Standing genetic variation and chromosome differences drove rapid ecotype formation in a major malaria vector

Scott T. Small[a,b,c1], Carlo Costantini[d,e], N'Fale Sagnon[d], Moussa W. Guelbeogo[d], Scott J. Emrich[b,f2], Andrew D. Kern[c], Michael C. Fontaine[e,g], Nora J. Besansky[a,b1].

Abstract

Local adaptation results in the formation of ecotypes—conspecific groups with heritable differences that increase fitness to distinct environments. The adaptive significance of ecotypes is widely recognized, but not its genetic basis. Here, focusing on a dominant pan-African malaria vector, we study two co-occurring isomorphic yet karyotypically differentiated populations from Burkina Faso reported to differ in larval ecology and epidemiologically relevant adult behaviors. Since their discovery in the 1990s, further investigation was precluded by lack of modern genomic resources. Here, we used deep whole-genome sequencing and analysis to test the hypothesis that these two morphologically cryptic groups are ecotypes differentially adapted to breeding in natural swamps versus irrigated rice fields. We conclusively demonstrate genome-wide differentiation despite extensive microsympatry, synchronicity, and occasional hybridization. Demographic inference supports a split only ~1,300 years ago, coinciding with the massive expansion of domesticated African rice cultivation, suggesting that a newly abundant anthropogenic habitat may have driven vector diversification. The genomic pattern of differentiation is heterogenous, and we infer that the regions of highest differentiation were under selection during lineage splitting, consistent with local adaptation. The origin of nearly all variation implicated in local adaptation, including chromosomal inversions, substantially predates the ecotype split, suggesting that rapid adaptation to rice fields was fueled mainly by standing genetic variation. Sharp differences in chromosomal inversion polymorphism between ecotypes likely promoted adaptive divergence, by maximizing recombination within the homokaryotypic standard rice ecotype, and suppressing admixture with the other ecotype, which carries inverted arrangements at high frequencies.

Introduction

Local adaptation is central to the evolutionary process (1). It occurs when species exposed to spatially varying selection in heterogeneous environments evolve ecotypes—intraspecific groups with a suite of heritable differences that increase fitness to local conditions. Pervasive across all domains of life (2), familiar examples include high altitude adaptation in humans (3), freshwater adaptation in three-spine sticklebacks (4), and adaptation to environmental background color in mice and peppered moths (5, 6). Local adaptation plays roles in shaping species diversity, response to climate change, range expansion, and potentially, ecological speciation (7, 8). Local adaptation also has underappreciated public health significance. Malaria, one of humankind's deadliest infectious diseases, is transmitted by highly evolvable *Anopheles* mosquito vectors (9) whose effective control may be complicated or even compromised by local adaptations that diversify the ecological and environmental range, and alter epidemiologically relevant behaviors.

First proposed forty years ago by Coluzzi (10), a hypothesis of local adaptation in anophelines envisioned chromosomal inversions as instruments of ecotypic differentiation. Coluzzi's theory was inspired by extensive cytogenetic data from laboratory colonies and natural populations of species in the *Anopheles gambiae* complex, suggesting that inversion polymorphisms are maintained by natural selection. In broad outline, the theory assumes a species whose range expands into an ecologically marginal zone, where local population isolates acquire mutations adaptive for the marginal conditions. Chance capture of the adaptive variants by a new chromosomal inversion would jointly protect the variants as a haplotype from recombination and homogenization with maladapted genetic backgrounds, leading to ecotypic divergence and potential speciation. In 2006, Kirkpatrick and Barton (11) proposed a mathematical model for the role of new inversions in local adaptation, similar in spirit to the verbal model proposed for anophelines. These models emphasize the role of *de novo* inversions in local adaptation, rather than the build-up of genic differences inside of established inversions (12).

Apart from their proposed role in the origin of ecotypes, inversion polymorphisms historically served as markers of population structure for anopheline mosquitoes, as molecular markers were limited in the pre-genomic era (13). In one of the most extensive applications, surveys of the *An. gambiae* complex in Mali conducted over a ~20 year period assessed chromosome 2R karyotype configurations in nearly 18,000 specimens from 76 sampling sites (14). A major highlight of this work emerged from population genetic analysis of polytene chromosome preparations from >1,500 mosquitoes collected from a single Malian village in the middle of one rainy season. The analysis revealed significant departures from Hardy-Weinberg equilibrium, marked by heterozygote deficits that could not be explained by the mixing of temporally independent samples or methodological sampling biases (14). These and other data supported the stable coexistence of synchronous and assortatively mating "chromosomal forms" characterized by different chromosome 2R karyotype configurations, larval breeding site associations, and ecoclimatic tolerances (14). Molecular genetic markers independent of inversions, and ultimately whole genome sequencing, eventually validated genetic differentiation within *An. gambiae* leading to the recognition of *An. coluzzii* as a new species (15-19). The original karyotype-based definitions were not

fully congruent with the genotypic groups, due to some sharing of both inversion polymorphisms and karyotype configurations (20-22). Nonetheless, historical polytene chromosome analysis provided strong and otherwise unobtainable evidence of cryptic population structure, both in this species group and the present study subject, the malaria vector *Anopheles funestus*.

Together with key members of the *An. gambiae* complex, *Anopheles funestus* contributes to highly efficient malaria transmission across its vast tropical African distribution (23, 24). Like *An. gambiae*, *An. funestus* harbors extensive amounts of nucleotide diversity and chromosomal inversion polymorphism, and displays weak genetic differentiation across the species range, consistent with large effective population size and high connectivity (25, 26). Nevertheless, reminiscent of the chromosomal forms of *An. gambiae* in Mali, extensive entomological surveys of *An. funestus* in Burkina Faso beginning in the 1990s revealed striking chromosomal inversion-based evidence of the stable coexistence of two strictly sympatric and synchronous assortatively mating chromosomal forms designated Kiribina and Folonzo (27, 28). Folonzo was hypothesized to be more closely related genetically to typical pan-African *An. funestus* populations, based on its high level of inversion polymorphism on chromosomes 2R and 3R, and its association with species-characteristic larval habitats—natural swamps and ditches with abundant aquatic vegetation. Kiribina was nearly depauperate of inversion polymorphisms, carrying mainly the standard chromosomal orientations, and was found in association with large-scale rice cultivation. The working hypothesis of differences in larval ecology between the forms was reinforced by intensive longitudinal studies of adults (polytene chromosome analyses are possible only at the adult stage). These revealed seasonal variation in the relative frequency of the two taxa, confirmed across six successive years (29). Only Folonzo abundance was correlated with climatic variables related to temperature and rainfall, as expected if the growth of aquatic vegetation is rainfall dependent in its natural larval habitats, but not the irrigated rice field sites of Kiribina (29). Behavioral and bionomic studies demonstrated that although their human biting behavior and malaria parasite infection rates were statistically indistinguishable and comparably high—indicating that both Kiribina and Folonzo are formidable malaria vectors—there were important epidemiologically relevant differences (30). In particular, Folonzo was significantly over-represented in indoor-resting collections and showed stronger

post-prandial endophily, while Kiribina predominated outdoors (30). This suggests that the two taxa may not be uniformly exposed to indoor-based vector control interventions. Mitochondrial and microsatellite markers outside of inversions indicated weak but significant differentiation between the groups (31). However, the absence of a molecular diagnostic assay and modern genomic tools (a high-quality reference genome and cost-effective DNA sequencing) posed insurmountable hurdles to further elucidation of this system.

