Locus- and Population-Specific Selection and Differentiation between Incipient Species of *Anopheles gambiae*

Thomas L. Turner and Matthew W. Hahn†*

*Center for Population Biology, University of California, Davis; and †Department of Biology and School of Informatics, Indiana University

Anopheles gambiae, the primary mosquito vector of malaria in sub-Saharan Africa, is divided into 2 sympatric incipient species known as *M* form and *S* form. Recent genomic analysis of each form revealed that differentiation between forms is clustered into 3 unlinked regions of the genome. Here, we expand the investigation of these "genomic islands of speciation" to multiple populations, including all of the genes across one of the islands. Differentiation between the *M* and *S* forms in 2 of the islands is complete across all individuals in all populations, confirming that the *M* and *S* forms are reproductively isolated taxa. Differentiation at the third island (on chromosome 2R) is limited to Cameroon populations. There is reduced variation in the *M* form in Cameroon at this location and increased divergence to the outgroup *Anopheles arabiensis*, supporting an association of adaptation with reproductive isolation.

Introduction

Patterns of molecular variation within populations are shaped by many processes, including genetic drift, natural selection, demographic history, and migration. Although selective forces vary greatly among loci, forces such as drift, demography, and migration are generally considered to act across the genome (Hudson et al. 1987; Begun and Whitley 2000; Glinka et al. 2003; Akey et al. 2004; Stajich and Hahn 2005). However, when migration occurs between incipient species, it may interact with selection to create different amounts of realized gene flow among loci (Machado et al. 2002; Emelianov et al. 2004; Payseur et al. 2004). This pattern can occur when hybrid genotypes at some loci are disadvantageous, whereas alleles at other loci are selectively neutral with respect to the different species or populations (Barton and Gale 1993; Wu and Ting 2004).

Previously, we found such a pattern of heterogeneous gene flow between 2 partially isolated, sympatric populations of *Anopheles gambiae* (Turner et al. 2005). Using Affymetrix microarrays, we were able to characterize differentiation across the genomes of the so-called *M* and *S* forms of *A. gambiae* in Cameroon at over 142,000 markers. We identified 3 unlinked regions that showed no evidence for gene flow (speciation islands), even as the other 99% of the genome showed little differentiation between the *M* and *S* forms (Turner et al. 2005). These differentiated regions are therefore expected to contain loci which are disadvantageous in a hybrid genetic background but advantageous within each form. However, we do not know which genes are the targets of selection or whether the alleles underlying these selective effects are found throughout the ranges of each form. Because *A. gambiae* is the primary vector of human malaria, understanding how *M* and *S* are different is also relevant to human health. These forms mate assortatively in nature (Tripet et al. 2001), implicating behavioral differences between them, but other phenotypic and ecological differences have not been conclusively demonstrated. Finding the genes that differentiate *M* and *S* may aid in this investigation by generating testable hypotheses from the bottom up.

Therefore, in this paper, we present an analysis of sequence variation and divergence between the *M* and *S* forms from Cameroon and Mali. We have also included individuals with and without the 2Rb inversion: this inversion is polymorphic in both *M* and *S* and spans one of the candidate speciation islands in Cameroon (Coluzzi et al. 2002). We have 2 main aims: 1) to investigate whether loci within the speciation islands show signatures of positive selection which could be related to speciation and 2) to determine whether the speciation islands identified previously in Cameroon are also present in Mali populations.

