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

Running Head (50 Chars): 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:

Species complexes are common, especially among insect disease vectors, and understanding how barriers to gene flow among these populations become established or violated is critical for implementation of vector-targeting disease control. *Anopheles gambiae*, the primary vector of human malaria in sub-Saharan Africa, exists as a series of ecologically specialized populations that are phylogenetically nested within a species complex. These populations 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 adaptive introgression of autosomal chromosomal regions among species and populations. The X chromosome, however, remains strongly differentiated among all of the subpopulations, 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 recently swept to fixation in the GOUNDRY subpopulation, which may be an inversion that serves as a partial barrier to gene flow. We also find that the GOUNDRY population is highly inbred, implying increased philopatry in this population. Our results show that ecological speciation in this species complex results in genomic mosaicism of divergence and adaptive introgression that creates a reticulate gene pool connecting vector populations across the speciation continuum with 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 populations, 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 isolation has led to genomic differentiation among mosquito species and populations, especially on the X chromosome. However, we also find evidence for exchange of genes on the autosomes that has provided the raw material for recent rapid adaptation. For example, we show that *A. arabiensis* has shared a mutation conferring insecticide resistance with two subgroups of *A. gambiae* within the last 60 years, illustrating the fluid nature of species boundaries among even more advanced species pairs. Our results underscore the expected challenges in deploying vector-based disease control strategies since many of the world's most devastating human pathogens are transmitted by arthropod species complexes.

Introduction:

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

The *Anopheles gambiae* species complex in sub-Saharan Africa includes several major vectors of malaria, which continues to place a devastating burden on local human populations [7]. Prior to the 1940s, *A. gambiae* was considered a single biologically variable species, but crossing studies and genetic analysis led to the subdivision of the species into a complex of nine morphologically similar named species that vary in their geographic distribution and ecology (reviewed in [8]). It is becoming increasingly appreciated from ecological distinctions and recent discoveries of additional genetic substructure that, even within species, *Anopheles* species frequently form partially reproductively isolated and differentiated subpopulations [9–11]. We term these "founder" populations to indicate their role in the early stages of speciation.

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

an often high level of local M-S hybridization[11], illustrating the fluid semispecies boundaries within the *Anopheles gambiae* group. It is not known how many cryptic subpopulations exist within *Anopheles*, or how much gene flow they share, but there is evidence that they may be common [11] as exemplified by the recently discovered GOUNDRY population of *A. gambiae* from Burkina Faso which shows considerable genetic distinction from other described *A. gambiae* and is particularly permissive for *Plasmodium* development [17].

Epidemiological modeling and vector-based malaria control strategies must explicitly consider unique aspects of subpopulations such as M, S, and GOUNDRY if they are to effectively predict disease dynamics and responses to intervention [18]. Because many of the world's most devastating human diseases are vectored by arthropods in species complexes, including *Anopheles* vectors of malaria, tsetse fly vectors of African sleeping sickness, and *Ixodes* tick vectors of Lyme disease [19–21], the dynamic evolution of phenotypic diversity and adaptive introgression among cryptic taxa in species complexes has serious implications for public health. In the specific case of malaria, the existence of partially differentiated but occasionally hybridizing subspecies could complicate malaria control efforts that rely on the spread of transgenes through mosquito populations or conventional controls that target specific aspects of mosquito ecology or life history [22].

Genomic dissection of species complexes also lends insight into processes of speciation. Studies have shown strong differences between sex chromosomes (X or Z) and autosomes in the rates of genetic divergence and introgression and therefore speciation among pairs of species in systems ranging from *Drosophila* to Hominids. Genetic divergence tends to be higher on sex chromosomes relative to autosomes and introgressed regions are underrepresented on the sex chromosomes [23–25]. Heterogeneity in levels of genetic divergence has also been shown among genomic regions that vary in rates of meiotic recombination, with elevated divergence in genomic regions with low recombination rates such as centromeric regions and chromosomal inversions [26,27]. Analysis of genetic divergence and introgression in the *Anopheles gambiae* species complex provides a valuable opportunity to test the generality of these patterns across the speciation continuum.

In the present study, we have completely sequenced the genomes of 32 field-captured *Anopheles* mosquitoes to determine the extent of genetic differentiation among species/subpopulations and to infer the role of natural selection in ecological specialization along the speciation continuum of the species complex. We sequenced *A. gambiae* GOUNDRY (n=12), *A. gambiae* M form (n=10), *A. gambiae* S form (n=1) and *Anopheles arabiensis* (n=9), and used publicly available genome sequences of *Anophleles merus* as an outgroup. We find evidence for strong isolation and adaptive differentiation among populations and species, with disproportionately high divergence along the length of the X chromosome and in regions of known chromosomal inversions. At the same time, we can identify examples of adaptive introgressions across partially isolated subspecies, including of alleles that potentially confer insecticide resistance. Each of the populations has a unique demographic history, which highlights the evolutionary complexity of closely related species groups.

Results and Discussion

Evidence for Extensive Adaptive Variation

Ecological speciation typically involves natural disruptive selection on ecologically relevant traits involving one or more genetic loci [2]. To identify loci putatively involved in ecological specialization and quantify the role of selection in adaptive differentiation, we scanned the genomes of GOUNDRY, M form, and *A. arabiensis* for signals of recent positive natural selection. Consistent with adaptation to distinct ecological niches, we find evidence for a number of recent selective sweeps that are private to each of the three populations [Figure 1, Table S1; Supplementary Material]. We found that the M form population has experienced more positive selection than GOUNDRY in the recent past [Supplementary Material; $P < 0.0001$]. We find evidence of 120 recent selective sweeps in the M form population, compared to only 67 in GOUNDRY and 33 in the *A. arabiensis* genome.

Two of the swept loci in the M form [*TEP1* and *Resistance to dieldrin* (*Rdl*)] have been identified previously using independent datasets and analytical approaches [28][13], providing an important validation of our analysis pipeline. These genes are involved in

immunity and resistance to insecticide, respectively. We additionally find a particularly intriguing novel sweep at a gene encoding a neuropeptide F (NPF; AGAP004122). In *Drosophila*, NPFs have been shown to control larval behavior and feeding[29], so it is tempting to speculate that adaptive differentiation of this gene may underlie the previously described unique larval bottom-feeding and extended predator avoidance behaviors of the M form [30].

The sweeps in the *A. arabiensis* genomes include genes encoding two proteins involved in development, Eclosion hormone and Frizzled-2, and the gene encoding the Sulfonylurea receptor. RNAi knockdown of this receptor in *Drosophila melanogaster* significantly increases susceptibility to viral infections [31].

One particularly notable sweep in GOUNDRY lies near the gene encoding the master gustatory and odorant receptor *Agam/Orco* (*AgOr7*) that is a required coreceptor for all odorant binding proteins [32]. Expression of *AgOr7* is predominantly localized in female antennae and palps and fluctuates with circadian cycles, leading to the hypothesis that this gene controls temporal shifts in olfactory sensitivity [33]. Adaptation at *AgOr7* could have important impacts on anthropophily, sensitivity to DEET [34], and may underlie exophilic behavior in GOUNDRY [17].

The most striking sweep in GOUNDRY covers 1.67 Mb on the X chromosome and results in nearly complete absence of polymorphism across this region [Figure S1, Supplementary Text]. The remarkably large size of the region devoid of diversity would imply exceptionally strong positive selection under standard rates of meiotic recombination. For comparison, previously identified strong sweeps associated with insecticide resistance span approximately 40kb and 100kb in freely recombining genomic regions of *Drosophila melanogaster* and *D. simulans,* respectively [35,36]. The swept region in GOUNDRY is marked by especially sharp boundaries relative to the other footprints of selection inferred from our data (Figure S1), implying that recombination has been suppressed in this region. Collectively, these observations suggest that the swept region may contain a small chromosomal inversion, which we have named X*h*. The region includes 92 predicted protein coding sequences (Table S2), including the *white* gene, two members of the gene family encoding the TWDL cuticular protein family (*TWDL8* and *TWDL9*), and five genes annotated with immune function (*CLIPC4,*

191 *CLIPC5, CLIPC6, CLIPC10, PGRPS1*). The remarkable lack of diversity in the region
192 implies that the presumed X*h* inversion has a single recent origin and was quickly swept
193 to fixation in GOUNDRY. Based on the very low number of mutations observed to have
194 arisen subsequent to the sweep, we infer that the haplotype reached fixation
195 approximately 78 years ago ($\sigma = 9.14$; assuming 10 generations per year; Supplementary
196 Text). Such extraordinarily recent adaptation in an otherwise old population is consistent
197 with the selection pressures related to 19th and 20th century human activity such as
198 insecticide pressure or widespread habitat modification.
199
200 *Introgression Facilitates Adaptation*
201 Natural selection is expected to remove most introgressed genetic material due to
202 ecological misfit or Dobzhanski-Muller Incompatibilities [37,38], especially between
203 more distant species, but some introgressed alleles may be selectively favored in the
204 recipient population. In line with previous studies [39–42], we find clear evidence for
205 introgression of the large 2L*a* inversion between *A. gambiae* and *A. arabiensis*. The local
206 frequencies of 2L*a* karyotypes are driven by environmental selection in West and Central
207 Africa [40]. We observe that genetic divergence between species is strongly and highly
208 significantly reduced across the entire inverted region relative to genome averages
209 (Figure S2), consistent with introgression of the inversion across species boundaries after
210 the majority of the genome had become differentiated.
211 To identify other putatively adaptive introgression events, we examined the
212 intersection between our scans for selection and for introgression (described below). We
213 found 17 selective events (10 in M form, 6 in GOUNDRY, 1 in *A. arabiensis*) that may
214 be due to selection on alleles that were originally introgressed from other populations or
215 species. One such example is the *Resistance to dieldrin* locus. In this case, the adaptive
216 allele that confers insecticide resistance swept first in *A. arabiensis* and more recently
217 swept in both M form and GOUNDRY (Fig. S3). In contrast to previous suggestion that
218 independent mutations at *Rdl* confer resistance in *A. arabiensis* and *A. gambiae* [43], we
219 find evidence for multiple introgression events at this locus. ABBA-BABA tests support
220 introgression from *A. arabiensis* into M form and subsequent sweep of the introgressed
221 allele within the M form (Figure S3). Moreover, we find that a large haplotype on 2L

overlapping the *Rdl* locus has been introgressed from *A. arabiensis* into GOUNDRY 2La^{+} chromosomes making GOUNDRY more closely related to *A. arabiensis* in part of this region (Figure S3), and that the *Rdl* locus was subsequently swept in GOUNDRY. Since the selective sweep in *A. arabiensis* was probably driven by insecticide pressure initiated in the 1950's [44], we conclude that introgressive hybridization has occurred among these taxa in very recent evolutionary history. This observation is consistent with other studies supporting adaptive introgression of insecticide resistant alleles at a separate locus among populations of *Anopheles* [45,46]. We additionally find a large region on chromosome 2L (~21.5-23.5 MB) that is enriched for introgression of derived mutations between GOUNDRY (2La^{a}) and *A. arabiensis* that harbors a recent sweep in *A. arabiensis* at the gene AGAP005855, which has no known function. Overall, our results indicate that sharing of advantageous genetic variants may be common in *Anopheles*. Importantly, we show that even rare hybridization events have important effects on contemporary adaptation in *Anopheles* populations to novel environmental conditions including anthropogenic modification like insecticides.

Introgression is Less Effective Common on the X Chromosome

Genetic introgression can impede ecological specialization among diverging populations, and patterns of differential introgression along the genome provides information about the location of genes involved in reproductive isolation and local adaptation. We hypothesized that differential introgression along the genome will be reflected in patterns of genetic divergence, since genetic divergence will be partially determined by the efficacy of introgression. As expected, we observe that genetic divergence (D_{xy}) among species and subspecies in the *Anopheles gambiae* species complex is higher than intra-population diversity in all comparisons (Figure 2), confirming that all groups are at least partially distinct. We compared genetic divergence in low-recombining pericentromeric regions, chromosomal inversions, and the X chromosome to the freely recombining autosomes in order to test the hypothesis that the speciation process establishes differentially among genomic regions of low meiotic recombination [23,25,47–49]. Since variation in mutation rate and the effects of linked selection along the genome would lead to variation in both intra-population genetic

253 diversity as well as inter-population genetic divergence among genomic regions even if
254 the rates of introgression were equal, we also evaluated heterogeneity in genetic diversity
255 among genomic regions.

256 We find that genetic divergence among *Anopheles* populations generally scales
257 with nucleotide diversity across the genome (Figure 2), consistent with broadly neutral
258 divergence and suggesting that the rate of introgression among populations is
259 qualitatively similar genome-wide. However, when divergence along the X chromosome
260 is explicitly scaled by the mutation rate inferred from levels of polymorphism (D_a), the
261 putatively adaptive X*h* inversion between GOUNDRY and M form is proportionally
262 much more divergent than is the remainder of the X chromosome ($P < 4.89\times10^{-08}$; Fig.
263 S4; Supplementary Text). Based on this and results from an additional permutation test
264 of divergence patterns along the X chromosome and coalescent-based model comparison
265 between the X and autosomes (Supplementary Text), we conclude that X*h* serves as an
266 additional partial barrier to introgression with the M form.

267 Hybrid male sterility maps to the X chromosome in *A. gambiae*-*A. arabiensis*
268 crosses and two large X-linked chromosomal inversion complexes [X*ag* and X*bcd*]
269 suppress recombination on the X in hybrids of these species, so we hypothesized that less
270 frequent introgression may lead to exceptionally high sequence divergence on the X
271 relative to other genomic regions. In support of this hypothesis, we found that nucleotide
272 diversity on the X is significantly lower than on the autosomes (Mann-Whitney test
273 (hereafter M-W) $P < 2.2\times10^{-16}$), but that genetic divergence is significantly higher on the
274 X (M-W $P < 2.2\times10^{-16}$; Figure 2). Moreover, genetic divergence is significantly higher
275 in the region harboring inversions (M-W $P < 2.2\times10^{-16}$) than in the surrounding
276 chromosome. Nucleotide diversity is also higher in this region relative to the centromere-
277 proximal region (M-W $P < 2.2\times10^{-16}$), where nucleotide diversity is especially low,
278 presumably due to the effects of linked selection on neutral genetic variation. While we
279 cannot formally rule out an elevated mutation rate inside the inverted region, there is no
280 reason to expect the inverted region to be more mutable. Our observations are consistent
281 with previous analyses of genetic differentiation among *A. arabiensis* and *A. gambiae*, as
282 well as with laboratory backcrossing experiments [41,50,51], indicating that introgression

is particularly inhibited on the X chromosome, and that the X chromosome plays a disproportionately large role in driving speciation.

The X chromosomes of the *A. gambiae* M and S forms are collinear with that of *A. merus*, so recombination is not expected to be suppressed in contemporary or historical hybrids among these groups. To test whether sequence divergence is exceptionally high on the X relative to the autosomes in these comparisons despite the lack of inverted regions, we compared genetic divergence and nucleotide diversity among genomic regions as described above. We find that genetic divergence between M and S scales with nucleotide diversity across the genomic regional classes (Figure 2). However, inspection of chromosomal distributions of divergence reveals a slight elevation in genetic divergence in a chromosomal region where nucleotide diversity is low relative to the rest of the chromosome (~16-20 MB; Figures S1 and S2) and recombination is known to be proportionally rare in M-S hybrids [52]. In contrast, genetic divergence between *A. gambiae* (M form) and *A. merus* does not scale with nucleotide diversity (Figure 2). Nucleotide diversity is lower on the X than on the autosomes (M-W $P < 2.2\times10^{-16}$), but genetic divergence is significantly higher on the X (M-W $P < 2.2\times10^{-16}$). Overall, these findings are consistent with a large role for the X in driving speciation even among populations with collinear X chromosomes.