Here, we leverage the recent chromosome-scale *An. funestus* assembly (32) and apply individual, whole-genome deep sequencing to investigate retrospective collections of Kiribina and Folonzo from Burkina Faso. We sought to establish the extent to which inversion-based differentiation extends to collinear genomic regions, and to explore whether the timing and pattern of genetic differentiation is consistent with local adaptation by Kiribina to breeding in rice fields. Our data conclusively demonstrate that Kiribina and Folonzo are differentiated genome-wide despite extensive sympatry and synchronicity. Because the genetic lineages are not fully congruent with the karyotype-based classifications, we follow precedent (15) in renaming these taxa K and F forms (henceforth, ecotypes) of *An. funestus*. Remarkably, the K-F split follows African rice domestication and the massive expansion of its cultivation ~1,850 years ago (33), suggesting that this newly abundant larval habitat may have been a driver of K diversification. As expected for local adaptation, the genomic pattern of differentiation is heterogenous, and we infer that the regions of highest differentiation were under selection during lineage splitting. The origin of nearly all variation implicated in local adaptation, including chromosomal inversions, substantially predates separation of the two groups, suggesting that rapid exploitation of a novel anthropogenic habitat was fueled mainly by standing genetic variation and not *de novo* mutation. Homozygosity for the standard arrangement of chromosomal inversions in K likely promoted adaptive divergence, by maximizing recombination within K and suppressing it between K and F, which carries inverted arrangements at high frequencies.

Results and Discussion

Genetic structure reveals two differentiated groups of *An. funestus* in Burkina Faso

We analyzed 168 *An. funestus* sampled from 11 villages spanning a ~300-km east–west transect of Burkina Faso, where Kiribina and Folonzo co-occur at the village level (Fig 1A; *SI Appendix*, Text, Fig. S1, Tables S1-S2). Individual whole genome sequencing yielded a mean coverage of 38X after filtering (*SI Appendix*, Text, Tables S3-S4). Following read mapping to the AfunF3 reference and further filtering (*SI Appendix*, Text), we identified 24,283,293 high quality single nucleotide polymorphisms (SNPs) among 114,270,430 accessible sites scored across all individuals. Using a subset of these SNPs, we estimated genetic population structure based on principle components analysis (PCA) and the individual ancestry-based algorithm implemented in ALStructure (34) (*SI Appendix*, Text).

To avoid the potentially confounding effects of autosomal inversion polymorphisms, our initial inference by PCA relied on the X chromosome. Given the pan-African distribution of *An. funestus*, previously published genomic variation data from Ghana and Uganda (26) were included to provide broader context (*SI Appendix*, Table S1). The analysis revealed two main clusters (Fig 1A). One contains all 70 of the Burkina Faso *An. funestus* that would have been classified as Folonzo by inversion karyotyping using the deterministic algorithm of Guelbeogo*, et al.* (28), as well as another 22 Burkina Faso *An. funestus* that would have been (mis)classified as Kiribina due to low inversion polymorphism (blue squares, Fig 1A; *SI Appendix*, Table S1). This same cluster also includes the geographically distant samples from both Ghana and Uganda (crosses, Fig 1A). The second cluster corresponds to 74 Burkina Faso *An. funestus* that would have been classified as Kiribina (28) (orange circles, Fig 1A). Only two Burkina Faso *An. funestus* mosquitoes were not included in either of the two clusters; these were considered unassigned (green triangles, Fig 1A).

PCA based on SNPs from individual autosome arms showed exactly the same two clusters and unassigned mosquitoes (*SI Appendix*, Fig S2). However, the pattern within clusters for autosome arms carrying polymorphic inversions was more complex, as expected, reflecting the inversion karyotype of the mosquito carriers (*SI Appendix*, Fig S3). Strikingly, mosquito carriers of inversion karyotypes that

segregate in both populations (*e.g.*, homokaryotypes for $3R+^{a}$, $3R+^{b}$, $2R+^{a}$, $2R+^{s}$, and 2Rs) cluster by population and not by inversion karyotype. There was no discernable contribution of village sampling location to the clustering patterns for any chromosome arm (*SI Appendix*, Fig S4).

Individual-based ancestry analysis of the Burkina Faso sample supported the existence of the same two genetically differentiated populations as those identified by PCA (Fig 1B). Notably, the two unassigned mosquitoes have admixture proportions suggestive of inter-population hybrids (rows enclosed by horizontal white lines in Fig 1B). Compelling evidence supporting this suggestion is presented below.

Taken together, these results are indicative of the existence of two genetically differentiated ecotypes of *An. funestus* in Burkina Faso. They closely correspond to the formerly recognized chromosomal forms Kiribina and Folonzo, but the correspondence is imperfect because the deterministic karyotypic classification system misclassifies a fraction of karyotypes that are shared between ecotypes (in particular, the standard karyotype characteristic of, but not exclusive to, Kiribina; *SI Appendix*, Table S1). Following precedent (15), we distinguish 'molecularly defined' forms from the prior chromosomal forms with the new designations K and F. The fact that the F ecotype from Burkina Faso clusters together with other *An. funestus* sampled from neighboring Ghana and distant Uganda (Fig 1A) is consistent with the hypothesis that F is genetically allied with *An. funestus* populations across tropical Africa, while K has a more recent and local origin (27).

Demographic inference of a K-F split 1,300 years ago correlates with expanded rice agriculture

To explore the history of K-F divergence, we used an approximate Bayesian computation (ABC) method (35) to compare five alternative demographic scenarios, involving different amounts and timing of gene flow over four epochs (*SI Appendix*, Text, Fig S5). Model selection in this statistical framework (*SI Appendix*, Text, Tables S5-S8) allowed us to reject histories of panmixia, allopatric isolation, and secondary contact. The most strongly supported scenario, chosen as the most likely, involved isolation with initial migration (IIM). However, its posterior probability (0.44) was only slightly higher than a model

of isolation with migration (0.40), reflecting the difficulty of distinguishing these two closely related models (*SI Appendix*, Table S7).

We estimated demographic parameters under the chosen IIM model using two independent methodologies, ABC and an alternative approach based on Generative Adversarial Networks (GAN) implemented in the software *pg-gan* (36). The latter uses real data to adaptively learn parameters capable of generating simulated data indistinguishable from real data by machine learning (36). It operates directly on genotype matrices and avoids reducing genotypes to summary statistics, unlike ABC. Importantly, genetic data simulated under the IIM model by *pg-gan* and ABC produced similar summary statistics (*SI Appendix*, Text, Table S8). If the chosen IIM model closely approximates the true demographic history, we expect that the observed data and data simulated under the IIM model should have comparable summary statistics. To verify this expectation, posterior predictive checking was performed by simulating 10,000 new datasets with population parameter values drawn from the posterior distributions of the population parameters estimated from the retained *pg-gan* runs. Summary statistics were calculated using noncoding regions only, under the assumption that noncoding regions are less affected by selection. The median of each observed summary statistic was compared with the distribution of the 10,000 simulated test statistics. We calculated the percentile for each median value within the distribution of the simulated data statistic. All observed medians fell within the 17% – 84% percentiles for K, and the 20% – 59% percentiles for F, providing confidence in the IIM model.