Materials and Methods

Anopheles gambiae samples were collected in the towns of Buea, Mutanguene, and Tiko in Cameroon in 2003 and in Selenkenyi in Mali in 2004 by the lab of G. C. Lanzaro (University of California, Davis), who graciously shared DNA from karyotyped mosquitoes for this study. Samples of *A. gambiae M* and *S* from Cameroon all possess the +/+ (standard) karyotype on chromosome 2R; Mali 2R karyotypes used in this study are as follows: Mopti (*M* form) individuals are *bc*/*bc*, *u*/+, and *bc*/+; Savanna (*S* form) individuals are *cu*/*cu* and *b*/*b*; Bamako (*S* form) individuals are *jcu*/*jcu*, *jbcu*/*jcu*, and *jbcu*/*jbcu* (following Touré et al. 1998). Samples of the congeneric species *Anopheles arabiensis* are from the SENN DDT colony, and *Anopheles quadriannulatus sp. A* are from the SKUQUA colony; both colonies are maintained in South Africa by M. Coetzee. These outgroup individuals were not karyotyped, but M. Coetzee (personal communication) reports that they both possess the "standard" karyotype for their species. DNA was extracted from *A. gambiae* following Post et al. (1993) and from *A. arabiensis* and *A. quadriannulatus sp. A* (hereafter *A. quadriannulatus*)using a Qiagen DNeasy kit; standard polymerase chain reaction (PCR) diagnostics were used to differentiate *A. gambiae* from *A. arabiensis* (Scott et al. 1993) and to differentiate *M* and *S* forms (Fanello et al. 2003). Polytene chromosome preparation and analysis of karyotypes (conducted in the lab of G. C. Lanzaro) are as described in Hunt (1973). Primers for PCR and sequencing were designed from the *A. gambiae* genome annotation (http://www.ensembl.org/Anopheles_gambiae/index.html) and are available in

Key words: speciation, mosquito, population genetics, natural selection.

E-mail: tlturner@ucdavis.edu.

Mol. Biol. Evol. 24(9):2132–2138. 2007
doi:10.1093/molbev/msm143
Advance Access publication July 17, 2007

supplementary table S1 (Supplementary Material online), which contains detailed information about each locus.

For 6 loci, on chromosome 2L (*LIM* and *Subtilase*) and 1 on chromosome 2R (*GPRor38*), sequence data from Cameroon populations of *A. gambiae* come from the previous study of Turner et al. (2005). We sequenced 3 of these loci and 1 additional locus (*MSH5*) from individuals in Mali as well as in the related species *A. arabiensis* and *A. quadriannulatus*. We also sequenced 5 additional loci on chromosome 2R, in populations from Cameroon. All new sequences have been deposited in GenBank under accession numbers DQ425111–DQ425208, DQ436826–DQ436911, and EF426141-EF426244. PCR products were directly sequenced in both directions for all loci except *NADH1b*, *pepM19*, *bkinase*, and *cd59*, which were cloned using an Invitrogen TOPO TA kit. Sequence chromatographs were assembled and edited in CodonCode (www.codoncode.com), which uses ABI quality scores and Phred/Phrap to call bases and find heterozygous single-nucleotide polymorphisms (SNPs). Sequences were aligned using ClustalW (Higgins et al. 1994). Upon further analysis, the locus *UNK1* reported in Turner et al. (2005) proved to be suspect due to close paralogous sequences in the *A. gambiae* genome, so this locus was not used in the current analysis.

Calculation of measures of population variation including Tajima's D, a measure of the skew of the frequency spectrum of mutations (Tajima 1989); π, the average number of differences between sequences (Tajima 1983); and F_{ST}, a measure of population differentiation (Wright 1951) were done using DNAsp v 4.00.5 (Rozas et al. 2003). All values of F_{ST} less than 0 were set to 0 in the text. Calculation of nonsynonymous and synonymous polymorphism and divergence—in order to perform the test of McDonald and Kreitman (1991)—and average levels of divergence, D_{xy}, were also done using DNAsp. HKA tests (Hudson et al. 1987) were performed using Jody Hey's HKA program available at http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm. Significance of HKA tests was determined by conducting 10,000 coalescent simulations as implemented in the HKA software (Kliman et al. 2000). These coalescent simulations were also used to assess the significance of Tajima's D values (Kliman et al. 2000).

Results

Summary statistics of levels of polymorphism and differentiation are shown in table 1 (for loci on 2L) and table 2 (for loci on 2R). Individuals of the *S* form from Mali include both the Bamako and Savanna "chromosomal forms" (sensu Touré et al. 1998), which are combined as "*S* form" in table 1; all *M* form individuals sequenced from Mali are the Mopti chromosomal form. Differentiation between *M* and *S* along the X chromosome has been well documented (Wang et al. 2001; Stump et al. 2005), so X-chromosome loci were not sequenced from Mali. Because all individuals are identified as *M* or *S* based on a PCR diagnostic which queries this region, however, the X-chromosome genotype is known. Loci on 2R and 2L were chosen to compare sequence variation within and outside the differentiated regions originally found in the Cameroon population.