Very recently diverged populations may still share polymorphism even if they are completely reproductively isolated in practice. To explicitly differentiate between incomplete lineage sorting in populations that have only recently become isolated versus extant introgression among populations that are only partially isolated, we tested for excess sharing of derived mutations relative to expectations under strict lineage sorting using the ABBA-BABA test [53]. We assigned GOUNDRY and M form as sister taxa and *A. merus* as the outgroup (as in Fig. 4) and asked whether the sister taxa share an excess of derived mutations with either *A. arabiensis* or S form. Amidst widespread divergence, we find significant evidence for introgression between GOUNDRY and both the S form population and *A. arabiensis* [$P < 0.0001$ for both tests, Figure 3; Supplementary Text]. After conservatively correcting for multiple testing genome-wide, we find that windows representing 3.2% of the GOUNDRY genome share a significant excess of derived mutations with the S form and 3.5% of the GOUNDRY genome shares

a significant excess with *A. arabiensis*. In contrast, we also find significant evidence of introgression between M form and both S form and *A. arabiensis*, but the proportions of the genome that are represented by significant windows are 1.1% and 3.6% with S form and *A. arabiensis,* respectively. Despite considerable evidence for introgression on the autosomes, we find no evidence of introgression of X chromosome sequence among any subgroups, which reinforces our interpretation of a disproportionately large role for the X in speciation.

It is clear that introgression has occurred in the evolutionarily recent past among the *Anopheles* taxa examined here, and we can also ask whether introgression has occurred via contemporary hybridization through comparisons of introgression with sympatric and allopatric populations. If introgression has occurred recently, we expect stronger affinity among sympatric relative to allopatric populations. We tested whether introgression between *A. gambiae* M form and *A. arabiensis* has been recent using the ABBA-BABA test with sympatric *A. arabiensis* versus allopatric *A. arabiensis* from Tanzania. We observe a significant excess of shared derived mutations among sympatric populations, which is consistent with recent gene flow (Supplementary Text). We find a significant excess of shared derived mutations between M form *A. gambiae* and sympatric *A. arabiensis* relative to allopatric *A. arabiensis* from Tanzania (D = -0.0542, Z-score = -13.1533, $P = 1.63\times10^{-39}$). Similarly, we find evidence for significant introgression between sympatric *A. arabiensis* and GOUNDRY (D = -0.0441, Z-score = -11.7559, $P = 6.58\times10^{-32}$). In contrast to expectations that introgression was the byproduct of historical hybridization between *A. arabiensis* and *A. gambiae*, our results provide strong evidence that introgression continues to occur via contemporary hybridization, with strong implication for the evolution of both ecologically and epidemiologically relevant traits.

The X chromosome has been shown to have disproportionately large effect in driving speciation in organisms ranging from *Drosophila* to mammals [23,25,54,55], yet a unifying explanation for this pattern has yet to emerge. One hypothesis to explain this pattern is that gene expression dosage is mis-regulated, causing sterility in hybrids [56]. A second hypothesis is that since the X is haploid in males, X-linked recessive mutations are exposed to selection, which can lead to faster adaptation on the X relative to the

autosomes ['faster-X'; [47]]. Higher rates of adaptive evolution on the X could partially explain the disproportionately large X effect if adaptive substitutions lead to Dobzhanski-Muller Incompatibilities in hybrids [37,38,57]. To determine whether this hypothesis could explain the large role of the X in speciation in *Anopheles*, we tested for higher rates of adaptive protein evolution on the X chromosome relative to autosomal proteins by comparing the proportion of non-synonymous fixations that were adaptive (α) in ~7,000 single-copy X-linked and autosomal genes as a class along each branch of the phylogeny in a McDonald-Kreitman test framework [58,59]. Power to estimate α is greatest on the long branches leading from the *A. gambiae* populations to *A. merus* and to *A. arabiensis*, and here we find that, although there is a strong excess of adaptive non-synonymous substitutions in both X-linked and autosomal loci on both branches (both tests $P < 10\times10^{-8}$, Figure S5), α for X-linked genes is lower (28%) than on the autosomes on both branches (53% for *A. arabiensis* and 31% for *A. merus*). Power is lower on branches among the *A. gambiae* subspecies due to the lower number of private substitutions (Figure S5) and evidence for faster X-linked protein evolution in this clade is equivocal. Overall, our data suggest that the disproportionately large X effect in driving speciation may not be explained by accumulation of DMIs due to 'faster X' adaptive protein evolution.

Origins of GOUNDRY, a New Cryptic Taxon

The GOUNDRY subgroup of *A. gambiae* was discovered in larval collections along a transect in Burkina Faso, but its relationship to the M and S molecular forms of *A. gambiae* remains unclear. We calculated genetic distance (D_{xy}) between all pairs of the M form, S form, and GOUNDRY populations using intergenic autosomal sites and found that GOUNDRY and the M form population are genetically more similar to each other ($D_{GM} = 0.0126$) than either is to the S form population ($D_{GS} = 0.0162$; $D_{MS} = 0.0162$; Figure S6). This is contrary to the previous placement of GOUNDRY as an outgroup to M and S based on microsatellite data [17], and indicates that GOUNDRY was founded as an offshoot from the M form after the M form had already differentiated from the S form. This population history exemplifies the serial founder model we believe is common in *Anopheles* populations.

It has been hypothesized that the advent of agriculture in sub-Saharan Africa ~5-10 kya played a role in driving diversification and expansion of *Anopheles* mosquitoes [60]. To test whether the origin of GOUNDRY could have been associated with habitat modification driven by agriculture, we fit three-epoch population historical models to the two-dimensional site frequency spectrum for GOUNDRY and M form *A. gambiae*. Although estimates of such old splits times inherently carry considerable uncertainty, the best-fitting model predicts that these populations diverged 111,266 ya (95% 96,718 – 125,010), followed by a 100-fold reduction in the size of both populations after isolation (Supplementary Material), and thus rejects any role of modern agriculture in subpopulation division. Our inferred model is consistent with habitat fragmentation and loss due to natural causes. The model supports a >500-fold population growth in M form and 19-fold growth in GOUNDRY with extensive gene flow between them 85,343 ya, consistent with a re-establishment of contiguous habitat and abundant availability of bloodmeal hosts. Hybridization related to secondary contact has not led to complete homogenization, however, as we conservatively identified nearly 9,000 fixed nucleotide differences distributed across the genomes of the two subgroups (Supplementary Material).

Overall, our data support a model where GOUNDRY was founded ~112 kya and existed in relative isolation prior to secondary contact with the M form. This secondary contact may have been related to the inferred large population growth estimated ~85 kya, and should have allowed more recent genetic exchange between GOUNDRY and M form *A. gambiae*. The X-linked inversion X*h* newly identified here now probably serves as a partial barrier to introgression from the M form.

Unexpectedly, we found that GOUNDRY exhibits a deficiency of heterozygotes relative to Hardy-Weinberg expectations and extensive regions of Identity-By-Descent (IBD), a pattern that is not observed in any of our other samples. Individual diploid GOUNDRY genomes are checkered with footprints of IBD, even though the genome as a whole harbors substantial genetic variation indicating a relatively large effective population size. The observation of stochastic tracts of IBD is most consistent with an unusually high rate of close inbreeding. Through careful analysis of mapping biases and variation in read depth, we can reject bioinformatic artifacts as possible explanations for

this signal (Supplementary Text). The specific chromosomal locations of the IBD regions are random and vary among the sequenced GOUNDRY individuals (Figure S7). We observe several tract lengths that span nearly 40% of the chromosome (Figure S8) suggesting the possibility of mating among half-siblings or first-cousins. IBD tract lengths vary substantially, often spanning very small regions, suggesting that inbreeding in GOUNDRY is not restricted to recent generations. For an independent estimate of deviations from Hardy-Weinberg equilibrium expectations, we used a maximum likelihood framework to infer inbreeding coefficients for each individual without explicitly calling genotypes and found that inbreeding coefficients range from 0 to nearly 70% per chromosomal arm (Figure S9), which are values comparable to some domesticated dog lineages [61].

We emphasize that the deficit of heterozygosity and presence of unusually long IBD tracts are not a function of small effective population size. The inbreeding that we see here is different from the strong drift that would be associated with small effective population sizes over many generations, and which would manifest as generally low levels of nucleotide diversity across the genome. Instead, the observed pattern indicates that some proportion of individuals in an otherwise large population tend to mate with closely related individuals. We hypothesize that GOUNDRY exists as a series of micro-populations, perhaps related to habitat fragmentation, where the likelihood of mating with a related individual is higher than that of larger populations such as the M and S molecular forms. Such dynamics have not been previously observed in mosquito populations, which are typically thought to be large and outbred.

Conclusions

We present the first analysis of complete genome sequences from multiple populations from the *Anopheles gambiae* species complex, including the newly discovered cryptic GOUNDRY subgroup, and demonstrate the power of multi-population analysis to comprehensively dissect the mosaic pattern of divergence, adaptation, and introgression across the genome at base-pair resolution. Mosquito-based efforts to control malaria, especially those based on transgenic manipulation, rely on accurate information about current gene flow among populations and an understanding of how

these dynamics vary across the genome. Our data point to a model where the speciation process is reticulate with offshoot founder populations and ancestral species remaining connected by periodic hybridization of differential efficiency across the genome that sometimes facilitates adaptation (Figure 4). The relationship between *A. gambiae* and *A. arabiensis* provides an extreme example of variation in the permeability of species boundaries, with evidence of autosomal introgression over the evolutionary history of these two species, including the sharing of an insecticide resistance allele in the last 60 years. In stark contrast to this pattern, the X chromosomes of these species remain highly diverged relative to the autosomes, implying that introgression has been ineffective on the X and that the X chromosome is the primary driver of speciation.

Ultimately, our results suggest that populations in the *A. gambiae* species complex comprise a diffuse and interconnected gene pool that confers access to beneficial genetic variants from a broad geographic and environmental range. Such genetic affinity has important implications for malaria control. On one hand, transgenes may spread more easily among populations and species of malaria vectors, which could reduce the effort needed to reach and manipulate all populations involved in disease transmission. On the other hand, our analysis suggests that ecological speciation is common and eradicating a population without removing the ecological niche may simply open the niche for a new population to invade with the potential to become a new and unknown vector. In both cases, our results underscore the complexities involved in vector control on a continental scale.

Methods and Supplementary Text

Mosquito samples

Mosquito sample collection and species/subgroup identification was previously described for the M form (recently re-named *A. coluzzii* [8]), GOUNDRY, and *A. arabiensis* samples [17]. Briefly, larvae and adults were collected from three villages in Burkina Faso in 2007 and 2008 (Table S4). Larvae were reared to adults in an insectary, and both field caught adults and reared adults were harvested and stored for DNA collection. One *A. gambiae* S form individual was also included in this study. This sample was collected indoors as an adult in the village of Korabo in the Kissidougou prefecture in Guinea in October 2012. Individuals were typed for species, molecular form and *2La* karyotype using a series of standard molecular diagnostics [62–64]. All M

form and *A. arabiensis* samples are $2La^{a/a}$ homokaryotypes and the S form sample typed as a heterokaryotype ($2La^{a/+}$). Eleven of the twelve GOUNDRY samples typed as $2La^{+/+}$ homokaryotype, but one sample (GOUND_0446) typed as a $2La^{a/a}$ homokaryotype.

DNA extractions and genome sequencing

DNA was extracted from female carcasses using standard protocols. Genomic DNA samples were diluted in purified water and sent to BGI (Shenzhen, China) for paired-end sequencing on the Illumina HiSeq2000 platform. We sequenced 12 females from the Goundry population, 10 females for the M form population, and 9 female *Anopheles arabiensis*. Sequence reads were filtered if they 1) contained more than 5% Ns or polyA structure, 2) contained 20% or more low quality (Q<20) bases, 3) contained adapter sequence, or 4) contained overlap between pairs. After quality filtering, each sample was represented by an average of 47.91 million reads with an expected insert size of 500 base pairs, except two samples that were sequence to deeper read depth (Supplemental Table S5).

In a separate sequencing effort, we also sequenced one individual *A. gambiae* S form female. The sequencing library was prepared using the Nextera kit (Illumina Inc., San Diego, CA) according to manufacturer's specifications and sequenced on the Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA) at the University of Minnesota Genomics Center core facility for a total of 125.88 million reads.

All Illumina reads newly generated for this study have been submitted to the Short Read Archive at NCBI under accession IDs ranging from XXXXXX-XXXXX.

We obtained publicly available *Anopheles merus* and *Anopheles arabiensis* short read Illumina data from NCBI SRA. We downloaded *A. merus* accessions ERR022713-8 that were generated and deposited by The Sanger Center. Individual accessions are paired-end 76 bp read fastq files representing whole genome sequence from single individual *A. merus* individuals collected Kenya sequenced on the Illumina GAII platform. In addition, we downloaded *A. arabiensis* accession SRX377561, representing an individual collected in Minepa, Tanzania [65]. These data are paired-end 100 bp reads generated on the Illumina HiSeq2000 platform. This sample is the individual sequenced to 'high coverage' from the Marsden et al. Tanzania population sample.

Illumina short read alignment and population specific references

At the time of this analysis, the only genome reference sequence with scaffolds mapped to chromosomes for any *Anopheles* mosquito was the *Anopheles gambiae* PEST AgamP3 assembly [[66]; vectorbase.org], so we used this reference for mapping reads from all groups and species in this study. Since an unknown subset of the short reads generated for this experiment were expected to be substantially diverged from the PEST reference, we conducted short read alignment iteratively in two steps. First, we conducted paired-end mapping of all available reads to the *Anopheles gambiae* PEST

reference genome [66] using the BWA *mem* algorithm [[67]; bio-bwa.sourceforge.net] with default settings except applying the –M flag, which marks shorter hits as secondary. We then generated new 'Population-Specific' references (hereafter SPEC) for *Anopheles gambie* M form, *Anopheles gambiae* GOUNDRY, *Anopheles arabiensis*, and *Anopheles merus* separately. To do so, we combined sequence reads for each population and generated a pseudo-consensus sequence for each population. We used the software package ANGSD (version 0.534; [68]; popgen.dk/wiki) to generate a read pileup, count reads and bases at each site, and identify the major allele at all sites. Then for every site that was covered by 1) at least 4 sequence bases that pass filtering, 2) fewer than a population-specific threshold number reads (mean read depth plus two standard deviations), and 3) a major allele that was segregating at a frequency of at least 0.5 across all reads, we assigned the new reference base to the major allele and to 'N' otherwise.

After alignment to the PEST reference, the mean total read depth for the autosomal arms was 136.14 for *A. arabiensis*, 136.63 for *A. gambiae* M form, 157.42 for *A. gambiae* GOUNDRY, and 52.94 for *A. merus*. Mean total read depth on the X chromosome was 124.27 for *A. arabiensis*, 145.63 for *A. gambiae* M form, and 168.75 for *A. gambiae* GOUNDRY, and 44.37 for *A. merus*. The proportion of unknown bases ('N's) increased in the new references (0.12, 0.11, 0.05, 0.14 for the autosomes of *A. arabiensis*, M form, GOUNDRY, and *A. merus*, respectively, and 0.22, 0.05, 0.06, 0.25 for the X) relative to PEST (0.02 for autosomes, 0.04 for X). The differences between populations in read depth and proportion of sites missing data can be attributed in part to differences in the number of individuals sequenced (n = 9,10,12, 6 for *A. arabiensis*, M form, GOUNDRY, *A. merus*, respectively). However, the differences between autosomes and the X likely reflect difficulties in mapping divergent reads to the PEST reference. The proportion of sites missing data were higher on the X relative to the autosomes for alignments of data from *A. arabiensis* and *A. merus* to both the PEST reference and the population specific references (Table S6). sTwo other studies found a similar discrepancy when mapping these species to the *A. gambiae* PEST reference [51,65], indicating that it is not unique to our pipeline and likely reflects the relatively higher divergence on the X that proves to be a higher barrier to read mapping.

Following the generation of new reference sequences for *A. gambiae* M form, GOUNDRY, *A. arabiensis,* and *A. merus*, short read datasets for each population were then aligned back to the new population specific (hereafter SPEC) reference using the BWA *mem* algorithm with default settings. The S form individual was aligned to the M form reference sequence. Local realignment around indels was then performed for each population separately using GATK [69]. Duplicates were removed using the SAMtools [70] *rmdup* function.

