are each challenging tasks in their own right, each with their own confounding issues.
807 Indeed, these two evolutionary processes have the potential to confound the population
808 signal of the other resulting in both false negative and false positive selection scan and
809 introgression results, although the test for selective sweeps used here was shown to be
810 robust to some plausible population genetic scenarios that included both migration and
811 selection [43]. Further exploration is needed to understand fully how these processes
812 interact and develop more sophisticated methods to explicitly test for adaptive
813 introgression.

814 The nature of species boundaries and divergence among species and subgroups of
815 *Anopheles* is not well understood, despite considerable effort [13,29,30,35,90]. The
816 original observation of centromeric 'islands of divergence' between the M and S
817 molecular forms led to the conclusion that gene flow among these species was extensive
818 and ongoing, but these islands remained differentiated for reasons presumably related to
819 their role in speciation among these taxa [29]. Subsequent studies observed especially
820 high levels of differentiation in the 'islands of divergence' against a background of
821 differentiation across most of the genome, leading to the conclusion reproductive
822 isolation among these taxa is more advanced than previously thought [13,30]. However,
823 these studies did not determine the relative roles of genetic drift, natural selection on
824 linked sites, and introgression in determining the observed levels of differentiation.

825 In this study, we use various methods that distinguish between these population
826 genetic processes and provide insight into whether introgression has occurred recently
827 and which genomic regions may be barriers to introgression. We show that chromosomal
828 segments have recently introgressed between the subgroups of *A. gambiae* as well as
829 between *A. gambiae* and *A. arabiensis* (Figure 6). Importantly, our analysis captures the
830 physical locations and genomic coverage of introgressed regions. However, we note that
831 these introgressed haplotypes were identified in a set of representative individuals and
832 therefore likely reflect a combination of stochastic drift processes as well as deterministic
833 selective forces following introgression. The specific composition of introgressed
834 haplotypes along the genome among individuals in each population likely differs to some
835 extent. Over the evolutionary time scales, we expect that natural selection will play a

836 large role relative to drift in determining the composition of ancestry along the genome
837 resulting in a mosaic of introgressed segments along the genome, as has been observed
838 for Neanderthal ancestry in modern humans [25]. To test for heterogeneity in ancestry
839 that may reflect the action of natural selection and therefore barriers to introgression, we
840 measured genetic sequence divergence among taxa and showed that the long-term
841 differential introgression has indeed shaped patterns of divergence among species and
842 subgroups of *A. gambiae* and *A. arabiensis* (Figure 9). Our divergence analysis points to
843 an exceptional history of autosomal introgression among *A. gambiae* and *A. arabiensis*
844 that has created dramatic differences along the genome in genetic affinity among these
845 species (Figure 9). Combined with the evidence that introgression has transferred
846 globally beneficial alleles, such as insecticide resistance alleles, it is clear that
847 introgression has played an important evolutionary role in this species complex and can
848 have large effects on public health.
849 We show that the X chromosome is the most diverged and harbors the least
850 evidence for introgression, suggesting that it is likely a barrier to introgression in multiple
851 *Anopheles* species (Figures 6 and 9). In some cases, chromosomal inversions seem to
852 play an important role. The observation that the X chromosome plays a
853 disproportionately large effect in driving speciation ('large-X') is in line with studies
854 from many organisms ranging from *Drosophila* to mammals [23,25,81,91], but a
855 unifying explanation for this pattern has yet to emerge. Multiple hypotheses have been
856 posited to explain the underlying evolutionary mechanisms underlying this pattern
857 including the 'faster-X' hypothesis, sex ratio meiotic drive, and mis-regulation of X-
858 linked genes in males (reviewed in [92]). The 'faster-X' hypothesis posits that X-linked
859 loci adapt faster than autosomal loci because X-linked recessive mutations are exposed to
860 selection in males that have only a single X chromosome, resulting in faster accumulation
861 of hybrid sterility factors [23,80]. Another popular hypothesis involves sex ratio meiotic
862 drive where species-specific sex ratio distorter suppressors are disrupted in hybrids
863 causing hybrid sterility [93,94]. A third hypothesis to explain this pattern is that gene
864 expression dosage compensation of X-linked genes is mis-regulated in hybrids, causing
865 sterility in hybrids [95]. Although data have accumulated in *Drosophila* allowing more

866 detailed speculation about the mechanisms underlying the large-X effect [92], more data
867 are needed to fully understand this pattern in *Anopheles*.
868 Interestingly, we also show that pericentromeric regions also harbor especially
869 high levels of divergence among more distantly related species pairs, suggesting a model
870 where the X chromosome is the first barrier to gene flow and these regions differentiate
871 secondarily. Implicit to this model is that different regions become barriers to
872 introgression in stages while introgression continues to homogenize the remaining
873 genomic regions as in standard speciation with gene flow models [96]. This pattern is
874 most apparent in the comparison between *A. gambiae* and *A. merus* (Figures 7 and 9) and
875 provides strong evidence that both *A. arabiensis* as well as *A. merus* diverged from *A.*
876 *gambiae* while continuing to hybridize at a non-negligible rate in at least the early stages
877 of speciation. In the case of *A. arabiensis*, we show that contemporary hybridization with
878 *A. gambiae* continues despite strong barriers to genetic introgression across the X
879 chromosome. Similar patterns of elevated divergence in lowly-recombining
880 pericentromeric and telomeric regions have also been observed in comparisons between
881 *Drosophila* species [26,28,97,98], but our results are the first demonstration of this
882 pattern of excess sequence divergence in *Anopheles*. It is important to note that our
883 results, especially the observation of excess divergence between *A. gambiae* and *A. merus*
884 in low-diversity pericentromeric regions, are robust to the issues confounding previous
885 observations of high differentiation in these regions, since our results derive from
886 absolute measures of divergence instead of relative divergence measures that are highly
887 sensitive to other population genetic processes including natural selection [32,33,35].
888 The ABBA-BABA test has become a preferred method for detecting introgression,
889 but there are several caveats and concerns relating to both the standard test as well as our
890 modified D_{BLOCK} test. First, it is possible that the signals of introgression we have
891 detected are not from the species used in the test, but in fact we have detected
892 introgression from an un-sampled, or 'ghost,' species. Durand et al. [56] showed that
893 such introgression can affect the results of ABBA-BABA tests. The presence of 'ghost'
894 *Anopheles* species hybridizing with the species sampled here is certainly a possibility and
895 could impact some of our results. However, the possibility of introgression from 'ghost
896 taxa' does not change our conclusion that introgression continues among *Anopheles*

897 species shaping patterns of divergence regardless of exactly which subgroup is the donor.
898 Second, a recent analysis showing that a similar block-wise test of introgression lacked
899 power to identifying introgressed haplotypes [99] raises questions about the robustness of
900 our D_{BLOCK} analysis. However, there are two important differences between our approach
901 and the one evaluated by Martin et al. [99]. While these authors implemented the test
902 based on constant-sized physical windows of the genome, we used constant-sized blocks
903 of informative sites that varied in their physical size. This is an important distinction
904 because our approach controls for the amount of information in each window, while the
905 number of ancestry informative sites is bound to vary greatly among physical windows in
906 the approach of Martin et al., which is likely to impact the sensitivity of this approach.
907 The second difference between the approaches is that our approach explicitly controlled
908 for linkage disequilibrium in the data. We chose block sizes of 500 informative sites
909 because this resulted in average physical window sizes of ~ 250 to 350 kb (see Methods),
910 which allowed each block to be divided into 100 segments larger in size than the
911 expected distance that LD decays to background levels in this system. As a result, we
912 believe that our approach is a robust approach for identifying introgressed genomic
913 regions and is not likely to suffer from the same concerns raised by Martin et al. [99].
914 Our data point to a model where the speciation process is reticulate with offshoot
915 populations and ancestral species remaining connected by periodic hybridization of
916 differential efficiency across the genome that sometimes facilitates adaptation (Figure 11).
917 Although the region of the genome that we identify with statistical evidence for recent
918 introgression is relatively small (~3%) among the individuals tested here, the effect on
919 the autosome is large over evolutionary time. The relationship between *A. gambiae* and
920 *A. arabiensis* provides an extreme example of variation in the permeability of species
921 boundaries, with evidence of autosomal introgression over the evolutionary history of
922 these two species, including the sharing of an insecticide resistance allele in the last 60
923 years. In stark contrast to this pattern, the X chromosomes of these species remain highly
924 diverged relative to the autosomes, implying that introgression has been ineffective on
925 the X and that the X chromosome is the primary driver of speciation. A recent
926 independent comparison between the *A. arabiensis* and *A. gambiae* genomes supports our

927 conclusion that introgression is a major factor in the evolutionary history of the *A.*
928 *gambiae* species complex [100].
929 In sum, our results suggest that species and subgroups in the *A. gambiae* species
930 complex comprise a diffuse and interconnected gene pool that confers access to
931 beneficial genetic variants from a broad geographic and environmental range. Such
932 genetic affinity has important implications for malaria control. On one hand, transgenes
933 may spread more easily among subgroups and species of malaria vectors, which could
934 reduce the effort needed to reach and manipulate all populations involved in disease
935 transmission. On the other hand, our analysis suggests that ecological specialization of
936 new subgroups is common and eradicating a population without removing the ecological
937 niche may simply open the niche for a new population to invade with potential a new and
938 unknown consequences for disease transmission. In both cases, our results underscore
939 the complexities involved in vector control on a continental scale.
940

941 **Materials and Methods**

942 ***Mosquito samples***
943 Mosquito sample collection and species/subgroup identification was previously
944 described for the M form (recently re-named *A. coluzzii* [8]), GOUNDRY, and *A.*
945 *arabiensis* samples [17]. Briefly, larvae and adults were collected from three villages in
946 Burkina Faso in 2007 and 2008 (Table S5). Larvae were reared to adults in an insectary,
947 and both field caught adults and reared adults were harvested and stored for DNA
948 collection. One *A. gambiae* S form individual was also included in this study. This
949 sample was collected indoors as an adult in the village of Korabo in the Kissidougou
950 prefecture in Guinea in October 2012. Individuals were typed for species, molecular
951 form and *2*L*a* karyotype using a series of standard molecular diagnostics [101–103]. All
952 M form and *A. arabiensis* samples are $2La^{a/a}$ homokaryotypes and the S form sample
953 typed as a heterokaryotype ($2La^{a/+}$). Eleven of the twelve GOUNDRY samples typed as
954 $2La^{+/+}$ homokaryotype, but one sample (GOUND_0446) typed as a $2La^{a/a}$
955 homokaryotype.
956

957 ***DNA extractions and genome sequencing***
958 DNA was extracted from female carcasses using standard protocols. Genomic
959 DNA samples were diluted in purified water and sent to BGI (Shenzhen, China) for
960 paired-end sequencing on the Illumina HiSeq2000 platform. We sequenced 12 females
961 from the GOUNDRY subgroup of *A. gambiae*, 10 females for the M form subgroup of *A.*
962 *gambiae*, and 9 female *Anopheles arabiensis*. Sequence reads were filtered if they 1)

963 contained more than 5% Ns or polyA structure, 2) contained 20% or more low quality
964 (Q<20) bases, 3) contained adapter sequence, or 4) contained overlap between pairs.
965 After quality filtering, each sample was represented by an average of 47.91 million reads
966 with an expected insert size of 500 base pairs, except two samples that were sequence to
967 deeper read depth (Table S1).
968 In a separate sequencing effort, we also sequenced one individual *A. gambiae* S
969 form female. The sequencing library was prepared using the Nextera kit (Illumina Inc.,
970 San Diego, CA) according to manufacturer's specifications and sequenced on the
971 Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA) at the University of
972 Minnesota Genomics Center core facility for a total of 125.88 million reads.
973 All Illumina reads newly generated for this study have been submitted to the
974 Short Read Archive at NCBI under accession IDs ranging from XXXXXX-XXXXX.
975 We obtained publicly available *Anopheles merus* and *Anopheles arabiensis* short
976 read Illumina data from NCBI SRA. We downloaded *A. merus* accessions ERR022713-8
977 that were generated and deposited by The Sanger Center. Individual accessions are
978 paired-end 76 bp read fastq files representing whole genome sequence from single
979 individual *A. merus* individuals collected Kenya sequenced on the Illumina GAII
980 platform. In addition, we downloaded *A. arabiensis* accession SRX377561, representing
981 an individual collected in Minepa, Tanzania [38]. These data are paired-end 100 bp reads
982 generated on the Illumina HiSeq2000 platform. This sample is the individual sequenced
983 to 'high coverage' from the Marsden et al. Tanzania population sample.
984
985 ***Illumina short read alignment and subgroup specific references***
986 At the time of this analysis, the only genome reference sequence with scaffolds
987 mapped to chromosomes for any *Anopheles* mosquito was the *Anopheles gambiae* PEST
988 AgamP3 assembly [[104]; vectorbase.org], so we used this reference for mapping reads
989 from all groups and species in this study. Since an unknown subset of the short reads
990 generated for this experiment were expected to be substantially diverged from the PEST
991 reference, we conducted short read alignment iteratively in two steps. First, we
992 conducted paired-end mapping of all available reads to the *Anopheles gambiae* PEST
993 reference genome [104] using the BWA *mem* algorithm [[105]; bio-bwa.sourceforge.net]
994 with default settings except applying the –M flag, which marks shorter hits as secondary.
995 We then generated new 'Subgroup-Specific' references (hereafter SPEC) for *Anopheles*
996 *gambie* M form, *Anopheles gambiae* GOUNDRY, *Anopheles arabiensis*, and *Anopheles*
997 *merus* separately. To do so, we combined sequence reads for each Subgroup and
998 generated a pseudo-consensus sequence for each Subgroup. We used the software
999 package ANGSD (version 0.534; [44]; popgen.dk/wiki) to generate a read pileup, count
1000 reads and bases at each site, and identify the major allele at all sites. Then for every site
1001 that was covered by 1) at least 4 sequence bases that pass filtering, 2) fewer than a
1002 Subgroup-specific threshold number reads (mean read depth plus two standard

1003 deviations), and 3) a major allele that was segregating at a frequency of at least 0.5 across
1004 all reads, we assigned the new reference base to the major allele and to 'N' otherwise.
1005 After alignment to the PEST reference, the mean total read depth for the
1006 autosomal arms was 136.14 for *A. arabiensis*, 136.63 for *A. gambiae* M form, 157.42 for
1007 *A. gambiae* GOUNDRY, and 52.94 for *A. merus*. Mean total read depth on the X
1008 chromosome was 124.27 for *A. arabiensis*, 145.63 for *A. gambiae* M form, and 168.75
1009 for *A. gambiae* GOUNDRY, and 44.37 for *A. merus*. The proportion of unknown bases
1010 ('N's) increased in the new references (0.12, 0.11, 0.05, 0.14 for the autosomes of *A.*
1011 *arabiensis*, M form, GOUNDRY, and *A. merus*, respectively, and 0.22, 0.05, 0.06, 0.25
1012 for the X) relative to PEST (0.02 for autosomes, 0.04 for X). The differences between
1013 Subgroups in read depth and proportion of sites missing data can be attributed in part to
1014 differences in the number of individuals sequenced (n = 9,10,12, 6 for *A. arabiensis*, M
1015 form, GOUNDRY, *A. merus*, respectively). However, the differences between
1016 autosomes and the X likely reflect difficulties in mapping divergent reads to the PEST
1017 reference. The proportion of sites missing data were higher on the X relative to the
1018 autosomes for alignments of data from *A. arabiensis* and *A. merus* to both the PEST
1019 reference and the Subgroup specific references (Table S6). Two other studies found a
1020 similar discrepancy when mapping these species to the *A. gambiae* PEST reference
1021 [37,38], indicating that it is not unique to our pipeline and likely reflects the relatively
1022 higher divergence on the X that proves to be a higher barrier to read mapping.
1023 Following the generation of new reference sequences for *A. gambiae* M form,
1024 GOUNDRY, *A. arabiensis,* and *A. merus*, short read datasets for each Subgroup were
1025 then aligned back to the new subgroup specific (hereafter SPEC) reference using the
1026 BWA *mem* algorithm with default settings. The S form individual was aligned to the M
1027 form reference sequence. Local realignment around indels was then performed for each
1028 subgroup separately using GATK [106]. Duplicates were removed using the SAMtools
1029 [107] *rmdup* function.
1030 Although the read mapping rate was consistently lower for the SPEC reference
1031 relative to rates when mapping to PEST, the proportion of mapped reads with a quality
1032 score of 20 increased for all groups except the M form Subgroup where it decreased
1033 slightly. The reduction in mapping rate when using the SPEC reference reflects the
1034 greater number of ambiguous bases ('N') in the SPEC reference relative to PEST and
1035 thus a smaller mapping target. Importantly, the proportion of mapped reads with quality
1036 score of 20 on the X chromosome increased from 0.6961 (PEST) to 0.7775 (SPEC) when
1037 mapping the *A. merus* data to the PEST and SPEC references. A similar increase was
1038 also observed for *A. arabiensis*, with the proportion of Q20 reads increasing from 0.7604
1039 (PEST) to 0.8241 (SPEC), indicating that the Subgroup specific reference is especially
1040 helpful for mapping X-chromosome reads. As discussed above, mapping non-*A.*
1041 *gambiae* reads to the PEST X chromosome is expected to be difficult. Mapping biases

may lead to underestimates of diversity and genetic divergence, but such downward biases do not change the main conclusions regarding *A. arabiensis* and *A. merus*.

Data Filtering

The data were filtered in two steps. First, we used SAMtools to generate a pileup and genotype likelihoods for each site. We calculated a series of alignment and read statistics and then applied a series of filters to obtain a set of sites considered reliable for downstream analysis. Filters were applied using the SNPcleaner Perl Script from the ngsTools package [46]. Those filters are as follows.

1. *Read distribution among individuals*: No more than one individual is allowed to have fewer than 2 reads covering the site.
2. *Maximum read depth*: The site must not be covered by more than 350, 350, and 400 reads for M form, *A. arabiensis*, and GOUNDRY, respectively, in each combined sample.
3. *Mapping Quality*: Only reads with a BWA mapping quality of at least 10 were included.
4. *Base Quality*: Only bases with Illumina base quality of 20 or more were included.
5. *Proper pairs*: Only reads the mapped in the proper paired-end orientation and within the expected distribution of insert lengths were included.
6. *Hardy-Weinberg proportions*: Expected genotype frequencies were calculated for each variable site based on allele frequencies based on genotypes called with SAMtools. Any site with an excess of heterozygotes were considered potential mapping errors and excluded.
7. *Heterozygous biases*: Sites with heterozygous genotype calls were evaluated with SAMtools for several biases using exact tests. If one of the two alternative alleles was biased with respect to the read base quality (minimum $P=1x10^{-100}$), read strand (minimum $P=1x10^{-4}$), or distance from the end of the read (minimum $P=1x10^{-4}$), the site was excluded.

The next filter was intended to exclude regions of the genome where short-read alignment may be compromised, such as low-complexity and regions with a high concentration of 'N' bases in the reference. The goal was to avoid edge effects and regions where mapping becomes ambiguous. To identify such regions, we scanned each Subgroup specific reference calculating the proportion of Ns in every 100-basepair window. All sites that fell within windows that contained 50 or more Ns were excluded. Since short read alignment is likely to be unreliable in highly repetitive genomic regions such as heterochromatic regions, we also excluded regions that have been identified as heterochromatic in *A. gambiae*, including both pericentric and intercalary heterochromatic regions [108] to be conservative in our analyses.