Under the IIM model, we estimated that the ancestral K-F lineage split from an outgroup population of *An. funestus* ~6,000 years before present (*SI Appendix*, Table S8), consistent with our previous inference of a range expansion of *An. funestus* to its present continent-wide distribution less than 13,000 years ago (26). The relatively steep decline of the ancestral K-F lineage to an N_e of ~11,000 about 2,000 years ago (Fig 2A) may reflect widespread unfavorable environmental conditions connected with the drying of the Sahara (37). Parallel climate-driven population reduction of wild African rice observed at the continental scale, beginning more than 10,000 years ago and reaching a minimum at ~3,400 years ago, is the hypothesized trigger for African rice domestication, which may have originated in the inner Niger delta of

West Africa (33) (but see ref 38). By ~1,850 years ago, there was a strong expansion of domesticated rice cultivation (33). Strikingly, we estimate that K and F split ~1,300 (credible interval ~800-1,800) years ago, which fits well with the timing of a newly abundant anthropogenic larval habitat (Fig 2A, *SI Appendix*, Table S8). Around the time of K-F splitting, we inferred an ancestral N_e of ~64,000. Since K-F divergence, F expanded to an N_e of ~3,250,000 while K maintained a steady population size (Fig 2A, *SI Appendix*, Table S8). Post-divergence K-F gene flow, initially mainly from F to K, slowed ~500 years ago and effectively ceased by ~100 years ago (Fig 2B).

A recent advance in genome-wide genealogy estimation, implemented in a new method called 'Relate', allows inference of genealogical trees across all loci (haplotypes) in a large sample of genomes (39). Using these genealogies directly, Relate infers recombination, mutational ages, and—based on the tracking of mutation frequencies through time—identifies loci under positive natural selection (see below). Applying Relate, we find that the first K-F cross-coalescent events, estimated from autosome arm genealogies, occurred with a mean timing of ~100 years ago (*SI Appendix*, Text, Fig S6), providing further evidence of recent gene flow cessation between K and F (*SI Appendix*, Text, Fig S6). Indeed, based on the observation that the most recent cross-coalescence is older on the X chromosome than for the autosomes (*SI Appendix*, Fig S6), these ecotypes may be in the early stages of incipient speciation, in line with the expectation that X-linked genes may disproportionately contribute to local adaptation and reproductive isolation (40, 41).

Genetic diversity and differentiation along the genome are heterogeneous between ecotypes

Overall, nucleotide diversity is comparable between K and F (genome-wide mean values of 0.603E-2 and 0.624E-2, respectively). The same was true in windows along the genome, except around chromosomal rearrangements (Fig 3). Inversions 2Ra, 3Ra and 3Rb segregate exclusively in the F sample at frequencies of 0.44, 0.73, and 0.40, respectively, while K carries only the corresponding standard orientation of those inversions. [Inversions 2Rs and 3La are not considered here, as the former is carried at low frequency by both K (0.07) and F (0.04), and the latter segregates only in F at a frequency of 0.05

(*SI Appendix*, Table S1)]. Unsurprisingly, nucleotide diversity appears slightly elevated at these rearrangements in F relative to K (Fig 3), particularly for inversions 2Ra and 3Rb whose frequencies are more intermediate than 3Ra. In addition, values of Tajima's D are less negative across the genome in K relative to F (genome-wide means, -2.031 and -2.408, respectively), likely owing to the rapid post-divergence population size expansion of F (Fig 2A).

We examined the genomic landscape of divergence using both absolute (d_{XY}) and relative (F_{ST}) measures, recognizing that relative measures of divergence may reflect reduced diversity instead of reduced gene flow (42). We found that the pattern of absolute K-F divergence closely corresponds to the pattern of nucleotide diversity in the ecotypes, probably owing to their recent split, and differs strikingly from the pattern of relative K-F divergence along the chromosomes (Fig 3). The relative measure clearly takes on much larger values in the rearranged versus collinear chromosomal regions (Fig 3). Indeed, the mean F_{ST} value for rearranged regions (0.045) is fourfold higher than for collinear regions (0.011), consistent with relatively high inversion frequencies in F and homozygosity for the standard arrangements in K. Although the corresponding d_{XY} values for these regions follow the same trend, being significantly larger for rearranged versus collinear regions (0.769E-2 and 0.620E-2; K-S test p-value < 9.56E-173), d_{XY} has less power to detect differences among loci due to the very recent K-F split, as new mutations must arise for d_{XY} to increase, while F_{ST} only requires changes in allele frequency (42). Genomic regions of high F_{ST} are most noticeable in rearranged regions, but additional peaks of F_{ST} also were identified in collinear regions, such as position ~13.8 Mb on the X chromosome (Fig 3).

Ancient standing variation and chromosome differences fueled rapid ecotype formation

To provide insight into the genomic underpinnings of local adaptation, we identified candidate targets of differential selection or linked selection using a two-step process. We began with an empirical outlier approach to identify exceptionally diverged genomic regions, under the expectation that loci responsible for local adaptation should be more diverged than neutral loci. Considering separately the collinear and rearranged genomic partitions due to large differences in mean F_{ST} between them, we computed average

normalized F_{ST} in 10-kb windows and found a total of 44 outlier windows, or 26 outlier blocks when consecutive windows were combined (*SI Appendix*, Text, Table S9). Because neutral loci can vary widely in levels of differentiation for reasons other than selection, we also used Relate (39) to infer genome-wide genealogies [1,484,174 genealogies with a mean length of 133 bp (SD = 1,337 bp)] and to identify those under positive selection. To arrive at a list of candidate targets of differential selection, we found the intersection between the 44 outlier F_{ST} windows and the SNPs from 40,453 genealogies inferred to be under selection in K and/or F. At this intersection were 37 outlier F_{ST} windows arranged in 22 genomic blocks, spanning 202 genealogies involving at least one SNP inferred to be under differential selection (262 SNPs in total; *SI Appendix*, Text, Figs S7-S8, Tables S9-S10). As candidate SNPs from the same genealogy are linked, we enumerate genealogies as targets of selection rather than component SNPs, and for simplicity, refer to them as 'loci', except where noted. We predicted that if these candidate loci contribute to local adaptation, they should show no evidence of migration between K and F. To test this prediction, we used a supervised machine learning approach implemented in the program FILET (Finding Introgressed Loci using Extra Trees Classifiers) (43) to identify 10 kb windows along the genome with a high probability of no-migration (*SI Appendix*, Text). All 37 outlier F_{ST} windows were classified with high confidence as no-migration windows (*SI Appendix*, Tables S9, S11), consistent with our expectation.