Supplementary Table S6 summarizes genome coverage, read and mapping statistics for alignments to the new specific references and the PEST reference. Although the read mapping rate was consistently lower for the SPEC reference relative to rates

when mapping to PEST, the proportion of mapped reads with a quality score of 20 increased for all groups except the M form population where it decreased slightly. The reduction in mapping rate when using the SPEC reference reflects the greater number of ambiguous bases ('N') in the SPEC reference relative to PEST and thus a smaller mapping target. Importantly, the proportion of mapped reads with quality score of 20 on the X chromosome increased from 0.6961 (PEST) to 0.7775 (SPEC) when mapping the *A. merus* data to the PEST and SPEC references. A similar increase was also observed for *A. arabiensis*, with the proportion of Q20 reads increasing from 0.7604 (PEST) to 0.8241 (SPEC), indicating that the population specific reference is especially helpful for mapping X-chromosome reads. As discussed above, mapping non-*A. 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 [71]. 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 any single individual.
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 population 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 [72] to be conservative in our analyses.

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

Site-frequency spectrum optimization

Population genetic inference from next-generation sequencing data was performed using a statistical framework implemented in the software package Analysis of Next-Generation Sequencing Data (ANGSD, [68], popgen.dk/wiki). Much of the BAM manipulation and read-filtering functionality implemented in SAMtools is also implemented in ANGSD and several functions were utilized in all of the following analyses with ANGSD. Minimum map quality and base quality thresholds of 10 and 20 were used. Probabilistic realignment for the computation of base alignment quality (BAQ) was enabled throughout with a downgrading coefficient of 50. Reads that were not aligned as proper pairs or had bitwise flags above 255 were excluded.

The first step in the pipeline was to infer the site-frequency spectrum (SFS) directly from genotype likelihoods estimated from the read data. The global SFS was estimated in two steps. The first step was to obtain a maximum likelihood estimate of per-site allele frequencies by calculating multi-sample genotype likelihoods using the SAMtools model ([70]; -GL 1 in ANGSD) and estimating the minor allele frequency using the –realSFS 1 function in ANGSD (Nielsen et al. 2012). Files containing the per-site estimates of allele frequencies were then used as input for optimization of the global SFS across all sites using a BFGS optimization algorithm implemented in the optimSFS program within ANGSD. Unfolded spectra were obtained by including ancestral polarization assigned by a synthetic ancestral sequence described below. Sites without ancestral assignment were excluded from SFS estimation. The optimized global SFS was

631 estimated for each chromosomal arm separately and the results from 2R, 3L, and 3R were
632 averaged to obtain an autosomal SFS. Autosomal and X chromosome site frequency
633 spectra were obtained for M form, GOUNDRY, and *A. arabiensis* separately. These
634 spectra were used as priors for downstream analyses within ANGSD and are shown in
635 Figure S10.
636
637 ***Ancestral sequence***
638 We generated a synthetic ancestral sequence by assigning ancestral and derived
639 alleles based on their presence in our M form, GOUNDRY, *A. arabiensis*, and *A. merus*
640 population samples. We first identified all alleles segregating at every site in the genome
641 of each group by calculating genotype likelihoods (-GL 1) and estimating the minor allele
642 frequency (–doMaf 2 in ANGSD). We used all variable sites except singletons. If all
643 groups shared the same allele at a site or if all groups were missing data at a site, the
644 conserved allele or an N was included in the new ancestral sequence, respectively. If
645 some groups lacked data at a site and only one allele was observed in the remaining
646 groups with data, this allele was included in the new sequence. Lastly, if an allele was
647 found to be segregating in at least three of the four groups and the alternate allele was
648 found in only one or two groups, the major allele was chosen as the ancestral allele. This
649 approach is conservative but also flexible in that it requires an allele to be found in either
650 1) three different species or 2) both subgroups of *A. gambiae* and one outgroup species.
651 It is possible that gene flow among these groups could introduce bias here, but it is
652 difficult to distinguish gene flow from lineage sorting at this stage in the analysis. We
653 compared the SFS spectra inferred using this ancestral sequence to those obtained using
654 *A. merus* as the ancestral sequence and found a larger enrichment of high frequency
655 derived alleles in the *A. merus*-derived spectra as would be expected if ancestral mis-
656 specification were common (not shown).
657
658 ***Genotype calling***
659 Genotypes were called in two steps at all sites that passed filtering. For both the
660 population samples (M form, GOUNDRY, *A. arabiensis, A. merus*) as well as the *A.*
661 *gambiae* S form individual, genotype likelihoods were calculated using the SAMtools
662 model as described above. Then genotype posterior probabilities were calculated for each
663 individual at each site. For the population samples, posterior probabilities were
664 calculated using maximum likelihood estimates of allele frequencies as a prior [73]. For
665 the single individual samples, a uniform prior was assumed for calculation of the
666 posterior probabilities. For all samples, diploid genotypes were called only at sites with
667 posterior probabilities of 0.9 or greater. Genotypes for GOUNDRY sample were
668 obtained using a slightly different approach since this population is partially inbred (see
669 Section S8 below).
670

671 ***Inbreeding analysis***

672 *Estimating inbreeding coefficients*

673 Initial estimates of the global site frequency spectrum (SFS) in GOUNDRY
674 produced distributions of allele frequencies that deviated substantially from standard
675 equilibrium expectations as well as from those observed in the *A. gambiae* M form and *A.*
676 *arabiensis* populations. Most notably, the proportion of doubletons was nearly equal to
677 that of singletons in *A. gambiae* GOUNDRY. This observation is consistent with
678 widespread inbreeding in the GOUNDRY population. We tested this hypothesis in two
679 ways, with the goals of both characterizing the pattern of inbreeding in this population as
680 well as obtaining inbreeding coefficients for each individual that could then be used as
681 priors for an inbreeding-aware genotype-calling algorithm. We used the method of
682 Vieira et al. [74], which estimates inbreeding coefficients in a probabilistic framework
683 taking uncertainty of genotype calling into account. This approach is implemented in a
684 program called ngsF (github.com/fgvieira/ngsF). ngsF estimates inbreeding coefficients
685 for all individuals in the sample jointly with the allele frequencies in each site using an
686 Expectation-Maximization (EM) algorithm [74]. We estimated minor allele frequencies
687 at each site (-doMaf 1) and defined sites as variable if their minor allele frequency was
688 estimated to be significantly different from zero using a minimum log likelihood ratio
689 statistic of 24, which corresponds approximately to a *P* value of 10^{-6}. Genotype
690 likelihoods were calculated at variable sites and used as input into ngsF using default
691 settings. For comparison, we estimated inbreeding coefficients for *A. gambiae* M form,
692 GOUNDRY, and *A. arabiensis* using data from each chromosomal arm separately
693 (Figure S9).

694 We observed strong evidence of inbreeding in the GOUNDRY sample, but not in
695 the other populations. Optimized individual inbreeding coefficients for chromosomal
696 arms ranged from 0 to nearly 0.7 in the *A. gambiae* GOUNDRY population sample,
697 where an inbreeding coefficient of 1 would indicate complete lack of heterozygosity.
698 Other than relatively high values on 3R for *A. arabiensis*, inbreeding coefficients for the
699 M form population and *A. arabiensis* were uniformly low. The underlying source of the
700 high values on 3R in *A. arabiensis* is not known. Overall, these results provide strong
701 evidence of an increased rate of inbreeding in GOUNDRY that is not observed in other
702 *Anopheles* populations.

703

704 *Recalibrating the site-frequency spectrum and genotype calls*

705 We used the inbreeding coefficients obtained above for the GOUNDRY sample
706 as priors to obtain a second set of inbreeding-aware genotype calls and an updated global
707 SFS. We used ANGSD to make genotype calls as described in Section S7. However, in
708 this case, we used the –indF flag within ANGSD, which takes individual inbreeding
709 coefficients as priors instead of the global SFS [74]. Similarly, we used the inferred
710 inbreeding coefficients to obtain an inbreeding-aware global SFS. We estimated the

global SFS from genotype probabilities using –realSFS 2 in ANGSD, which is identical to –realSFS 1 [68] except that it uses inbreeding coefficients as priors for calculations of posterior probabilities [74].

IBD tracts

We examined the effects of inbreeding within diploid individuals by calculating the proportion of called heterozygous genotypes within sliding10kb windows along the genome. We observed many contiguous windows with levels of individual heterozygosity near 0, indicating that these chromosomal regions are Identical-By-Descent. For visual demonstration of this pattern, we present chromosomal patterns of homozygosity in 6 representative individuals for chromosomal arms 3L and X (Figure S7). Clusters of contiguous 10kb windows with nearly complete homozygosity were observed in all 12 GOUNDRY individuals, but not in any M form or *A. arabiensis* individuals. To identify approximate boundaries of these IBD tracts, we generated LOESS-smoothed heterozygosity curves using the loess.smooth function in R [75] with a span of 0.01 and a degree of 2. We considered a point on the curve to be within an IBD tract heterozygosity fell below 0.002 on the LOESS curve. We then identified points on the curve that switch from less than 0.002 heterozygosity to 'normal' levels of heterozygosity above 0.002 or *vice versa* and calculated the distance between pairs of switches to obtain IBD tract lengths. The 'normal', outbred chromosomal regions resemble very closely the levels of nucleotide diversity found in our M form sample (Figure S11).

Analysis of bioinformatic artifacts

An alternative explanation for deficits of heterozygosity relative to Hardy-Weinberg expectations could be that heterozygous sites are not being accurately recovered due to either read mapping biases or insufficient read depth in the GOUNDRY sample. Since we did not find evidence of inbreeding in the M form population, we compared read mapping statistics in M form with those of GOUNDRY among a random set of 10,000 sites on 3L to determine whether such biases can explain the inbreeding signals in GOUNDRY. To determine whether read mapping is biased in GOUNDRY such that reads carrying non-reference bases are less frequently mapped, we compared the proportion of reference bases mapped at heterozygous sites in both GOUNDRY and M form. The proportion of reference bases at heterozygous sites should be distributed with a mean of 0.5 if mapping is unbiased. We find that the mean proportion of reference bases at heterozygous sites is 0.4893 ($\sigma = 0.1646$) in M form and 0.4757 ($\sigma = 0.1581$) in GOUNDRY indicating very similar distributions in these populations (Fig S12). Although both populations show a tiny deviation from 0.5, this deviation cannot explain large regions of homozygosity in GOUNDRY. To assess whether insufficient read depths could explain the lack of heterozygosity, we compared read depth at homozygous

sites and heterozygous sites in M form and GOUNDRY. If true heterozygous calls are erroneously being called homozygous due to insufficient read depth, we would expect to find a larger difference between the distribution of read depth at these two classes of sites in GOUNDRY relative to M form. We find that the mean read depth is 10.4773 (σ = 4.7555) at homozygous reference sites, 11.0082 (σ = 4.1849) at homozygous alternative sites, and 10.6660 (σ = 4.7755) in M form, indicating that the distribution of read depth is very similar between all three classes (FigS13). We find a similar pattern in GOUNDRY with the mean read depth is 12.3569 (σ = 5.3917) at homozygous reference sites, 12.2156 (σ = 5.1235) at homozygous alternative sites, and 12.6871 (σ = 5.5163) at heterozygous sites. In both the M form and GOUNDRY populations, the distributions of read depths at heterozygous sites and homozygous sites are very similar (Fig S13). Therefore, bioinformatics artifacts cannot explain the excess homozygosity and IBD tracts observed in GOUNDRY.

It is important to consider the potential impact of inbreeding in GOUNDRY on other features of the data such as estimates of divergence, nucleotide diversity, and signatures of selective sweeps. Our estimates of genetic divergence are based on pairwise differences between alleles from different populations (D_{xy}), which correspond to average coalescent times for alleles from two populations. Since inbreeding is an intrapopulation dynamic that affects the relationship between alleles within a population but not between populations, our estimates of genetic divergence are robust to inbreeding. Similarly, the signatures of selective sweeps that we observe in GOUNDRY cannot be explained by and are not affected by inbreeding. Inbreeding results in IBD tracts as a result of matings between relatively closely related individuals who carry chromosomal segments with very recent coalescence. Such matings manifest in high levels of homozygosity in some genomic regions. However, even if a historically outcrossing population goes through a single generation of full-sibling mating, this would not result in complete homozygosity resembling a sweep. Instead, the offspring would all exhibit excess homozygosity but for alternative alleles. For a region to be 'swept' of all diversity, all or most of the parental generation would have to be homozygous for the same allele in a chromosomal region. This does, however, imply that estimates of nucleotide diversity will be affected by inbreeding, especially estimates based on pairwise differences among alleles within a population (π). Since we use π as our estimator for nucleotide diversity, we avoid using estimates of nucleotide diversity in GOUNDRY in other tests. Instead, we use estimates from the M form population as a proxy for a number of analyses including the comparison between nucleotide diversity and divergence (see Section S11) as well as McDonald-Kreitman tests (see Section S18).

Recombination categories and Linkage disequilibrium

Defining genomic regions based on recombination rates

To test hypotheses related to the role of recombination in determining the genomic architecture of reproductive isolation in this system, we divided the genome into regions based on expected levels of recombination in hypothetical hybrids. A fine-scale genetic map is not yet available for *Anopheles* mosquitoes, but it has been shown in *Drosophila* that recombination rates approach zero within several megabases on each side of the centromere and also near the telomeres [76,77]. Although patterns of linkage disequilibrium (LD) are also affected by processes other than local rates of recombination, estimated recombination rate should give a rough approximation of expected LD across the genome. In fact, patterns of LD have been used to define genetic maps in some vertebrates and correspond approximately to genetic maps based on experimental crosses [78,79]. We measured background LD (see below for details) in our M form and *A. arabiensis* samples, taking average r^2 values within 10 kb physical windows across the genome. We found that LD was relatively constant across the genome except for large increases near the autosomal centromeres and smaller increases near the telomeres (Figure S14). Based on this pattern and the assumption that recombination rates in *Anopheles* correspond approximately to the *Drosophila* genetic map, we defined several broad recombinational categories for analysis. We first defined the 'Pericentromeric-Telomeric' regions of the autosomes to be all windows within 10 MB on either side of the centromere or within 1 MB from the telomere. It should be noted that we assumed that the starting and ending coordinates of the PEST reference chromosomal sequences were reliable indicators for distance from centromeres and telomeres. Unless a chromosomal inversion was present, all remaining regions on the autosomes were assigned to the 'Freely Recombining' category. For the comparison between *A. gambiae* and *A. merus,* we assigned all windows inside of the 2Rop chromosomal inversion complex to the 'Autosomal-Inversion' category. We used the outer coordinates for 2Ro and 2Rp breakpoint regions estimated by Kamali et al. [80].

The X chromosome was categorized for each comparison, according to species-specific conditions. We did not define a general 'Pericentromeric-Telomeric' category for two reasons: 1) We did not observe an increase in LD in the euchromatic regions near centromeres and telomeres (Supp Fig S14) similar to increases observed on the autosomes. 2) In *Drosophila* [76,77], the pericentromeric reduction in recombination affects a relatively small region on the X relative to the autosomes, and we have excluded a large heterochromatic region around the centromere that likely encompasses the effected region in *Anopheles.* For the comparison between *A. gambiae* and *A. merus*, and the comparison between the M and S molecular forms, the entire euchromatic region on the X was considered 'Freely Recombining' since no inversions differentiate these groups. For the comparison between *A. gambiae* and *A. arabiensis*, we assigned the entire euchromatic region of the X as 'X-Inversion', since these species are differentiated across nearly 75% of the entire chromosome and introgression rates have been estimated to be 0 in laboratory crosses [50]. For the comparison between GOUNDRY and the

molecular forms of *A. gambiae*, the entire euchromatic X was categorized as 'Freely-Recombining' except for the region spanning 8.47 MB to 10.1 MB, which was categorized as 'X-Inversion'. As described below, we were not able to identify inversion breakpoints for the GOUNDRY inversion, but these coordinates correspond to the outer boundaries of the region with reduced nucleotide diversity (Figure S2).