1082 A final list of sites-to-exclude was compiled for each subgroup-specific reference
1083 that included any site excluded by any of the above filters. This resulted in a different
1084 number and set of sites available for downstream analysis for each Subgroup (Table S7).
1085 In addition to these excluded sites, we identified a series of regions that exhibited
1086 exceptionally high nucleotide diversity even after front-end filtering. These sites were
1087 not excluded from ANGSD analyses (see below), but were excluded from all population
1088 genetic analyses (Table S7).
1089

1090 ***Site-frequency spectrum inference***

1091 Population genetic inference from next-generation sequencing data was
1092 performed using a statistical framework implemented in the software package Analysis of
1093 Next-Generation Sequencing Data (ANGSD, [44], popgen.dk/wiki). Much of the BAM
1094 manipulation and read-filtering functionality implemented in SAMtools is also
1095 implemented in ANGSD and several functions were utilized in all of the following
1096 analyses with ANGSD. Minimum map quality and base quality thresholds of 10 and 20
1097 were used. Probabilistic realignment for the computation of base alignment quality
1098 (BAQ) was enabled throughout with a downgrading coefficient of 50. Reads that were
1099 not aligned as proper pairs or had bitwise flags above 255 were excluded.
1100 The first step in the pipeline was to infer the site-frequency spectrum (SFS)
1101 directly from genotype likelihoods estimated from the read data. The global SFS was
1102 estimated in two steps. The first step was to obtain a maximum likelihood estimate of
1103 per-site allele frequencies by calculating multi-sample genotype likelihoods using the
1104 SAMtools model ([107]; -GL 1 in ANGSD) and estimating the minor allele frequency
1105 using the –realSFS 1 function in ANGSD [44]. Files containing the per-site estimates of
1106 allele frequencies were then used as input for optimization of the global SFS across all
1107 sites using a BFGS optimization algorithm implemented in the optimSFS program within
1108 ANGSD. Unfolded spectra were obtained by including ancestral polarization assigned by
1109 a synthetic ancestral sequence described below. Sites without ancestral assignment were
1110 excluded from SFS estimation. The optimized global SFS was estimated for each
1111 chromosomal arm separately and the results from 2R, 3L, and 3R were averaged to obtain
1112 an autosomal SFS. Autosomal and X chromosome site frequency spectra were obtained
1113 for M form, GOUNDRY, and *A. arabiensis* separately. These spectra were used as priors
1114 for downstream analyses within ANGSD and are shown in Figure S6.
1115

1116 ***Ancestral sequence***

1117 We generated a synthetic ancestral sequence by assigning ancestral and derived
1118 alleles based on their presence in our M form, GOUNDRY, *A. arabiensis*, and *A. merus*
1119 population samples. We first identified all alleles segregating at every site in the genome
1120 of each group by calculating genotype likelihoods (-GL 1) and estimating the minor allele
1121 frequency (–doMaf 2 in ANGSD). We used all variable sites except singletons. If all

1122 groups shared the same allele at a site or if all groups were missing data at a site, the
1123 conserved allele or an N was included in the new ancestral sequence, respectively. If
1124 some groups lacked data at a site and only one allele was observed in the remaining
1125 groups with data, this allele was included in the new sequence. Lastly, if an allele was
1126 found to be segregating in at least three of the four groups and the alternate allele was
1127 found in only one or two groups, the major allele was chosen as the ancestral allele. This
1128 approach is conservative but also flexible in that it requires an allele to be found in either
1129 1) three different species or 2) both subgroups of *A. gambiae* and one outgroup species.
1130 It is possible that gene flow among these groups could introduce bias here, but it is
1131 difficult to distinguish gene flow from lineage sorting at this stage in the analysis. We
1132 compared the SFS spectra inferred using this ancestral sequence to those obtained using
1133 *A. merus* as the ancestral sequence and found a larger enrichment of high frequency
1134 derived alleles in the *A. merus*-derived spectra as would be expected if ancestral mis-
1135 specification were common (not shown).
1136
1137 ***Genotype calling***
1138 Genotypes were called in two steps at all sites that passed filtering. For both the
1139 population samples (M form, GOUNDRY, *A. arabiensis, A. merus*) as well as the *A.*
1140 *gambiae* S form individual, genotype likelihoods were calculated using the SAMtools
1141 model as described above. Then genotype posterior probabilities were calculated for each
1142 individual at each site. For the population samples, posterior probabilities were
1143 calculated using maximum likelihood estimates of allele frequencies as a prior [109]. For
1144 the single individual samples, a uniform prior was assumed for calculation of the
1145 posterior probabilities. For all samples, diploid genotypes were called only at sites with
1146 posterior probabilities of 0.9 or greater. Genotypes for GOUNDRY sample were
1147 obtained using a slightly different approach since this Subgroup is partially inbred (see
1148 below).
1149
1150 ***Inbreeding analysis***
1151 *Estimating inbreeding coefficients*
1152 Initial estimates of the global site frequency spectrum (SFS) in GOUNDRY
1153 produced distributions of allele frequencies that deviated substantially from standard
1154 equilibrium expectations as well as from those observed in the *A. gambiae* M form and *A.*
1155 *arabiensis* groups. Most notably, the proportion of doubletons was nearly equal to that of
1156 singletons in *A. gambiae* GOUNDRY. This observation is consistent with widespread
1157 inbreeding in the GOUNDRY subgroup. We tested this hypothesis in two ways, with the
1158 goals of both characterizing the pattern of inbreeding in this subgroup as well as
1159 obtaining inbreeding coefficients for each individual that could then be used as priors for
1160 an inbreeding-aware genotype-calling algorithm. We used the method of Vieira et al.
1161 [110], which estimates inbreeding coefficients in a probabilistic framework taking

1162 uncertainty of genotype calling into account. This approach is implemented in a program
1163 called ngsF (github.com/fgvieira/ngsF). ngsF estimates inbreeding coefficients for all
1164 individuals in the sample jointly with the allele frequencies in each site using an
1165 Expectation-Maximization (EM) algorithm [110]. We estimated minor allele frequencies
1166 at each site (-doMaf 1) and defined sites as variable if their minor allele frequency was
1167 estimated to be significantly different from zero using a minimum log likelihood ratio
1168 statistic of 24, which corresponds approximately to a P value of 10^{-6}. Genotype
1169 likelihoods were calculated at variable sites and used as input into ngsF using default
1170 settings. For comparison, we estimated inbreeding coefficients for *A. gambiae* M form,
1171 GOUNDRY, and *A. arabiensis* using data from each chromosomal arm separately
1172 (Figure S7).
1173 We inspected the distribution of homozygous regions likely related to inbreeding
1174 dynamics and found that these regions are randomly distributed and different among
1175 individuals. A more thorough analysis and discussion of these patterns, possible causes,
1176 and possible implications will be presented in a separate manuscript.
1177
1178 *Recalibrating the site-frequency spectrum and genotype calls*
1179 We used the inbreeding coefficients obtained above for the GOUNDRY sample
1180 as priors to obtain a second set of inbreeding-aware genotype calls and an updated global
1181 SFS. We used ANGSD to make genotype calls as described above. However, in this
1182 case, we used the –indF flag within ANGSD, which takes individual inbreeding
1183 coefficients as priors instead of the global SFS [110]. Similarly, we used the inferred
1184 inbreeding coefficients to obtain an inbreeding-aware global SFS. We estimated the
1185 global SFS from genotype probabilities using –realSFS 2 in ANGSD, which is identical
1186 to –realSFS 1 [44] except that it uses inbreeding coefficients as priors for calculations of
1187 posterior probabilities [110].
1188
1189 ***Genetic divergence and neighbor-joining tree***
1190 *D_{xy} calculations*
1191 We measured absolute genetic divergence as the average number of pairwise
1192 differences between alleles from different subgroups, or D_{xy} [111]. We calculated D_{xy} as

1193 $$D_{xy} = \frac{h+2H}{2L}$$

1194 where h is the number of sites where one or both individuals carry heterozygous
1195 genotypes, H is the number of sites where the two individuals are homozygous for
1196 different alleles, and L is the number of sites where both individuals have called
1197 genotypes. As with our estimates of within-population diversity, we excluded all sites
1198 within 200 bp of the outer coordinates of annotated protein coding sequences in the
1199 AgamP3.8 gene set (vectorbase.org). For subgroups where we have sequences from
1200 multiple individuals (M form, GOUNDRY, *A. arabiensis, A. merus*), we estimated D_{xy}

1201 calculated from the number of pairwise differences between each individual from every
1202 other subgroup and then averaged across comparisons for each subgroup to obtain an
1203 average D_{xy} value for each pair of subgroups. For comparisons made to the S form, D_{xy}
1204 was calculated between the individual S form sequence and the individual sequence with
1205 the highest average read depth from each of the other subgroups.
1206 For chromosomal plots (Figure 7) and boxplot analysis (Figure 9), each
1207 chromosomal arm was divided into 10 kb windows and D_{xy} was calculated across each
1208 window. Windows with fewer than 200 sites with data for comparison were excluded.
1209 Chromosomal curves were LOESS-smoothed using a span of 1% and degree of 2 with
1210 the loess.smooth function in R [112]. For the boxplot analysis, chromosome 2L was
1211 excluded to avoid the large effects of the *2L a* inversion. Comparisons were made
1212 between categories using a Mann-Whitney test implemented in the wilcox.test function in
1213 R [112].
1214
1215 *Neighbor-joining trees*
1216 We visualized average genetic relatedness across the genome among individuals
1217 and species using neighbor-joining distance trees based on pairwise values of D_{xy}. We
1218 calculated D_{xy} across chromosomes 2R, 3R, 3L, and X, but excluded 2L to avoid the
1219 effects of the 2L*a* inversion. D_{xy} was calculated across intergenic sites at least 200 bp
1220 from protein coding sequences in euchromatic genomic regions. Neighbor-joining trees
1221 were generated using the *ape* [113] package in R and drawn using the Geneious software
1222 (www.geneious.com). We used a bootstrap analysis to estimate confidence in the major,
1223 internal node topology (i.e. species relationships) by dividing the dataset into 1 MB
1224 physical windows and sampling with replacement 1,000 new datasets. Distance
1225 neighbor-joining trees were generated for each bootstrap replicate and compared to the
1226 tree made from the true data to determine consistency of major clade placement.
1227
1228 ***Demographic inference***
1229 *2D-spectra and model fitting*
1230 We fit population historical models to the two-dimensional site frequency
1231 spectrum for GOUNDRY and M form subgroups using the program *dadi* [42]. We
1232 obtained the unfolded 2D spectrum for chromosomal arms 2R, 3R and 3L by first
1233 estimating allele frequencies at each site for M form and GOUNDRY separately, then
1234 calculating posterior probabilities of allele frequencies at each site using the global SFS
1235 priors, and finally calculating the 2D spectrum. Posterior probabilities were calculated
1236 using the program sfstools distributed with ANGSD. The program 2Dsfs included in the
1237 software package ngsTools (https://github.com/mfumagalli/ngsTools; [46]) was used to
1238 calculate the 2D spectra with -maxlike set to 1. Per-site allele frequencies were estimated
1239 for each subgroup as described above, except GOUNDRY was handled differently to
1240 accommodate the inbreeding signal. Since GOUNDRY is partially inbred, we randomly

1241 sampled one allele at each position in the genome to be included for analysis thereby
1242 eliminating the effects of inbreeding within a diploid genome on estimates of allele
1243 frequency. This resulted in a sample size of 12 chromosomes for GOUNDRY. Initial
1244 testing suggested that optimization in *dadi* performed better when sample sizes were
1245 equal, so we projected the M form sample from 20 chromosomes down to 12 to make the
1246 spectra symmetric using the project function with *dadi*. As for before, all
1247 heterochromatic regions and low-complexity regions were excluded. We also excluded
1248 all sites within 200 bp of exons annotates in the *A. gambiae* PEST AgamP3 gene set. We
1249 calculated the 2D spectrum for each of the chromosomal arms separately and summed
1250 them to represent the autosome (Figure 2). Chromosome 2L was excludee to avoid the
1251 effects of the large segregating 2L*a* inversion. The final spectrum contained 61,254,546
1252 sites, 3,189,751 of which were variable in our samples.
1253 We defined the demographic function to include three epochs (Figure 2). As with
1254 all models in *dadi*, the model begins with an ancestral population and proceeds forward
1255 in time. In the first epoch, the ancestral population is split into two populations. Each of
1256 the three epochs is defined by five parameters: 1) the duration of the epoch, 2) the
1257 effective size of M form, 3) the effective size of GOUNDRY, 4) the rate of migration
1258 from GOUNDRY into M form, and 5) the rate of migration from M form into
1259 GOUNDRY. We defined the demographic function using the following syntax:
1260

```
def Three_epoch_complx4(params,ns,pts):
        """
        T1 = duration of first epoch
        T2 = duration of second epoch
        T3 = duration of third epoch

        nuM1 = size of pop1 (M form) in first epoch after split
        nuM2 = size of pop1 (M form) in second epoch
        nuM3 = size of pop1 (M form) in second epoch
        nuG1 = size of pop2 (GOUNDRY) in first epoch after split
        nuG2 = size of pop2 (GOUNDRY) in second epoch
        nuG3 = size of pop2 (GOUNDRY) in second epoch

        mMG1 = migration from GOUNDRY into M form during first epoch
        mMG2 = migration from GOUNDRY into M form during second epoch
        mMG3 = migration from GOUNDRY into M form during third epoch
        mGM1 = migration from M form into GOUNDRY during first epoch
        mGM2 = migration from M form into GOUNDRY during second epoch
        mGM3 = migration from M form into GOUNDRY during third epoch

        """
        T1,T2,T3,nuM1,nuM2,nuM3,nuG1,nuG2,nuG3,mMG1,mMG2,mMG3,mGM1,mGM2,mGM3=pa
        rams;
        ## specify grid
        xx = dadi.Numerics.default_grid(pts)
        ## ancestral equilibrium pop
        phi = dadi.PhiManip.phi_1D(xx);
        ## split pops
        phi = dadi.PhiManip.phi_1D_to_2D(xx,phi);
        ## Epoch1
        phi = dadi.Integration.two_pops(phi,xx,T1,nu1=nuM1,nu2=nuG1,m12=mMG1,m21=mGM1);
        ## Epoch2
        phi = dadi.Integration.two_pops(phi,xx,T2,nu1=nuM2,nu2=nuG2,m12=mMG2,m21=mGM2);
        ## Epoch3
        phi = dadi.Integration.two_pops(phi,xx,T3,nu1=nuM3,nu2=nuG3,m12=mMG3,m21=mGM3);
        fs = dadi.Spectrum.from_phi(phi,ns,(xx,xx));
        return fs
```

We used the multinomial likelihood approach for evaluating the fit between the data and the 2D spectrum predicted by the model, and conducted inference in two steps. First, we set large boundaries for each parameter and conducted 466 replicate optimizations, perturbing the starting parameters for each run by a factor of 1.5 and allowing 300 iterations per optimization. This step allowed us to manually evaluate the likelihood surface for each parameter and identify regions of the parameter space with high likelihoods for further optimization. Second, to ensure that the optimizations

1308 reached best-fit parameter values, we narrowed the parameter search boundaries and
1309 conducted a second set of 1,016 optimizations, perturbing the starting values by a factor
1310 of 0.5 and allowing 750 iterations. Models that include strong bottlenecks are extremely
1311 slow to optimize and not likely to be biologically plausible for systems like mosquitoes
1312 with typically large census sizes, so we set lower boundaries of 0.01 for all subgroups
1313 sizes for computational efficiency. After ensuring that the optimized values were not
1314 hitting boundaries and parameters were converging over replicate optimizations, we
1315 chose the optimization with the highest likelihood as the point estimate. To obtain an
1316 estimate of the ancestral effective population size, we estimated the population scaled
1317 mutation rate ($4N_A\mu$) that fits the data best using the optimal_sfs_scaling function in *dadi*
1318 and solved for N_A, assuming a mutation rate of $\mu = 7.0 \times 10^{-9}$. The spontaneous mutation
1319 rate for mosquitoes has not been estimated to date, so we used a mutation rate estimated
1320 from *Drosophila melanogaster* as our closest approximation for the mutation rate in
1321 *Anopheles*. Multiple estimates of the spontaneous mutation rate for *D. melanogaster* are
1322 available from studies differing in their approach, and the estimates vary by an order of
1323 magnitude from 2.8×10^{-9} to 1.1×10^{-8} [114–118]. We chose an intermediate value of
1324 7.0×10^{-9}.
1325
1326 *Parameter confidence intervals*
1327 We estimated confidence intervals for each parameter using a nonparametric
1328 bootstrap approach. We generated bootstrapped genomes by first concatenating
1329 chromosomal arms 2R, 3R, and 3L, and then dividing this 'genome' into physical 200 kb
1330 regions. These regions were sampled with replacement to obtain 100 new genomes with
1331 lengths equal to that of the true genome. Since repeating the two-step optimization
1332 described above would be computationally infeasible, we conducted optimizations for
1333 each bootstrap replicate with a different approach. For each replicate, we conducted 100
1334 optimizations with wide boundaries, allowing 750 iterations and perturbing starting
1335 values by a factor of 0.5. We used the maximum likelihood values obtained for the true
1336 genome as pre-perturbation starting values. Then for each bootstrap replicate, we chose
1337 the optimization with the maximum likelihood. Approximate 95% confidence intervals
1338 were calculated for each model parameter as the mean of all replicate values +/- 1.96
1339 standard deviations of all replicate values. All bootstrap optimizations were run on the
1340 Stampede compute cluster at University of Texas, Austin through the support of the
1341 XSEDE program. Optimized parameter values and confidence intervals are reported in
1342 Table 1.
1343
1344 ***Fixed differences***
1345 We identified fixed differences between M form and GOUNDRY using the
1346 following criteria. A substitution was inferred if
1347 1) At least (n-2) individuals were represented at a site.