Chromosome 2L

Data from 2 loci on chromosome 2L were sequenced in Mali: *LIM* was previously found to lie within the region of no gene flow, and *Subtilase* serves as a nearby control locus. The pattern of differentiation in Mali and Cameroon is the same, with fixed differences and no shared polymorphisms between *M* and *S* at the *LIM* locus, and shared polymorphisms without fixed differences at the *Subtilase* locus (table 1). The 4 SNPs fixed between forms in Cameroon are fixed between forms in all individuals sequenced regardless of geography or chromosomal form, indicating perfect correspondence of these SNPs with the *M* and *S* form diagnostic region on the X chromosome. The difference in number of fixed and shared polymorphisms between the differentiated and undifferentiated regions is highly significant in both Cameroon and Mali (Fisher's exact test; Cameroon: $P = 0.0010$; Mali: $P = 0.0014$). Although this difference could, in principle, be due to lineage sorting between loci after the cessation of gene flow, the inference of locus-specific selection against gene flow within the 2L speciation island is supported by the observation of continued hybridization in Mali (Tripet et al. 2001) and by the coalescent simulations of Turner et al. (2005). As expected from the pattern of fixed and shared polymorphisms, differentiation is high between *M* and *S* at the *LIM* locus ($F_{ST} = 0.950$ in Mali, 0.909 in Cameroon) and much lower at the *Subtilase* locus ($F_{ST} = 0.110$ in both populations). Differentiation within each form between countries is low (table 3), with F_{ST} values ranging from 0.091 to 0 at both loci on chromosome 2L.

Levels of polymorphism are also similar between Cameroon and Mali: the *LIM* locus shows 10–12 times lower values of π than the *Subtilase* locus in both populations (table 1). The much lower levels of polymorphism within nonintrogressing regions are one possible indication of ongoing natural selection against hybrid individuals at a linked locus. Another potential indication of linked selection is the significantly negative values of Tajima's D at the *LIM* locus in the Cameroon population of the *S* form (-1.83), which is the only locus in the differentiated region with more than 2 polymorphisms in either population (we did not calculate Tajima's D for loci with 2 or fewer polymorphisms). To further investigate possible patterns of natural selection, we sequenced *LIM* and *Subtilase* in 2 additional species of the *A. gambiae* sensu lato complex: *A. arabiensis* and *A. quadriannulatus*. Using the outgroup sequences, we find that of the 4 nt fixed between *M* and *S*, 2 are derived in *M*, 1 is derived in *S*, and 1 is ambiguous (*A. gambiae M* matches the *A. quadriannulatus* genotype and *A. gambiae S* matches *A. arabiensis*). We compared polymorphism and divergence at the *LIM* and *Subtilase* loci using the HKA test (Hudson et al. 1987). This test would have low power in a comparison of the *M* and *S* forms because of extremely low divergence between forms (Ford and Aquadro 1996); instead, we compared each form with the closest outgroup, *A. arabiensis*. The test was significant for both the *M* form ($P = 0.021$) and the *S* form ($P = 0.032$). It should be noted, however, that the proximity of *LIM* locus to the centromere—where there is likely reduced recombination (Turner et al. 2005)—could result in

Table 1
DNA Variation and Differentiation in 2 Loci on Chromosome 2L

Location	Gene	Population	Form	n[a]	l[b]	S[c]	π[d]	S_{priv}[e]	Fixed:Shared[f]	F_{ST}	Tajima's D
2L 0.75	*LIM*	Cameroon[g]	*M*	12	508	1	0.001	1	4:0	0.909	—
			S	16	508	4	0.001	4			−1.831*, **
		Mali	*M*	8	433	0	0	0	4:0	0.950	—
			S	16	433	1	0.001	1			—
2L 4.18	*Subtilase*	Cameroon[g]	*M*	26	429	11	0.010	1	0:10	0.110	1.615
			S	26	429	25	0.012	15			−0.760
		Mali	*M*	10	429	12	0.012	3	0:9	0.110	0.878
			S	18	429	15	0.012	6			0.834

NOTE.—*$P < 0.05$, **$P < 0.05$ based on coalescent simulation (not reported for loci with fewer than 3 polymorphisms).
[a] Number of chromosomes sequence.
[b] Length of sequenced region in base pairs.
[c] Number of polymorphisms.
[d] Heterozygosity.
[e] Number of private polymorphisms.
[f] Ratio of fixed differences to shared polymorphisms.
[g] Data from Turner et al. (2005).

reduced polymorphism and a skew in the frequency spectrum due to linked selection on genes unrelated to reproductive isolation (Maynard Smith and Haigh 1974; Kaplan et al. 1989).