Under the hypothesized scenario that K recently adapted to exploit a novel habitat for growth and development of its immature stages, we expect asymmetries between ecotypes in the number of candidate loci and the timing when selection was first inferred by Relate. In particular, we predict an excess of candidate loci, and a more recent onset of selection, in K versus F. Consistent with these predictions, we detect more candidate loci in K: 123 compared to only 79 in F (with 5 loci shared between ecotypes) (*SI Appendix*, Table S10). Furthermore, considering the set of individual SNPs inferred to be under differential or linked selection in K and F, the timing of first selection is more recent for K (Kruskal/Wallace p-value=0.00015). Shown in Fig 4 as overlapping histograms along the right-hand Y-axis, the mean time in years before present was 1,867 for K (0.05–0.95 quantile, 243–4,708) and 2,283 for F (1,192–4,708) (see also *SI Appendix*, Fig S9). These histograms suggest that K experienced more selection than F following their split ~1,300 years ago (indicated in Fig 4 as a horizontal black line against

the right-hand Y-axis). Indeed, 107 SNPs in K versus 15 in F were inferred to have come under selection since that time (see *Data Availability*).

The recency and rapidity of local adaptation to a new habitat in K is difficult to reconcile with new mutation, because fitness-relevant heritable variation is expected to be highly polygenic (44). To gain insight into the origin of selected variation in K and F, we focused on the same set of individual SNPs inferred to be under differential or linked selection, and estimated their frequency at the earliest detected onset of selection using Relate. We found that most selected variants (89% in K; 93% in F) were segregating at frequencies >5% at this time, implying that selection must have arisen from standing genetic variation (*SI Appendix*, Fig S10). To investigate further, we estimated the mutation age of each variant (*SI Appendix*, Text). Shown in Fig 4 as overlapping histograms along the upper X-axis, the mean mutation age in years before present was 37,058 for K (0.05–0.95 quantile, 9,240–61,993) and 40,687 for F (12,153–67,594). These mutation ages not only greatly predate the K-F split, they overlap the split of *An. funestus* from its sister species, *An. funestus-like* (~38,000 years ago; 95% CI, 31,000 – 45,000 years ago) (26). Taken together, Fig 4 and Fig S10 (*SI Appendix*) indicate that the candidate loci implicated in local adaptation were present as standing variation for substantial lengths of time and at intermediate frequencies prior to the onset of selection, factors that could have facilitated the rapid adaptation by K to a human-modified habitat (45).

Both the Relate method and our simulations under the chosen IIM model of K-F divergence suggest that there was migration between these ecotypes until the last ~100 years. We explored the genetic architecture of candidate loci in the K ecotype, to test the hypothesis that chromosomal inversions help maintain divergence through suppressed recombination despite gene flow. Although K and F are not distinguished by fixed differences between alternative chromosomal arrangements, three inversions on chromosome arms 2R and 3R segregate in F at high frequencies but are absent (or undetectable due to very low frequency) in K. These inversions (2Ra, 3Ra, and 3Rb) did not arise *de novo* near the time of the K-F split, as our estimates of inversion age suggest that all three are relatively ancient, well-predating that split and approaching or exceeding the age of the split between *An. funestus* and its sister species

An. funestus-like >30,000 years ago (*SI Appendix*, Text, Fig S11). Nevertheless, three observations implicate inversion differences in maintaining K-F divergence. First, whether considering the set of candidate loci or the set of individual candidate SNPs in K and F (*SI Appendix*, Table S10), the number that map to rearranged versus collinear genomic regions is significantly greater in K (Pearson Chi-square $p<0.00047$; *SI Appendix*, Fig S12). Second, considering K alone, 5,483 of 28,749 (19%) selected loci across the genome map to genomic regions rearranged between ecotypes (red dot in Fig 5A). Drawing 1000 random samples of equal size (5,483) from putatively neutral SNPs across the genome to create a 'null' distribution, the fraction of neutral SNPs in K mapping to rearranged regions did not overlap the fraction of candidate loci that did so from any one the draws (box plot in Fig 5A), suggesting that within K there is a significant enrichment of candidate loci inside regions that are rearranged between ecotypes. The third observation exploits the fact that the reduction in recombination in genomic regions rearranged between ecotypes is expected only between chromosomes with opposite orientations. Although inversions segregate at high frequency in the F ecotype, the standard configuration is present (*SI Appendix*, Table S1), and can recombine with the corresponding standard configuration in K, as is evident from Fig S13 (*SI Appendix*). We considered the set of 168 candidate SNPs under differential selection in K, and compared their frequencies in K to the frequency of the same SNPs in F, after partitioning F into two groups by karyotype: homozygous standard (denoted F_S) and all remaining F (*i.e.*, those carrying at least one chromosomal inversion, denoted F_I). In Fig 5B, we show that recombination suppression between F and K due to opposite inversion configurations is sufficient to preserve higher differentiation between K and F in rearranged genomic regions, despite the potential for gene flow between standard chromosomes in these regions.

The set of candidate SNPs under differential selection (or linked selection) in the ecotypes map to 25 annotated genes, of which 9 are implicated in both, 13 are unique to K, and 3 are unique to F (*SI Appendix*, Table S10). Highly incomplete functional annotation of the *An. funestus* FUMOZ reference renders functional inference particularly fraught. Using orthology to the better-annotated *Drosophila melanogaster* genes in Flybase (flybase.org) where possible, we attribute putative functions to 15 genes (*SI Appendix*, Table S10). In broad categories, these include pattern formation during development (e.g.,

Wnt and Egfr signaling), metabolism (gluconeogenesis, steroid and neurotransmitter biosynthesis), reproduction (male fertility, sperm storage), and nervous system function (phospholipase C-based and GPCR signaling). Of special note in the last category are two diacylglycerol kinase genes on chromosomes 2 and X. The X-linked gene, an ortholog of *D. melanogaster retinal degeneration A*, has been implicated in pleiotropic functions ranging from visual, auditory, and olfactory signal processing to starvation resistance and ecological adaptation in *Drosophila* and *An. gambiae* (46-49). As the ortholog of this gene in *An. gambiae* was recently identified as the target of a selective sweep in populations from Uganda (50), further investigation may be warranted despite the significant challenges of definitively linking candidate targets of natural selection to adaptive consequences (51).

Prospects for a molecular diagnostic tool for ecotype identification

The K and F ecotypes are morphologically indistinguishable, and we have shown here that chromosome inversion-based identification tools are imprecise at best. Both future evolutionary ecology studies, and vector surveillance for malaria epidemiology and control, will critically depend upon molecular diagnostics, which are lacking.

Motivated by this goal, we began by assembling a complete ribosomal DNA (rDNA) gene unit by leveraging the long single molecule (PacBio) database developed during construction of the *An. funestus* FUMOZ AfunF3 reference genome (32) (*SI Appendix*, Text). Noncoding spacer regions of rDNA have historically proven fruitful targets for identifying fixed SNP differences between closely related anopheline sibling species (52), owing to the uniquely rapid evolutionary dynamics of the tandemly arrayed rDNA genes. Despite near-complete success including a substantial amount of intergenic spacer upstream and downstream of the transcription unit (*SI Appendix*, Table S12; GenBank accession 'TBD'), alignment of genomic sequences from K and F to the rDNA assembly yielded no evident fixed differences (*SI Appendix*, Fig S14), consistent with their early stage of divergence.