Measuring linkage disequilibrium

We measured LD using Haploview [81] applied to genotype and SNP calls made using ANGSD as described above for M form, GOUNDRY, and *A. arabiensis*. LD was calculated independently on each chromosomal arm and comparisons were made only between SNPs separated by 10 kb or less. For all analyses, we used the r^2 statistic generated by Haploview. For computational tractability, we reduced the number of SNP-by-SNP comparisons by randomly sampling comparisons down to either 1%, 10% or 20% of the total number of comparisons depending on the total number of comparisons for each population. To obtain LD decay curves, we binned r^2 values for each population based on physical distance in the PEST reference, and then averaged within each bin and plotted as a function of physical distance (Figures S15). We combined measures for chromosomal arms 2R, 3R, and 3L to obtain an autosomal curve while avoiding the effects of the 2L*a* inversion. We estimated a curve for the X chromosome separately. In all three populations, LD decays to background levels within several hundred basepairs on both the autosomes and on the X, consistent with previous analyses in these species [65,82,83]. Interestingly, background levels of LD differ substantially among populations. Background LD is slightly higher in *A. arabiensis* than in M form, which is consistent with a small effective population size and perhaps some greater degree of population substructure in *A. arabiensis* than in M form. GOUNDRY exhibits very high levels of background LD, but this is due to the high levels of inbreeding-related homozygosity. In all cases, LD decay curves suggest bootstrapping can be conducted on 200 kb regions with confidence.

We generated background LD chromosomal plots (Figure S14), by dividing each chromosome into 10kb windows, identifying all SNPs inside each window, and taking the average r^2 value for all comparisons after thinning (see above) between SNPs inside the window and SNPs greater than 1 kb but less than 10 kb away on the PEST reference sequence. Only windows with at least 100 comparisons were included. Background LD was measured for M form and *A. arabiensis* only since inbreeding has skewed measures of LD in GOUNDRY. LOESS-smoothed curves were generated using a span of 1% and a degree of 2 in the loess.smooth function in R [75].

Nucleotide diversity

To enrich our data set for neutrally evolving sequence, we excluded sites within 200 bp of a coding sequence annotated in the AgamP3.8 gene set for the *A. gambiae*

PEST reference available on VectorBase.org from our estimates of nucleotide diversity. If only a single individual was available for a group (*A. gambiae* S form), nucleotide diversity was estimated simply as the fraction of sites with genotypes that were called as heterozygous. If a population sample was available (GOUNDRY, M form and *A. arabiensis, A. merus*), we estimated nucleotide diversity directly from the read data using a maximum likelihood approach based on posterior probabilities of per-site allele frequencies [84]. The method is implemented under the –doThetas function within ANGSD and takes a global SFS as a prior to calculate posterior probabilities of allele frequencies. We chose to use Tajima's π statistic [85] as our estimate for nucleotide diversity. For all downstream analyses, nucleotide diversity was calculated as an average within 10 kb non-overlapping windows. The comparison of nucleotide diversity between genomic regions (see below) included only 10 kb windows containing at least 500 sites from 2R, 3L, 3R, and X chromosomal arms. The distribution of nucleotide diversity was compared between chromosomal regions (Figure S16) using a Mann-Whitney test implemented in the wilcox.test function in R [75].

Genetic divergence

D_{xy} calculations

We measured absolute genetic divergence as the average number of pairwise differences between alleles from different populations, or D_{xy} [86]. We calculated D_{xy} as

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

where h is the number of sites where one or both individuals carry heterozygous genotypes, H is the number of sites where the two individuals are homozygous for different alleles, and L is the number of sites where both individuals have called genotypes. As with our estimates of within-population diversity, we excluded all sites within 200 bp of the outer coordinates of annotated protein coding sequences in the AgamP3.8 gene set (vectorbase.org). For populations where we have sequences from multiple individuals (M form, GOUNDRY, *A. arabiensis, A. merus*), we estimated D_{xy} calculated from the number of pairwise differences between each individuals from every other population and then averaged across comparisons for each population to obtain an average D_{xy} value for each pair of populations. For comparisons made to the S form, D_{xy} was calculated between the individual S form sequence and the individual sequence with the highest average read depth from each of the other populations.

For chromosomal plots (Figure S4) and boxplot analysis (Figures 2 and S16), each chromosomal arm was divided into 10 kb windows and D_{xy} was calculated across each window. Windows with fewer than 200 sites with data for comparison were excluded. Chromosomal curves were LOESS-smoothed using a span of 1% and degree of 2 with the loess.smooth function in R [75]. For the boxplot analysis, chromosome 2L was excluded to avoid the large effects of the *2La* inversion. Comparisons were made

between categories using a Mann-Whitney test (hereafter M-W) implemented in the wilcox.test function in R [75].

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. ΔD_{xy} was positive for every window in the comparisons between M and GOUNDRY. We tested this pattern statistically using a permutation analysis. 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. Both an uncorrected *P* value of 0.05 as well as a Bonferroni-correct *P* value of 0.05 are indicated log-transformed in Figure S17. 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.

Lowly recombining regions

Lowly recombining pericentromeric and telomeric regions and autosomal chromosomal inversions are also especially diverged in comparisons between *A. gambiae, A. arabiensis*, and *A. merus*. We generally observe a reduction in nucleotide diversity

within groups in pericentromeric and telomeric regions relative to the freely recombining regions of the autosomes (M-W $P < 2.2 \times 10^{-16}$ for all species; Figure S16). Similar patterns are observed in many other species (reviewed in [87]), and they are generally attributed to the effects on effective population size of selection on positively or negatively selected mutations [88,89]. Genetic divergence between M form and *A. merus* is increased in these regions relative to freely recombining autosomal regions (M-W $P < 2.2 \times 10^{-16}$, Figure 2). Moreover, M form ($2R+^{op}$) and *A. merus* (2R*op*) are fixed for alternative forms of the overlapping 2R*o* and 2R*p* inversions and genetic divergence between these species is even higher inside this region compared to the freely recombining genome (M-W $P < 2.2 \times 10^{-16}$; Figure 2). Within-population nucleotide diversity is slightly higher within the inverted region, but only without correcting for multiple testing (M-W uncorrected $P = 0.0203$; Figure S16). In the comparison between M form and *A. arabiensis,* we generally find no evidence for elevated divergence in telomeric and pericentromeric regions. However, inspection of the genomic distribution of divergence and nucleotide diversity (Figures S4 and S2, respectively) reveals that the difference between nucleotide divergence and inter-species diversity is especially high in the pericentromeric region of chromosome 3. This region also harbors evidence for multiple recent selective sweeps, including one that involves the gene encoding *Frizzled-2*, a member of the *Wnt* receptor signaling pathway involved in development [90]. The collective data indicated that this region may have entered the very early stages of reproductive isolation. Overall, the comparisons with *A. arabiensis* and *A. merus* both point to a model in which the X chromosome plays a disproportionately large role in the establishment and/or maintenance of reproductive isolation, with isolation factors accumulating slowly and secondarily in the lowly recombining regions of the autosome.

Neighbor-joining trees

We represented species and population relationships using neighbor-joining distance trees based on pairwise values of D_{xy} (Figure S1). For the autosomal tree,we calculated D_{xy} across chromosomes 2R, 3R, and 3L to avoid the effects of the 2L*a* inversion. D_{xy} was calculated across the euchromatic region on the X for the X tree. Neighbor-joining trees were generated and drawn using the *ape* [91] package in R. We used a bootstrap analysis to estimate confidence in the autosomal tree and the X chromosome tree. We divided each dataset into 200 kb physical windows and sampled with replacement 1,000 new datasets. Distance neighbor-joining trees were generated for each bootstrap replicate and compared to the tree made from the true data. The same tree topology was recovered in all 1,000 bootstrap replicates, resulting in block bootstrap values of 1.0 for all nodes. It has been suggested that *A. merus* is sister to *A. gambiae* based on their sharing of the *Xag* chromosomal inversion [40,92]. However, we recover a tree that suggests *A. gambiae* and *A. arabiensis* are more closely related to each other than either is with *A. merus*. This topology is consistent with a parsimony-based

chromosomal phylogeny of the *Anopheles gambiae* species complex estimated using chromosomal inversions [80] as well as a recent D_{xy} based assessment of relationships between *A. gambiae*, *A. arabiensis*, and *A. merus* [51].

Fixed differences

We identified fixed differences between M form and GOUNDRY using the following criteria. A substitution was inferred if

1) At least (n-2) individuals were represented at a site.
2) The site was considered monomorphic based on allele frequencies inferred using ANGSD being either less than 1/2n or greater than (1-(1/2n)).
3) GOUNDRY and M form harbor different major alleles.

Demographic inference

2D-spectra and model fitting

We fit population historical models to the two-dimensional site frequency spectrum for GOUNDRY and M form populations using the program *dadi* [93]. We obtained the unfolded 2D spectrum for chromosomal arms 2R, 3R and 3L by first estimating allele frequencies at each site for M form and GOUNDRY separately, then calculating posterior probabilities of allele frequencies at each site using the global SFS priors, and finally calculating the 2D spectrum. Posterior probabilities were calculated using the program sfstools distributed with ANGSD. The program 2Dsfs included in the software package ngsTools (https://github.com/mfumagalli/ngsTools; [71]) was used to calculate the 2D spectra with -maxlike set to 1. Per-site allele frequencies were estimated for each population as described above in Section S5, except GOUNDRY was handled differently to accommodate the inbreeding signal. Since GOUNDRY is partially inbred, we randomly sampled one allele at each position in the genome to be included for analysis thereby eliminating the effects of inbreeding within a diploid genome on estimates of allele frequency. This resulted in a sample size of 12 chromosomes for GOUNDRY. Initial testing suggested that optimization in *dadi* performed better when sample sizes were equal, so we projected the M form sample from 20 chromosomes down to 12 to make the spectra symmetric using the project function with *dadi*. As for before, all heterochromatic regions and low-complexity regions were excluded. We also excluded all sites within 200 bp of exons annotates in the *A. gambiae* PEST AgamP3 gene set. We calculated the 2D spectrum for each of the chromosomal arms separately and summed them to represent the autosome (Figure S19). Chromosome 2L was excludee to avoid the effects of the large segregating 2L*a* inversion. The final spectrum contained 61,254,546 sites, 3,189,751 of which were variable in our samples.

We defined the demographic function to include three epochs (Figure S18). As with all models in *dadi*, the model begins with an ancestral population and proceeds forward in time. In the first epoch, the ancestral population is split into two populations.