Chromosome 2R

In contrast to the highly differentiated 2L and X islands, the region identified on chromosome 2R was only marginally significant in the previous whole-genome microarray analysis (Turner et al. 2005). A differentiated region was predicted on chromosome 2R between megabase 24.847 and megabase 24.884—a 37-kb span—but confidence intervals on this estimate are unknown.

To precisely define this region, we sequenced 10 loci across the boundaries of the previously defined island of speciation in populations from Cameroon (table 2). The sequenced loci include portions of all annotated genes

Table 2
DNA Variation and Differentiation in 10 Loci on Chromosome 2R

Location		Gene	Population	Form	n[a]	l[b]	S[c]	π[d]	S_{priv}[e]	Fixed:Shared[f]	F_{ST}	Tajima's D
2R	24.62	*NADH1b*	Cameroon	*M*	7	696	37	0.021	7	0:30	0.138	−0.238
				S	13	696	61	0.026	33			−0.545
2R	24.77	*pepM19*	Cameroon	*M*	9	743	49	0.025	33	0:18	0.067	−0.510
				S	6	743	30	0.020	13			−0.020
2R	24.81	*GPRgr13*	Cameroon[g]	*M*	20	231	5	0.004	3	0:2	0.504	−0.946
				S	10	231	8	0.012	6			−0.110
2R	24.85	*GPRor39*	Cameroon[g]	*M*	24	419	9	0.005	1	0:8	0.110	0.387
				S	22	419	16	0.100	8			−0.204
2R	24.86	*GPRor38*	Cameroon[g]	*M*	24	516	12	0.003	12	0:0	0.712	−1.620**
				S	24	516	28	0.012	28			−0.590
			Mali	*M*	12	322	9	0.009	5	0:4	0.133	−0.241
				S	22	322	17	0.013	8			1.001
2R	24.89	*i8*	Cameroon	*M*	12	430	3	0.002	3	2:0	0.643	−0.379
				S	18	430	10	0.007	10			0.228
2R	24.89	*MSH5*	Cameroon	*M*	28	414	1	0.001	1	3:0	0.626	1.36
				S	22	414	25	0.013	25			−0.857
			Mali	*M*	12	428	23	0.02	13	0:11	0.109	0.323
				S	20	428	22	0.011	12			−1.13
2R	24.91	*cd59*	Cameroon	*M*	9	981	43	0.017	31	0:10	0.248	0.11
				S	9	981	49	0.016	38			−0.714
2R	24.97	*bkinase*	Cameroon	*M*	10	511	17	0.013	11	0:6	0.413	0.457
				S	10	511	16	0.009	10			−0.81
2R	25.20	*FAC3C*	Cameroon[g]	*M*	24	464	7	0.005	5	0:2	0.030	−0.268
				S	20	464	8	0.004	6			−0.264

NOTE.—*$P < 0.05$, **$P < 0.05$ based on coalescent simulation (not reported for loci with fewer than 3 polymorphisms).
[a] Number of chromosomes sequence.
[b] Length of sequenced region in base pair.
[c] Number of polymorphisms.
[d] Heterozygosity.
[e] Number of private polymorphisms.
[f] Ratio of fixed differences to shared polymorphisms.
[g] Data from Turner et al. (2005).

Table 3
Differentiation between Cameroon and Mali

	F_{ST} between Mali and Cameroon	
	M	*S*
LIM (2L)	0.091	0.022
Subtilase (2L)	0	0
GPRor38 (2R)	0.199	0.136
MSH5 (2R)	0.271	0.066

between megabase 24.850 and megabase 24.908, which includes all of the genes within the boundaries of the previously identified speciation island. As shown in figure 1, there is a small region of differentiation coincident with reduced variation in the *M* form very close to the predicted island. Of the 10 loci, 2 show fixed differences with no shared polymorphisms between *M* and *S* individuals (*MSH5* and *i8*), whereas a third (*GPRor38*) has no fixed differences but no shared polymorphisms either. All 3 of these loci have average values of F_{ST} above 0.6 between *M* and *S* forms. One of these highly differentiated loci lies within the island of differentiation defined in Turner et al. (2005); the other 2 lie just outside the previously defined boundaries (fig. 1). The differences in the fixed:shared polymorphism ratio between these 3 loci and the adjacent 7 loci are highly significant (Fisher's exact test; $P = 3.9 \times 10^{-8}$).