1348 2) The site was considered monomorphic based on allele frequencies inferred
1349 using ANGSD being either less than 1/2n or greater than (1-(1/2n)).
1350 3) GOUNDRY and M form harbor different major alleles.
1351
1352 ***Linkage disequilibrium***
1353 We measured LD using Haploview [119] applied to genotype and SNP calls made
1354 using ANGSD as described above for M form, GOUNDRY, and *A. arabiensis*. LD was
1355 calculated independently on each chromosomal arm and comparisons were made only
1356 between SNPs separated by 10 kb or less. For all analyses, we used the r^2 statistic
1357 generated by Haploview. For computational tractability, we reduced the number of SNP-
1358 by-SNP comparisons by randomly sampling comparisons down to either 1%, 10% or
1359 20% of the total number of comparisons depending on the total number of comparisons
1360 for each subgroup. To obtain LD decay curves, we binned r^2 values for each subgroup
1361 based on physical distance in the PEST reference, and then averaged within each bin and
1362 plotted as a function of physical distance (Figure S8). We combined measures for
1363 chromosomal arms 2R, 3R, and 3L to obtain an autosomal curve while avoiding the
1364 effects of the 2L*a* inversion. We estimated a curve for the X chromosome separately. In
1365 all three subgroups, LD decays to background levels within several hundred basepairs on
1366 both the autosomes and on the X, consistent with previous analyses in these species
1367 [38,120,121]. Interestingly, background levels of LD differ substantially among
1368 subgroups. Background LD is slightly higher in *A. arabiensis* than in M form, which is
1369 consistent with a small effective population size and perhaps some greater degree of
1370 population substructure in *A. arabiensis* than in M form. GOUNDRY exhibits very high
1371 levels of background LD, but this is due to the high levels of inbreeding-related
1372 homozygosity. In all cases, LD decay curves suggest bootstrapping can be conducted on
1373 200 kb regions with confidence.
1374 We generated background LD chromosomal plots (Figure S2), by dividing each
1375 chromosome into 10kb windows, identifying all SNPs inside each window, and taking
1376 the average r^2 value for all comparisons after thinning (see above) between SNPs inside
1377 the window and SNPs greater than 1 kb but less than 10 kb away on the PEST reference
1378 sequence. Only windows with at least 100 comparisons were included. Background LD
1379 was measured for M form and *A. arabiensis* only since inbreeding has skewed measures
1380 of LD in GOUNDRY. LOESS-smoothed curves were generated using a span of 1% and
1381 a degree of 2 in the loess.smooth function in R [112].
1382
1383 ***Scan for recent positive selection***
1384 *Genome scans*
1385 We conducted full genome scans for recent complete selective sweeps in *A.*
1386 *gambiae* M form and GOUNDRY as well as *A. arabiensis* using SweepFinder, an
1387 implementation of the composite-likelihood test that compares the likelihood of allele

frequencies and their physical distribution along a chromosome under a sweep model to the likelihood of the data, given a neutral spectrum of allele frequencies [43]. To assemble input data, we estimated unfolded allele frequencies from genotype likelihoods using the method of Kim ([109]; -doMafs 5 in ANGSD) with ancestral state assigned based on the ancestral sequence constructed as described above. We converted ANGSD output to SweepFinder input files and ran SweepFinder using a grid size that corresponds starting points every 1kb. Instead of using SweepFinder to estimate the global site frequency spectrum, we provided SweepFinder with spectra estimated independently using realSFS in describing SFS inference above.

Neutral simulations in GOUNDRY and M form

To establish critical thresholds for test statistics, we simulated population samples of 50 kb M form and GOUNDRY haplotypes using coalescent simulations under the demographic model estimated using *dadi* (described above). We used the scaled mutation rate, θ**,** estimated in the *dadi* inference. Recombination rates are not known in this system, so we conducted a limited set of coalescent simulations under a range of scaled recombination rates. LD decay curves from simulated datasets were compared to LD decay curves in the M form data. We found that increasing the scaled recombination rate by a factor of 5, 6, or 7 relative to the scaled mutation rate produced LD decay curves comparable to the M form data. These factors correspond to per site scaled recombination rates of 0.0177, 0.0212, and 0.0247 compared to a scaled mutation rate of 0.0035. To allow for some variation in recombination rate, we used a mixed distribution of scaled recombination rates with probability of 0.5, 0.25, and 0.25 for 0.0177, 0.0212, and 0.0247, respectively. We conducted all simulations using MaCS ([122]; version 0.5d) with –h equal to 1 to take full advantage of the Markovian approximation of the coalescent implemented in this method. Consistent with the analytical correspondence between the approximate coalescent in MaCS and the full coalescent, simulations have shown that data generated under this approximation compare quite well to data generated under standard full coalescent models [122]. Accordingly, the two dimensional site frequency spectrum calculated from the simulated data closely resembles both the spectra from our data and the maximum likelihood model inferred by *dadi* (Figure S9). The LD decay curves from the simulated data also follow a similar decay rate as the real data (Figure S9).

We used the following MaCS command for autosomal loci:

```
./macs 44 50000 -h 1 -t 0.00353506596129744 -r 0.0247454617290821 -I 2 20 24 -n 1 12.34317 -n 2
0.782097 -m 1 2 0.02965848 -m 2 1 1.9131664 -en 0.19650155 1 5.748946 -en 0.19650255 2 0.1944363 -
em 0.19650355 1 2 0.2238564 -em 0.19650455 2 1 0.0005993458 -en 1.68994305 1 0.01000816 -en
1.68994405 2 0.01010736 -em 1.68994505 1 2 0.9882836 -em 1.68994605 2 1 7.880658 -ej 2.20323955 2
1 | ./msformatter > sim.out
```

1429
1430 For autosomal loci, we simulated 792,800 50 kb regions corresponding to 200
1431 complete autosomes and applied SweepFinder to the M form and GOUNDRY simulated
1432 haplotypes separately. We used a grid size of 50 for SweepFinder to be consistent with
1433 the 1kb scale used for the real data. The background site frequency spectrum was
1434 estimated by SweepFinder for each 50kb region and used for that region. To simulate
1435 data resembling the X chromosome, we simulated 150,000 50 Kb regions under the same
1436 demographic model, but applied a ¾ correction to the ancestral effective size inferred
1437 from *dadi* and adjusted simulation parameters accordingly.
1438 To determine critical thresholds for a 50 Kb region (hereafter referred to as the
1439 per-locus threshold), we found the maximum log likelihood ratio (LLR) found in each 50
1440 kb region for M form and GOUNDRY separately (Figure S10). Then, to obtain per-locus
1441 *P* values, the maximum LLRs for all M form and GOUNDRY selective sweep windows
1442 in the real data were compared to the corresponding distribution of LLRs calculated from
1443 simulated regions. To obtain a genome wide critical threshold corrected for multiple
1444 testing, we assembled autosomal 50 Kb regions and X-linked 50 kb regions into synthetic
1445 218.23 Mb genomes and identified the highest LLR within each simulated genome
1446 (Figure S10). After identifying windows with LLR values significant at the genome level,
1447 we then identified the highest point in each string of continuous significant windows to
1448 find the 'peak of the peak'.
1449 We found that peaks were unexpectedly clustered across the genome, suggesting
1450 that some peaks may be fractured evidence of the same selective event. Since
1451 SweepFinder has substantially more power when monomorphic sites are included in the
1452 analysis [123], we re-ran SweepFinder using both monomorphic and variable sites within
1453 clustered regions. We identified clusters of peaks where significant peaks were found
1454 within 100 Kb of each. To determine whether multiple peaks separated by windows with
1455 non-significant LLR values were actually part of a single large peak, we asked whether
1456 more than four continuous windows with LLR values less than 20 separated the peaks in
1457 the 'all-sites' analysis. SweepFinder is prohibitively slow when monomorphic sites are
1458 included in the analysis, so we could not conduct this version across the whole genome.
1459 In some cases, this high-resolution analysis revealed a single continuous peak. We
1460 present one representative example in Figure S11 where non-significant windows in the
1461 variable-site-only analysis separate a number of significant peaks, but the all-site analysis
1462 includes only a single low-LLR window separating the peaks. To be conservative,
1463 clusters of peaks satisfying these criteria were collapsed into a single peak and the
1464 window with the highest LLR was used for annotation. In other cases, however, the all-
1465 sites analysis did not provide evidence a single continuous peak, despite an overall
1466 peaked shape of the significant windows in the region (Figure S11). The final set of
1467 selective sweeps after collapsing clusters is presented in Table S2.
1468

1469 *More selection in GOUNDRY or M form?*
1470 Selective events may not always be identified as single events by SweepFinder
1471 and thus our count of independent selective events may be biased, we chose to quantify
1472 the proportion of megabases (n = 230) in each genome that harbors at least one selective
1473 sweep (i.e. one significant window). To determine whether the difference between M
1474 and GOUNDRY is statistically significant, we randomly permuted subgroup assignments
1475 of each selective sweep in our data 10^4 times and conducted the same analysis. We could
1476 not make the same comparison with *A. arabiensis* because a different method was used to
1477 identify significant signals of selection for that species.
1478
1479 *Identifying peaks in* A. arabiensis
1480 We did not estimate a demographic model for *A. arabiensis*, so we used a
1481 different approach to identify credible signals of recent selective sweeps. We searched
1482 the LLR surface for peaks where the peak included at least two adjacent windows with
1483 LLRs in the top 0.1% genome wide. The 0.1% threshold corresponds to a value of 15.88
1484 in this dataset. After finding all such cases, we used the window with the highest LLR in
1485 the cluster to represent the peak. This approach identifies 34 distinct peaks across the
1486 autosomes and none on the X chromosome.
1487
1488 *Annotating peaks*
1489 For each selective sweep, we found the protein coding sequence most closely
1490 associated with the highest point on the LLR peak. If the maximum LLR window fell
1491 within a gene in the Agam3.8 gene set, information from this gene was included in the
1492 annotation database. If the maximum LLR window fell outside known genes,
1493 information from the gene with the nearest 3' or 5' boundary was used for the peak. To
1494 annotate each sweep, we downloaded information from Vectorbase.org for AgamP3.8
1495 gene set from the *A. gambiae* PEST genome, including gene names, membership in
1496 ImmunoDB [124], best GO annotations, and best KOG annotations. Information for all
1497 selective sweeps can be found in Tables S2 and S3.
1498
1499 ***Nucleotide diversity***
1500 To enrich our data set for neutrally evolving sequence, we excluded sites within
1501 200 bp of a coding sequence annotated in the AgamP3.8 gene set for the *A. gambiae*
1502 PEST reference available on VectorBase.org from our estimates of nucleotide diversity.
1503 If only a single individual was available for a group (*A. gambiae* S form), nucleotide
1504 diversity was estimated simply as the fraction of sites with genotypes that were called as
1505 heterozygous. If a population sample was available (GOUNDRY, M form and *A.*
1506 *arabiensis, A. merus*), we estimated nucleotide diversity directly from the read data using
1507 a maximum likelihood approach based on posterior probabilities of per-site allele
1508 frequencies [45]. The method is implemented under the –doThetas function within

1509 ANGSD and takes a global SFS as a prior to calculate posterior probabilities of allele
1510 frequencies. We chose to use Tajima's π statistic [125] as our estimate for nucleotide
1511 diversity. For all downstream analyses, nucleotide diversity was calculated as an average
1512 within 10 kb non-overlapping windows. The comparison of nucleotide diversity between
1513 genomic regions (see below) included only 10 kb windows containing at least 500 sites
1514 from 2R, 3L, 3R, and X chromosomal arms. The distribution of nucleotide diversity was
1515 compared between genomic regions using a Mann-Whitney test implemented in the
1516 wilcox.test function in R [112].
1517
1518 ***Introgression analysis and* D *statistics***
1519 We used permutation and jackknife analyses to conduct two ABBA-BABA-based
1520 tests for introgression. For the set of tests, we divided the genome into blocks of 500
1521 informative sites (i.e. ABBAs and BABAs). These genomic blocks were then divided
1522 into 100 segments of 5 informative sites. We chose this block size because this number
1523 of sites corresponded to a physical size of ~100 x L_{LD}, where L_{LD} is the physical distance
1524 at which point linkage disequilibrium decays to background levels. As a result, we could
1525 then divide the genomic blocks into 100 segments that would be larger in physical size
1526 than the L_{LD}. Since introgressed haplotypes will be highly correlated over exceptionally
1527 high physical distances relative to ancestral haplotypes, we expect that the size of
1528 introgressed haplotypes would exceed L_{LD} while ancestral haplotypes would not. For the
1529 tests involving S form and *A. arabiensis* as the H3 taxon, the mean genomic block length
1530 was ~250 Kb and ~350 Kb, respectively. Since L_{LD} is approximately 200 bp in this
1531 system (Figure S8), and segments within the blocks were approximately 2.5 or 3.5 Kb in
1532 size, segments exceeded L_{LD} by a factor of more than 10x. These analyses were
1533 conducted using pseudo-haploidized genomes from the individuals from each
1534 population/species with the highest short-read coverage. In general, sites involved in
1535 calculating *D* are not entirely independent due to linkage disequilibrium, so permutation
1536 and jackknife analyses are necessary to properly test for significance. We have used
1537 these corrected tests for significance in the following ways.
1538 For the first test, we calculated the *D* statistic calculated for each genomic block
1539 (hereafter D_{BLOCK}) and the variance among D_{BLOCK}, hereafter Var[D_{BLOCK}], to test for an
1540 excess in variance among genomic blocks that would be consistent with the presence of
1541 correlated genomic segments (haplotypes) with shared derived mutations. Under the null
1542 hypothesis of no introgression, Var[D_{BLOCK}] among true genomic blocks derives largely
1543 from relatively small ancestral haplotypes such that random permutation of segments
1544 among blocks will not affect the variance among blocks. If the genome contains
1545 introgressed haplotypes that are larger than the segments, Var[D_{BLOCK}] will be larger
1546 when these haplotypes are intact in the empirical data and smaller after random
1547 permutation that dissolves correlations among segments. We calculated Var[D_{BLOCK}]
1548 among genomic blocks in the empirical data. Then, for each comparison, we permuted

1549 the internals ~2.5 Kb segments among genomic blocks and re-calculated Var[D_{BLOCK}] for
1550 each permuted genome. Then Var[D_{BLOCK}] from the empirical data was compared to the
1551 distribution of Var[D_{BLOCK}] from permuted genomes to ask whether variance among
1552 genomic blocks is higher in the empirical data where true correlations remain intact
1553 relative to the permuted genomes where variance in *D* will be driven largely by
1554 segregating ancestral haplotypes.
1555 For the second test, establish confidence intervals for estimates of the D_{BLOCK}
1556 statistic in order to identify individual genomic blocks with significant evidence for
1557 introgression. To do so, we conducted block jackknife analyses within each genomic
1558 block [55] by dropping each genomic segment within a given block in turn and
1559 recalculating D_{BLOCK}. We calculated 95% confidence intervals for each genomic block
1560 using variance estimated from this jackknife procedure. These confidence intervals are
1561 presented as ribbons in Figure 6.
1562 For the third test, we established genome wide thresholds corrected for multiple
1563 testing in order identify genomic blocks exceeding these thresholds consistent with recent
1564 introgression. We conducted the permutation of segments within blocks procedure as
1565 above, but for each permuted genome, we calculated D_{BLOCK} and retained the maximum
1566 and minimum values of D_{BLOCK}. To determine whether any individual true genomic
1567 blocks showed evidence of significant excess sharing of derived alleles, we established
1568 95% critical thresholds (Table 2) from this permutation procedure and compared the
1569 value of D_{BLOCK} among true blocks. These genome-wide critical thresholds are presented
1570 as dashed lines in Figure 6.
1571 To determine whether introgression has been very recent between *A. arabiensis*
1572 and either M form or GOUNDRY, we compared the proportion of the genome in
1573 windows with significant *D* values between sympatric *A. arabiensis* from Burkina Faso
1574 and allopatric *A. arabiensis* from Tanzania [38]. Since the standard assumption of
1575 introgression with only one of the two sister taxa holds for this test, we calculated the
1576 standard error of *D* for each comparison using the block jackknife approach and used a Z-
1577 test to assess significance [55,56].
1578
1579 ***Comparing genetic divergence among genomic regions***
1580 To test hypotheses related to the role of recombination in determining the
1581 genomic architecture of reproductive isolation in this system, we divided the genome into
1582 regions based on expected levels of recombination in hypothetical hybrids. A fine-scale
1583 genetic map is not yet available for *Anopheles* mosquitoes, but it has been shown in
1584 *Drosophila* that recombination rates approach zero within several megabases on each
1585 side of the centromere and also near the telomeres [126,127]. Although patterns of
1586 linkage disequilibrium (LD) are also affected by processes other than local meiotic
1587 recombination rates, estimated recombination rate should give a rough approximation of
1588 expected LD across the genome. In fact, patterns of LD have been used to define genetic

1589 maps in some vertebrates and correspond approximately to genetic maps based on
1590 experimental crosses [128,129]. We measured background LD (see description of LD
1591 measurement above for details) in our M form and *A. arabiensis* samples, taking average
1592 r^2 values within 10 kb physical windows across the genome. We found that LD was
1593 relatively constant across the genome except for large increases near the autosomal
1594 centromeres and smaller increases near the telomeres (Figure S2). Based on this pattern
1595 and the assumption that recombination rates in *Anopheles* correspond approximately to
1596 the *Drosophila* genetic map, we defined several broad recombinational categories for
1597 analysis. We first defined the 'Pericentromeric-Telomeric' regions of the autosomes to
1598 be all windows within 10 MB on either side of the centromere or within 1 MB from the
1599 telomere. It should be noted that we assumed that the starting and ending coordinates of
1600 the PEST reference chromosomal sequences were reliable indicators for distance from
1601 centromeres and telomeres. Unless a chromosomal inversion was present, all remaining
1602 regions on the autosomes were assigned to the 'Freely Recombining' category. For the
1603 comparison between *A. gambiae* and *A. merus,* we assigned all windows inside of the
1604 2R*op* chromosomal inversion complex to the 'Autosomal-Inversion' category. We used
1605 the outer coordinates for 2R*o* and 2R*p* breakpoint regions estimated by Kamali et al. [40].
1606 The X chromosome was categorized for each comparison, according to species-
1607 specific conditions. We did not define a general 'Pericentromeric-Telomeric' category
1608 for two reasons: 1) We did not observe an increase in LD in the euchromatic regions near
1609 centromeres and telomeres (Figure S2) similar to increases observed on the autosomes.
1610 2) In *Drosophila* [126,127], the pericentromeric reduction in recombination affects a
1611 relatively small region on the X relative to the autosomes, and we have excluded a large
1612 heterochromatic region around the centromere that likely encompasses the effected
1613 region in *Anopheles.* For the comparison between *A. gambiae* and *A. merus*, and the
1614 comparison between the M and S molecular forms, the entire euchromatic region on the
1615 X was considered 'Freely Recombining' since no inversions differentiate these groups.
1616 For the comparison between *A. gambiae* and *A. arabiensis*, we assigned the entire
1617 euchromatic region of the X as 'X-Inversion', since these species are differentiated across
1618 nearly 75% of the entire chromosome and introgression rates have been estimated to be 0
1619 in laboratory crosses [82]. For the comparison between GOUNDRY and the molecular
1620 forms of *A. gambiae*, the entire euchromatic X was categorized as 'Freely-Recombining'
1621 except for the region spanning 8.47 MB to 10.1 MB, which was categorized as 'X-
1622 Inversion'. As described below, we were not able to identify inversion breakpoints for
1623 the GOUNDRY inversion, but these coordinates correspond to the outer boundaries of
1624 the region with reduced nucleotide diversity (Figure 4).
1625 To identify regions that are barriers to introgression in the *Anopheles gambiae*
1626 species complex, we compared genetic divergence among the four genomic categories
1627 using the following logic. Since genome-wide variation in mutation rate and the effects
1628 of linked selection could also lead to genomic variation in divergence among species