As an alternative, we screened the genomic sequence data for fixed differences between K and F, using variation (VCF) files compiled separately for K and F samples. This resulted in a total of seven fixed SNP differences, all of which mapped to intronic positions of a single gene (the GPCR AFUN019981) on chromosome 2R, spanning ~15 kb (*SI Appendix*, Fig S15; for alignment, see *Data Availability*). The 5'-end of the predicted mRNA lies only ~2.5 kb outside the proximal breakpoint of inversion 2Ra, a location that is likely protected from gene flux (recombination and gene conversion) between the inverted and standard orientations, potentially explaining—at least in part—the heightened differentiation. As four of the fixed SNP differences between K and F are separated by only 600 bp, it appears feasible that a simple PCR amplicon strategy could be developed. How widely applicable this result might be remains to be investigated, and the inability to detect additional fixed differences across the accessible genome despite deep sequencing is remarkable in itself. It is noteworthy that the two putative K-F hybrids revealed by individual ancestry analysis (Fig 1B) are detectably heterozygous at four or five of the six diagnostic positions that could be unambiguously genotyped, consistent with contemporary K-F hybridization.

Conclusions

Ecotypic differentiation in a heterogeneous environment implies some form of resource partitioning. Despite their genetic divergence, K and F female adult stages share the same specialization on human blood meal hosts (30). Adults of both ecotypes can be found in strict sympatry host seeking and resting inside houses from the same village, although K is less likely than F to rest indoors after a blood meal (30). This behavioral difference is epidemiologically if not ecologically relevant, as an exophilic fraction of the malaria vector population is expected to reduce overall exposure to indoor-based vector control measures. Evidence for ecological divergence between the immature stages of K and F is circumstantial thus far, based on differences in the relative abundance of adult stages of K and F across breeding seasons, correlated with changes in temperature, rainfall, and land use at a local scale (29). In the better-studied Afrotropical *An. gambiae* complex, resource partitioning of the larval habitat has long been appreciated. The most striking ecological difference among *An. gambiae s.l.* taxa is the larval habitat, a

characteristic correlated with chromosome structural differences involving the central part of chromosome 2R in this group (13). It was the advent of molecular diagnostics for *An. gambiae s.l.* that enabled field studies of the larval adaptations and the ecological conditions that promote divergent selection in alternative larval habitats (53). This is precisely the next step for advancing the study of ecotypic differentiation in *An. funestus,* a step made more promising with our discovery of multiple fixed SNP differences over a short stretch of AFUN019981. The relative neglect of *An. funestus* in contemporary vector research—understandable in terms of the more modest genomic tools and resources available to date—is disproportional to its dominant role in malaria transmission on the African continent. Importantly, our study reveals a powerful advantage of the *An. funestus* K-F system for studying the genomic and phenotypic basis of local adaptation, because the remarkable recency and microgeographic proximity of the process avoids the accumulated genetic divergence that can obscure directly relevant differences in older and more dispersed systems.

We have shown here that standing genetic variation within *An. funestus* provides a basis for adaptation, adding to a growing literature from diverse organisms (45, 54, 55). Of particular interest is the role of established chromosomal inversions in this process. Both the ancestral *An. funesus* lineage, and F itself, carry several relatively old balanced inversion polymorphisms. The inverted orientations of rearrangements on chromosome arms 2R and 3R in F (2Ra, 3Ra, 3Rb) segregate at relatively high frequencies, while K is monomorphic for the standard orientations ($2R+^{a}$, $3R+^{a}$, $3R+^{b}$). If irrigated rice fields can be considered ecologically marginal environments for a species that normally exploits natural swamps, this pattern echoes a phenomenon observed for nearly all chromosomally polymorphic species of *Drosophila* studied, in which inversion polymorphism declines in marginal environments (56). Chromosomally monomorphic marginal populations reflect the sieving of ancestral balanced polymorphism during local adaptation, either through random genetic drift or by selection favoring the fixation of the standard orientation (56, 57). Genetic models of adaptation in peripheral populations suggest that directional selection may be the more likely force (58). In the 1950s, Carson proposed the homoselection-heteroselection hypothesis to explain the patterns seen in chromosomally polymorphic *Drosophila* species, whereby central populations in optimal habitats are favored by heterotic structural

heterozygosity while marginal populations in sub-optimal environments are exposed to strong directional selection and are favored by structural homozygosity, allowing for maximal recombination (59). This hypothesis has been empirically supported in *Drosophila* (60, 61). The notion that structural homozygosity favors ecotypic formation by facilitating recombination is not at odds with the models of Coluzzi and others (10, 11), whose focus is on the role of *de novo* inversions in preventing recombination between the emerging ecotype and ancestral populations with different genetic backgrounds; these are two sides of the same coin. Indeed, in the K-F system, it appears that the relatively high frequencies of three established inversions in F populations, and their absence in sympatric K populations, facilitated ecotype formation by suppressing inter-ecotype gene flow and homogenization inside these chromosomal rearrangements.

Materials and Methods

Please see *SI Appendix* Text for detailed materials and methods.

Acknowledgments

We thank Claudia Witzig, and previous members of the Emrich lab for their numerous intellectual and analytical contributions to an earlier conception of this project based on population pool sequencing and an unanchored genome assembly. We thank Matt Hahn, members of the Besansky lab, and members of the Kern-Ralph Co-lab for discussion, Leo Speidel for help with implementing *Relate*, and developers of *tskit* and *tsdate* for discussion and coding suggestions. We gratefully acknowledge Jeff Powell for pointing out the homoselection hypothesis of Carson.

Funding

This work was supported by a grant from the Bill & Melinda Gates Foundation and Open Philanthropy.

Author contributions

Conceptualization and design: S.T.S, N.J.B.; Planning, conduct, supervision of field collections and contribution of study materials: C.C., N'F.S., M.W.G.; Analysis and visualization: S.T.S.; Contribution to analysis: M.C.F., S.J.E.; Funding and supervision: N.J.B., A.D.K.; Wrote the paper: S.T.S., N.J.B.; Edited and reviewed the paper: all authors.

Competing interests

The authors declare no competing interests.

Data and materials availability

Genomic resources used in this study can be downloaded from VectorBase: (i) *An. funestus* reference assembly, www.vectorbase.org/common/downloads/Legacy%20VectorBase%20Files/Anopheles-funestus/Anopheles-funestus-FUMOZ_CHROMOSOMES_AfunF3.fa.gz; (ii) repeat annotations used for masking, www.vectorbase.org/common/downloads/Legacy%20VectorBase%20Files/Anopheles-funestus/Anopheles-funestus-FUMOZ_REPEATFEATURES_AfunF3.gff3.gz; (iii) gene annotations, www.vectorbase.org/common/downloads/Legacy%20VectorBase%20Files/Anopheles-funestus/Anopheles-funestus-FUMOZ_BASEFEATURES_AfunF3.1.gff3.gz. All sequencing data are available as aligned bam-formatted files in NCBI Sequence Read Archives under accessions listed in Table S1. Custom scripts used for analyses are available at GitHub: www.github.com/stsmall/abc_scripts2 and www.github.com/stsmall/Kiribina_Folonzo. The latter link also contains a Jupyter Notebook (www.jupyter.org/) that can be used to recreate figures in the manuscript. Data files are available in the Figshare repository, including: (i) PCA-associated Zarr files for use in scikit-allel (10.6084/m9.figshare.21585543); (ii) admixture proportions (10.6084/m9.figshare.21585555); (iii) diversity and divergence (10.6084/m9.figshare.21585561); (iv) recombination (10.6084/m9.figshare.21585537); (v) input file for Stairway Plot 2 (10.6084/m9.figshare.21585540); (vi) FILET analyses (10.6084.m9.figshare.21585741); (vii) selection analyses (10.6084/m9.figshare.21585534); (viii) rDNA and GPCR alignments (10.6084/m9.figshare.21585573).