In order to examine the geographical extent of differentiation on chromosome 2R, we sequenced *GPRor38* and *MSH5* in populations from Mali. In contrast to the patterns found in Cameroon, these 2 loci have no polymorphisms fixed between forms and share many polymorphisms across *M* and *S* in Mali (table 2; dividing the Mali sample into groups with and without the 2Rb inversion does not change this result). Along with the lack of fixed differences, there is an approximately 6-fold reduction in the values of F_{ST} between forms across both loci. As expected from the patterns just described, F_{ST} between Mali and Cameroon is higher at loci on 2R than at loci on 2L (table 3).

Levels of polymorphism at the loci in and around the 2R speciation island are much lower in the *M* form than in the *S* form in Cameroon, suggesting that a selective sweep occurred within the *M* form (fig. 1). To test for the signature of a selective sweep in this region, we again conducted multilocus HKA tests to the closest outgroup, *A. arabiensis*, using all 4 loci we have sequences for in *A. arabiensis* (*LIM*, *Subtilase*, *GPRor38*, and *MSH5*). An HKA test between the *M* form and *A. arabiensis* is significant ($P = 0.014$), but the same test comparing the *S* form and *A. arabiensis* is not significant ($P = 0.122$), as would be expected if there has been a selective sweep only in the *M* form lineage since its split with the *S* form. We also conducted the *M* form HKA test excluding the *LIM* locus (which may be subject to linked selection; see above), and the test remained significant ($P = 0.017$). The significant result is therefore due to low polymorphism in *M* at *GPRor38* (0.003) and *MSH5* (0.001) compared with *Subtilase* (0.012) and all the 3 loci in *S* (0.010–0.013). High divergence at *GPRor38* (5.7%) may also contribute to the significant result (all other loci in *M* and *S* = 1.1%–2.3%). Although the Tajima's *D* value of *GPRor38* in the *M* form in Cameroon was only marginally significant ($0.05 < P < 0.10$) in a model-based test (Tajima 1989), it was found to be significant based on coalescent simulations ($P = 0.036$), providing further support for a selective sweep near this locus in the *M* form. The low values of heterozygosity in the *M* form in Cameroon are not seen in Mali, suggesting that the selective sweep has occurred only in the Cameroon population of the *M* form.

The percent divergence (D_{xy}) between the *M* form in Cameroon and the 2 outgroups at *GPRor38* (5.7% to *A. arabiensis*, 7.2% to *A. quadriannulatus*) is much greater than the range of divergence from the other 3 loci in *M* for which we have data (1.4–2.1% to *A. arabiensis*, 1.3–2.9% to *A. quadriannulatus*). We used the test suggested by McDonald and Kreitman (1991) to determine whether the high levels of divergence at the *GPRor38* locus in