1629 even in the absence of introgression, we jointly analyzed absolute genetic divergence
1630 (D_{xy}) and nucleotide diversity (π) in 10 kb non-overlapping windows and asked whether
1631 differences in genetic divergence among genomic regions are observed that cannot be
1632 explained by differences in nucleotide diversity (where diversity approximates the effects
1633 of linked selection or variation in mutation rate). D_{xy}, is an estimate of $2\mu t + 4N_e\mu$ where
1634 N_e is the effective size of the ancestral population, μ is the mutation rate, and t is the
1635 divergence time. π which provides an estimate of $4N_e\mu$ where N_e is the effective size of
1636 the current population. By jointly considering these, we can make comparisons among
1637 genomic regions that differ substantially in $4N_e\mu$ such that differences in D_{xy} largely
1638 reflect differences in $2\mu t$. We note that this analysis assumes that estimates of $4N_e\mu$ from
1639 current populations (or in most cases, average of the estimates from the two subgroups)
1640 are reliable estimates of $4N_e\mu$ in the ancestral population, and we believe this to be a
1641 reasonable assumption given the recency of radiation of the *A. gambiae* species complex.
1642 Under scenarios with gene flow among diverging subgroups or species, t will be smaller
1643 in regions that have introgressed, so regional barriers to gene flow are expected to have
1644 especially large values of t, and therefore values of D_{xy} that are larger than regions with
1645 similar $4N_e\mu$ but more introgression. In contrast, under a model of divergence in
1646 allopatry with no introgression, t is approximately equal among genomic regions such
1647 that divergence should be determined largely by $4N_e\mu$, and thus correlate well with
1648 nucleotide diversity, even if elevated. We made comparisons among genomic regions
1649 using this framework and assuming that most freely recombining autosomal regions will
1650 have been introgressed at some point in the history of divergence since high rates of
1651 recombination inhibit associations between barriers to introgression (e.g. hybrid sterility
1652 factors) and surrounding chromosomal regions. Therefore, we compared other genomic
1653 regions to freely recombining autosomal regions to with respect to both divergence and
1654 nucleotide diversity to ask whether these other regions harbor excess divergence
1655 consistent with less introgression in these regions. Comparisons of distributions of
1656 genetic divergence and nucleotide diversity were made using the Mann-Whitney test in R
1657 [112].
1658 For this analysis, we used only 10 Kb windows at least 600 sites with data for the
1659 taxa involved. We chose to use a focal species comparison to minimize the number of
1660 comparisons for concise presentation, but we obtain similar results in other comparisons
1661 (GOUNDRY vs. *A. arabiensis,* for example) since most variation in divergence depends
1662 on variance in coalescence in the ancestral population. We note that estimates of
1663 nucleotide diversity in GOUNDRY are not reliable estimates of the ancestral population
1664 since GOUNDRY is partially inbred and has the large swept region on the X, so we used
1665 estimates of nucleotide diversity from the M form for this analysis.
1666
1667 ***Estimating the age of the GOUNDRY X-linked selective sweep***

1668 We estimated the number of generations since the fixation of the haplotype inside
1669 the putative X*h* inversion region in GOUNDRY in the following way. We assumed that
1670 no new mutations increased to have a frequency > 50% after the selective sweep. Under
1671 this assumption, we can estimate the mean time since the most recent common ancestor
1672 of the sampled haplotypes, representing the time of the sweep, by assembling a consensus
1673 sequence among the haplotypes to represent the common ancestor and counting the
1674 number of mutations separating each haplotype from the ancestor. Then we can calculate
1675 the number of mutations divided by the total haplotype size and divide this number by the
1676 mutation rate to obtain the number of generations separating haplotype from the
1677 consensus. We summed the number of mutations across all diploid sequences, counting
1678 1 for genotypes called heterozygotes and 2 for homozygotes, and divided by 2 times the
1679 total number of sites passing filters. This grand total was then divided by the mutation
1680 rate to get the number of mutations. This approach can be simplified and stated as the
1681 following where the age to the most recent common ancestor is

$$\frac{\left[\dfrac{\sum_{j=1}^{s} f_j}{2\sum_{i=1}^{n} L_i}\right]}{\mu}$$

1682

1683 where f_j is the derived allele count at variable site j, s is the number of variable sites, n is
1684 the number of diploid individuals, L_i is the number of sites with called genotypes passing
1685 filters for individual i, and μ is the mutation rate, again assumed to be $7.0\text{x}10^{-9}$ (See
1686 demographic inference section above).
1687 Genotypes for 12 diploid GOUNDRY individuals were queried across a region
1688 consisting of 1,372,538 sites after initial filtering. A total of 1,052 variable sites were
1689 recorded prior to additional read depth filtering. Since this estimate is highly sensitive to
1690 the number of mutations included, we estimated the number of generations using a series
1691 of filters varying in stringency. First, we removed clusters of variable sites, since these
1692 are likely to represent errors, by dividing the inversion region into 50 bp windows and
1693 excluding any windows with more than two variable sites (14 out of 33,478 windows).
1694 This resulted in a total of 989 variable sites across 1,367,970 sites in total. We also
1695 excluded sites with read depth in the top 5% for each individual according to each
1696 individual's read depth distribution. Then we counted mutations for each individual
1697 using different thresholds such that sites passed filter if that individual was covered by at
1698 least 6, 8, 9, 10, 11, or 12 reads. We found that too few sites passed filtering to be
1699 informative with a 15-read cutoff. After conservatively filtering the data to minimize the
1700 effect of errors, the majority of the remaining variance in our estimates derives from
1701 variance in the number of mutations per haplotype. If we assume that the number of

1702 mutations per haplotype is Poisson distributed, we can calculate the standard deviation
1703 for read-filter point estimates by taking the square root of the point estimate, as the
1704 variance for a Poisson is equal to its mean.
1705 Age estimates varied by 2.5 fold depending on the minimum read depth filter
1706 (Figure S12), decreasing from 1,975 generations to 776 generations with 6 and 12-read
1707 minimum cutoffs, respectively. However, excluding the 6 read cutoff, the remaining five
1708 age estimates varied by less than 1.4 fold, ranging from 776 to 1,079. Standard
1709 deviations largely overlap among estimates from the 8-12-read filters, suggesting that the
1710 estimates are quite similar in this range. If we examine how the depth filters affect
1711 mutation counts per site for each individual separately, the 6-read filter results in an
1712 increase in mutations for all individuals (Figure S12). The mutation counts per site
1713 remain relatively flat for the remaining filters with only a slight systematic downward
1714 trend with the 12-read filter. The estimated age drops substantially between the 6 and 8-
1715 read filters and continues to decrease slowly from there. This strongly suggests that the
1716 proportion of variable sites that are genotype-calling errors is relatively high for this
1717 region when low read-depth minimums are used. To determine whether genotype-calling
1718 errors may be inflating the estimated number of mutations, we calculated the minor allele
1719 count at each variable site for each of the read-depth thresholds and see that, when a
1720 minimum of 6 reads is used, there is an excess of doubletons compared to higher read
1721 depths (Figure S13), consistent with a high proportion of true heterozygous sites being
1722 called as homozygous for the minor allele. The ratio of doubletons to singletons
1723 decreases substantially when the minimum number of reads is increased to 8, implying
1724 that many erroneous homozygous-alternative genotype calls are converted to
1725 heterozygous with this threshold (Figure S13). This improvement does not fully explain
1726 the drop in the number of observed mutations, however.
1727 We can leverage the extra sequencing invested in one individual to gauge the
1728 effect of our filtering process. Our sequencing effort was not evenly distributed among
1729 individual GOUNDRY samples since we sequenced one individual (GOUND-0446) to
1730 average 20.03x read depth, while the remaining individuals were sequenced to an average
1731 of 10.82x read depth. Since the accuracy of genotype calling is correlated with read
1732 depth, we expect the genotype calls made for this individual will harbor the fewest errors.
1733 When considering how the read-depth thresholds affect estimates from individual
1734 samples, GOUND-0446 showed patterns that differ from the other samples (Fig S12).
1735 Specifically, while the number of sites passing filter changes less for this individual with
1736 increasing read-depth minimums, the number of mutations observed in this individual is
1737 also less sensitive to minimum read-depth, suggesting that genotype calling is
1738 substantially more error prone in the lower-read depth individuals. The variance in the
1739 number of observed mutations among individuals reduces substantially when the 12-read
1740 minimum is implemented with GOUND-0446 falling in the middle of the distribution,
1741 despite variance in the number of sites passing filter remaining large. This suggests that

many of the errors have been removed using this filter. We, therefore, convert the point estimate from this filter and the standard deviation from this estimate (see above) to years, conservatively assuming 10 generations per year [130], to obtain the estimated age.

Permutation test of M-GOUNDRY divergence on the X

To specifically test whether the large sweep region on the X chromosome of GOUNDRY individuals is significantly more diverged than similarly sized windows on the X, we defined a new statistic to compare the average D_{xy} among 10 kb windows within the putative inverted region (n = 166) with the average D_{xy} among windows in every other 1.67 MB window on the X that doesn't overlap with the inverted region (n = 1,499). The statistic ΔD_{xy} is defined as

$$\Delta D_{XY} = \frac{1}{w_I} \sum_{1}^{w_I} D_{XY}I - \frac{1}{w_O} \sum_{1}^{w_O} D_{XY}O$$

where w_I is the number of windows inside the inversion, $D_{xy}I$ is a D_{xy} value for a 10kb window inside the inversion, w_O is the number of windows outside the inversion, $D_{xy}O$ is a D_{XY} value for a 10kb window outside the inversion. A positive ΔD_{xy} indicates that the average value inside the inversion is greater than the average value in the window outside the inversion. A negative value indicates that the average value is higher in the window outside the inversion. We calculated this statistic between M and GOUNDRY. For every window outside the swept region, we calculated ΔD_{xy} with the true window assignments and then randomly permuted 10 Kb window assignments to the inversion or the outside window and recalculated ΔD_{xy}. Each window comparison was permuted 10^5 times.

One potential concern with this approach is that the windows inside of the inversion are part of a haplotype with little recombination while windows outside of the inversion represent many independent genealogical histories. However, this is a consideration only for very recent sequence evolution and most of the patterns of D_{xy} are determined by coalescence in the ancestral population prior to the origin of the inversion. After the inversion becomes relative common in the population, it will recombine normally with other inverted chromosomes in the population, so the comparison between windows inside and outside of the swept region should not be biased.

Relative Node Depth expectation modeling

To explicitly test whether such a pattern could be obtained under a pure split model with no gene flow, we obtained expected values of Relative Node Depth (RND) by assuming a phylogeny where M form and GOUNDRY form a clade with S form as the

1777 outgroup and using coalescent theory under this model to calculate expected RND. We
1778 calculated the ratio of expected values of D_{GM} and D_{GS} as
1779

$$\frac{D_{GM}}{D_{GS}} = \frac{\mu \times (2t_{GM} + 4N_{GM} - e^{\frac{t_{GM} - t_{GS}}{2N_{GM}}} (N_{GM} - N_{GS}))}{\mu \times (2t_{GS} + 4N_{GS})}$$

1780 where μ is the mutation rate, N_{GM} and N_{GS} are the effective sizes of the M-GOUNDRY
1781 ancestral population and the M-S-GOUNDRY ancestral population, respectively, t_{GM} is
1782 the number of generations since the split between GOUNDRY and M form, and t_{GS} is the
1783 number of generations since the split between S form and the M-GOUNDRY. The
1784 denominator is the standard definition for D_{xy} [111], but we adjusted the equation in the
1785 numerator using standard coalescent theory to accommodate the probability that two the
1786 M and GOUNDRY lineages may either coalesce in the M-GOUNDRY ancestor or in the
1787 M-S-GOUNDRY ancestor that may differ in size, and therefore probability of
1788 coalescence, from the M-GOUNDRY ancestor (that may differ in size).
1789 We calculated expected values of RND under a range of parameterizations using
1790 the M-GOUNDRY split time of 1,112,660 generations obtained from the demographic
1791 inference described above. For all calculations, we varied the M-GOUNDRY ancestral
1792 size from 100 to 10^6, including 20,000 grid points. In the first set of calculations, we
1793 assumed that the split time between S and M-GOUNDRY was 1.1 times longer than the
1794 M-GOUNDRY split time and varied the relative sizes of the ancestral species such that
1795 the M-S-GOUNDRY ancestor such that the M-GOUNDRY ancestor varied from 1% to
1796 100% the size of the M-S-GOUNDRY ancestor (Figure 10). We conducted a second set
1797 of calculations that were identical to the first except that we assumed that the difference
1798 in split times was 1.5 times longer instead of 1.1 (Figure 10). In the last set of
1799 calculations, we set the sizes of the two ancestors to be equal and varied the difference in
1800 split times such that the split time between S form and M-GOUNDRY was 1%, 10%,
1801 20%, 40%, 70%, and 100% longer than the split time for M-GOUNDRY (Figure 10).
1802
1803 ***Putative GOUNDRY Inversion breakpoint mapping***
1804 The dramatic reduction in nucleotide diversity across a 1.67MB region on the X
1805 chromosome in our GOUNDRY sample led us to hypothesize the existence of a novel
1806 chromosomal inversion in this region. Read mapping and depth is normal across the
1807 region containing the reduction of diversity (Figure S3). The exceptionally low level of
1808 nucleotide diversity in this region suggest a strong selective sweep recently fixed in this
1809 region, but such a wide footprint of the sweep and the remarkably rapid recovery to
1810 background diversity levels outside the sweep are both unexpected under normal
1811 recombination conditions. The existence of a new inversion that suppresses
1812 recombination with the alternative, presumably ancestral, form is the most likely scenario.
1813 We observe in our data the fixation of a 1.67 MB haplotype implying that this haplotype

1814 was maintained without recombination during the majority of the selective sweep. Since
1815 recombination should be normal among chromosomes carrying the inverted form, this
1816 suggests that the inverted form of the inversion was extremely rare in the population at
1817 the beginning of the sweep
1818 Several issues inherent to this system and our data complicate identifying the
1819 breakpoints of this putative inversion. The first is that the inversion was recently under
1820 positive selection resulting in patterns of nucleotide diversity, linkage disequilibrium,
1821 homozygosity, and allele frequency differentiation that may extend beyond the
1822 breakpoints. There are clear changes in these population genetic signals in the putative
1823 region, but we can only conclude that the inversion lies somewhere inside of this
1824 footprint. The second issue is that GOUNDRY is inbred and harbors long tracts of IBD,
1825 some of which overlap the putative inverted region, making it difficult to distinguish
1826 homozygosity related to the sweep from that resulting from inbreeding. The third issue is
1827 that the X chromosome is repetitive and harbors many transposable elements resulting in
1828 a reference sequence riddled with gaps. Inversion breakpoints that have been
1829 characterized in *Anopheles* [131] often lie inside or near such lowly-complex, repetitive
1830 regions. Our short paired-end read data has insert sizes of approximately 470bps, so
1831 mapping the presumed breakpoints is expected to be challenging if they are in repetitive
1832 DNA, and we are unlikely to be able to map read pairs across these regions. It should be
1833 noted that very few chromosomal inversions identified with cytogenetics have been
1834 characterized molecularly, even in systems with reference sequences [40,103,132].
1835 With these challenges in mind, we attempted to map the breakpoints of the
1836 inversion using a variety of approaches to collect candidate sites that could be assayed
1837 using PCR amplification. We manually inspected the short read data in this region using
1838 IGV [133] and located the edges of the swept region at positions 8,462,788 on the left
1839 and 10,137,178 on the right, according to coordinates in the AgamP3 PEST reference
1840 sequence. Since the selective sweep likely purged nucleotide diversity from the region
1841 surrounding the inverted haplotype as well, we attempted to identify the true breakpoints
1842 assuming that true breakpoints fall just inside these boundaries. For all of the following
1843 analyses except the read-depth analysis, we used reads that were trimmed using
1844 SolexaQA (v.2.1; [134]) with default settings and 50bp minimum size. We attempted a
1845 large series of *de novo* assemblies using the multi-kmer approach in SOAPdenovo2 [135]
1846 with –R –F and –p 4. We first attempted a de novo assembly of just reads that mapped to
1847 the inversion region on the X and obtained many contig sets based on variety of subsets
1848 of the data. None of the contigs showed the 'T' shape expected when the contig is
1849 aligned to the PEST reference and part of the contig is inverted. We also looked for cases
1850 of consensus where a variety of de novo assemblies using subsets of the data or all
1851 GOUNDRY reads all failed to assemble across a point on the reference, assuming that we
1852 could rule out regions that were properly assembled and aligned and focus on regions that
1853 fail to assemble as candidates. In another approach, we measured mapped read depth in

100bp windows across the region in GOUNDRY, M form, and *A. arabiensis*. We then systematically searched for windows where M form and *A. arabiensis* showed normal read depths while GOUNDRY reads failed to map in any sample. We also took a similar mapping-based approach where we identified all reads where the pair was mapped in the incorrect orientation or only one of the two mates mapped. We then searched for local enrichments of these mate-pair violations. Lastly, we used two computational implementations of algorithms specifically designed to identify structural variants such as inversions. We used PINDEL [136] with a window size of 2 million and a minimum inversion size of 2000. We also used GASV [132] with default settings.

From each of these approaches, we identified a series of candidate breakpoints. Several of the candidates were supported by multiple approaches. For these candidates, we designed primers on either side and attempted to PCR amplify across the candidate region in samples of GOUNDRY, M form, and *A. arabiensis*, with the assumption that a true breakpoint should amplify in M form and *A. arabiensis*, but not in GOUNDRY. However, all of the candidate PCR reactions either amplified in all three groups or failed to reliably amplify in any group, ultimately preventing the molecular characterization of this inversion.

Acknowledgements

We thank Matteo Fumagalli, Filipe Vieira, and Tyler Linderoth for assistance with next generation sequence data analyses and ANGSD. We thank members of the Nielsen group for helpful discussions on various aspects of this work. We also thank Russ Corbett-Detig and three anonymous reviewers for helpful comments on an earlier version of this manuscript. We are thankful for the use of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

1885 **References:**

1886 1. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, et al. (2007) Cryptic species
1887 as a window on diversity and conservation. Trends Ecol Evol 22: 148–155.
1888 doi:10.1016/j.tree.2006.11.004.

1889 2. Schluter D (2001) Ecology and the origin of species. Trends Ecol Evol 16: 372–
1890 380.

1891 3. Harrison RG, Larson EL (2014) Hybridization, introgression, and the nature of
1892 species boundaries. J Hered 105 Suppl 1: 795–809.
1893 doi:10.1093/jhered/esu033.

1894 4. Servedio MR, Hermisson J, van Doorn GS (2013) Hybridization may rarely
1895 promote speciation. J Evol Biol 26: 282–285. doi:10.1111/jeb.12038.

1896 5. Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, et al. (2012)
1897 Adaptive introgression across species boundaries in Heliconius butterflies.
1898 PLoS Genet 8: e1002752. doi:10.1371/journal.pgen.1002752.

1899 6. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, et al. (2014) Altitude
1900 adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature
1901 advance online publication. Available:
1902 http://www.nature.com/nature/journal/vaop/ncurrent/full/nature13408.ht
1903 ml. Accessed 24 July 2014.

1904 7. WHO (2013) World Malaria Report.

1905 8. Coetzee M, Hunt RH, Wilkerson R, Torre AD, Coulibaly MB, et al. (2013)
1906 Anopheles coluzzii and Anopheles amharicus , new members of the Anopheles
1907 gambiae complex. Zootaxa 3619: 246–274. doi:10.11646/zootaxa.3619.3.2.

1908 9. Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, et al. (2009) Living at
1909 the edge: biogeographic patterns of habitat segregation conform to speciation
1910 by niche expansion in Anopheles gambiae. BMC Ecol 9: 16. doi:10.1186/1472-
1911 6785-9-16.

1912 10. Gnémé A, Guelbéogo WM, Riehle MM, Sanou A, Traoré A, et al. (2013) Equivalent
1913 susceptibility of Anopheles gambiae M and S molecular forms and Anopheles
1914 arabiensis to Plasmodium falciparum infection in Burkina Faso. Malar J 12:
1915 204. doi:10.1186/1475-2875-12-204.