References

1. Kawecki Tj & Ebert D (2004) Conceptual issues in local adaptation. *Ecol. Lett.* 7:1225-1241.
2. Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. *Am Nat* 173:579-588.
3. Hall JE, Lawrence ES, Simonson TS, & Fox K (2020) Seq-ing Higher Ground: Functional Investigation of Adaptive Variation Associated With High-Altitude Adaptation. *Front Genet* 11:471.
4. Jones FC*, et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484:55-61.
5. Van't Hof AE*, et al.* (2016) The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534:102-105.
6. Linnen CR*, et al.* (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339:1312-1316.
7. Savolainen O, Lascoux M, & Merila J (2013) Ecological genomics of local adaptation. *Nat Rev Genet* 14:807-820.

8. Tiffin P & Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* 29:673-680.
9. Neafsey DE*, et al.* (2015) Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* 347:1258522.
10. Coluzzi M (1982) Spatial distribution of chromosomal inversions and speciation in anopheline mosquitoes. *Mechanisms of Speciation*, ed Barigozzi C (Alan R. Liss, Inc., New York), pp 143-153.
11. Kirkpatrick M & Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics* 173:419-434.
12. Navarro A & Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution* 57:447-459.
13. 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.
14. Toure YT*, et al.* (1998) The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia* 40:477-511.
15. della Torre A*, et al.* (2001) Molecular evidence of incipient speciation within *Anopheles gambiae s.s.* in West Africa. *Insect Mol. Biol.* 10:9-18.
16. Coetzee M*, et al.* (2013) *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa* 3619:246–274.
17. Lawniczak MK*, et al.* (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330:512-514.
18. Neafsey DE*, et al.* (2010) SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science* 330:514-517.
19. Love RR*, et al.* (2016) Chromosomal inversions and ecotypic differentiation in Anopheles gambiae: the perspective from whole-genome sequencing. *Mol. Ecol.* 25:5889-5906.
20. della Torre A, Tu ZJ, & Petrarca V (2005) On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms. *Insect Biochem. Mol. Biol.* 35:755-769.
21. Costantini C*, 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.
22. Simard F*, et al.* (2009) Ecological niche partitioning between the M and S molecular forms of *Anopheles gambiae* in Cameroon: the ecological side of speciation. *BMC Ecol.* 9:17.
23. Coetzee M & Koekemoer LL (2013) Molecular systematics and insecticide resistance in the major African malaria vector *Anopheles funestus*. *Annu. Rev. Entomol.* 58:393-412.
24. Dia I, Guelbeogo MW, & Ayala D (2013) Advances and perspectives in the study of the malaria mosquito *Anopheles funestus*. *Anopheles mosquitoes - New insights into malaria vectors*, ed Manguin S (InTech, DOI: 10.5772/55389).
25. Michel AP*, et al.* (2005) Rangewide population genetic structure of the African malaria vector *Anopheles funestus*. *Mol. Ecol.* 14:4235-4248.
26. Small ST*, et al.* (2020) Radiation with reticulation marks the origin of a major malaria vector. *Proc. Natl. Acad. Sci. U. S. A.* 117:31583–31590.
27. Costantini C, Sagnon NF, Ilboudo-Sanogo E, Coluzzi M, & Boccolini D (1999) Chromosomal and bionomic heterogeneities suggest incipient speciation in *Anopheles funestus* from Burkina Faso. *Parassitologia* 41:595-611.
28. Guelbeogo WM*, et al.* (2005) Chromosomal evidence of incipient speciation in the Afrotropical malaria mosquito *Anopheles funestus*. *Med. Vet. Entomol.* 19:458-469.
29. Guelbeogo WM*, et al.* (2009) Seasonal distribution of *Anopheles funestus* chromosomal forms from Burkina Faso. *Malar. J.* 8:239.

30. Guelbeogo WM, Sagnon N, Liu F, Besansky NJ, & Costantini C (2014) Behavioural divergence of sympatric *Anopheles funestus* populations in Burkina Faso. *Malar. J.* 13:65.
31. Michel AP*, et al.* (2005) Molecular differentiation between chromosomally defined incipient species of *Anopheles funestus*. *Insect Mol. Biol.* 14:375-387.
32. Ghurye J*, et al.* (2019) A chromosome-scale assembly of the major African malaria vector *Anopheles funestus*. *GigaScience* 8:1-8.
33. Cubry P*, et al.* (2018) The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Curr. Biol.* 28:2274-2282 e2276.
34. Cabreros I & Storey JD (2019) A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. *Genetics* 212:1009-1029.
35. Beaumont MA, Zhang W, & Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025-2035.
36. Wang Z*, et al.* (2021) Automatic inference of demographic parameters using generative adversarial networks. *Mol. Ecol. Resour.* 21:2689-2705.
37. Kropelin S*, et al.* (2008) Climate-driven ecosystem succession in the Sahara: the past 6000 years. *Science* 320:765-768.
38. Choi JY*, et al.* (2019) The complex geography of domestication of the African rice Oryza glaberrima. *PLoS Genet* 15:e1007414.
39. Speidel L, Forest M, Shi S, & Myers SR (2019) A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* 51:1321-1329.
40. Presgraves DC (2008) Sex chromosomes and speciation in Drosophila. *Trends Genet.* 24:336-343.
41. Lasne C, Sgro CM, & Connallon T (2017) The Relative Contributions of the X Chromosome and Autosomes to Local Adaptation. *Genetics* 205:1285-1304.
42. Cruickshank TE & Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23:3133-3157.
43. Schrider DR, Ayroles J, Matute DR, & Kern AD (2018) Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS genetics* 14:e1007341.
44. Pritchard JK, Pickrell JK, & Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20:R208-215.
45. Barrett RD & Schluter D (2008) Adaptation from standing genetic variation. *Trends Ecol. Evol.* 23:38-44.
46. Kain P*, et al.* (2008) Reduced odor responses from antennal neurons of G(q)alpha, phospholipase Cbeta, and rdgA mutants in Drosophila support a role for a phospholipid intermediate in insect olfactory transduction. *J. Neurosci.* 28:4745-4755.
47. Senthilan PR*, et al.* (2012) Drosophila auditory organ genes and genetic hearing defects. *Cell* 150:1042-1054.
48. Nelson CS*, et al.* (2016) Cross-phenotype association tests uncover genes mediating nutrient response in Drosophila. *BMC Genomics* 17:867.
49. Cheng C, Tan JC, Hahn MW, & Besansky NJ (2018) A systems genetic analysis of inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U. S. A.* 115:E7005-E7014.
50. Bergey CM*, et al.* (2020) Assessing connectivity despite high diversity in island populations of a malaria mosquito. *Evol Appl* 13:417-431.
51. Barrett RD & Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* 12:767-780.
52. Collins FH & Paskewitz SM (1996) A review of the use of ribosomal DNA (rDNA) to differentiate among cryptic Anopheles species. *Insect Mol. Biol.* 5:1-9.