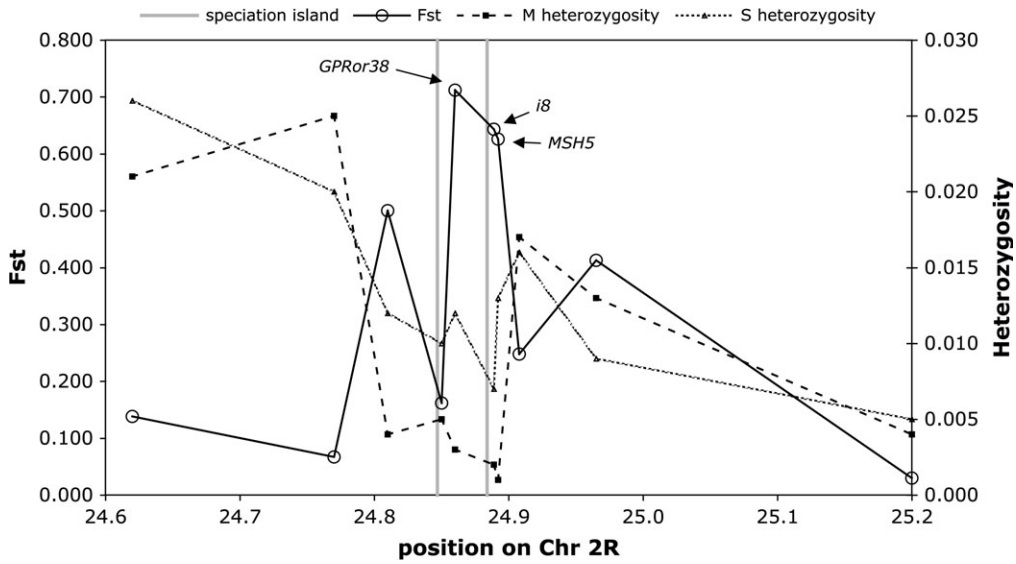

FIG. 1.—Variation and differentiation at 10 loci sequenced in Cameroon. The region predicted to be differentiated by Turner et al. (2005) is shown between the gray lines.

the Cameroon *M* form could be due to directional selection on the protein sequence of this olfactory receptor. Using *A. arabiensis* as an outgroup, this test revealed a marginally significant departure from neutral expectations, despite low power, with only 13 sites considered (synonymous: 7 polymorphic, 1 fixed; nonsynonymous: 1 polymorphic, 4 fixed; 1-tailed $P = 0.032$). The same test to the more distant outgroup *A. quadriannulatus* was not significant (synonymous: 7 polymorphic, 7 fixed; nonsynonymous: 1 polymorphic, 5 fixed; 1-tailed $P = 0.185$). Because the comparison to the more distant *A. quadriannulatus* includes the continued fixation of synonymous mutations, any signal of recent positive selection would be harder to detect. Together, the tests presented here suggest that adaptive natural selection has acted on the protein sequence of this highly differentiated olfactory receptor. There is no evidence of this adaptation in the S form in Cameroon (synonymous: 7 polymorphic, 1 fixed; nonsynonymous: 5 polymorphic, 1 fixed; $P = 0.987$).

Discussion

Some models of speciation with gene flow predict that the loci responsible for ecological, behavioral, and other isolating traits will not introgress freely between diverging taxa (Barton and Gale 1993; Wu and Ting 2004). Within *A. gambiae*, we have the opportunity to test this prediction with multiple sympatric populations of the *M* and *S* forms. We previously found 3 regions of the genome that were highly differentiated between *M* and *S* in Cameroon: one each on chromosomes 2L, 2R, and X (Turner et al. 2005). The differentiation in these speciation islands indicates that gene flow between forms is reduced relative to the rest of the genome, so these regions are hypothesized to contain genes involved in reproductive isolation (i.e., genes that have different fitness effects in the genomic backgrounds of *M* and *S*). We have now found that the island on chromosome 2L is present in both Mali and Cameroon: the 4 nt differences at the *LIM* locus on chromosome 2L are fixed between *M* and *S* in all 52 chromosomes sequenced, including those from both Mali and Cameroon. Because all of these individuals were identified as *M* or *S* based on their X-chromosome genotype, this means that the X and 2L regions are in complete association in our sample. This is significant, as the existence of 2 differentiated regions between *M* and *S*, each on an independently assorting chromosome, provides strong support for the existence of 2 partially isolated taxa within *A. gambiae*. The control locus just outside the island on 2L remains undifferentiated in both populations, providing further evidence of locus-specific selection against gene flow across *A. gambiae*.

We infer the action of positive selection in the 2L island as evidenced by reduced polymorphism and significant HKA tests in both forms, and a significant value of Tajima's *D* at the only locus with more than 2 polymorphisms; similar evidence has previously been found in the X-chromosome island (Stump et al. 2005). In both cases, however, the linked selection may be unrelated to speciation: both of these islands are near centromeres and are likely to have reduced recombination. Regions of reduced recombination show lower levels of polymorphism and/or an excess of low-frequency mutations in almost every organism that has been examined (e.g., Begun and Aquadro 1992; Stephan and Langley 1998; Nachman 2001; Tenaillon et al. 2001; Stajich and Hahn 2005). It is thought that this pattern is due to linked positive selection (Kaplan et al. 1989) and/or background selection (Charlesworth et al. 1993). Therefore, our results on chromosome 2L, as well as those of others on the X chromosome (Stump et al. 2005), should not be taken as evidence for selection directly associated with reproductive isolation between *M* and *S* in *A. gambiae*.