1916 11. Lee Y, Marsden CD, Norris LC, Collier TC, Main BJ, et al. (2013) Spatiotemporal
1917 dynamics of gene flow and hybrid fitness between the M and S forms of the
1918 malaria mosquito, Anopheles gambiae. Proc Natl Acad Sci: 201316851.
1919 doi:10.1073/pnas.1316851110.

1920 12. Della Torre A, Fanello C, Akogbeto M, Dossou-yovo J, Favia G, et al. (2001)
1921 Molecular evidence of incipient speciation within Anopheles gambiae s.s. in
1922 West Africa. Insect Mol Biol 10: 9–18.

1923 13. Lawniczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, et al. (2010)
1924 Widespread divergence between incipient Anopheles gambiae species
1925 revealed by whole genome sequences. Science 330: 512–514.
1926 doi:10.1126/science.1195755.

1927 14. Lehmann T, Diabate A (2008) The molecular forms of Anopheles gambiae: a
1928 phenotypic perspective. Infect Genet Evol J Mol Epidemiol Evol Genet Infect
1929 Dis 8: 737–746. doi:10.1016/j.meegid.2008.06.003.

1930 15. Slotman MA, Mendez MM, Torre AD, Dolo G, Touré YT, et al. (2006) Genetic
1931 differentiation between the BAMAKO and SAVANNA chromosomal forms of
1932 Anopheles gambiae as indicated by amplified fragment length polymorphism
1933 analysis. Am J Trop Med Hyg 74: 641–648.

1934 16. Slotman MA, Tripet F, Cornel AJ, Meneses CR, Lee Y, et al. (2007) Evidence for
1935 subdivision within the M molecular form of Anopheles gambiae. Mol Ecol 16:
1936 639–649. doi:10.1111/j.1365-294X.2006.03172.x.

1937 17. Riehle MM, Guelbeogo WM, Gneme A, Eiglmeier K, Holm I, et al. (2011) A cryptic
1938 subgroup of Anopheles gambiae is highly susceptible to human malaria
1939 parasites. Science 331: 596–598. doi:10.1126/science.1196759.

1940 18. Griffin JT, Hollingsworth TD, Okell LC, Churcher TS, White M, et al. (2010)
1941 Reducing Plasmodium falciparum Malaria Transmission in Africa: A Model-
1942 Based Evaluation of Intervention Strategies. PLoS Med 7: e1000324.
1943 doi:10.1371/journal.pmed.1000324.

1944 19. Xu G, Fang QQ, Keirans JE, Durden LA (2003) Molecular phylogenetic analyses
1945 indicate that the Ixodes ricinus complex is a paraphyletic group. J Parasitol 89:
1946 452–457. doi:10.1645/0022-3395(2003)089[0452:MPAITT]2.0.CO;2.

1947 20. Harbach RE (2004) The classification of genus Anopheles (Diptera: Culicidae): a
1948 working hypothesis of phylogenetic relationships. Bull Entomol Res 94: 537–
1949 553.

1950 21. Dyer NA, Furtado A, Cano J, Ferreira F, Odete Afonso M, et al. (2009) Evidence
1951 for a discrete evolutionary lineage within Equatorial Guinea suggests that the
1952 tsetse fly Glossina palpalis palpalis exists as a species complex. Mol Ecol 18:
1953 3268–3282. doi:10.1111/j.1365-294X.2009.04265.x.

1954 22. The malERA Consultative Group on Vector Control (2011) A Research Agenda
1955 for Malaria Eradication: Vector Control. PLoS Med 8: e1000401.
1956 doi:10.1371/journal.pmed.1000401.

1957 23. Coyne JA, Orr HA (1989) Patterns of speciation in Drosophila. Evolution: 362–
1958 381.

1959 24. Geraldes A, Ferrand N, Nachman MW (2006) Contrasting Patterns of
1960 Introgression at X-Linked Loci Across the Hybrid Zone Between Subspecies of
1961 the European Rabbit (Oryctolagus cuniculus). Genetics 173: 919–933.
1962 doi:10.1534/genetics.105.054106.

1963 25. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, et al. (2014) The
1964 genomic landscape of Neanderthal ancestry in present-day humans. Nature
1965 507: 354–357. doi:10.1038/nature12961.

1966 26. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, et al. (2007) Population
1967 Genomics: Whole-Genome Analysis of Polymorphism and Divergence in
1968 Drosophila simulans. PLoS Biol 5: e310. doi:10.1371/journal.pbio.0050310.

1969 27. McGaugh SE, Noor MAF (2012) Genomic impacts of chromosomal inversions in
1970 parapatric Drosophila species. Philos Trans R Soc Lond B Biol Sci 367: 422–
1971 429. doi:10.1098/rstb.2011.0250.

1972 28. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. (2012) The
1973 Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.
1974 doi:10.1038/nature10811.

1975 29. Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in
1976 Anopheles gambiae. PLoS Biol 3: e285. doi:10.1371/journal.pbio.0030285.

1977 30. Neafsey DE, Lawniczak MKN, Park DJ, Redmond SN, Coulibaly MB, et al. (2010)
1978 SNP genotyping defines complex gene-flow boundaries among African malaria
1979 vector mosquitoes. Science 330: 514–517. doi:10.1126/science.1193036.

1980 31. Weetman D, Wilding CS, Steen K, Pinto J, Donnelly MJ (2012) Gene flow-
1981 dependent genomic divergence between Anopheles gambiae M and S forms.
1982 Mol Biol Evol 29: 279–291. doi:10.1093/molbev/msr199.

1983 32. Charlesworth B (1998) Measures of divergence between populations and the
1984 effect of forces that reduce variability. Mol Biol Evol 15: 538–543.

1985 33. Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert?
1986 Examining the role of restricted recombination in maintaining species.
1987 Heredity 103: 439–444. doi:10.1038/hdy.2009.151.

1988 34. Hahn MW, White BJ, Muir CD, Besansky NJ (2012) No evidence for biased co-
1989 transmission of speciation islands in Anopheles gambiae. Philos Trans R Soc
1990 Lond B Biol Sci 367: 374–384. doi:10.1098/rstb.2011.0188.

1991 35. Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of
1992 speciation are due to reduced diversity, not reduced gene flow. Mol Ecol 23:
1993 3133–3157. doi:10.1111/mec.12796.

1994 36. Wang-Sattler R, Blandin S, Ning Y, Blass C, Dolo G, et al. (2007) Mosaic Genome
1995 Architecture of the Anopheles gambiae Species Complex. PLoS ONE 2: e1249.
1996 doi:10.1371/journal.pone.0001249.

1997 37. O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, et al. (2014) Genomic
1998 analyses of three malaria vectors reveals extensive shared polymorphism but
1999 contrasting population histories. Mol Biol Evol 31: 889–902.
2000 doi:10.1093/molbev/msu040.

2001 38. Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, et al. (2014) Diversity,
2002 differentiation, and linkage disequilibrium: prospects for association mapping
2003 in the malaria vector Anopheles arabiensis. G3 Bethesda Md 4: 121–131.
2004 doi:10.1534/g3.113.008326.

2005 39. Korneliussen T, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next
2006 Generation Sequencing Data. BMC Bioinformatics 15: 356.
2007 doi:10.1186/s12859-014-0356-4.

2008 40. Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A new chromosomal phylogeny
2009 supports the repeated origin of vectorial capacity in malaria mosquitoes of the
2010 Anopheles gambiae complex. PLoS Pathog 8: e1002960.
2011 doi:10.1371/journal.ppat.1002960.

2012 41. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene
2013 chromosome analysis of the Anopheles gambiae species complex. Science 298:
2014 1415–1418. doi:10.1126/science.1077769.

2015 42. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring
2016 the Joint Demographic History of Multiple Populations from Multidimensional
2017 SNP Frequency Data. PLoS Genet 5: e1000695.
2018 doi:10.1371/journal.pgen.1000695.

2019 43. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans
2020 for selective sweeps using SNP data. Genome Res 15: 1566.

2021 44. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling,
2022 genotype calling, and sample allele frequency estimation from New-
2023 Generation Sequencing data. PloS One 7: e37558.
2024 doi:10.1371/journal.pone.0037558.

2025 45. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of
2026 Tajima's D and other neutrality test statistics from low depth next-generation

2027 sequencing data. BMC Bioinformatics 14: 289. doi:10.1186/1471-2105-14-
2028 289.

2029 46. Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for
2030 population genetics analyses from next-generation sequencing data.
2031 Bioinforma Oxf Engl 30: 1486–1487. doi:10.1093/bioinformatics/btu041.

2032 47. Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, et al. (2009) Ecological niche
2033 partitioning between Anopheles gambiae molecular forms in Cameroon: the
2034 ecological side of speciation. BMC Ecol 9: 17.

2035 48. White BJ, Lawniczak MKN, Cheng C, Coulibaly MB, Wilson MD, et al. (2011)
2036 Adaptive divergence between incipient species of Anopheles gambiae
2037 increases resistance to Plasmodium. Proc Natl Acad Sci U S A 108: 244–249.
2038 doi:10.1073/pnas.1013648108.

2039 49. Wu Q, Wen T, Lee G, Park JH, Cai HN, et al. (2003) Developmental Control of
2040 Foraging and Social Behavior by the Drosophila Neuropeptide Y-like System.
2041 Neuron 39: 147–161. doi:10.1016/S0896-6273(03)00396-9.

2042 50. Gimonneau G, Pombi M, Dabire RK, Diabate A, Morand S, et al. (2012)
2043 Behavioural responses of Anopheles gambiae sensu stricto M and S molecular
2044 form larvae to an aquatic predator in Burkina Faso. Parasit Vectors 5: 65.
2045 doi:10.1186/1756-3305-5-65.

2046 51. Croker B, Crozat K, Berger M, Xia Y, Sovath S, et al. (2007) ATP-sensitive
2047 potassium channels mediate survival during infection in mammals and insects.
2048 Nat Genet 39: 1453–1460. doi:10.1038/ng.2007.25.

2049 52. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical membrane
2050 topology and heteromeric function of Drosophila odorant receptors in vivo.
2051 PLoS Biol 4: e20. doi:10.1371/journal.pbio.0040020.

2052 53. Rund SSC, Hou TY, Ward SM, Collins FH, Duffield GE (2011) Genome-wide
2053 profiling of diel and circadian gene expression in the malaria vector
2054 Anopheles gambiae. Proc Natl Acad Sci 108: E421–E430.
2055 doi:10.1073/pnas.1100584108.

2056 54. Ditzen M, Pellegrino M, Vosshall LB (2008) Insect Odorant Receptors Are
2057 Molecular Targets of the Insect Repellent DEET. Science 319: 1838–1842.
2058 doi:10.1126/science.1153121.

2059 55. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft
2060 Sequence of the Neandertal Genome. Science 328: 710–722.
2061 doi:10.1126/science.1188021.

2062 56. Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture
2063 between closely related populations. Mol Biol Evol 28: 2239–2252.
2064 doi:10.1093/molbev/msr048.

2065 57. Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from
2066 the lengths of migrant tracts. Genetics 181: 711–719.
2067 doi:10.1534/genetics.108.098095.

2068 58. Gravel S (2012) Population genetics models of local ancestry. Genetics 191: 607–
2069 619. doi:10.1534/genetics.112.139808.

2070 59. Nosil P, Crespi BJ, Sandoval CP (2003) Reproductive isolation driven by the
2071 combined effects of ecological adaptation and reinforcement. Proc Biol Sci
2072 270: 1911–1918. doi:10.1098/rspb.2003.2457.

2073 60. Grant PR, Grant BR, Petren K (2005) Hybridization in the recent past. Am Nat
2074 166: 56–67. doi:10.1086/430331.

2075 61. Weetman D, Steen K, Rippon EJ, Mawejje HD, Donnelly MJ, et al. (2014)
2076 Contemporary gene flow between wild An. gambiae s.s. and An. arabiensis.
2077 Parasit Vectors 7: 345. doi:10.1186/1756-3305-7-345.

2078 62. Clarkson CS, Weetman D, Essandoh J, Yawson AE, Maslen G, et al. (2014)
2079 Adaptive introgression between Anopheles sibling species eliminates a major
2080 genomic island but not reproductive isolation. Nat Commun 5: 4248.
2081 doi:10.1038/ncomms5248.

2082 63. Ffrench-Constant RH, Rocheleau TA, Steichen JC, Chalmers AE (1993) A point
2083 mutation in a Drosophila GABA receptor confers insecticide resistance. Nature
2084 363: 449–451. doi:10.1038/363449a0.

2085 64. Du W, Awolola TS, Howell P, Koekemoer LL, Brooke BD, et al. (2005)
2086 Independent mutations in the Rdl locus confer dieldrin resistance to
2087 Anopheles gambiae and An. arabiensis. Insect Mol Biol 14: 179–183.
2088 doi:10.1111/j.1365-2583.2005.00544.x.

2089 65. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA (1979) Chromosomal
2090 differentiation and adaptation to human environments in the Anopheles
2091 gambiae complex. Trans R Soc Trop Med Hyg 73: 483–497.

2092 66. Della Torre A, Merzagora L, Powell JR, Coluzzi M (1997) Selective introgression
2093 of paracentric inversions between two sibling species of the Anopheles
2094 gambiae complex. Genetics 146: 239–244.

2095 67. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, et al. (2003)
2096 Semipermeable species boundaries between Anopheles gambiae and

2097 Anopheles arabiensis: evidence from multilocus DNA sequence variation. Proc
2098 Natl Acad Sci U S A 100: 10818–10823. doi:10.1073/pnas.1434337100.

2099 68. White BJ, Hahn MW, Pombi M, Cassone BJ, Lobo NF, et al. (2007) Localization of
2100 candidate regions maintaining a common polymorphic inversion (2La) in
2101 Anopheles gambiae. PLoS Genet 3: e217. doi:10.1371/journal.pgen.0030217.

2102 69. Payne F (1924) Crossover Modifiers in the Third Chromosome of DROSOPHILA
2103 MELANOGASTER. Genetics 9: 327–342.

2104 70. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and
2105 speciation. Genetics 173: 419–434. doi:10.1534/genetics.105.047985.

2106 71. Hoffmann AA, Rieseberg LH (2008) Revisiting the Impact of Inversions in
2107 Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and
2108 Speciation? Annu Rev Ecol Evol Syst 39: 21–42.
2109 doi:10.1146/annurev.ecolsys.39.110707.173532.

2110 72. Dobzhansky TG (1937) Genetics and the Origin of Species. New York, NY:
2111 Columbia University Press. 364 p.

2112 73. Muller HJ (1940) Bearing of the Drosophila work on systematics. The new
2113 systematics (ed. J.S. Huxley). Oxford, UK: Clarendon Press. pp. 185–268.

2114 74. Cheng C, White BJ, Kamdem C, Mockaitis K, Costantini C, et al. (2012) Ecological
2115 Genomics of Anopheles Gambiae Along a Latitudinal Cline: A Population-
2116 Resequencing Approach. Genetics 190: 1417–1432.
2117 doi:10.1534/genetics.111.137794.

2118 75. Elliott R, Ramakrishna V (1956) Insecticide resistance in Anopheles gambiae
2119 Giles. Nature 177: 532–533.

2120 76. Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and
2121 the reproductive isolation of species. Proc Natl Acad Sci U S A 98: 12084–
2122 12088. doi:10.1073/pnas.221274498.

2123 77. Rieseberg LH (2001) Chromosomal rearrangements and speciation. Trends Ecol
2124 Evol 16: 351–358.

2125 78. Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in
2126 parapatry: a new twist on chromosomal speciation. Evol Int J Org Evol 57:
2127 447–459.

2128 79. Butlin RK (2005) Recombination and speciation. Mol Ecol 14: 2621–2635.
2129 doi:10.1111/j.1365-294X.2005.02617.x.

2130 80. Charlesworth B, Coyne JA, Barton NH (1987) The Relative Rates of Evolution of
2131 Sex Chromosomes and Autosomes. Am Nat 130: 113. doi:10.1086/284701.

2132 81. Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, et al. (2012) Genome
2133 sequencing reveals complex speciation in the Drosophila simulans clade.
2134 Genome Res. Available:
2135 http://genome.cshlp.org/content/early/2012/04/13/gr.130922.111.
2136 Accessed 15 October 2012.

2137 82. Slotman MA, Della Torre A, Calzetta M, Powell JR (2005) Differential
2138 introgression of chromsomal regions between Anopheles gambiae and An.
2139 arabiensis. Am J Trop Med Hyg 73: 326–335.

2140 83. Chen CM, Struhl G (1999) Wingless transduction by the Frizzled and Frizzled2
2141 proteins of Drosophila. Dev Camb Engl 126: 5441–5452.

2142 84. Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a
2143 transposon insertion in Drosophila simulans. Proc Natl Acad Sci U S A 101:
2144 1626–1631. doi:10.1073/pnas.0303793101.

2145 85. Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via
2146 transposition-mediated adaptive gene truncation in Drosophila. Science 309:
2147 764–767. doi:10.1126/science.1112699.

2148 86. Corbett-Detig RB, Hartl DL (2012) Population genomics of inversion
2149 polymorphisms in Drosophila melanogaster. PLoS Genet 8: e1003056.
2150 doi:10.1371/journal.pgen.1003056.

2151 87. Feder JL, Xie X, Rull J, Velez S, Forbes A, et al. (2005) Mayr, Dobzhansky, and
2152 Bush and the complexities of sympatric speciation in Rhagoletis. Proc Natl
2153 Acad Sci U S A 102 Suppl 1: 6573–6580. doi:10.1073/pnas.0502099102.

2154 88. Slotman MA, Reimer LJ, Thiemann T, Dolo G, Fondjo E, et al. (2006) Reduced
2155 recombination rate and genetic differentiation between the M and S forms of
2156 Anopheles gambiae s.s. Genetics 174: 2081–2093.
2157 doi:10.1534/genetics.106.059949.

2158 89. Dao A, Yaro AS, Diallo M, Timbiné S, Huestis DL, et al. (2014) Signatures of
2159 aestivation and migration in Sahelian malaria mosquito populations. Nature.
2160 Available:
2161 http://www.nature.com/nature/journal/vaop/ncurrent/fig_tab/nature1398
2162 7_SF2.html. Accessed 3 December 2014.

2163 90. Turner TL, Hahn MW (2010) Genomic islands of speciation or genomic islands
2164 and speciation? Mol Ecol 19: 848–850. doi:10.1111/j.1365-
2165 294X.2010.04532.x.

2166 91. Geraldes A, Basset P, Gibson B, Smith KL, Harr B, et al. (2008) Inferring the
2167 history of speciation in house mice from autosomal, X-linked, Y-linked and
2168 mitochondrial genes. Mol Ecol 17: 5349–5363. doi:10.1111/j.1365-
2169 294X.2008.04005.x.

2170 92. Presgraves DC (2008) Sex chromosomes and speciation in Drosophila. Trends
2171 Genet TIG 24: 336–343. doi:10.1016/j.tig.2008.04.007.

2172 93. Frank SA (1991) Divergence of Meiotic Drive-Suppression Systems as an
2173 Explanation for Sex- Biased Hybrid Sterility and Inviability. Evolution 45: 262.
2174 doi:10.2307/2409661.