Each of the three epochs is defined by five parameters: 1) the duration of the epoch, 2) the effective size of M form, 3) the effective size of GOUNDRY, 4) the rate of migration from GOUNDRY into M form, and 5) the rate of migration from M form into GOUNDRY. We defined the demographic function using the following syntax:

```
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 reached best-fit parameter values, we narrowed the parameter search boundaries and conducted a second set of 1,016 optimizations, perturbing the starting values by a factor of 0.5 and allowing 750 iterations. Models that include strong bottlenecks are extremely slow to optimize and not likely to be biologically plausible for systems like mosquitoes with typically large census sizes, so we set lower boundaries of 0.01 for all populations sizes for computational efficiency. After ensuring that the optimized values were not hitting boundaries and parameters were converging over replicate optimizations, we chose the optimization with the highest likelihood as the point estimate. To obtain an estimate of the ancestral effective population size, we estimated the population scaled mutation rate ($4N_A\boldsymbol{\mu}$) that fits the data best using the optimal_sfs_scaling function in *dadi* and solved for N_A, assuming a mutation rate of $\boldsymbol{\mu} = 7.0 \times 10^{-9}$. The spontaneous mutation rate for mosquitoes has not been estimated to date, so we used a mutation rate estimated from *Drosophila melanogaster* as our closest approximation for the mutation rate in *Anopheles*. Multiple estimates of the spontaneous mutation rate for *D. melanogaster* are available from studies differing in their approach, and the estimates vary by an order of magnitude from 2.8×10^{-9} to 1.1×10^{-8} [94–98]. We chose an intermediate value of 7.0×10^{-9}.

The resulting 2D spectra from the data and the best-fit model are presented in Figure S19. Residuals indicating differences between the model and data are also presented. In general the fit is quite good. Most of the weight in the 2D spectrum from the data is matched in the 2D spectrum from the model and the residual plot reveals that in general the model captures most features of the true spectrum. Several exceptions exist, however. The number of SNPs that are absent in GOUNDRY and fixed in M form is underestimated in the model. Additionally, the number of SNPs that are private in GOUNDRY are overestimated by the model, both when those SNPs are intermediate frequency and when they are singletons.

Parameter confidence intervals

We estimated confidence intervals for each parameter using a nonparametric bootstrap approach. We generated bootstrapped genomes by first concatenating chromosomal arms 2R, 3R, and 3L, and then dividing this 'genome' into physical 200 kb regions. These regions were sampled with replacement to obtain 100 new genomes with lengths equal to that of the true genome. Since repeating the two-step optimization described above would be computationally infeasible, we conducted optimizations for each bootstrap replicate with a different approach. For each replicate, we conducted 100 optimizations with wide boundaries, allowing 750 iterations and perturbing starting values by a factor of 0.5. We used the maximum likelihood values obtained for the true genome as pre-perturbation starting values. Then for each bootstrap replicate, we chose the optimization with the maximum likelihood. Approximate 95% confidence intervals

were calculated for each model parameter as the mean of all replicate values +/- 1.96 standard deviations of all replicate values. All bootstrap optimizations were run on the Stampede compute cluster at University of Texas, Austin through the support of the XSEDE program. Optimized parameter values and confidence intervals are reported in Table S8.

Relative Node Depth expectation modeling

The effects of the large sweep region on the GOUNDRY X chromosome as a barrier to gene flow may extend beyond the boundaries of the inverted region, resulting in very recent accumulation of divergence on the X chromosome that is confounded in our analysis with inherent evolutionary differences between the autosomes and the X chromosome. One approach to estimate differences in divergence is to compare divergence between a focal pair of populations (GOUNDRY and M form) to divergence between one of the focal groups and an outgroup (GOUNDRY and S form) in order to scale by differences in mutation rate and other population genetic parameters among regions. This approach estimates what is known as Relative Node Depth (RND = D_{GM}/D_{GS}), where subscripts G, M, and S indicate GOUNDRY, M form, and S form respectively, and a higher RND indicates greater divergence between the focal groups [99]. We find that RND is 0.7797 on the autosomes and 0.8058 on the X, indicating lower relative genetic divergence between GOUNDRY and M form on the autosomes than on the X. 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 outgroup and using coalescent theory under this model to calculate expected RND. We calculated the ratio of expected values of D_{GM} and D_{GS} as

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

where μ is the mutation rate, N_{GM} and N_{GS} are the effective sizes of the M-GOUNDRY ancestral population and the M-S-GOUNDRY ancestral population, respectively, t_{GM} is the number of generations since the split between GOUNDRY and M form, and t_{GS} is the number of generations since the split between S form and the M-GOUNDRY. The denominator is the standard definition for D_{xy} [86], but we adjusted the equation in the numerator using standard coalescent theory to accommodate the probability that two the M and GOUNDRY lineages may either coalesce in the M-GOUNDRY ancestor or in the M-S-GOUNDRY ancestor that may differ in size, and therefore probability of coalescence, from the M-GOUNDRY ancestor (that may differ in size).

We calculated expected values of RND under a range of parameterizations. We used the M-GOUNDRY split time of 1,112,660 generations obtained from the

demographic inference described above. For all calculations, we varied the M-GOUNDRY ancestral size from 100 to 10^6, including 20,000 grid points. In the first set of calculations, we assumed that the split time between S and M-GOUNDRY was 1.1 times longer than the M-GOUNDRY split time and varied the relative sizes of the ancestral populations such that the M-S-GOUNDRY ancestor such that the M-GOUNDRY ancestor varied from 1% to 100% the size of the M-S-GOUNDRY ancestor (Figure S20). We conducted a second set of calculations that were identical to the first except that we assumed that the difference in split times was 1.5 times longer instead of 1.1 (Figure S20). In the last set of calculations, we set the sizes of the two ancestors to be equal and varied the difference in split times such that the split time between S form and M-GOUNDRY was 1%, 10%, 20%, 40%, 70%, and 100% longer than the split time for M-GOUNDRY (Supp Fig S20).

This analysis assumes a simple split model with no migration and provides expected ratios of divergence under the given population tree. We observe in our data that RND is smaller on the autosomes than it is on the X. Our analytical results indicate that under some parameter combinations (Fig S20), RND decreases with increasing effective M-GOUNDRY effective population size, which could result in a smaller RND value on the autosomes since the autosomes should have an effective size at least as big as the X. However, most parameter combinations suggest that this pattern is unexpected (i.e. most regions of the curves predict that RND should increase with increasing effective population size), and the estimate for the ancestral effective size of M-GOUNDRY we obtained in a separate demographic analysis above suggests that these populations exist in a parameter space where the RND function is consistently increasing with increasing effective sizes. These results support the hypothesis that D_{GM} is downwardly biased relative to D_{GS} on the autosomes as a result of higher rates of gene flow on the autosomes relative to the X. Taken together with the demographic inference, the above results suggest that, after having initially diverged ecologically approximately 100,000 years ago, speciation has commenced among these populations within the last 100 to 200 years, presumably owing to the accumulation and extended effects of locally adapted loci or genetic incompatibility factors within large swept region on the GOUNDRY X chromosome.

Estimating the age of the GOUNDRY X-linked selective sweep

We estimated the number of generations since the fixation of the haplotype inside the putative GOUNDRY large swept region on the X in the following way. We assumed that no new mutations increased to have a frequency > 50% after the selective sweep. Under this assumption, we can estimate the mean time since the most recent common ancestor of the sampled haplotypes, representing the time of the sweep, by assembling a consensus sequence among the haplotypes to represent the common ancestor and counting the number of mutations separating each haplotype from the ancestor. Then we

can calculate the number of mutations divided by the total haplotype size and divide this number by the mutation rate to obtain the number of generations separating haplotype from the consensus. We summed the number of mutations across all diploid sequences, counting 1 for genotypes called heterozygotes and 2 for homozygotes, and divided by 2 times the total number of sites passing filters. This grand total was then divided by the mutation rate to get the number of mutations. This approach can be simplified and stated as the following where the age to the most recent common ancestor is

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

where f_j is the derived allele count at variable site j, s is the number of variable sites, n is the number of diploid individuals, L_i is the number of sites with called genotypes passing filters for individual i, and μ is the mutation rate, again assumed to be 7.0×10^{-9} (See Section S12).

Genotypes for 12 diploid individuals were queried across a region consisting of 1,372,538 sites after initial application of low-complexity filters as described above. A total of 1,052 variable sites were recorded prior to additional read depth filtering. Since this estimate is highly sensitive to the number of mutations included, we estimated the number of generations using a series of filters varying in stringency. First, we removed clusters of variable sites, since these are likely to represent errors, by dividing the inversion region into 50 bp windows and excluding any windows with more than two variable sites (14 out of 33,478 windows). This resulted in a total of 989 variable sites across 1,367,970 sites in total. We also excluded sites with read depth in the top 5% for each individual according to each individual's read depth distribution. Then we counted mutations for each individual using different thresholds such that sites passed filter if that individual was covered by at least 6, 8, 9, 10, 11, or 12 reads. We found that too few sites passed filtering to be informative with a 15-read cutoff. After conservatively filtering the data to minimize the effect of errors, the majority of the remaining variance in our estimates derives from variance in the number of mutations per haplotype. If we assume that the number of mutations per haplotype is Poission distributed, we can calculate the standard deviation for read-filter point estimates by taking the square root of the point estimate, as the variance for a Poisson is equal to its mean.

Age estimates varied by 2.5 fold depending on the minimum read depth filter (Figure S22), decreasing from 1,975 generations to 776 generations with 6 and 12-read minimum cutoffs, respectively. However, excluding the 6 read cutoff, the remaining five age estimates varied by less than 1.4 fold, ranging from 776 to 1,079. Standard

deviations largely overlap among estimates from the 8-12-read filters, suggesting that the estimates are quite similar in this range. If we examine how the depth filters affect mutation counts per site for each individual separately, the 6-read filter results in an increase in mutations for all individuals (Figure S22). The mutation counts per site remain relatively flat for the remaining filters with only a slight systematic downward trend with the 12-read filter. The estimated age drops substantially between the 6 and 8-read filters and continues to decrease slowly from there. This strongly suggests that the proportion of variable sites that are genotype-calling errors is relatively high for this region when low read-depth minimums are used. To explore this pattern further, we calculated the minor allele count at each variable site for each of the read-depth thresholds and see that, when a minimum of 6 reads is used, there is an excess of doubletons compared to higher read depths (Figure S23), consistent with a high proportion of true heterozygous sites being called as homozygous for the minor allele. The ratio of doubletons to singletons decreases substantially when the minimum number of reads is increased to 8, implying that many erroneous homozygous-alternative genotype calls are converted to heterozygous with this threshold (Figure S23). This improvement does not fully explain the drop in the number of observed mutations, however. Our sequencing effort was not evenly distributed among individual GOUNDRY samples since we sequenced one individual (GOUND-0446) to average 20.03x read depth while the remaining individuals were sequenced to an average of 10.82x read depth. Since the accuracy of genotype calling is correlated with read depth, we expect the genotype calls made for this individual will harbor the fewest errors. When considering how the read-depth thresholds affect estimates from individual samples, GOUND-0446 showed patterns that differ from the other samples (Fig S22). Specifically, while the number of sites passing filter changes less for this individual with increasing read-depth minimums, the number of mutations observed in this individual is also less sensitive to minimum read-depth. This suggests that genotype calling is substantially more error prone in the lower-read depth individuals. The variance in the number of observed mutations among individuals reduces substantially when the 12-read minimum is implemented with GOUND-0446 falling in the middle of the distribution, 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 [100], and estimate the age of the haplotype inside the sweep region to be 78 years with a standard deviation of 9.15. This age is extraordinarily recent for a chromosomal inversion and implies that speciation between GOUNDRY and the molecular forms of *A. gambiae* initiated in very recent history.

Introgression **analysis and D statistics**

Background and genomic test for introgression

To test for admixture among species and subspecies of *Anopheles* mosquitoes, we used Patterson's *D* statistic [53,101]. This statistic compares the distribution of alleles on the four taxon tree ((H1,H2),H3),O), where H1 and H2 are sister taxa and O is the outgroup. Under the null hypothesis of no gene flow, the number of derived mutations that are only shared between the genomes of H2 and H3 (ABBA) is expected to equal the number of those that are only shared between H1 and H3 (BABA). *D* is then calculated as the standardized difference between the numbers of ABBAs and the number of BABAs such that the expectation of *D* is zero under the null hypothesis [53]. Significant excess sharing of derived alleles will result in a non-zero *D* and can be interpreted as evidence of admixture.

We tested for admixture using four taxon trees with M form and GOUNDRY as the sister taxa (H1 and H2, respectively), *A. merus* as the outgroup (O), and either S form or *A. arabiensis* as H3. The ABBA-BABA test was originally conceived for an African-European-Neandertal-chimp framework in which admixture was thought to be impossible between Neandertal and Africans [53]. In our case, however, introgression is possible between all ingroups, reducing the power of the test. For example, similar but independent gene flow events from *A. arabiensis* into M form and GOUNDRY would increase both the number of ABBAs as well as BABAs, in effect canceling each other out and obscuring both events, especially when calculating *D* over large genomic regions. To avoid this confounding effect, we tested for evidence of gene flow as a significant excess variance in *D* across genomic blocks. Since we expect introgressed nucleotides to be physically clustered, we can compare the variance in *D* among true genomic blocks with the variance in *D* among genomic blocks where internal segments within the blocks have been permuted between blocks. Under the null hypothesis, the variance in *D* among true genomic blocks is not expected to be larger than the variance among blocks of permuted segments. We calculated *D* in blocks of 500 informative sites (i.e. ABBAs or BABAs) after arbitrarily resolving diploid genomes into haplotypes. For each population, we used the genome from the individual with the highest mean read depth to minimize genotyping errors. We chose a block size of 500 sites as this corresponds to a physical size that is 1000 times the physical distance at which linkage disequilibrium decays to background in this system (200 bp; see Section S9). This block size allowed jackknife analysis within each genomic block using segments of length 2000 bp, which corresponds to 10 times the rate of LD decay.

We permuted segments within blocks 10^4 times, calculated *D* for each block, and calculated variance across the all genomic blocks. We compared the variance in *D* among true genomic blocks to the distribution of estimates of variance from permuted genomes to determine whether the true genomic blocks were over-dispersed. We interpreted significant excess variance among true genomic blocks as evidence for introgression between H3 and either one or both H1 and H2. We present the comparisons

of variance among *D* values across genomic blocks in Table S9. We conducted three different analyses using different populations and find that the variance in *D* among genomic blocks is significantly higher than any permuted genome in our analysis ($P < 0.0001$; Table S9). These results are consistent with significant introgression between the two ingroups (M form and GOUNDRY) and both S form and *A. arabiensis*.

Genomic block confidence intervals and critical values

We used permutation and jackknife analyses to conduct two additional tests. First, we conducted block jackknife analyses within each genomic block of 500 informative sites [53]. We divided each block into 100 segments of 5 informative sites, dropped each segment in turn and recalculated *D*. We calculated 95% confidence intervals for each genomic block using variance estimated from this jackknife procedure. Second, we established genome wide thresholds corrected for multiple testing. We conducted the permutation of segments within blocks procedure as above, but for each permuted genome, we calculated *D* for each block and retained the maximum and minimum values of *D*. To determine whether any individual true genomic blocks showed evidence of significant excess sharing of derived alleles, we established 95% critical thresholds (Table S10) from this permutation procedure and compared the value of *D* among true blocks.

We identified genomic blocks with values of *D* that were more extreme than the genome 95% thresholds. To put these windows in genomic context, we calculated the cumulative length of significant windows and compared this value to the total length of the genome to obtain a proportion. We find that the proportion of the genome that falls within significant windows ranges from 1.1% to 3.6% (Table S10).

To determine whether introgression has been recent between *A. arabiensis* and either M form or GOUNDRY, we compared the proportion of the genome in windows with significant *D* values between sympatric *A. arabiensis* from Burkina Faso and allopatric *A. arabiensis* from Tanzania (Marsden et al. 2013). Since the standard assumption of introgression with only one of the two sister taxa holds for this test, we calculated the standard error of *D* for each comparison using the block jackknife approach and used a Z-test to assess significance [53,101]. We find evidence for higher affinity between sympatric *A. arabiensis* and both M form and GOUNDRY relative to allopatric *A. arabiensis* (Table S10), consistent with recent introgression.

Scan for recent positive selection

Genome scans

We conducted full genome scans for recent complete selective sweeps in *A. gambiae* M form and GOUNDRY as well as *A. arabiensis* using SweepFinder, an 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 [102]. To

assemble input data, we estimated unfolded allele frequencies from genotype likelihoods using the method of Kim (2011; -doMafs 5 in ANGSD) with ancestral state assigned based on the ancestral sequence constructed in Section S6. 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* in Section S12. 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 ([103]; 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 [103]. 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 S24). The LD decay curves from the simulated data also follow a similar decay rate as the real data (Figure S24).

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
```

For autosomal loci, we simulated 792,800 50 kb regions corresponding to 200 complete autosomes and applied SweepFinder to the M form and GOUNDRY simulated haplotypes separately. We used a grid size of 50 for SweepFinder to be consistent with the 1kb scale used for the real data. The background site frequency spectrum was estimated by SweepFinder for each 50kb region and used for that region. To simulate data resembling the X chromosome, we simulated 150,000 50 kb regions under the same demographic model, but applied a ¾ correction to the ancestral effective size inferred from *dadi* and adjusted simulation parameters accordingly.

To determine critical thresholds for a 50 kb region (hereafter referred to as the per-locus threshold), we found the maximum log likelihood ratio (LLR) found in each 50 kb region for M form and GOUNDRY separately (Figure S25). Then, to obtain per-locus *P* values, the maximum LLRs for all M form and GOUNDRY selective sweep windows in the real data were compared to the corresponding distribution of LLRs calculated from simulated regions. To obtain a genome wide critical threshold corrected for multiple testing, we assembled autosomal 50 kb regions and X-linked 50 kb regions into synthetic 218.23 Mb genomes and identified the highest LLR within each simulated genome (Figure S26). After identifying windows with LLR values significant at the genome level, we then identified the highest point in each string of continuous significant windows to find the 'peak of the peak'.

We found that peaks were unexpectedly clustered across the genome, suggesting that some peaks may be fractured evidence of the same selective event. Since SweepFinder has substantially more power when monomorphic sites are included in the analysis [104], we re-ran SweepFinder using both monomorphic and variable sites within clustered regions. We identified clusters of peaks where significant peaks were found within 100 kb of each. To determine whether multiple peaks separated by windows with non-significant LLR values were actually part of a single large peak, we asked whether more than four continuous windows with LLR values less than 20 separated the peaks in the 'all-sites' analysis. SweepFinder is prohibitively slow when monomorphic sites are included in the analysis, so we could not conduct this version across the whole genome. In some cases, this high-resolution analysis revealed a single continuous peak. We present one representative example in Figure S27 where non-significant windows in the variable-site-only analysis separate a number of significant peaks, but the all-site analysis includes only a single low-LLR window separating the peaks. To be conservative, clusters of peaks satisfying these criteria were collapsed into a single peak and the window with the highest LLR was used for annotation. In other cases, however, the all-sites analysis did not provide evidence a single continuous peak, despite an overall peaked shape of the significant windows in the region (Figure S28). The final set of selective sweeps after collapsing clusters is presented in Table S1.

Is there proportionally more selection in GOUNDRY or M form?

We were interested in determining whether GOUNDRY or M form has experienced significantly more selection than the other. Since we were not able to satisfactorily resolve all clusters of peaks and thus our count of independent selective events may be biased, we chose to quantify the proportion of megabases in each genome that harbors at least one selective sweep (i.e. one significant window). We scanned each genome for selective windows and found that, of the 230 possible megabase windows, 36 windows harbored at least one selective sweep in GOUNDRY while 58 harbored sweeps in M form. To determine whether this difference is statistically significant, we randomly permuted population assignments of each selective sweep in our data 10^4 times and conducted the same analysis. We found that this difference was larger than differences found in any of our permuted datasets (Fig S29), indicating that M form has experienced significantly more positive selection in recent evolutionary history.

Identifying peaks in A. arabiensis

We did not estimate a demographic model for *A. arabiensis*, so we used a different approach to identify credible signals of recent selective sweeps. We searched the LLR surface for peaks where the peak included at least two adjacent windows with LLRs in the top 0.1% genome wide. The 0.1% threshold corresponds to a value of 15.88 in this dataset. After finding all such cases, we used the window with the highest LLR in the cluster to represent the peak. This approach identifies 34 distinct peaks across the autosomes and none on the X chromosome.

Annotating peaks

For each selective sweep, we found the protein coding sequence most closely associated with the highest point on the LLR peak. If the maximum LLR window fell within a gene in the Agam3.8 gene set, information from this gene was included in the annotation database. If the maximum LLR window fell outside known genes, information from the gene with the nearest 3' or 5' boundary was used for the peak. To annotate each sweep, we downloaded information from Vectorbase.org for AgamP3.8 gene set from the *A. gambiae* PEST genome, including gene names, membership in ImmunoDB [105], best GO annotations, and best KOG annotations. Information for all selective sweeps can be found in Table S1.

Adaptive introgression

Background and approach

Natural selection is expected to remove most introgressed genetic material due to ecological misfit or Bateson-Dobzhansky-Muller incompatibilities (BDMIs; [37,38,57]), especially from more distant species, but it some variants may be selectively favored in the recipient population. To determine whether any selective sweeps observed in our

data may have involved selection on introgressed alleles, we asked whether any selective sweep regions are located within genomic windows that showed significant excess sharing of derived mutations between populations in the ABBA-BABA test (see Section S15). We find 10 M form sweeps that fell within windows significant in test with the S form and *A. arabiensis* populations, respectively. We find six sweeps in GOUNDRY that fall within windows with significant derived allele sharing with *A. arabiensis*. Lastly, we find one sweep in *A. arabiensis* that falls within windows with significant *D* values with GOUNDRY. The identities of these sweeps putatively involving introgressed alleles are indicated in Table S1. Although it is difficult to formally test whether these sweeps involved introgressed mutations without knowing which mutations are adaptive, the physical proximity between the locations of selective sweeps and regions harboring excess shared derived mutations suggests that allele sharing between these populations may facilitate adaptation.