53. Lehmann T & Diabate A (2008) The molecular forms of *Anopheles gambiae*: a phenotypic perspective. *Infect. Genet. Evol.* 8:737-746.
54. Schrider DR & Kern AD (2017) Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Mol. Biol. Evol.* 34:1863-1877.
55. Louis M*, et al.* (2021) Selection on ancestral genetic variation fuels repeated ecotype formation in bottlenose dolphins. *Sci Adv* 7:eabg1245.
56. Hardie DC & Hutchings JA (2010) Evolutionary ecology at the extremes of species' ranges. *Environmental Reviews* 18:1-20.
57. Guerrero RF & Hahn MW (2017) Speciation as a sieve for ancestral polymorphism. *Mol. Ecol.* 26:5362-5368.
58. Garcia-Ramos G & Kirkpatrick M (1997) Genetic Models of Adaptation and Gene Flow in Peripheral Populations. *Evolution* 51:21-28.
59. Carson HL (1956) Marginal Homozygosity for Gene Arrangement in Drosophila robusta. *Science* 123:630-631.
60. Carson HL (1958) Response to selection under different conditions of recombination in Drosophila. *Cold Spring Harb. Symp. Quant. Biol.* 23:291-306.
61. Tabachnick WJ & Powell JR (1977) Adaptive Flexibility of "Marginal" Versus "Central" Populations of Drosophila Willistoni. *Evolution* 31:692-694.

Figures and Tables

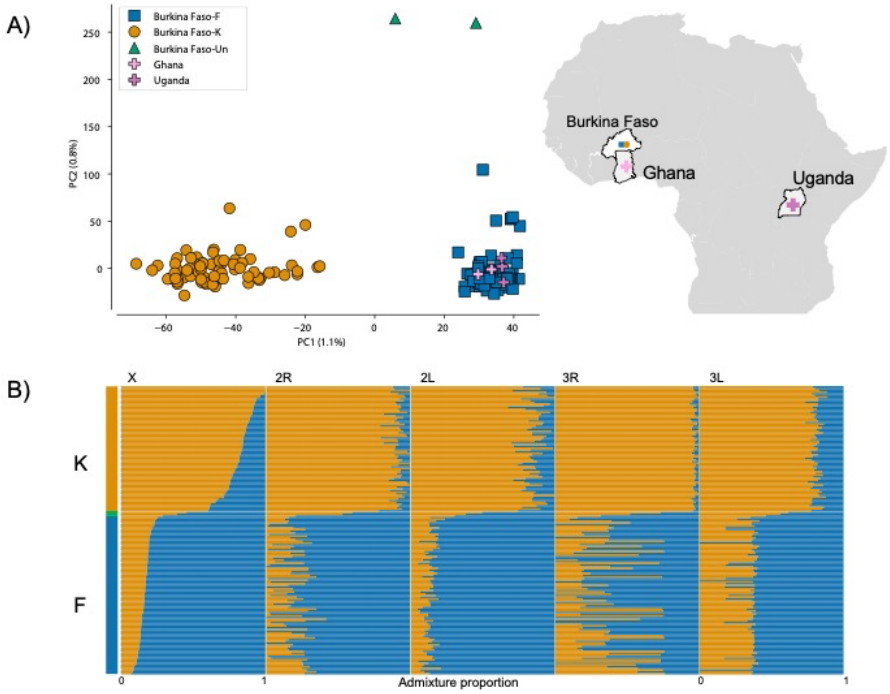

Figure 1. **Genetic structure of *An. funestus* mosquitoes sampled from Burkina Faso**. A) Plot of the first two principal components from the X chromosome PCA that included *An. funestus* reference samples from Ghana and Uganda. Orange circles and blue squares correspond to Burkina Faso genotypes assigned to ecotypes K and F, respectively. Green triangles represent two unclustered Burkina Faso genotypes that were unassigned ('Burkina Faso-Un'). B) Estimated individual ancestry proportions for Burkina Faso *An. funestus*. Y-axis color bar at left reflects individual assignments based on the X chromosome PCA (K, orange; F, blue; unassigned, green). Individuals are represented as thin horizontal lines, grouped by assignment and displayed across the plot in descending order of ancestry proportion to K for the X chromosome. Different chromosome arms (X, 2R, 2L, 3R, 3L), designated by columns across the top of the plot, are separated by white lines.

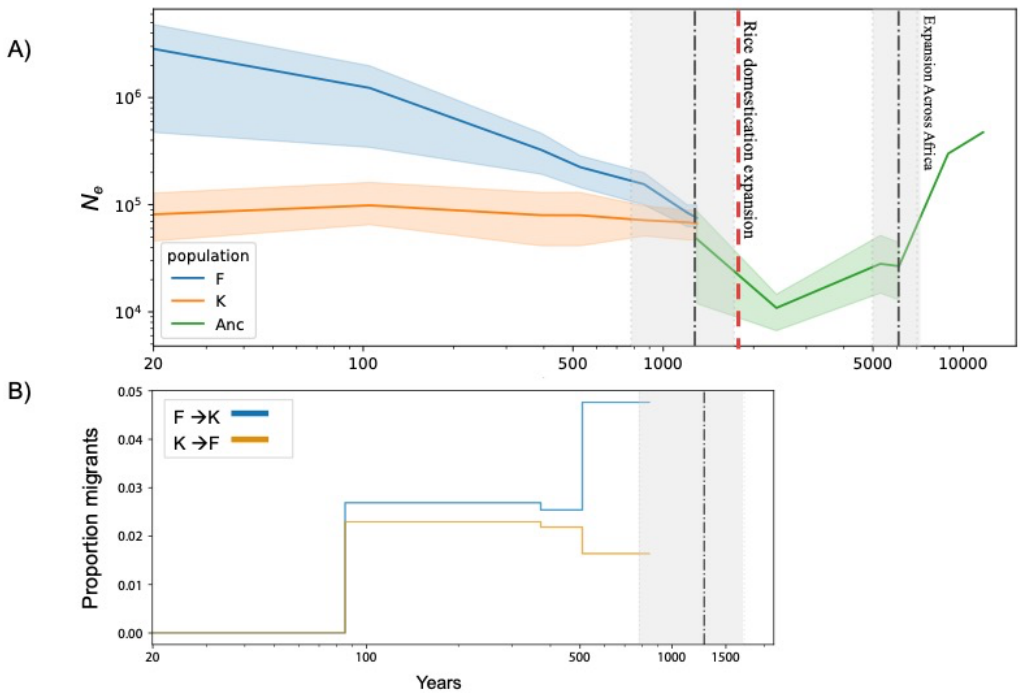

Figure 2. Demographic history inferred with ABC for *An. funestus* ecotypes K and F. A) Effective population size (N_e) trajectories and population split times in years before present, assuming 11 generations/year. Median N_e values and 95% credible intervals are plotted as colored lines and corresponding shading. Median split times and 95% credible intervals are represented as dash-dotted black vertical lines with gray shading. The dashed red vertical line represents the proposed time (~1,850 years) of a strong domesticated rice expansion in the Niger River delta (33). B) K-F migration rates through time in years before present, subsequent to their split (dash-dotted black vertical line). The proportion of migrants declined to 0 in both directions at ~100 years.