2175 94. Hurst LD, Pomiankowski A (1991) Causes of sex ratio bias may account for
2176 unisexual sterility in hybrids: a new explanation of Haldane's rule and related
2177 phenomena. Genetics 128: 841–858.

2178 95. Lifschytz E, Lindsley DL (1972) The role of X-chromosome inactivation during
2179 spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-
2180 dosage compensation). Proc Natl Acad Sci U S A 69: 182–186.

2181 96. Wu C-I (2001) The genic view of the process of speciation. J Evol Biol 14: 851–
2182 865. doi:10.1046/j.1420-9101.2001.00335.x.

2183 97. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, et al. (2012) Genomic
2184 Variation in Natural Populations of Drosophila melanogaster. Genetics 192:
2185 533–598. doi:10.1534/genetics.112.142018.

2186 98. Garrigan D, Kingan SB, Geneva AJ, Vedanayagam JP, Presgraves DC (2014)
2187 Genome diversity and divergence in Drosophila mauritiana: multiple
2188 signatures of faster X evolution. Genome Biol Evol 6: 2444–2458.
2189 doi:10.1093/gbe/evu198.

2190 99. Martin SH, Davey JW, Jiggins CD (2014) Evaluating the Use of ABBA-BABA
2191 Statistics to Locate Introgressed Loci. Mol Biol Evol.
2192 doi:10.1093/molbev/msu269.

2193 100. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, et al. (2014)
2194 Extensive introgression in a malaria vector species complex revealed by
2195 phylogenomics. Science. doi:10.1126/science.1258524.

2196 101. Fanello C, Santolamazza F, della Torre A (2002) Simultaneous identification of
2197 species and molecular forms of the Anopheles gambiae complex by PCR-RFLP.
2198 Med Vet Entomol 16: 461–464.

2199 102. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, et al. (2008) Insertion
2200 polymorphisms of SINE200 retrotransposons within speciation islands of

2201 Anopheles gambiae molecular forms. Malar J 7: 163. doi:10.1186/1475-2875-
2202 7-163.

2203 103. White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, et al. (2007)
2204 Molecular karyotyping of the 2La inversion in Anopheles gambiae. Am J Trop
2205 Med Hyg 76: 334–339.

2206 104. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The
2207 genome sequence of the malaria mosquito Anopheles gambiae. Science 298:
2208 129–149. doi:10.1126/science.1076181.

2209 105. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs
2210 with BWA-MEM. ArXiv13033997 Q-Bio. Available:
2211 http://arxiv.org/abs/1303.3997. Accessed 19 May 2014.

2212 106. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A
2213 framework for variation discovery and genotyping using next-generation DNA
2214 sequencing data. Nat Genet 43: 491–498. doi:10.1038/ng.806.

2215 107. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence
2216 Alignment/Map format and SAMtools. Bioinforma Oxf Engl 25: 2078–2079.
2217 doi:10.1093/bioinformatics/btp352.

2218 108. Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, et al. (2010)
2219 Genome mapping and characterization of the Anopheles gambiae
2220 heterochromatin. BMC Genomics 11: 459. doi:10.1186/1471-2164-11-459.

2221 109. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, et al. (2011)
2222 Estimation of allele frequency and association mapping using next-generation
2223 sequencing data. BMC Bioinformatics 12: 231. doi:10.1186/1471-2105-12-
2224 231.

2225 110. Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding
2226 coefficients from NGS data: Impact on genotype calling and allele frequency
2227 estimation. Genome Res 23: 1852–1861. doi:10.1101/gr.157388.113.

2228 111. Nei M (1987) Molecular Evolutionary Genetics. New York, NY: Columbia
2229 University Press. 526 p.

2230 112. R Development Core Team (2011) R: A language and Environment for
2231 Statistical Computing. Available: http://www.R-project.org/.

2232 113. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and
2233 Evolution in R language. Bioinforma Oxf Engl 20: 289–290.

2234 114. Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly
2235 (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21: 36–44.
2236 doi:10.1093/molbev/msg236.

2237 115. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, et al. (2007)
2238 Direct estimation of per nucleotide and genomic deleterious mutation rates in
2239 Drosophila. Nature 445: 82–85. doi:10.1038/nature05388.

2240 116. Schrider DR, Houle D, Lynch M, Hahn MW (2013) Rates and genomic
2241 consequences of spontaneous mutational events in Drosophila melanogaster.
2242 Genetics 194: 937–954. doi:10.1534/genetics.113.151670.

2243 117. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, et al. (2009) Analysis
2244 of the genome sequences of three Drosophila melanogaster spontaneous
2245 mutation accumulation lines. Genome Res 19: 1195–1201.
2246 doi:10.1101/gr.091231.109.

2247 118. Keightley PD, Ness RW, Halligan DL, Haddrill PR (2014) Estimation of the
2248 spontaneous mutation rate per nucleotide site in a Drosophila melanogaster
2249 full-sib family. Genetics 196: 313–320. doi:10.1534/genetics.113.158758.

2250 119. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and
2251 visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
2252 doi:10.1093/bioinformatics/bth457.

2253 120. Harris C, Lambrechts L, Rousset F, Abate L, Nsango SE, et al. (2010)
2254 Polymorphisms in Anopheles gambiae immune genes associated with natural
2255 resistance to Plasmodium falciparum. PLoS Pathog 6: e1001112.
2256 doi:10.1371/journal.ppat.1001112.

2257 121. Weetman D, Wilding CS, Steen K, Morgan JC, Simard F, et al. (2010)
2258 Association Mapping of Insecticide Resistance in Wild Anopheles gambiae
2259 Populations: Major Variants Identified in a Low-Linkage Disequilbrium
2260 Genome. PLoS ONE 5: e13140. doi:10.1371/journal.pone.0013140.

2261 122. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA
2262 sequence data. Genome Res 19: 136–142. doi:10.1101/gr.083634.108.

2263 123. Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive
2264 selection in whole-genome SNP data from nonequilibrium populations.
2265 Genetics 185: 907–922. doi:10.1534/genetics.110.116459.

2266 124. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, et al. (2007)
2267 Evolutionary dynamics of immune-related genes and pathways in disease-
2268 vector mosquitoes. Science 316: 1738–1743. doi:10.1126/science.1139862.

2269 125. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis
2270 by DNA polymorphism. Genetics 123: 585.

2271 126. Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA (2010) Drosophila
2272 melanogaster recombination rate calculator. Gene 463: 18–20.
2273 doi:10.1016/j.gene.2010.04.015.

2274 127. Chan AH, Jenkins PA, Song YS (2012) Genome-Wide Fine-Scale Recombination
2275 Rate Variation in Drosophila melanogaster. PLoS Genet 8: e1003090.
2276 doi:10.1371/journal.pgen.1003090.

2277 128. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-
2278 scale structure of recombination rate variation in the human genome. Science
2279 304: 581–584. doi:10.1126/science.1092500.

2280 129. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, et al. (2013) Genetic
2281 Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS
2282 Genet 9: e1003984. doi:10.1371/journal.pgen.1003984.

2283 130. Lehmann T, Hawley WA, Grebert H, Collins FH (1998) The effective population
2284 size of Anopheles gambiae in Kenya: implications for population structure.
2285 Mol Biol Evol 15: 264–276.

2286 131. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, et al. (2006)
2287 Breakpoint structure reveals the unique origin of an interspecific
2288 chromosomal inversion (2La) in the Anopheles gambiae complex. Proc Natl
2289 Acad Sci U S A 103: 6258–6262. doi:10.1073/pnas.0509683103.

2290 132. Sindi S, Helman E, Bashir A, Raphael BJ (2009) A geometric approach for
2291 classification and comparison of structural variants. Bioinformatics 25: i222–
2292 i230. doi:10.1093/bioinformatics/btp208.

2293 133. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics
2294 Viewer (IGV): high-performance genomics data visualization and exploration.
2295 Brief Bioinform 14: 178–192. doi:10.1093/bib/bbs017.

2296 134. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality
2297 assessment of Illumina second-generation sequencing data. BMC
2298 Bioinformatics 11: 485. doi:10.1186/1471-2105-11-485.

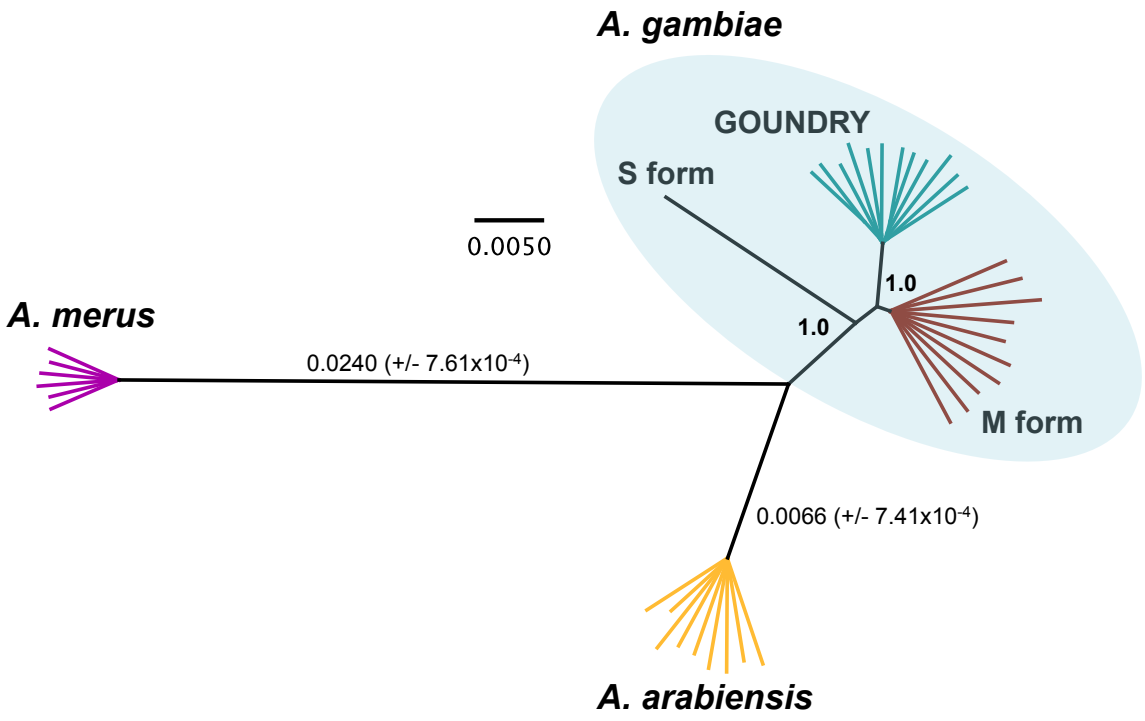
2299 135. Luo R, Liu B, Xie Y, Li Z, Huang W, et al. (2012) SOAPdenovo2: an empirically
2300 improved memory-efficient short-read de novo assembler. GigaScience 1: 18.
2301 doi:10.1186/2047-217X-1-18.

2302 136. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth
2303 approach to detect break points of large deletions and medium sized

2304 insertions from paired-end short reads. Bioinforma Oxf Engl 25: 2865–2871.
2305 doi:10.1093/bioinformatics/btp394.

2306

2307 **Figure Legends:**
2308
2309 **Figure 1: Average genetic relationships among species and subgroups in *Anopheles***
2310 ***gambiae* species complex.** Unrooted neighbor-joining tree calculated with the *ape*
2311 package in R and drawn with Geneious software. Branches indicate genetic distance
2312 (D_{xy}) calculated using intergenic sites (see Methods) with scale bar for reference.
2313 Bootstrap support indicated for nodes leading to *A. gambiae* clade and GOUNDRY-M
2314 form clade. Branch lengths and 95%CIs indicated for branches leading to *A. merus* and *A.*
2315 *arabiensis*.
2316
2317



2318
2319
2320
2321
2322

Figure 2: Demographic history of GOUNDRY and M form *A. gambiae*. A) Three-epoch demographic model. *N* parameters indicate effective population sizes. The duration of each epoch is indicated with the *t* parameters. Migration parameters ($2Nm$) are included as functions of the ancestral effective size. We included separate migration parameters for M into GOUNDRY migration ($2N_A m_{GM}$) and GOUNDRY into M ($2N_A m_{GM}$). **B)** Autosomal two-dimensional site-frequency spectra for GOUNDRY and M form for both the data and model. Residuals are calculated as the normalized difference between the model and the data (model – data), such that red colors indicate an excess number of SNPs predicted by the model.

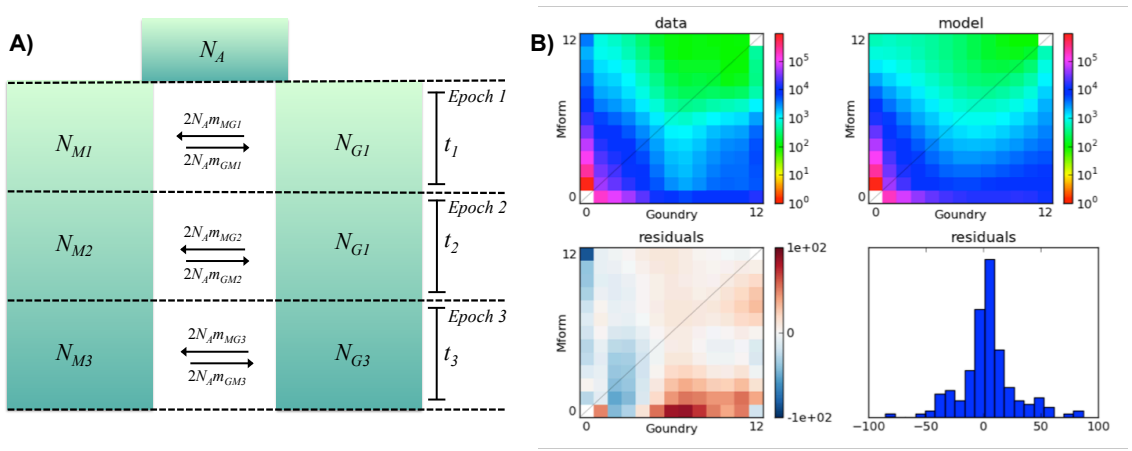
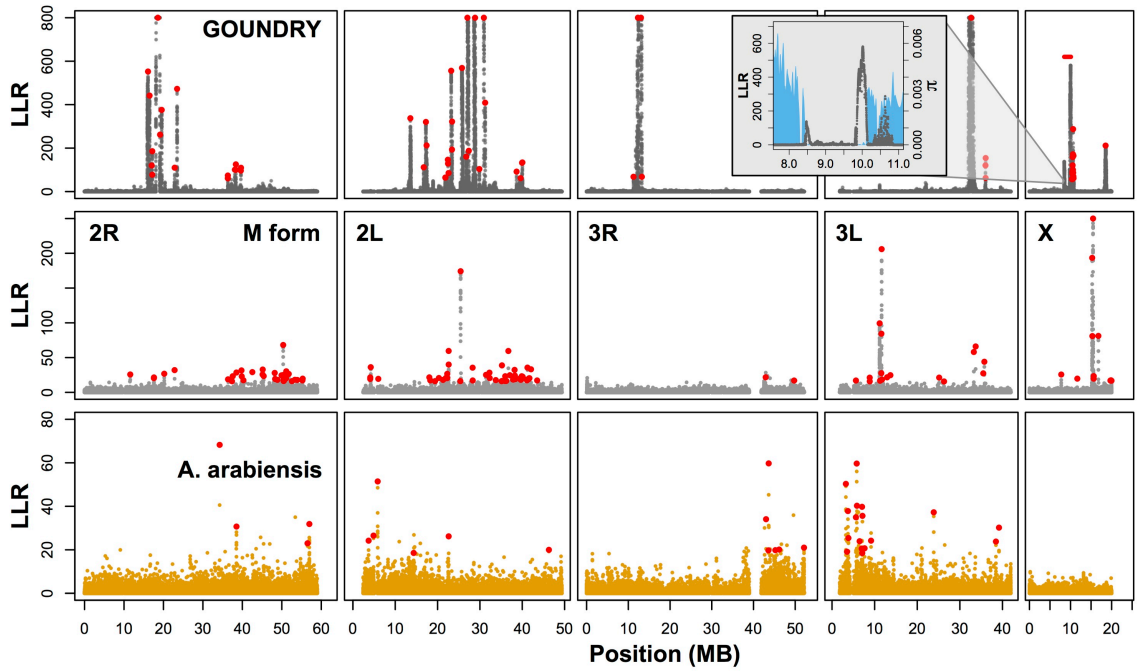
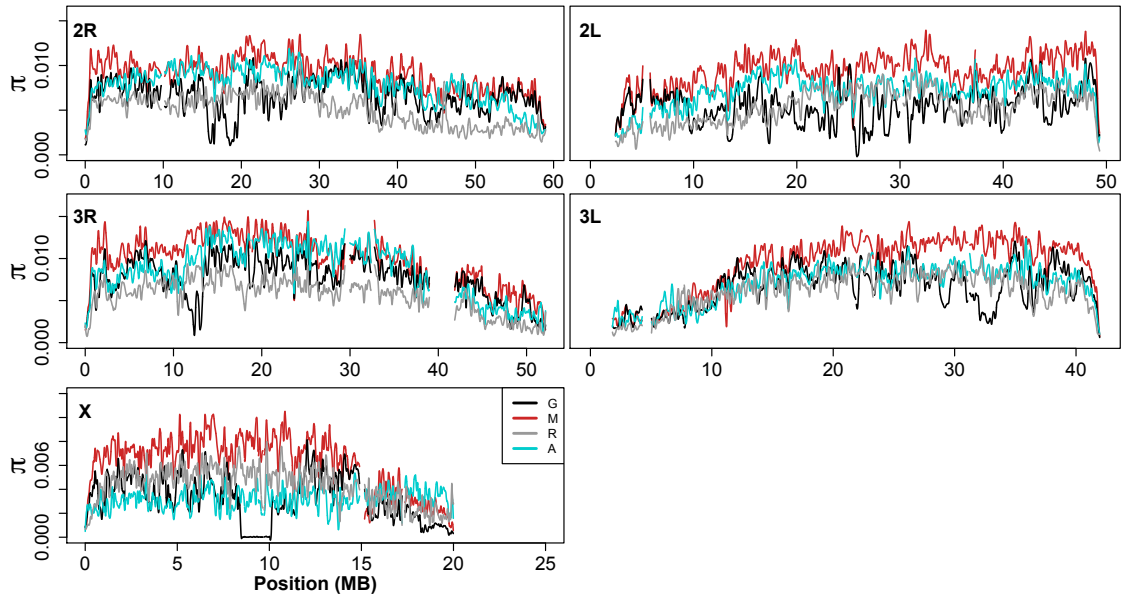


Figure 3: Maps of recent positive selection indicate adaptive differentiation in GOUNDRY (top row), M form (middle row), and *A. arabiensis* (bottom row). Recent selective sweeps are private and distributed across the genome of each group. Positive selection is significantly more prevalent in M form than GOUNDRY ($P < 0.0001$). Each point indicates the Log-Likelihood Ratio (LLR) value for a selective sweep at a given genomic position. Credible peaks are indicated with red dots. The large sweep region on the X chromosome in GOUNDRY is indicated with a horizontal red line, and the inset shows both LLRs and nucleotide diversity (blue). GOUNDRY windows with LLR values > 800 are truncated for presentation. Low-complexity regions were excluded. A full list of inferred targets of selection is given in Tables S2 and S3.