Rdl locus

The most striking example of these putative adaptive introgression events involves allele sharing at the GABA receptor on 2L known to be involved in insecticide resistance in both *A. gambiae* and *A. arabiensis* ([43]; Figure S6). Our scan for selection points to a selective sweep at the boundary of the coding sequence for this gene in the M form population ($P_{gen} < 0.005$, $P_{loc} < 1.26\text{x}10^{-06}$). However, we also find evidence, albeit below the significance threshold at the genome level, for selection in both GOUNDRY and *A. arabiensis* (Figure S6). There is an independent selection peak in GOUNDRY ($P_{gen} > 0.05$, $P_{loc} < 2.67\text{x}10^{-04}$) upstream of the M form peak. We do not see elevated LLR values indicating recent selection in *A. arabiensis*, but we do see a local elevation in LD and a local reduction of diversity consistent with a historical sweep at this locus (Figure S6).

It is necessary to consider the genomic background when interpreting data at the *Rdl* locus since if falls within the 2L*a* chromosomal inversion. *A. arabiensis* is fixed for the inverted 2La^a arrangement that is also nearly fixed in most West Africa populations of *A. gambiae* (i.e. the M and S forms; [40]). However, both forms of the inversion (2La^a and 2La^+) are segregating in GOUNDRY [17] and we analyzed both GOUNDRY forms here. Consistent with previous studies [39–42,106], we find strong evidence that the inverted form has introgressed between *A. gambiae* and *A. arabiensis.* There is a reduction of divergence between *A. arabiensis* and *A. gambiae* 2La^a chromosomes as well as a large reduction in *D* when the GOUNDRY 2La^+ chromosome is used (Figures S4 and 3). Conversely, there is exceptionally high genetic divergence between GOUNDRY 2La^+ chromosomes and the M form population and *A. arabiensis* chromosomes (Figure S4). As a result, this region harbors genomic patterns of divergence and allele sharing that deviate from the genome at large in a comparison-

specific fashion. Local deviations from this pattern provide evidence for recent introgression events that are independent from the inversion introgression event.

Evidence from ABBA-BABA tests and patterns of pairwise genetic divergence in the *Rdl* region suggest that both GOUNDRY and M form have independently received introgressed material in this region from *A. arabiensis* in a *2La*-inversion karyotype dependent fashion. In the ABBA-BABA test involving the tree ((GOUNDRY 2La^a, M form 2La^a), *A. arabiensis* 2La^a), *A. merus* 2La^a), we find evidence for significant allele sharing between *A. arabiensis* and M form at this locus. Moreover, we see a reduction in genetic divergence between these two populations relative to immediately surrounding regions. Together, these results provide strong evidence for introgression between M form and *A. arabiensis* at the *Rdl* locus. However, we also see evidence for an independent introgression event between *A. arabiensis* and the population of 2La^+ chromosomes in GOUNDRY. In the ABBA-BABA test based on the tree ((GOUNDRY 2La^+, M form 2La^a), *A. arabiensis* 2La^a), *A. merus* 2La^a), the region inside the *2*L*a* inversion is almost entirely negative and nearly -1 indicating strong excess of allele sharing between the 2La^a chromosomes from *A. arabiensis* and M form. However, amidst this pattern, we see a large region of positive *D* statistics at the *Rdl* locus indicating excess derived mutation sharing between *A. arabiensis* and GOUNDRY. Consistent with this result, we also see a large and extended reduction in genetic divergence between GOUNDRY 2La^+ and both M form and *A. arabiensis*, indicating recent introgression between GOUNDRY and one of these populations. Genetic divergence is exceptionally low between GOUDNRY and *A. arabiensis* across the entire putatively introgressed region while divergence is reduced between M form and GOUNDRY only where genetic divergence is also low between M form and *A. arabiensis*. These observations provide strong evidence that an introgression event occurred first from *A. arabiensis* into M form followed by a second independent introgression event from *A. arabiensis* into GOUNDRY. Together with the evidence for positive selection in these populations at this locus, we hypothesize that a mutation conferring insecticide resistance arose and adaptively fixed first in *A. arabiensis* and then was shared successively with the M form and GOUNDRY populations where is underwent positive selection in these populations as well.

Test for 'faster-X' protein evolution

One hypothesis to explain higher divergence and reduced gene flow on the X relative to the autosomes, or the 'large-X' effect, stems from the fact that adaptive mutations are more efficiently exposed to selection when they reside on the X because of hemizygosity in males [56]. It is therefore expected that adaptive protein evolution would occur at a faster rate on the X relative the autosomes if adaptive mutations tend to be recessive ('faster-X' ; [47]). If adaptive fixations function as BDMIs [37,38,57], introgression would be less effective in the genomic regions linked to these substitutions

and could explain the observed 'large-X' effect on divergence. One approach for comparing the rates of adaptive evolution is to calculate the proportion of amino-acid changing substitutions that are adaptive, also called α [58,59]. This proportion is calculated using a McDonald-Kreitman test [107] framework as:

$$\alpha = 1 - \frac{N_s P_r}{N_r P_s}$$

where N_s is the number of substitutions that are silent (i.e. not amino-acid changing), N_r is the number of substitutions that are amino-acid changing, P_s is the number of polymorphisms that are silent, and P_r is the number of polymorphisms that are amino-acid changing. Since we are interested in comparing autosomal loci and X-linked loci as classes, we calculated α using the total number of substitutions and polymorphisms in single-copy genes in each class. We conservatively restricted our analysis to single-copy genes identified using OrthoDB [108] to avoid possible mapping errors involving gene families or paralogs and substitutions that were considered private on each branch of the species tree. Specifically, for each group that was represented by a population sample (i.e. S form was excluded), we identified substitutions as sites where either all individuals were represented (*A. merus*) or all but one individual were represented (M form, GOUNDRY, *A. arabiensis*). Substitutions were considered private if one population was fixed for one allele while all three others were fixed for the alternative allele. Moreover, to minimize the effects of sites falsely counted as polymorphic and the effects of weakly-deleterious mutations that can bias estimates of α [109,110], we excluded all singleton sites. Polymorphic sites were identified as described in Section S10 above for *A. merus, A. arabiensis*, and M form. High rates of inbreeding precluded robust estimates of polymorphism in GOUNDRY, so we used the ratio of polymorphic sites from the M form population for comparison with GOUNDRY substitutions. We annotated substitutions and polymorphic sites using the SNPeff program [111] and the AgamP3.7 gene set database distributed with SNPeff.

For this analysis, we used only GOUNDRY individuals homozygous for 2La^+, which includes in a large number of private substitutions on this chromosome relative to other populations that are all fixed for the alternative 2La^a arrangement. While the evolutionary biology of this inversion is interesting, it has the potential to bias our genome-wide estimate of α in GOUNDRY. Therefore, to be conservative on all branches, we excluded genes in all autosomal chromosomal inversions fixed or segregating on each branch. We did not exclude the X-linked inversions as this would result in exclusion of nearly the entire X chromosome in *A. gambiae* and *A. arabiensis.*

We summarize the results for single-copy genes on each branch in Table S11. We find a strong and significant excess of replacement substitutions among single-copy autosomal loci on all external branches as well as X-linked loci on all external branches except GOUNDRY (Figure S5). We also find that all branches and classes show values of α above 0.28 consistent with rampant positive selection in this system. For

comparison, estimates of α are similar in *Drosophila* and have been interpreted as evidence for substantial positive selection across the genome [112,113]. However, there is no discernable evidence for 'faster-X' protein evolution in our data. On the branches leading from *A. gambiae* to *A. arabiensis* and *A. merus* where the number of substitutions is high and thus power to estimate α is high, the proportion of adaptively fixed substitutions is higher among autosomal loci than among X-linked loci. Relative estimates between Autosomal and X-linked loci vary in direction on the other branches, presumably due to variance in the estimates of α stemming from low numbers of substitutions. Although power is generally low since not much time has passed for substitution of adaptive mutations among these taxa, these results do not provide evidence for a 'faster-X' effect in these species, suggesting that the 'large-X' effect apparent in estimates of introgression and genetic divergence cannot be attributed to DMIs differentiating these populations due to positive selection.

Putative GOUNDRY Inversion breakpoint mapping

The dramatic reduction in nucleotide diversity across a 1.67MB region on the X chromosome in our GOUNDRY sample led us to hypothesize the existence of a novel chromosomal inversion in this region. Read mapping and depth is normal across the region containing the reduction of diversity (Figure S21). The exceptionally low level of nucleotide diversity in this region suggest a strong selective sweep recently fixed in this region, but such a wide footprint of the sweep and the remarkably rapid recovery to background diversity levels outside the sweep are both unexpected under normal recombination conditions. The existence of a new inversion that suppresses recombination with the alternative, presumably ancestral, form is the most likely scenario. We observe in our data the fixation of a 1.67 MB haplotype implying that this haplotype was maintained without recombination during the majority of the selective sweep. Since recombination should be normal among chromosomes carrying the inverted form, this suggests that the inverted form of the inversion was extremely rare in the population at the beginning of the sweep

Several issues inherent to this system and our data complicate identifying the breakpoints of this putative inversion. The first is that the inversion was recently under positive selection resulting in patterns of nucleotide diversity, linkage disequilibrium, homozygosity, and allele frequency differentiation that may extend beyond the breakpoints. There are clear changes in these population genetic signals in the putative region, but we can only conclude that the inversion lies somewhere inside of this footprint. The second issue is that GOUNDRY is inbred and harbors long tracts of IBD, some of which overlap the putative inverted region, making it difficult to distinguish homozygosity related to the sweep from that resulting from inbreeding. The third issue is that the X chromosome is repetitive and harbors many transposable elements resulting in a reference sequence riddled with gaps. Inversion breakpoints that have been

characterized in *Anopheles* [114] often lie inside or near such lowly-complex, repetitive regions. Our short paired-end read data has insert sizes of approximately 470bps, so mapping the presumed breakpoints is expected to be challenging if they are in repetitive DNA, and we are unlikely to be able to map read pairs across these regions. It should be noted that very few chromosomal inversions identified with cytogenetics have been characterized molecularly, even in systems with reference sequences [64,80,115].

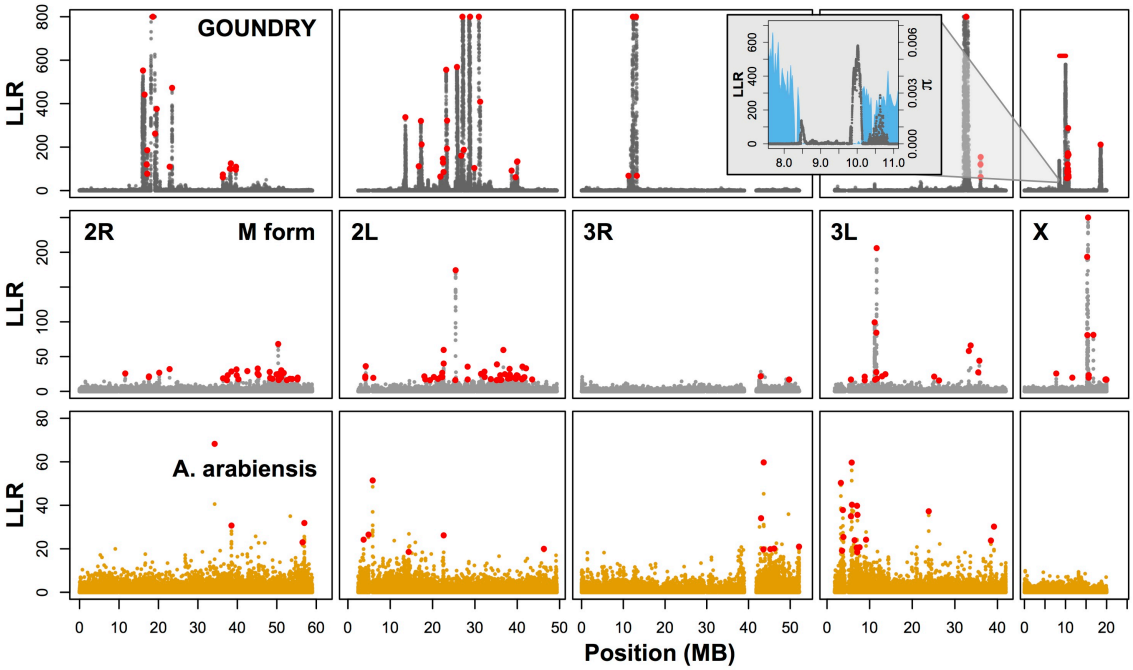
With these challenges in mind, we attempted to map the breakpoints of the inversion using a variety of approaches to collect candidate sites that could be assayed using PCR amplification. We manually inspected the short read data in this region using IGV [116] and located the edges of the swept region at positions 8,462,788 on the left and 10,137,178 on the right, according to coordinates in the AgamP3 PEST reference sequence. Since the selective sweep likely purged nucleotide diversity from the region surrounding the inverted haplotype as well, we attempted to identify the true breakpoints assuming that true breakpoints fall just inside these boundaries. For all of the following analyses except the read-depth analysis, we used reads that were trimmed using SolexaQA (v.2.1; [117]) with default settings and 50bp minimum size. We attempted a large series of *de novo* assemblies using the multi-kmer approach in SOAPdenovo2 [118] with –R –F and –p 4. We first attempted a de novo assembly of just reads that mapped to the inversion region on the X and obtained many contig sets based on variety of subsets of the data. None of the contigs showed the 'T' shape expected when the contig is aligned to the PEST reference and part of the contig is inverted. We also looked for cases of consensus where a variety of de novo assemblies using subsets of the data or all GOUNDRY reads all failed to assemble across a point on the reference, assuming that we could rule out regions that were properly assembled and aligned and focus on regions that 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 [119] with a window size of 2 million and a minimum inversion size of 2000. We also used GASV [115] 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

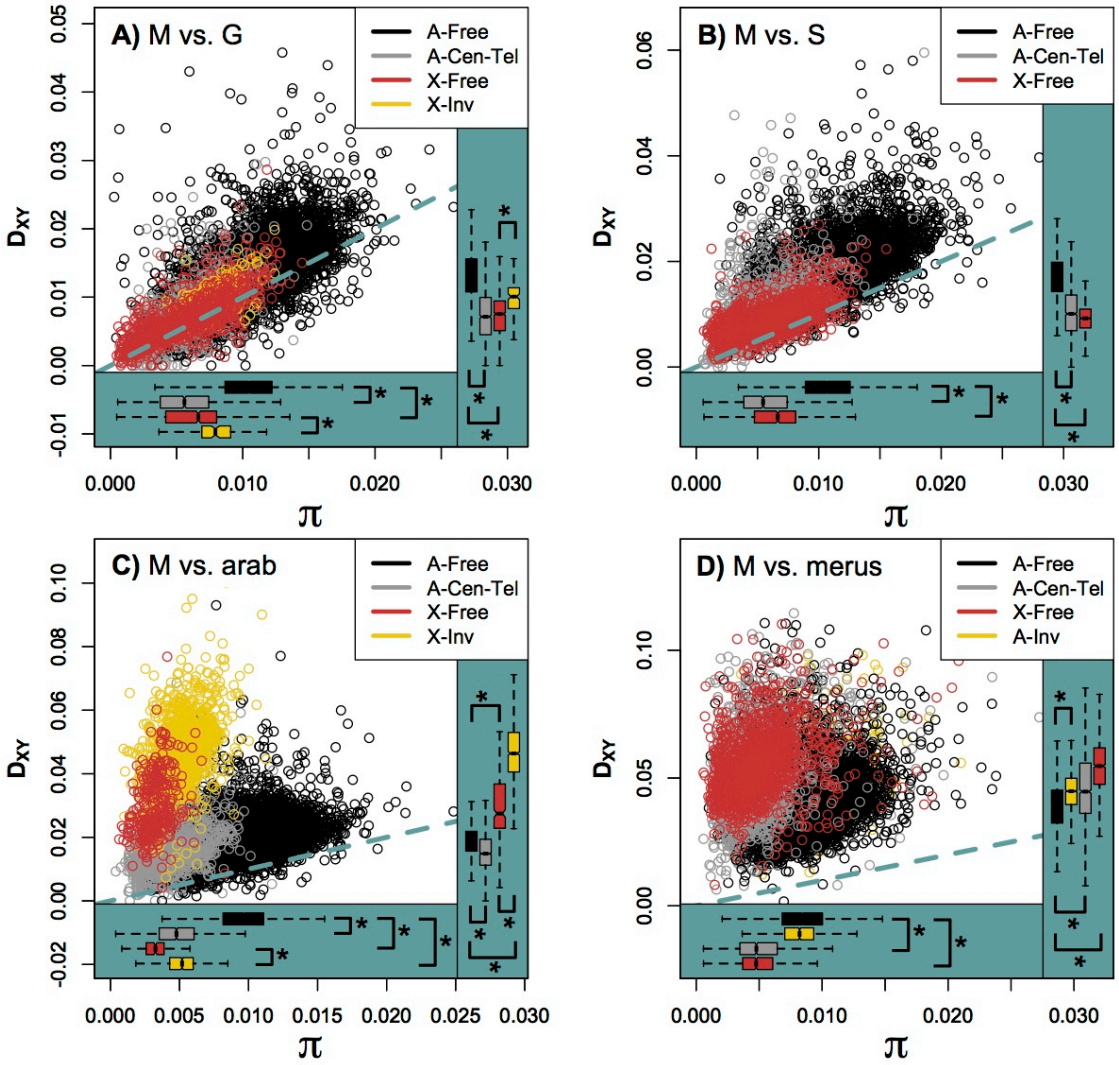
to reliably amplify in any group, ultimately preventing the molecular characterization of this inversion.