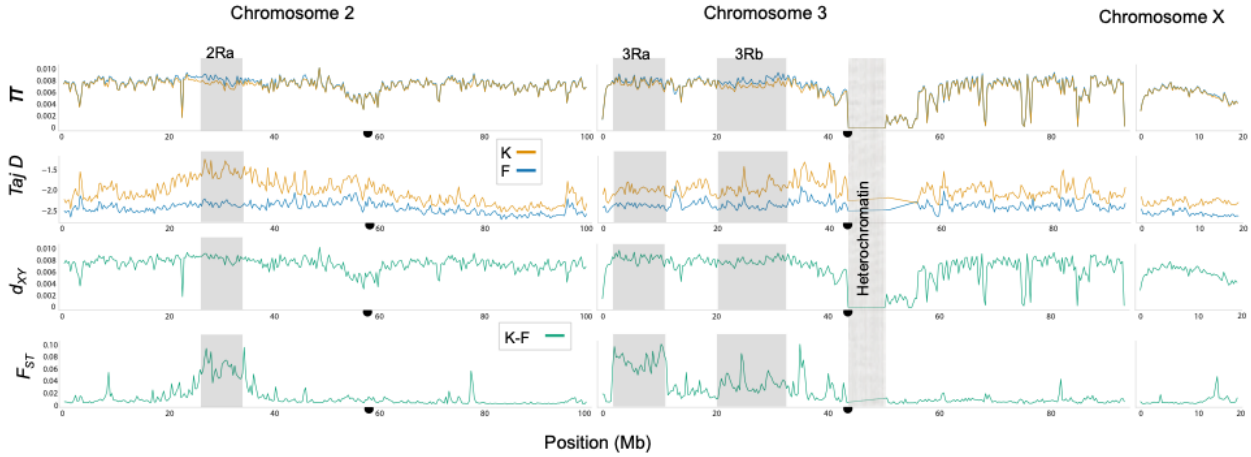

Figure 3. Whole-genome scans of diversity and divergence for K and F. Each statistic was calculated in 10-kb non-overlapping windows and smoothed with a moving average over 10 windows for plotting. From top to bottom, panels display mean values of nucleotide diversity (π), Tajima's *D* (Taj D), absolute divergence (d_{XY}), and differentiation (F_{ST}) at genomic positions (in millions of bases, Mb) along chromosomes 2, 3, and X. Rearranged and heterochromatic (no data) regions are shaded and labeled. Centromeres are indicated as black half circles.

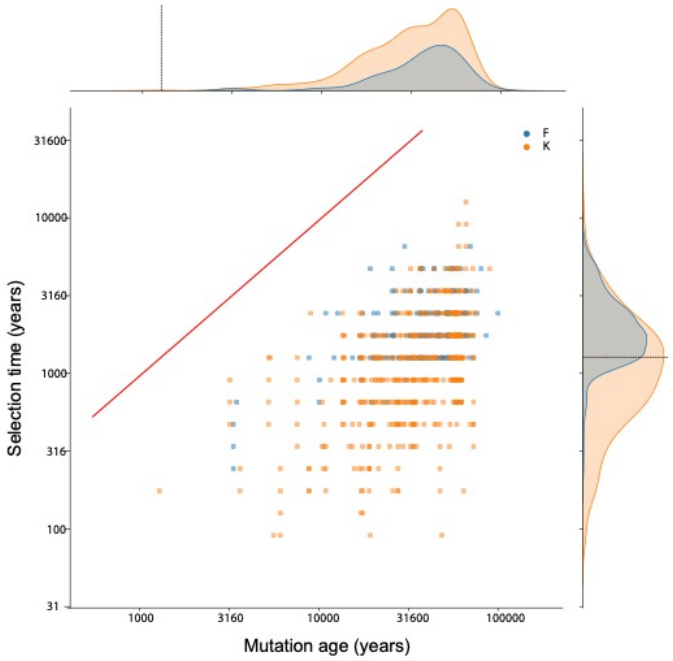

Figure 4. Age of mutation versus time at onset of selection. In the central panel, SNPs under selection in K (orange) or F (blue) are plotted as dots according to their estimated age (X axis, in years) against the inferred time of first selection (Y axis, in years), as inferred by *Relate* (39). The red line, included for reference, indicates a hypothetical 1:1 correspondence between mutation age and onset of selection. Time in years assumes a generation time of 11 per year. Against the top-most X axis and the right-hand Y axis and are the corresponding densities of mutation ages and selection times, respectively. Black lines on the density plots mark the split time of K and F.

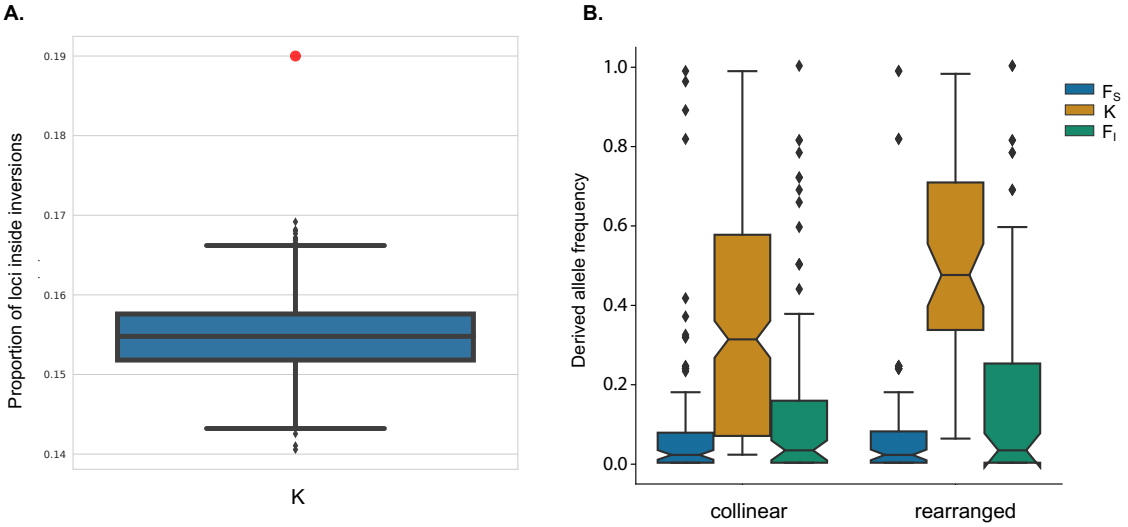

Figure 5. Selected loci in K are enriched within and protected by chromosomal rearrangements. A) Proportion of SNPs inside inversions. The red dot represents the observed fraction of candidate loci in the K genome that map inside rearranged regions. The boxplot beneath represents the percentage in 1,000 random sample sets of the same number of putatively neutral SNPs. B) Boxplot of allele frequencies for individual candidate SNPs in ecotype K, and their corresponding frequencies in F, in collinear versus rearranged genomic compartments. Ecotype F was partitioned by genotype, where F_S represents mosquitoes carrying the homozygous standard karyotype with respect to all three focal inversions (2Ra, 3Ra, 3Rb) and F_I represents all other karyotypic combinations. Boxplots whose notches do not overlap another boxplot are considered significantly different.