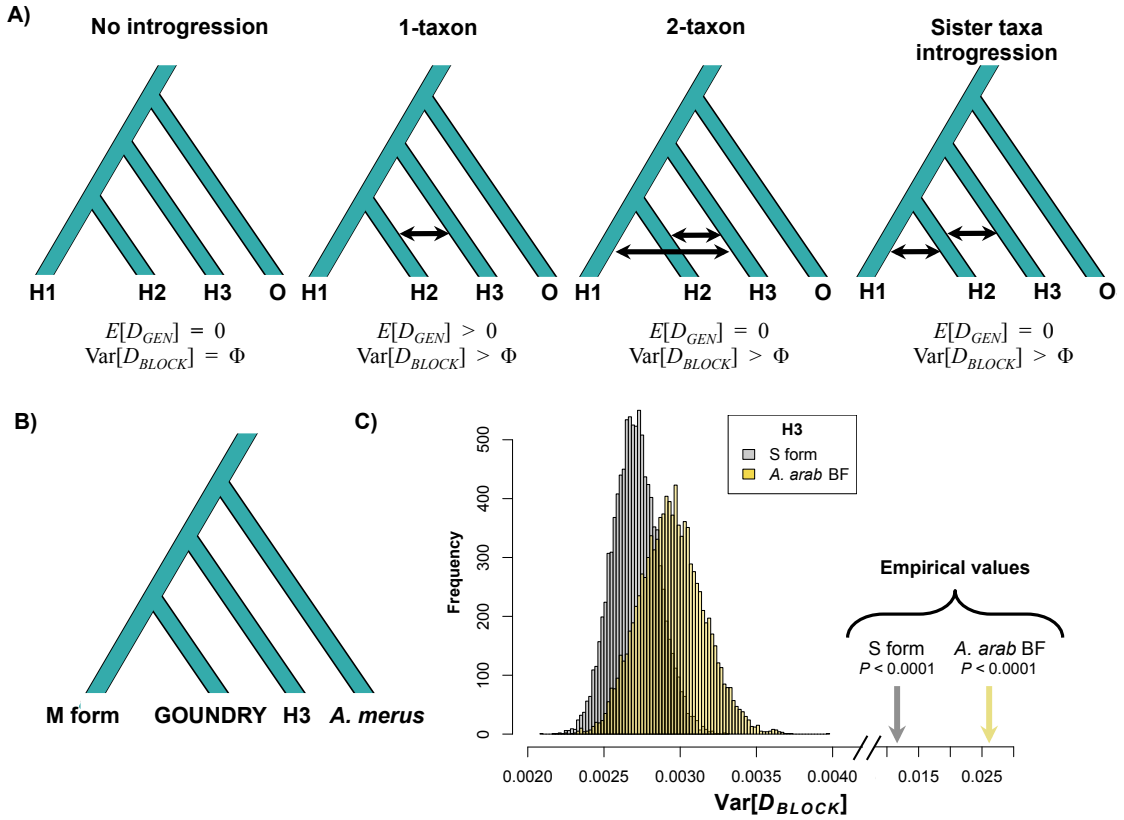


2354 **Figure 4:** Chromosomal distributions of nucleotide diversity (π) at inter-genic sites
2355 (LOESS-smoothed with span of 1% using 10 kb non-overlapping windows). Low
2356 complexity and heterochromatic regions were excluded. M = *A. gambiae* M form; G = *A.*
2357 *gambiae* GOUNDRY; R = *A. merus*; A = *A. arabiensis.*

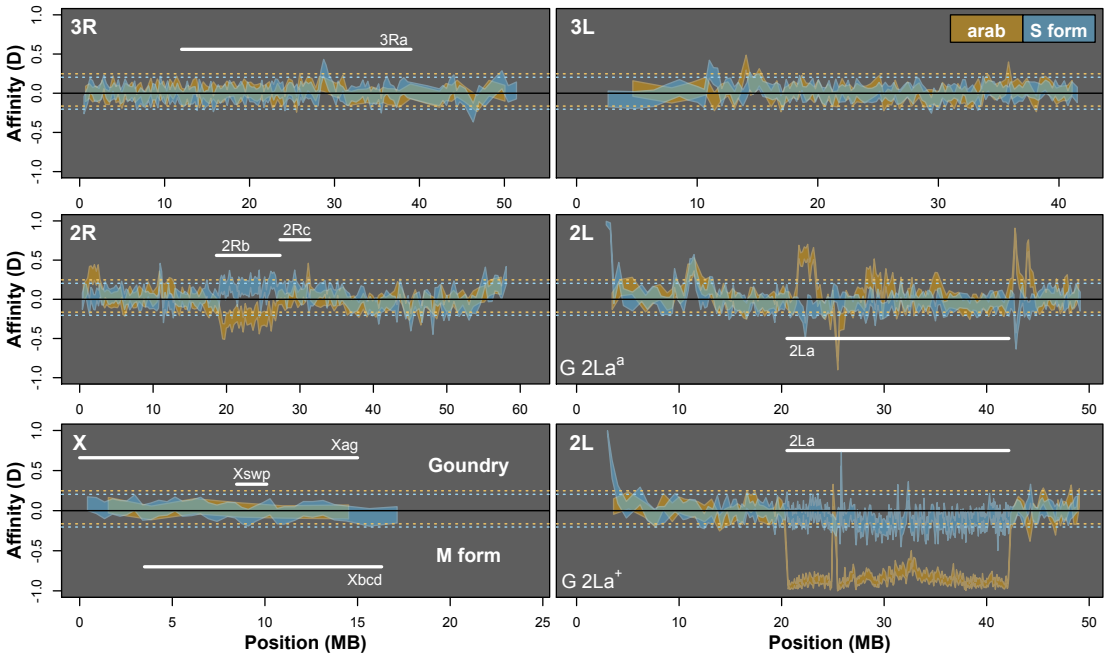


2358

Figure 5: Excess variance among in D_{BLOCK} indicates recent introgression. A) Four-taxon trees used in ABBA-BABA tests with four alternative introgression models. The expected genome-wide value of the D statistic ($E[D_{GEN}]$) is presented below in addition to the expected variance among D statistics calculated in genomic blocks ($\mathrm{Var}[D_{BLOCK}]$). $\mathrm{Var}[D_{BLOCK}]$ under the 'No introgression' model is unknown and indicated here by Φ for comparison in other models. The '2-taxon' and 'Sister taxa introgression' models may result in $E[D_{GEN}]$ of 0, but is expected to have increased $\mathrm{Var}[D_{BLOCK}]$ relative to the 'No introgression' model providing a test for introgression. **B)** The four-taxon test tree used for analysis. **C)** The distributions of $\mathrm{Var}[D_{BLOCK}]$ calculated from 10^4 permuted genomes (see Methods) for test trees with *A. gambiae* S form (Grey) and *A. arabiensis* from Burkina Faso (Yellow) as the H3 outgroup are presented. The true $\mathrm{Var}[D_{BLOCK}]$ values from each empirical data set, presented on a broken x-axis for comparison, are greater than all permuted genomes in each case consistent with the presence of introgressed haplotypes in these genomes.



2380 **Figure 6: Significant autosomal introgression between *Anopheles* species and**
2381 **subspecies.** ABBA-BABA statistics were calculated in non-overlapping windows of 500
2382 informative sites using *A. merus* as the outgroup. Blue ribbon indicates 95% confidence
2383 region for introgression between S form ($2La^{a/+}$) and GOUNDRY (positive *D*; $2La^{a/a}$ and
2384 $2La^{+/+}$; *3R+; Xag*) and M form (negative *D*; $2La^{a/a}$; *3R+*; *Xag*). Orange ribbon indicates
2385 95% confidence region for introgression between *A. arabiensis* ($2La^{a/a}$; *3Ra*; *Xbcd*) and
2386 GOUNDRY (positive *D*) and M form (negative *D*). Horizontal dotted lines (orange = *A.*
2387 *arabiensis*; blue = S form) indicate genome wide significance level after correction for
2388 multiple testing. Positions of relevant chromosomal inversions indicated with horizontal
2389 white lines. A full list of genes within significant windows is given in Table S4.



2390
2391

Figure 7: Patterns of divergence among subgroups of *A. gambiae* follow similar curves (LOESS-smoothed with span of 1% using 10 kb non-overlapping windows), although differing slightly in magnitude, except increases in pericentromeric region in the M vs. S comparison and inside the 2L*a* inversion where these populations differ in karyotype (G-2L$a^{+/+}$, M-2L$a^{a/a}$, S-2L$a^{a/+}$). Divergence between the M form and *A. arabiensis* and *A. merus* is enriched on the X chromosome, especially inside the inverted X*ag* and *Xbcd* region (M vs. *arab*) and in pericentromeric regions (M vs. *merus*). Grey bars indicate locations of differentially fixed chromosomal inversions as well as the 2L*a* inversion and the large sweep on the GOUNDRY X (X*h*). Low complexity and heterochromatic regions were excluded.

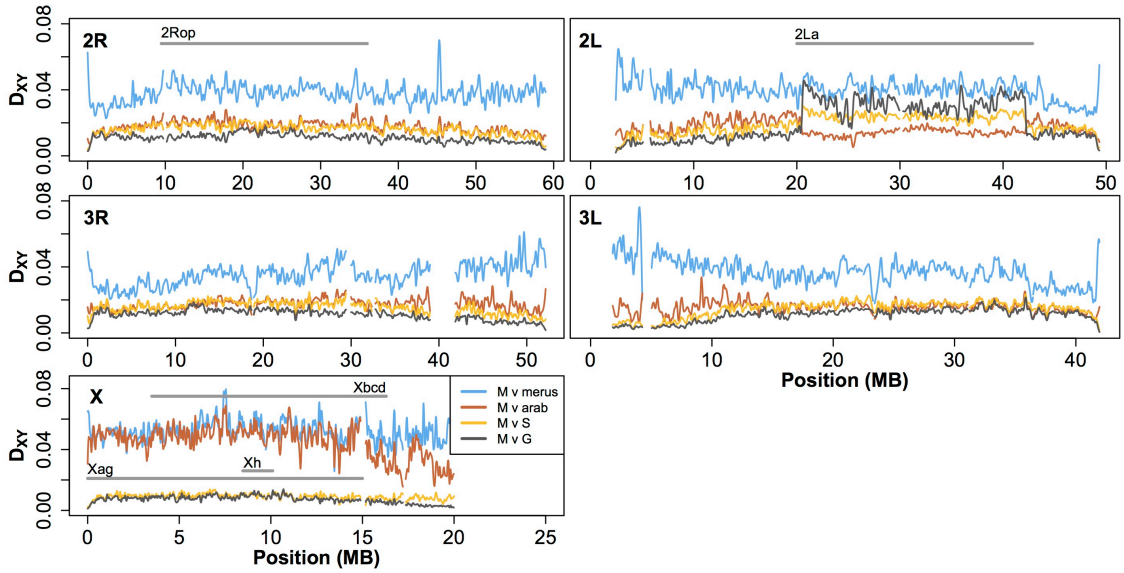
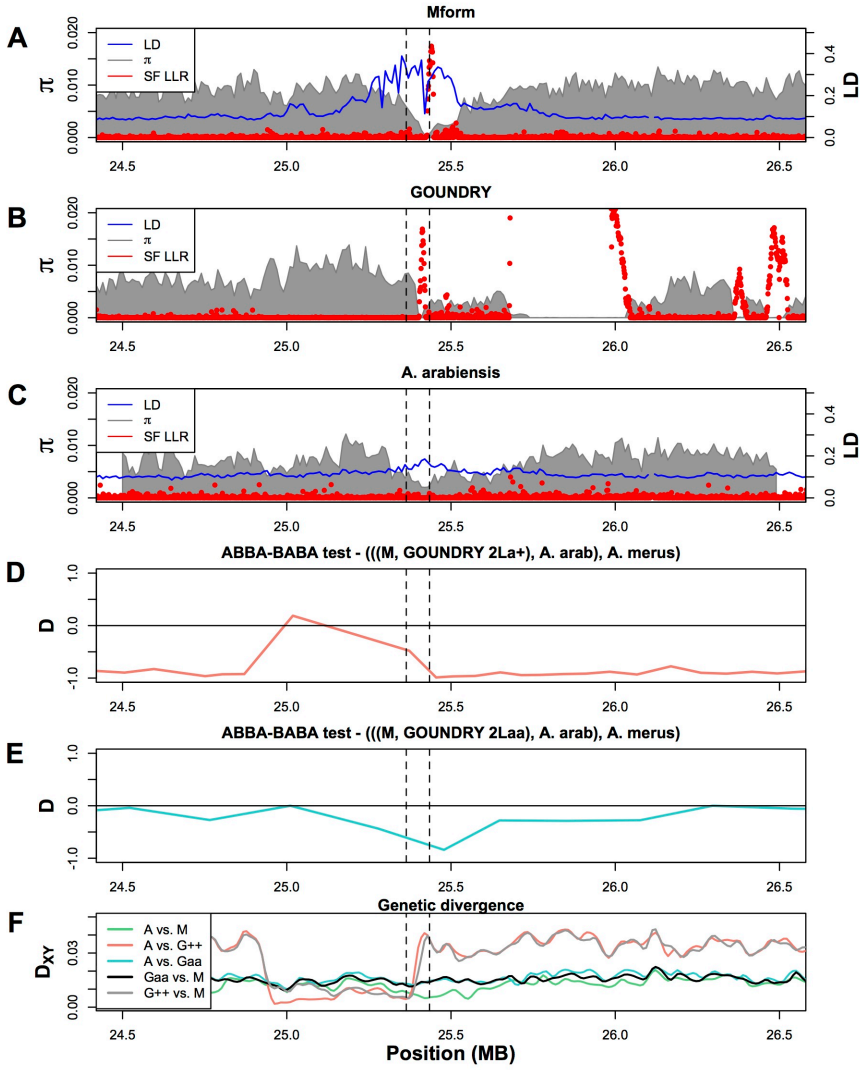
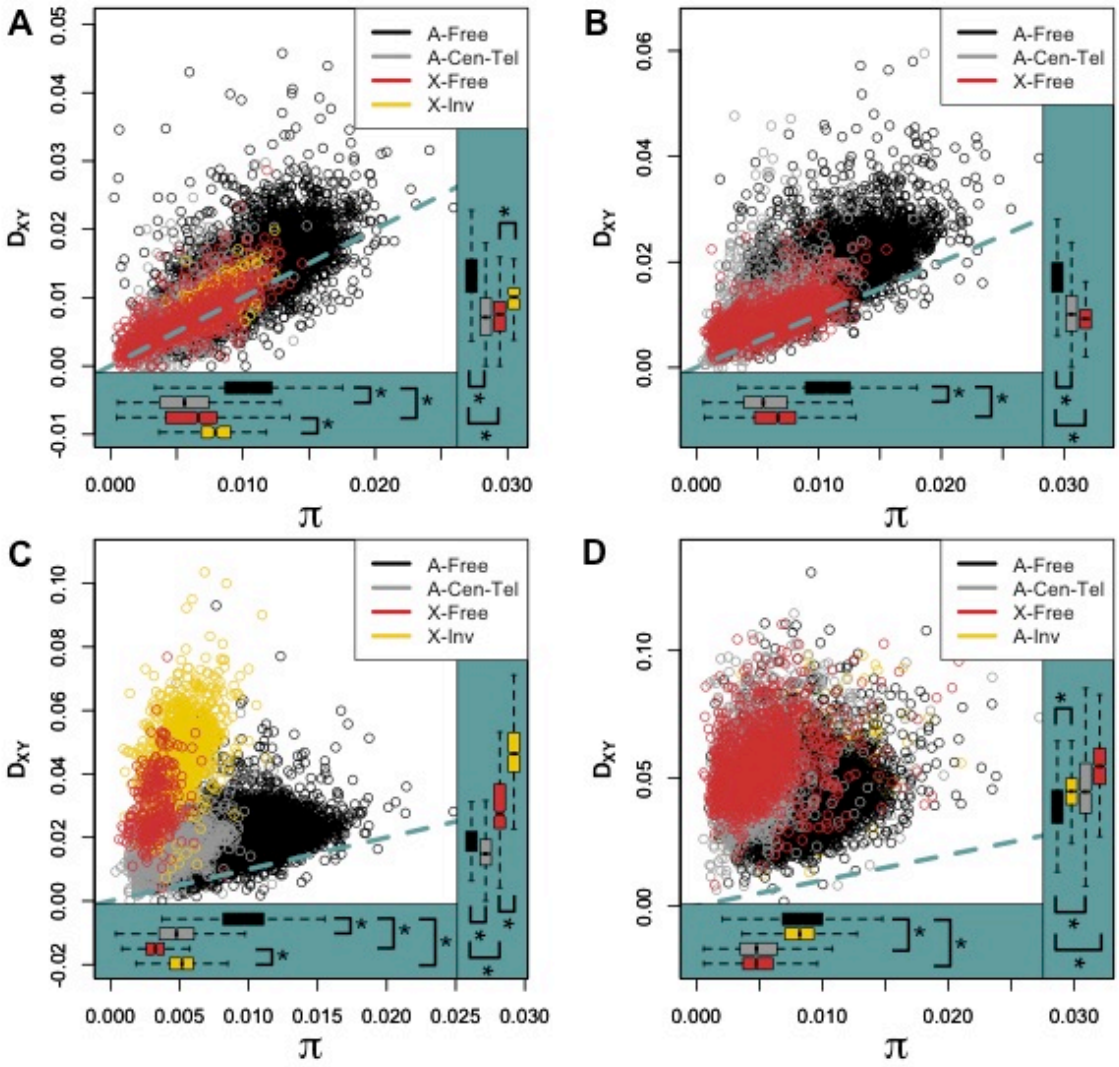


Figure 8: Patterns of nucleotide diversity (π, pi), background linkage disequilibrium (r^2 SNPs > 1 kb apart; LD), genetic divergence (D_{xy}), introgression (*D*), and selective sweeps (LLR) at the *Rdl* locus in *A. gambiae* M form, *A. gambiae* GOUNDRY, and *A. arabiensis*. Nucleotide diversity, LD, and SweepFinder LLRs are presented together for **A**) M form, **B**) GOUNDRY, and **C**) *A. arabiensis.* Y-axes are omitted for the SF LLR statistic. Linkage disequilibrium is not included for GOUNDRY because it could not be estimated robustly for GOUNDRY (see Methods). ABBA-BABA test statistics presented for the trees **D)** (((M form 2La^a,GOUNDRY 2La^+), *A. arabiensis* 2La^a), *A. merus*) and **E**) (((M form 2La^a,GOUNDRY 2La^a), *A. arabiensis* 2La^a), *A. merus*). **F)** Pairwise estimates of genetic divergence among *A. arabiensis* (A), M form (M), GOUNDRY 2L$a^{a/a}$ (Gaa), and GOUNDRY 2L$a^{+/+}$ (G++). Outer boundaries of the *Rdl* coding sequence are indicated with vertical dashed lines.