Figures:

Figure 1:

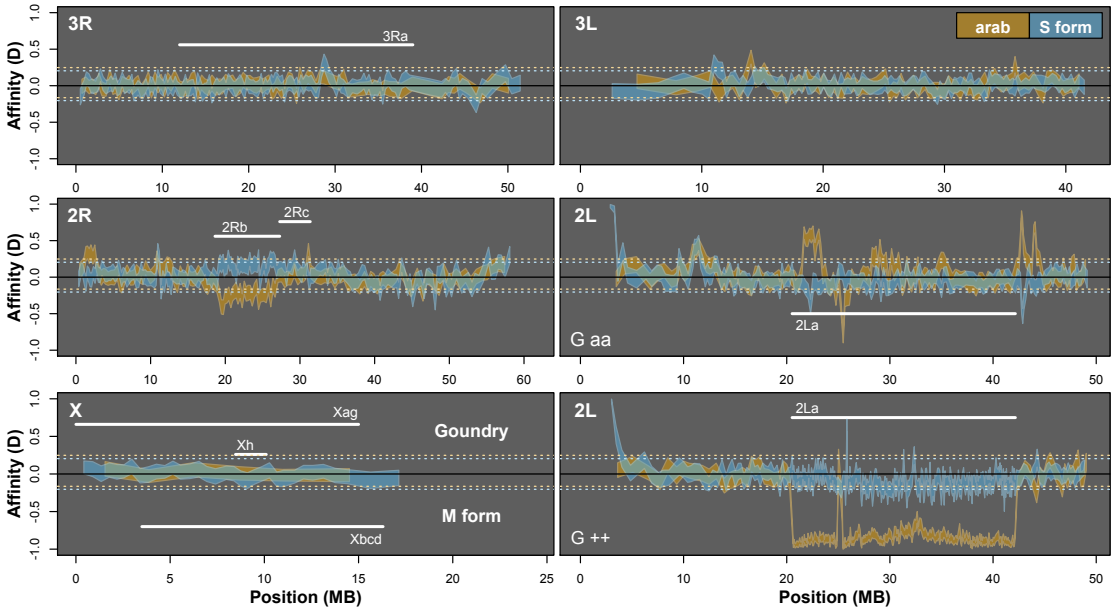


1677 **Figure 2:**

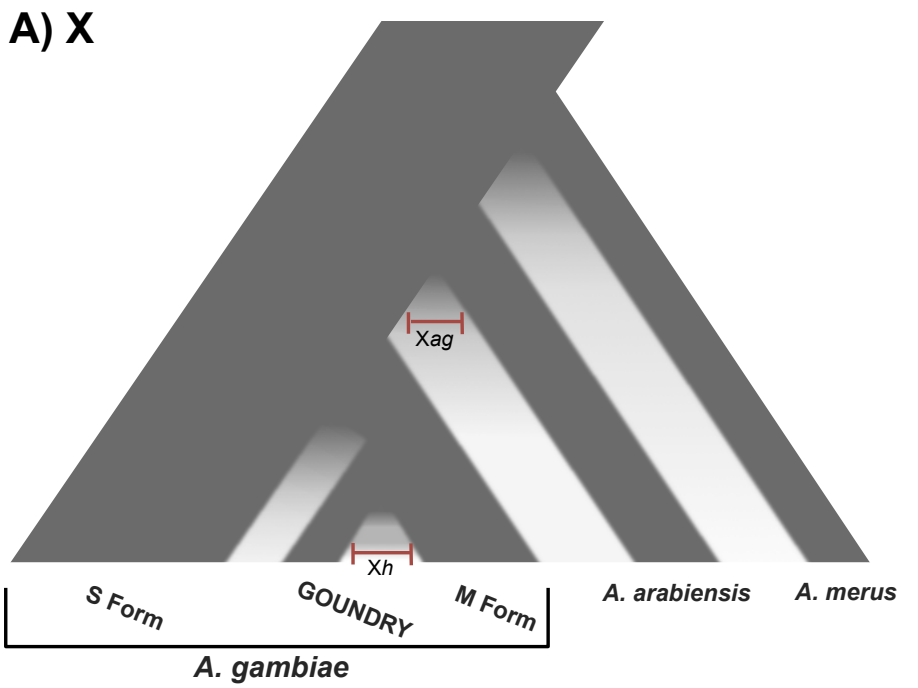


1678
1679

Figure 3:



1683 **Figure 4:**



1684
1685
1686

Figure Legends:

Figure 1: 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 populations. 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 were truncated for presentation. Low-complexity regions were excluded. A full list of inferred targets of selection is given in Table S1.

Figure 2: Patterns of genetic divergence (D_{xy}) between populations as a function of nucleotide diversity (π) reveal differential gene flow during speciation. Genomic regions defined by expected rates of recombination in hybrids (Supplementary Text) differ in their distributions of nucleotide diversity and genetic divergence, but not always in the same direction, suggesting that gene flow has been restricted on the X and lowly recombining regions in some cases. **A)** M form vs. GOUNDRY, **B)** M form vs. S form, **C)** M form vs. *A. arabiensis* **D)** M form vs. *A. merus*. Panel legends indicate colors corresponding to genomic location of each 10 kb window where 'Free' indicates freely recombining regions, 'Cen-Tel' indicates centromeric and telomeric autosomal regions, and 'Inv' indicates chromosomal inversions. 'A-' and 'X-' indicate autosomal or X chromosome. Dashed blue-green line indicates perfect correlation. Asterisks indicate Mann-Whitney tests with P values < 3.92×10^{-05} for comparisons indicated with brackets. Note that the y-axis scale differs among panels.

Figure 3: Significant autosomal introgression between *Anopheles* species and subspecies. ABBA-BABA statistics were calculated in non-overlapping windows of 500 informative sites using *A. merus* as the outgroup. Blue ribbon indicates 95% confidence region for introgression between S form ($2La^{a/+}$) and GOUNDRY (positive D; $2La^{a/a}$ and $2La^{+/+}$; *3R+; Xag*) and M form (negative D; $2La^{a/a}$; *3R+*; *Xag*). Orange ribbon indicates 95% confidence region for introgression between *A. arabiensis* ($2La^{a/a}$; *3Ra*; *Xbcd*) and GOUNDRY (positive D) and M form (negative D). Horizontal dotted lines (orange = *A. arabiensis*; blue = S form) indicate genome wide significance level after correction for multiple testing. Positions of relevant chromosomal inversions indicated with horizontal white lines. A full list of genes within significant windows is given in Table S3.

Figure 4: 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*).

Supplementary Table Legends:

Table S1: Inferred selective sweeps in GOUNDRY, M form and *A. arabiensis*. Chr: chromosomal location of sweep; Peak Pos: location of window with highest LLR; LLR – log likelihood ratio from SweepFinder; P_{gen} – genome-wide *P* value, 0 values indicate *P* values less than 0.005; P_{loc} – per-locus *P* value, 0 values correspond to *P* values less than 1.26×10^{-6} for autosomes and less than 6.67×10^{-6} for the X chromosome. Gene name, ImmunoDB ID, and Gene description from Vectorbase.org.

Table S2: Genes located inside large swept region (~8.47 – 10.1 MB) on X chromosome in GOUNDRY subgroup of *A. gambiae*. Chr: chromosomal location of sweep; Gene name, ImmunoDB ID, and Gene description from Vectorbase.org.

Table S3: Genes located inside introgressed windows with genome-wide significant values of *D*. Introg - Introgression between indicated populations; G_Abf – Introgression between GOUNDRY and Burkina Faso population of *A. arabiensis*; M_Abf - Introgression between M form and Burkina Faso population of *A. arabiensis*; M_Sform - Introgression between M and S forms; G_Sform - Introgression between GOUNDRY and S form; Chr – chromosomal location; Gene start bp – basepair coordinate of start in coding gene sequence; Gene end bp – basepair coordinate of end in coding gene sequence; Gene name, ImmunoDB ID, and Gene description from Vectorbase.org.

Table S4: Collection site and date information for mosquito samples.

Table S5: Next-generation sequencing statistics for mosquito samples.

Table S6: Summary statistics for mapping to *A. gambiae* PEST reference and population SPEC reference.

Table S7: Sites included and excluded from analysis in all populations.

Table S8: Optimized parameter values and confidence intervals from demographic inference. See Figure S18 for parameter descriptions.

Table S9: Variance-based test for introgression.

Table S10: Genome-wide critical values for ABBA-BABA test of introgression.

Table S11: McDonald-Kreitman tests and proportion of substitutions that are adaptive.

Supplementary Figure Legends:

Figure S1: Chromosomal distributions of nucleotide diversity (π) at inter-genic sites (LOESS-smoothed with span of 1% using 10 kb non-overlapping windows). Low complexity and heterochromatic regions were excluded. M = *A. gambiae* M form; G = *A. gambiae* GOUNDRY; R = *A. merus*; A = *A. arabiensis.*

Figure S2: Patterns of divergence among subgroups of *A. gambiae* follow similar curves (LOESS-smoothed with span of 1% using 10kb 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 S3: 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 these statistics. 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.

Figure S4: Relative genetic divergence (D_a) between GOUNDRY and M form. Da plotted as a function of nucleotide diversity (M form) using only intergenic sites in non-overlapping 10kb windows. Low complexity and heterochromatic regions were excluded. Genomic regions were defined based on predicted rates of recombination in hybrids (see Supplementary Methods) and compared using non-parametric Mann-Whitney tests. X-Free: freely recombining regions on X chromosome. X-Inv: region inside putative X*h* chromosomal inversion. Asterisks indicate P values less than 4.896e-08.

Figure S5: Adaptive protein evolution is rampant in *Anopheles*, but the proportions of substitutions that are adaptive (α) on the X and the autosomes do not support 'faster-X' hypothesis. Private substitutions in single-copy genes on each branch and polymorphisms were used to calculate α for the X and autosomes separately (SuppInfo). The numbers of replacement (amino-acid changing) and silent substitutions are presented

(D_r/D_s) for Autosomal ('A') and X-linked loci ('X') on each branch and asterisks indicate significant excess of replacement substitutions for that class and branch. Autosomal inversion regions were excluded to avoid large *2La* effect.

Figure S6: Genetic distance (D_{xy}) based neighbor-joining trees for A) the autosomes (chromosomal arms 2R, 3L, 3R) and B) X chromosome. Intergenic sites were used for distance calculations excluding low complexity and heterochromatic regions. Note different scale bars. Am: *A. merus*, Aa: *A. arabiensis*, Ag-S: *A. gambiae* S form, Ag-M: *A. gambiae* M form, Ag-G: *A. gambiae* GOUNDRY. All nodes supported by 1000 out of 1000 block bootstrap replicates.

Figure S7: Heterozygosity across 3L for 6 representative GOUNDRY individuals. Heterozygosity was measured using genotype calls in 10kb windows and plotted as a function of chromosomal position. These plots show neither the most extreme case of inbreeding nor the most extreme case of outbreeding, but represent representative "average" state. Plots for other autosomal chromosome arms and for other individuals are similar.

Figure S8: Histogram of IBD tract lengths in GOUNDRY individuals presented as proportion of chromosome reveals small number of large tracts indicating matings between closely related individuals. Inset shows only IBD tracts with size > 0.1.

Figure S9: Inbreeding coefficients for each individual and each chromosomal arm. Inbreeding coefficients were estimated directly from genotype likelihoods for each chromosomal arm separately using Vieira et al. 2013 (see Section S8 for more info). Each panel corresponds to a chromosomal arm. Each bar represents an individual mosquito sample. Colors indicate population assignment according to the legend.

Figure S10: Autosomal 1D unfolded site frequency spectra (SFS) for intergenic variable sites in each population. The SFS expected under standard coalescent equilibrium conditions is presented in black. The SFS was inferred for M form and *A. arabiensis* using –realSFS 1 (see Section S5), but the SFS for GOUNDRY was inferred using the inbreeding-aware version (-realSFS 2) after estimating inbreeding coefficients for this population (see Section S8).

Figure S11: Nucleotide diversity among parental chromosomes in seven representative GOUNDRY individuals (3L diploid sequence presented) is reduced in long Identity-By-Descent (IBD) tracts, suggesting GOUNDRY is inbred. Outbred regions bear resemblance to M form population levels of nucleotide diversity (red line).

Figure S12: Analysis of reference (REF) read proportions at heterozygous sites in GOUNDRY (bottom) and M form (top). Histograms of the proportions of reads carrying the REF allele are presented for sites called heterozygous in each population. In the absence of read mapping biases, we expect the distribution of proportions of reads carrying the reference and alternative bases at heterozygous sites to be centered on 0.5. The distributions for both M form and GOUNDRY have a mean very close to 0.5.

Distributions of read proportions look very similar between M form and GOUNDRY, suggesting that read mapping biases are not likely to explain excess homozygosity in GOUNDRY.

Figure S13: Analysis of read depth distributions at homozygous and heterozygous sites in GOUNDRY (right column) and M form (left column). Histograms of read depth are presented for sites called as (top row) homozygous for the alternative allele, (middle row) homozygous for reference allele, and (bottom row) heterozygous. Red lines indicate mean value for each class and population. Distributions of read depths look very similar between M form and GOUNDRY, suggesting that read mapping bias or lack of read depth are not likely to explain excess homozygosity in GOUNDRY.

Figure S14: Background Linkage Disequilibrium (LD) for each chromosomal arm in *A. gambiae* Mform and *A. arabiensis*. LD (r^2) between SNPs separated by at least 1kb (10 kb maximum) was averaged in 10 kb non-overlapping windows and LOESS-smoothed using a span of 1%. Low complexity and heterochromatic regions have been excluded. The large spikes on 2L and X (noted with vertical blue shaded bars) also coincide with reductions in nucleotide diversity and are thus likely to be recent selective sweeps. Additionally, there are a number of long distance increases (noted with horizontal grey bars) in LD in both M form and *A. arabiensis* that coincide approximately with the locations of known chromosomal inversions segregating in these populations (Coluzzi et al. 2002).

Figure S15: Decay of Linkage Disequilibrium (LD) in *A. gambiae* M form and *A. arabiensis*. LD (r^2) between SNPs separated by no more than 5 kb binned, averaged, and plotted as a function of physical distance. Low complexity regions were excluded. Chromosomal arms 2R, 3L, and 3R were included for the autosome curves and the X plotted separately. Note different X and autosome y-axis scales

Figure S16: Distribution of nucleotide diversity among genomic regions. Nucleotide diversity was calculated either from population samples (M form and *A. arabiensis*) using allele frequencies estimated from genotype likelihoods in ANGSD, or as the proportion of heterozygous genotype calls within a single diploid genome sequence (S form and *A. merus*). In both cases, diversity was estimated using only intergenic sites in non-overlapping 10kb windows. Low complexity and heterochromatic regions were excluded. Genomic regions were defined based on predicted rates of recombination in hybrids (see Supplementary Methods) and compared using non-parametric Mann-Whitney tests. Auto-Free: Freely recombining autosomal regions, Cen-Tel: Pericentromeric and telomeric regions, X-Free: freely recombining regions on X chromosome. Inv: regions inside known or suspected chromosomal inversions on the X (X-Inv) or autosomes (A-Inv). Asterisks indicate P values less than 2.2e-16.