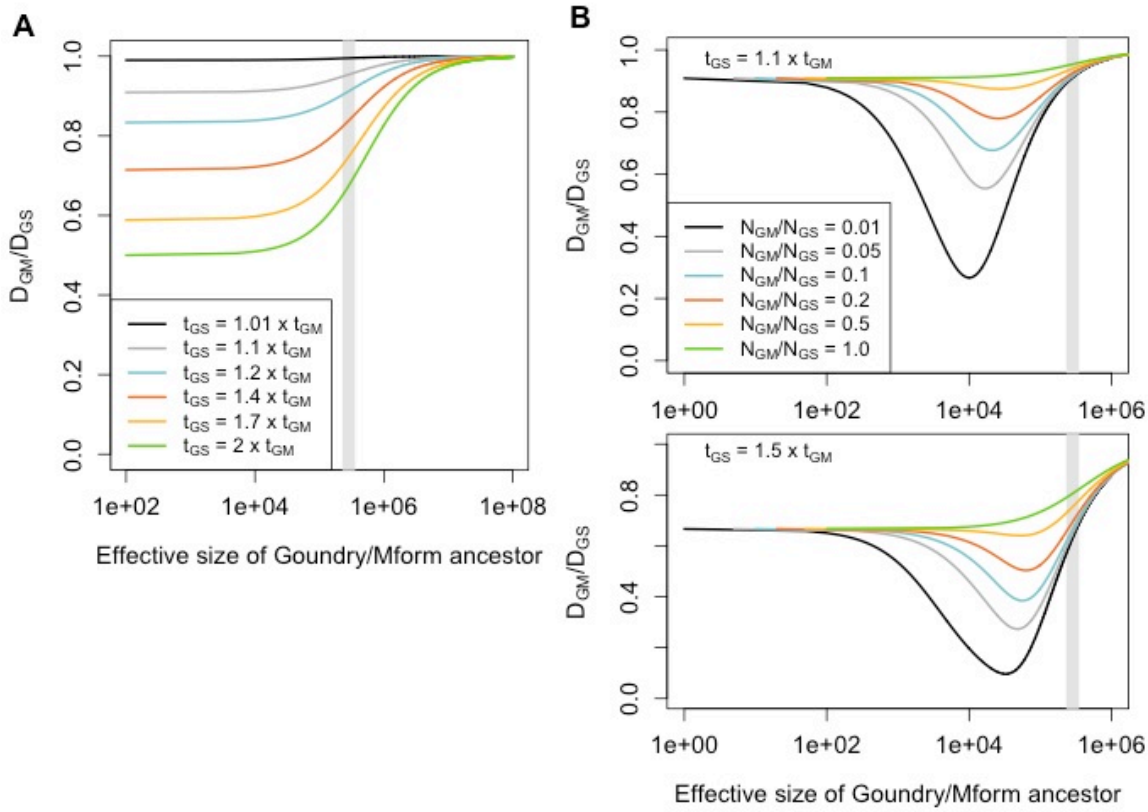


2424
2425
2426 **Figure 9: Patterns of genetic divergence (D_{xy}) between populations as a function of**
2427 **nucleotide diversity (π) reveal differential gene flow during speciation.** Genomic
2428 regions defined by expected rates of recombination in hybrids (see Methods) differ in
2429 their distributions of nucleotide diversity and genetic divergence, but not always in the
2430 same direction, suggesting that gene flow has been restricted on the X and lowly
2431 recombining regions in some cases. **A)** M form vs. GOUNDRY, **B)** M form vs. S form,
2432 **C)** M form vs. *A. arabiensis* **D)** M form vs. *A. merus*. Panel legends indicate colors
2433 corresponding to genomic location of each 10 Kb window where 'Free' indicates freely
2434 recombining regions, 'Cen-Tel' indicates centromeric and telomeric autosomal regions,
2435 and 'Inv' indicates chromosomal inversions. 'A-' and 'X-' indicate autosomal or X
2436 chromosome. Dashed blue-green line indicates perfect correlation. Asterisks indicate
2437 Mann-Whitney tests with P values < 3.92×10^{-05} for comparisons indicated with brackets.
2438 Note that the y-axis scale differs among panels.
2439



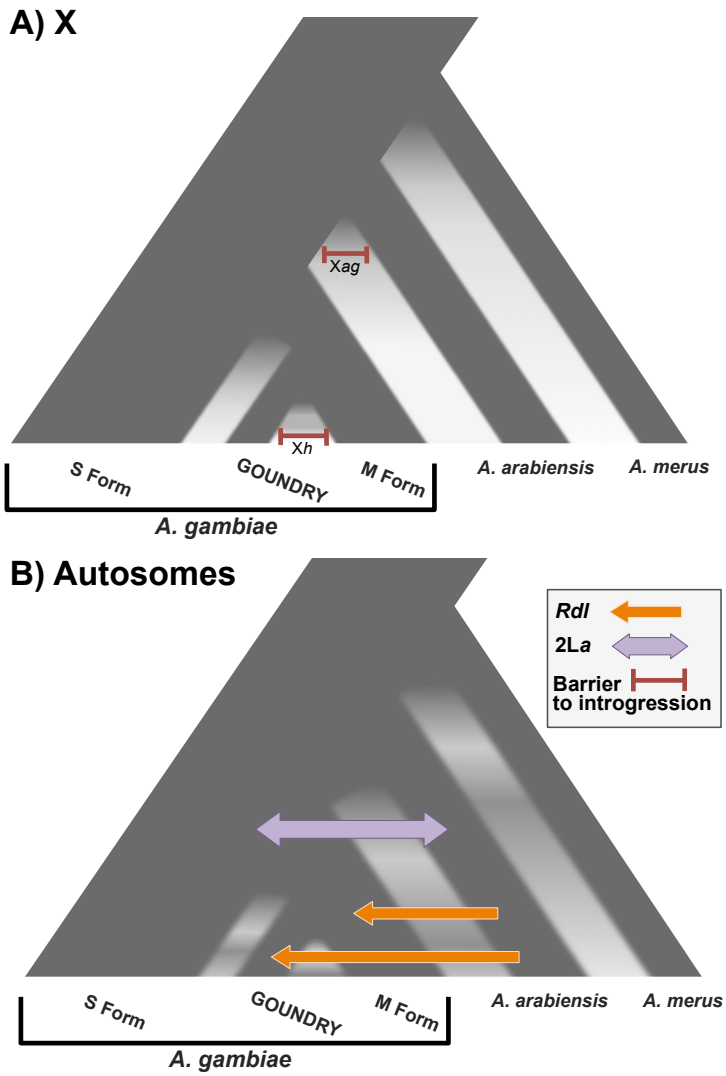
2440

2441 **Figure 10: Modeling expected values of Relative Node Depth (D_{GM}/D_{GS}). A)**
2442 Expected values of RND when ancestral population sizes are assumed to be equal.
2443 Colors indicate the expectations under different relative split times. **B)** Expected values
2444 with t_{GS} split time fixed to 1.1 (top) times the split time between GOUNDRY and M
2445 (t_{GM}) or 1.5 times (bottom). Colors indicated relative effective sizes of ancestral
2446 populations. Values are plotted as a function of the GOUNDRY-M effective size (x-axis).
2447 Grey bar indicates 95% confidence interval demographic estimate for GOUNDRY-M
2448 ancestral size (see Methods).
2449



2450

Figure 11: Evolution of founder populations in a species complex. Reproductive isolation among ecologically specialized founder populations and their ancestral population establishes differentially across the genome. **A)** Introgression is ineffective on the X chromosome in general, and chromosomal inversions may serve as barriers in some cases. **B)** Amongst a background of genetic divergence and adaptive differentiation on the autosomes, introgression maintains genetic connectivity across the speciation continuum occasionally allowing rapid adaptation to novel environments in the recipient population, including arid habitats (2L*a*) and insecticides (*Rdl*).



Tables:

Table 1: Optimized parameter values and confidence intervals from best fit demographic model for GOUNDRY and M form. See Figure S2 for parameter descriptions.

Parameter	**Optimized Value**	**95% Confidence Interval**	
		Lower	Upper
θ ($4N_A\mu L$)[1]	180,914		
N_A[2]	126,252	88,916	150,976
Split Times			
t_1	259,220	114,087	396,171
t_2	754,204	741,946	766,215
t_3	99,235	93,113	105,754
t_{TOT}	1,112,660	967,181	1,250,106
Population sizes			
N_{M1}/N_A	0.01	0.01	0.01
N_{M2}/N_A	5.74	4.58	7.65
N_{M3}/N_A	12.34	9.54	16.61
N_{G1}/N_A	0.01	0.00*	0.08
N_{G2}/N_A	0.19	0.11	0.32
N_{G3}/N_A	0.78	0.62	1.05
Migration rates (4Nm)[3]			
G1 into M1	0.010	0.0047	0.02
G2 into M2	1.29	1.25	1.33
G3 into M3	0.37	0.08	0.74
M1 into G1	0.080	0.00*	0.64
M2 into G2	1.17×10^{-4}	0.00*	0.0015
M3 into G3	1.50	1.44	1.55

1 – Instead of estimating a confidence interval for θ which itself is not model parameter, we solved for N_A and calculated a confidence for this parameter. In the implementation of *dadi*, N_A is used to scale all other parameters in the model.

2 – Ancestral population size was calculated from the estimate of θ by dividing this value by 4 times the number of sites times the mutation rate (see Methods above).

3 – Values of $4Nm$ were calculated by multiplying the migration rate reported by *dadi* ($2N_Am$) by 2 times the ratio of the effective size of the recipient population (e.g. N_{G1}) over N_A.

* indicates cases where the lower bound of the 95% CI was negative. This is not meaningful, so we set these values to 0.

2471
2472
2473 **Table 2:** Modified block-based ABBA-BABA test of introgression. Genome-wide 95%
2474 thresholds and the proportions of the genome in significant windows are presented for
2475 three comparisons.

H1[a]	H2[a]	H3[a]	Upper 95%[b]	Lower 95%[b]	Prop genome in sig M form Windows[c]	Prop genome in sig GOUNDRY Windows[c]
M form	GOUNDRY	S form	0.204	-0.204	0.0114	0.0325
M form	GOUNDRY	*A. arabiensis* (Burkina Faso)	0.248	-0.164	0.0364	0.0354
M form	GOUNDRY	*A. arabiensis* (Tanzania)	0.256	-0.156	0.0323	0.0312

a – Taxonomic assignment in the ABBA-BABA test tree ((H1,H2),H3),O).
b – Boundaries of the genome wide 95% threshold region estimated using block jackknife estimates of standard error within genomic regions after permutation (n = 10^4 replicates).
c – For each comparison, windows exceeding the 95% thresholds were identified, and the sum of their length was compared to the total length of the *A. gambiae* PEST reference.

2476
2477
2478 **Supplementary Table Legends:**
2479
2480 **Table S1**: Next-generation sequencing statistics for mosquito samples.
2481
2482 **Table S2**: Inferred selective sweeps in GOUNDRY, M form and *A. arabiensis*. Chr:
2483 chromosomal location of sweep; Peak Pos: location of window with highest LLR; LLR –
2484 log likelihood ratio from SweepFinder; P_{gen} – genome-wide *P* value, 0 values indicate *P*
2485 values less than 0.005; P_{loc} – per-locus *P* value, 0 values correspond to *P* values less than
2486 1.26×10^{-6} for autosomes and less than 6.67×10^{-6} for the X chromosome. Gene name,
2487 ImmunoDB ID, and Gene description from Vectorbase.org.
2488
2489 **Table S3:** Genes located inside large swept region (~8.47 – 10.1 MB) on X chromosome
2490 in GOUNDRY subgroup of *A. gambiae*. Chr: chromosomal location of sweep; Gene
2491 name, ImmunoDB ID, and Gene description from Vectorbase.org.
2492
2493 **Table S4:** Genes located inside introgressed windows with genome-wide significant
2494 values of *D*. Introg - Introgression between indicated subgroups; G_Abf – Introgression
2495 between GOUNDRY and Burkina Faso population of *A. arabiensis*; M_Abf -
2496 Introgression between M form and Burkina Faso population of *A. arabiensis*; M_Sform -
2497 Introgression between M and S forms; G_Sform - Introgression between GOUNDRY
2498 and S form; Chr – chromosomal location; Gene start bp – basepair coordinate of start in
2499 coding gene sequence; Gene end bp – basepair coordinate of end in coding gene
2500 sequence; Gene name, ImmunoDB ID, and Gene description from Vectorbase.org.
2501

2502 **Table S5:** Collection site and date information for mosquito samples.
2503
2504 **Table S6**: Summary statistics for mapping to *A. gambiae* PEST reference and subgroup
2505 SPEC reference.
2506
2507 **Table S7**: Sites included and excluded from analysis in all subgroups.
2508
2509
2510 **Supplementary Figure Legends:**
2511
2512 **Figure S1:** Test for differences in the amount of recent selection between the
2513 GOUNDRY versus M form subgroups. We tested whether the number of megabases
2514 harboring at least a single significant selective sweep window differed between the
2515 subgroups by subtracting the number of megabases in M form from the number of
2516 megabases in GOUNDRY. The red line indicates the true difference. The histogram
2517 presents the same calculation made using 10^4 datasets with subgroup assignments
2518 randomly permuted. No permuted genome produced the same or larger difference
2519 between the two subgroups.
2520
2521 **Figure S2:** Background Linkage Disequilibrium (LD) for each chromosomal arm in *A.*
2522 *gambiae* M form and *A. arabiensis*. LD (r^2) between SNPs separated by at least 1kb (10
2523 kb maximum) was averaged in 10 kb non-overlapping windows and LOESS-smoothed
2524 using a span of 1%. Low complexity and heterochromatic regions have been excluded.
2525 The large spikes on 2L and X (noted with vertical blue shaded bars) also coincide with
2526 reductions in nucleotide diversity and are thus likely to be recent selective sweeps.
2527 Additionally, there are a number of long distance increases (noted with horizontal grey
2528 bars) in LD in both M form and *A. arabiensis* that coincide approximately with the
2529 locations of known chromosomal inversions segregating in these subgroups [41].
2530
2531 **Figure S3:** Mean total read depth for GOUNDRY X chromosome sweep region. Mean
2532 total read depth across all GOUNDRY samples (n=12) for sites within non-overlapping
2533 500 bp windows and plotted as a function of chromosomal position (megabases). The
2534 position of large GOUNDRY X sweep region is shown with grey bar.
2535
2536 **Figure S4:** Relative genetic divergence (D_a) between GOUNDRY and M form. *Da*
2537 plotted as a function of M form nucleotide diversity using only intergenic sites in non-
2538 overlapping 10kb windows. Low complexity and heterochromatic regions were excluded.
2539 Genomic regions were defined based on predicted rates of recombination in hybrids (see
2540 Methods) and compared using non-parametric Mann-Whitney tests. X-Free: freely
2541 recombining regions on X chromosome. X-Inv: region inside putative X*h* chromosomal
2542 inversion. Asterisks indicate P values less than 4.896e-08.
2543
2544 **Figure S5:** ΔD_{xy} between *A. gambiae* subgroups GOUNDRY and M form plotted for
2545 windows across the X. Permutation test *P*-values for GOUNDRY vs. M form
2546 comparisons presented on log-scale in bottom panel with standard and Bonferonni-
2547 corrected thresholds (Methods). Grey bar indicates inverted X*h* chromosomal region.

2548
2549 **Figure S6:** Autosomal 1D unfolded site frequency spectra (SFS) for intergenic variable
2550 sites in each subgroup. The SFS expected under standard coalescent equilibrium
2551 conditions is presented in black. The SFS was inferred for M form and *A. arabiensis*
2552 using –realSFS 1 (see Methods), but the SFS for GOUNDRY was inferred using the
2553 inbreeding-aware version (-realSFS 2) after estimating inbreeding coefficients for this
2554 subgroup (see Methods).
2555
2556 **Figure S7:** Inbreeding coefficients for each individual and each chromosomal arm.
2557 Inbreeding coefficients were estimated directly from genotype likelihoods for each
2558 chromosomal arm separately using Vieira et al. 2013 (see Methods). Each panel
2559 corresponds to a chromosomal arm. Each bar represents an individual mosquito sample.
2560 Colors indicate subgroup assignment according to the legend.
2561
2562 **Figure S8:** Decay of Linkage Disequilibrium (LD) in *A. gambiae* M form and *A.*
2563 *arabiensis*. LD (r^2) between SNPs separated by no more than 5 kb binned, averaged, and
2564 plotted as a function of physical distance. Low complexity regions were excluded.
2565 Chromosomal arms 2R, 3L, and 3R were included for the autosome curves and the X
2566 plotted separately. Note different X and autosome y-axis scales
2567
2568 **Figure S9:** Comparison between the data, the best-fit model from *dadi*, and neutral
2569 coalescent simulations using MaCS. Site-frequency spectra were calculated from the **A)**
2570 true autosomal data, **B)** best-fit autosomal demographic model inferred using *dadi*, and
2571 **C)** polymorphism data generated using neutral coalescent simulations in MaCS. LD
2572 decay curves from true data (red line) are compared to decay curves inferred from **D)**
2573 coalescent simulations for X chromosome (grey), and **E)** coalescent simulations for
2574 autosomal data.
2575
2576 **Figure S10:** A) Distribution of maximum SweepFinder log-likelihood ratio statistics in a
2577 50 kb region of neutral polymorphism data simulated under the demographic model
2578 inferred for GOUNDRY (top) and M form (bottom). See Methods for simulation details.
2579 B) Distribution of maximum SweepFinder log-likelihood ratio statistics in a full synthetic
2580 genome of neutral polymorphism data simulated under the demographic model inferred
2581 for GOUNDRY (top) and M form (bottom). See Methods for simulation details.
2582
2583 **Figure S11:** A) A representative cluster of significant selection peaks from Sweepfinder
2584 in GOUNDRY that were collapsed. The top panel shows the LLR profile when
2585 SweepFinder was applied to only variable sites (Red = genome wide significance; Gold =
2586 per-locus significant; Grey = not significant). The bottom panel presents the LLR profile
2587 when SweepFinder was applied to all sites. Critical thresholds were not established for
2588 'all-site' analyses. The different y-axis values reflect the different sets of data considered
2589 in the composite likelihood function. The dotted line on the bottom panel indicates our
2590 threshold for 'low' values (LLR = 20). Under our clustering approach, this cluster was
2591 collapsed into a single peak. B) A representative cluster of significant selection peaks
2592 from Sweepfinder in GOUNDRY that were not collapsed. The top panel shows the LLR
2593 profile when SweepFinder was applied to only variable sites (Red = genome wide

significance; Gold = per-locus significant; Grey = not significant). The bottom panel presents the LLR profile when SweepFinder was applied to all sites. Critical thresholds were not established for 'all-site' analyses. The dotted line on the bottom panel indicates our threshold for 'low' values (LLR = 20). The different y-axis values reflect the different sets of data considered in the composite likelihood function. Under our clustering approach, this cluster was not collapsed into a single peak.

Figure S12: Estimating the age of the selective sweep inside the putative (X*h*) inversion on the X chromosome in GOUNDRY. **A)** Each line shows the number of sites in millions that pass filtering according to different minimum read depth filters. **B)** The number of mutations counted for each individual according to minimum read filters. **C)** The average age in generations of the two chromosomes in each diploid individual as function of minimum read depth. **D)** Estimates for age of the most recent common ancestor as a function of minimum read depth. Bars indicate standard deviations as calculated from the average number mutations per haplotype. Colors correspond to individuals and are the same among panels A-C, with GOUND-0446 indicated in light-blue.

Figure S13: Minor allele counts at variable sites within the large X-linked sweep region (X*h*) in GOUNDRY genomes. Counts were made using called genotypes. Each panel presents a histogram distribution of allele counts at sites covered by at least 6-12 reads per individual (see Methods).