Figure S17: ΔD_{xy} between *A. gambiae* subgroups GOUNDRY and M form plotted for windows across the X. Permutation test *P*-values for GOUNDRY vs. M form comparisons presented on log-scale in bottom panel with standard and Bonferonni-

corrected thresholds (Supplementary Text). Grey bar indicates inverted X*h* chromosomal region.

Figure S18: 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 ($2N_Am_{GM}$) and GOUNDRY into M ($2N_Am_{GM}$).

Figure S19: 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 S20: Modeling expected values of Relative Node Depth (D_{GM}/D_{GS}). **A)** Expected values of RND when ancestral population sizes are assumed to be equal. Colors indicate the expectations under different relative split times. **B)** Expected values with t_{GS} split time fixed to 1.1 (top) times the split time between GOUNDRY and M (t_{GM}) or 1.5 times (bottom). Colors indicated relative effective sizes of ancestral populations. Values are plotted as a function of the GOUNDRY-M effective size (x-axis). Grey bar indicates 95% confidence interval demographic estimate for GOUNDRY-M ancestral size (see Section S13).

Figure S21: Mean total read depth for GOUNDRY X chromosome sweep region. Mean total read depth across all GOUNDRY samples (n=12) for sites within non-overlapping 500 bp windows and plotted as a function of chromosomal position (megabases). The position of large GOUNDRY X sweep region is shown with grey bar.

Figure S22: Estimating the age of the selective sweep inside the putative 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 S23: Minor allele counts at variable sites within the large X-linked sweep region 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 Supplementary Text).

Figure S24: Comparison between the data, the best-fit model from *dadi*, and neutral coalescent simulations using MaCS. Site-frequency spectra were calculated from the **A)** true autosomal data, **B)** best-fit autosomal demographic model inferred using *dadi*, and

C) polymorphism data generated using neutral coalescent simulations in MaCS. LD decay curves from true data (red line) are compared to decay curves inferred from **D)** coalescent simulations for X chromosome (grey), and **E)** coalescent simulations for autosomal data.

Figure S25: Distribution of maximum SweepFinder log-likelihood ratio statistics in a 50 kb region of neutral polymorphism data simulated under the demographic model inferred for GOUNDRY (top) and M form (bottom). See Section S16 for simulation details.

Figure S26: Distribution of maximum SweepFinder log-likelihood ratio statistics in a full synthetic genome of neutral polymorphism data simulated under the demographic model inferred for GOUNDRY (top) and M form (bottom). See Section S16 for simulation details.

Figure S27: A representative cluster of significant selection peaks from Sweepfinder in GOUNDRY that were collapsed. The top panel shows the LLR 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 different y-axis values reflect the different sets of data considered in the composite likelihood function. The dotted line on the bottom panel indicates our threshold for 'low' values (LLR = 20). Under our clustering approach, this cluster was collapsed into a single peak.

Figure S28: A representative cluster of significant selection peaks from Sweepfinder in GOUNDRY that were not collapsed. The top panel shows the LLR 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 S29: Test for differences in the amount of recent selection between the GOUNDRY versus M form populations. We tested whether the number of megabases harboring at least a single significant selective sweep window differed between the populations by subtracting the number of megabases in M form from the number of megabases in GOUNDRY. The red line indicates the true difference. The histogram presents the same calculation made using 10^4 datasets with population assignments randomly permuted. No permuted genome produced the same or larger difference between the two populations.

References:

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

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

3. Harrison RG, Larson EL (n.d.) Hybridization, introgression, and the nature of species boundaries. Heredity In Press.

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

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

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

7. WHO (2013) World Malaria Report.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2071 23. Coyne JA, Orr HA (1989) Patterns of speciation in Drosophila. Evolution: 362–
2072 381.

2073 24. Geraldes A, Ferrand N, Nachman MW (2006) Contrasting Patterns of
2074 Introgression at X-Linked Loci Across the Hybrid Zone Between Subspecies of
2075 the European Rabbit (Oryctolagus cuniculus). Genetics 173: 919–933.
2076 doi:10.1534/genetics.105.054106.

2077 25. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, et al. (2014) The
2078 genomic landscape of Neanderthal ancestry in present-day humans. Nature
2079 507: 354–357. doi:10.1038/nature12961.

2080 26. Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF (2008) Fine-scale mapping
2081 of recombination rate in Drosophila refines its correlation to diversity and
2082 divergence. Proc Natl Acad Sci 105: 10051–10056.
2083 doi:10.1073/pnas.0801848105.

2084 27. McGaugh SE, Noor MAF (2012) Genomic impacts of chromosomal inversions in
2085 parapatric Drosophila species. Philos Trans R Soc Lond B Biol Sci 367: 422–
2086 429. doi:10.1098/rstb.2011.0250.

2087 28. White BJ, Lawniczak MKN, Cheng C, Coulibaly MB, Wilson MD, et al. (2011)
2088 Adaptive divergence between incipient species of Anopheles gambiae
2089 increases resistance to Plasmodium. Proc Natl Acad Sci U S A 108: 244–249.
2090 doi:10.1073/pnas.1013648108.

2091 29. Wu Q, Wen T, Lee G, Park JH, Cai HN, et al. (2003) Developmental Control of
2092 Foraging and Social Behavior by the Drosophila Neuropeptide Y-like System.
2093 Neuron 39: 147–161. doi:10.1016/S0896-6273(03)00396-9.

2094 30. Gimonneau G, Pombi M, Dabire RK, Diabate A, Morand S, et al. (2012)
2095 Behavioural responses of Anopheles gambiae sensu stricto M and S molecular
2096 form larvae to an aquatic predator in Burkina Faso. Parasit Vectors 5: 65.
2097 doi:10.1186/1756-3305-5-65.

2098 31. Croker B, Crozat K, Berger M, Xia Y, Sovath S, et al. (2007) ATP-sensitive
2099 potassium channels mediate survival during infection in mammals and insects.
2100 Nat Genet 39: 1453–1460. doi:10.1038/ng.2007.25.

2101 32. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical membrane
2102 topology and heteromeric function of Drosophila odorant receptors in vivo.
2103 PLoS Biol 4: e20. doi:10.1371/journal.pbio.0040020.

2104 33. Rund SSC, Hou TY, Ward SM, Collins FH, Duffield GE (2011) Genome-wide
2105 profiling of diel and circadian gene expression in the malaria vector
2106 Anopheles gambiae. Proc Natl Acad Sci 108: E421–E430.
2107 doi:10.1073/pnas.1100584108.

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

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

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

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

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

39. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, et al. (2003) Semipermeable species boundaries between Anopheles gambiae and Anopheles arabiensis: evidence from multilocus DNA sequence variation. Proc Natl Acad Sci U S A 100: 10818–10823. doi:10.1073/pnas.1434337100.

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

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

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

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

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

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

2144 46. Weill M, Chandre F, Brengues C, Manguin S, Akogbeto M, et al. (2000) The kdr
2145 mutation occurs in the Mopti form of Anopheles gambiaes. s. through
2146 introgression. Insect Mol Biol 9: 451–455.

2147 47. Charlesworth B, Coyne JA, Barton NH (1987) The Relative Rates of Evolution of
2148 Sex Chromosomes and Autosomes. Am Nat 130: 113. doi:10.1086/284701.

2149 48. Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and
2150 the reproductive isolation of species. Proc Natl Acad Sci U S A 98: 12084–
2151 12088. doi:10.1073/pnas.221274498.

2152 49. Rieseberg LH (2001) Chromosomal rearrangements and speciation. Trends Ecol
2153 Evol 16: 351–358.

2154 50. Slotman MA, Della Torre A, Calzetta M, Powell JR (2005) Differential
2155 introgression of chromsomal regions between Anopheles gambiae and An.
2156 arabiensis. Am J Trop Med Hyg 73: 326–335.

2157 51. O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, et al. (2014) Genomic
2158 analyses of three malaria vectors reveals extensive shared polymorphism but
2159 contrasting population histories. Mol Biol Evol 31: 889–902.
2160 doi:10.1093/molbev/msu040.

2161 52. Slotman MA, Reimer LJ, Thiemann T, Dolo G, Fondjo E, et al. (2006) Reduced
2162 recombination rate and genetic differentiation between the M and S forms of
2163 Anopheles gambiae s.s. Genetics 174: 2081–2093.
2164 doi:10.1534/genetics.106.059949.

2165 53. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft
2166 Sequence of the Neandertal Genome. Science 328: 710–722.
2167 doi:10.1126/science.1188021.

2168 54. Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, et al. (2012) Genome
2169 sequencing reveals complex speciation in the Drosophila simulans clade.
2170 Genome Res. Available:
2171 http://genome.cshlp.org/content/early/2012/04/13/gr.130922.111.
2172 Accessed 15 October 2012.

2173 55. Geraldes A, Basset P, Gibson B, Smith KL, Harr B, et al. (2008) Inferring the
2174 history of speciation in house mice from autosomal, X-linked, Y-linked and
2175 mitochondrial genes. Mol Ecol 17: 5349–5363. doi:10.1111/j.1365-
2176 294X.2008.04005.x.

2177 56. Presgraves DC (2008) Sex chromosomes and speciation in Drosophila. Trends
2178 Genet TIG 24: 336–343. doi:10.1016/j.tig.2008.04.007.

57. Bateson W (1909) Heredity and variation in modern lights. Darwin and modern science. Cambridge, UK: Cambridge University Press. pp. 85–101.

58. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.

59. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022–1024. doi:10.1038/4151022a.

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

61. Calboli FCF, Sampson J, Fretwell N, Balding DJ (2008) Population structure and inbreeding from pedigree analysis of purebred dogs. Genetics 179: 593–601. doi:10.1534/genetics.107.084954.

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

63. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, et al. (2008) Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. Malar J 7: 163. doi:10.1186/1475-2875-7-163.

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

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

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

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

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

2215 69. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A
2216 framework for variation discovery and genotyping using next-generation DNA
2217 sequencing data. Nat Genet 43: 491–498. doi:10.1038/ng.806.

2218 70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence
2219 Alignment/Map format and SAMtools. Bioinforma Oxf Engl 25: 2078–2079.
2220 doi:10.1093/bioinformatics/btp352.

2221 71. Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for
2222 population genetics analyses from next-generation sequencing data.
2223 Bioinforma Oxf Engl 30: 1486–1487. doi:10.1093/bioinformatics/btu041.

2224 72. Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, et al. (2010)
2225 Genome mapping and characterization of the Anopheles gambiae
2226 heterochromatin. BMC Genomics 11: 459. doi:10.1186/1471-2164-11-459.

2227 73. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, et al. (2011)
2228 Estimation of allele frequency and association mapping using next-generation
2229 sequencing data. BMC Bioinformatics 12: 231. doi:10.1186/1471-2105-12-
2230 231.

2231 74. Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding
2232 coefficients from NGS data: Impact on genotype calling and allele frequency
2233 estimation. Genome Res 23: 1852–1861. doi:10.1101/gr.157388.113.

2234 75. R Development Core Team (2011) R: A language and Environment for Statistical
2235 Computing. Available: http://www.R-project.org/.

2236 76. Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA (2010) Drosophila
2237 melanogaster recombination rate calculator. Gene 463: 18–20.
2238 doi:10.1016/j.gene.2010.04.015.

2239 77. Chan AH, Jenkins PA, Song YS (2012) Genome-Wide Fine-Scale Recombination
2240 Rate Variation in Drosophila melanogaster. PLoS Genet 8: e1003090.
2241 doi:10.1371/journal.pgen.1003090.

2242 78. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-
2243 scale structure of recombination rate variation in the human genome. Science
2244 304: 581–584. doi:10.1126/science.1092500.

2245 79. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, et al. (2013) Genetic Recombination
2246 Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genet 9: e1003984.
2247 doi:10.1371/journal.pgen.1003984.

2248 80. Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A new chromosomal phylogeny
2249 supports the repeated origin of vectorial capacity in malaria mosquitoes of the

Anopheles gambiae complex. PLoS Pathog 8: e1002960. doi:10.1371/journal.ppat.1002960.

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

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

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

84. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 14: 289. doi:10.1186/1471-2105-14-289.

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

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

87. Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet 14: 262–274. doi:10.1038/nrg3425.

88. Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356: 519–520. doi:10.1038/356519a0.

89. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

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

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

92. García BA, Caccone A, Mathiopoulos KD, Powell JR (1996) Inversion monophyly in African anopheline malaria vectors. Genetics 143: 1313–1320.

2285 93. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring
2286 the Joint Demographic History of Multiple Populations from Multidimensional
2287 SNP Frequency Data. PLoS Genet 5: e1000695.
2288 doi:10.1371/journal.pgen.1000695.

2289 94. Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly
2290 (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21: 36–44.
2291 doi:10.1093/molbev/msg236.

2292 95. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, et al. (2007) Direct
2293 estimation of per nucleotide and genomic deleterious mutation rates in
2294 Drosophila. Nature 445: 82–85. doi:10.1038/nature05388.

2295 96. Schrider DR, Houle D, Lynch M, Hahn MW (2013) Rates and genomic
2296 consequences of spontaneous mutational events in Drosophila melanogaster.
2297 Genetics 194: 937–954. doi:10.1534/genetics.113.151670.

2298 97. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, et al. (2009) Analysis of
2299 the genome sequences of three Drosophila melanogaster spontaneous
2300 mutation accumulation lines. Genome Res 19: 1195–1201.
2301 doi:10.1101/gr.091231.109.

2302 98. Keightley PD, Ness RW, Halligan DL, Haddrill PR (2014) Estimation of the
2303 spontaneous mutation rate per nucleotide site in a Drosophila melanogaster
2304 full-sib family. Genetics 196: 313–320. doi:10.1534/genetics.113.158758.

2305 99. Feder JL, Xie X, Rull J, Velez S, Forbes A, et al. (2005) Mayr, Dobzhansky, and
2306 Bush and the complexities of sympatric speciation in Rhagoletis. Proc Natl
2307 Acad Sci U S A 102 Suppl 1: 6573–6580. doi:10.1073/pnas.0502099102.

2308 100. Lehmann T, Hawley WA, Grebert H, Collins FH (1998) The effective population
2309 size of Anopheles gambiae in Kenya: implications for population structure.
2310 Mol Biol Evol 15: 264–276.

2311 101. Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient
2312 admixture between closely related populations. Mol Biol Evol 28: 2239–2252.
2313 doi:10.1093/molbev/msr048.

2314 102. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic
2315 scans for selective sweeps using SNP data. Genome Res 15: 1566.

2316 103. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA
2317 sequence data. Genome Res 19: 136–142. doi:10.1101/gr.083634.108.

2318 104. Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive
2319 selection in whole-genome SNP data from nonequilibrium populations.
2320 Genetics 185: 907–922. doi:10.1534/genetics.110.116459.

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

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

107. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654. doi:10.1038/351652a0.

108. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res 41: D358–365. doi:10.1093/nar/gks1116.

109. Fay JC, Wyckoff GJ, Wu C-I (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415: 1024–1026. doi:10.1038/4151024a.

110. Charlesworth J, Eyre-Walker A (2008) The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol 25: 1007–1015. doi:10.1093/molbev/msn005.

111. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80–92. doi:10.4161/fly.19695.

112. Baines JF, Sawyer SA, Hartl DL, Parsch J (2008) Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in Drosophila. Mol Biol Evol 25: 1639–1650. doi:10.1093/molbev/msn111.

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

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

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

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

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

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

119. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinforma Oxf Engl 25: 2865–2871. doi:10.1093/bioinformatics/btp394.

Acknowledgements:

We thank Matteo Fumagalli, Filipe Vieira, and Tyler Linderoth for assistance with next generation sequence data analyses and ANGSD. We also thank Russ Corbett-Detig for helpful comments on an earlier version of this manuscript. J.E.C. was funded by a graduate fellowship from Cornell Center for Comparative and Population Genomics at Cornell University, a Genentech Innovation Postdoctoral Fellowship in the Center for Computational Biology at University of California, Berkeley, and a Ruth L. Kirschstein National Research Service Award from the National Institutes of Health. This work was supported by NIH grant R01 AI062995 and made use of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

Sequence data from this study has been archived at NCBI's Short Read Archive under accession numbers XXXXX-XXXXX.