The third island, on chromosome 2R, shows a quite different pattern. This region is unremarkable in the Mali population, with levels of π, F_{ST}, and divergence that are similar to the *Subtilase* control locus and other loci from chromosomes 2R (this study) and 3R (Turner et al. 2005). In the Cameroon population, however, several observations implicate this region as being involved in reproductive isolation. We had previously discovered fixed differences on 2R at the *UNK1* locus (Turner et al. 2005), but these data are difficult to interpret given that this locus may be a multicopy-transposable element. However, we have now discovered 5 more fixed differences at the *MSH5* and *i8* loci—which lie just outside the previously defined speciation island—confirming that this region is strongly differentiated between *M* and *S* in Cameroon (table 2).

Reduced polymorphism, significant Tajima's *D*, and significant HKA test all point to a selective sweep in the *M* form. A significant McDonald–Kreitman test indicates that directional selection on the protein sequence of *GPRor38* could be the target of this selective sweep, though this result is tentative because we have sequenced only approximately a third of this gene. Despite considerable effort, we have been unable to clone to rest of this locus from our samples; because the *MSH5* locus has also been resistant to sequencing, we suspect that there may be differences between our samples and the reference assembly in this region. In the v36 genome assembly, the only genes predicted to lie in or near the differentiated region are 5; we have sequence data from *GPRgr13*, *GPRor38*, *GPRor39*, *MSH5*, and *CD59* and 2 predicted genes, which appear to be transposons (data not shown). In release 42 of the genome annotation, these 2 predicted transposons were removed, and a small protease inhibitor (locus *i8*) was further predicted, which we have also sequenced. Despite having sequenced portions of all the genes in this predicted island and successfully defined the extent of differentiation, strong inferences about the target of selection would require resequencing of an approximately 100-kb region (which might reveal differences with the current assembly), and functional studies of interesting candidate substitutions.

Unlike the 2L and X islands, the 2R island is not expected to lie in a region of reduced recombination. The observation of selection at the *GPRor38* locus is therefore more likely to be related to reproductive isolation. It is very interesting that differentiation and reduced variation on chromosome 2R are not seen in Mali: this seems to indicate that an isolating barrier was lost in Mali or gained in Cameroon without spreading to Mali. Whatever isolating mechanism is maintained in this chromosomal region in Cameroon is therefore not necessary to maintain isolation

in Mali. Hybridization between forms has been directly estimated to occur approximately 1% of the time in Mali (Tripet et al. 2001), and hybrid larvae and adults have been found (della Torre et al. 2005). No *M*/*S* hybrids have been found in Cameroon despite significant sampling effort, perhaps because reproductive isolation is stronger between forms in Cameroon (della Torre et al. 2005). Previous studies have found polymorphic isolating barriers between allopatric species, and between species that are sympatric over only some of their range (Ortiz-Barrientos et al. 2004; Reed and Markow 2004; Kopp and Frank 2005). The present results should be taken as further evidence for the lability of isolating barriers across a species range, even between 2 populations that are both sympatric (although this conclusion will remain tentative until additional evidence implicates this region in reproductive isolation). In the speciation model described by Wu and Ting (2004), incipient species become reproductively isolated locus by locus; because of selection against hybrid genotypes at 1 or a few loci, realized gene flow is reduced across the genome, shifting the balance between selection and gene flow toward differentiation at other loci. It is possible that the differentiated region found only in Cameroon on chromosome 2R is such an emerging speciation island, whose differentiation has been facilitated because of selection against hybrids at the 2L and X islands.

Although support for the existence of isolated groups within *A. gambiae* has existed for some time, the current work helps to clarify what exactly defines these groups. Much confusion has arisen from the switch in focus from inversion frequencies (chromosomal forms sensu Touré et al. 1998) to SNPs (molecular forms, Favia et al. 2001), and some doubt has remained about whether either form represents a valid taxonomic unit. It was the chromosomal forms that were discovered first; the discovery of marker SNPs on the X chromosome followed, and these SNPs were subsequently investigated throughout the species range (della Torre et al. 2001; Gentile et al. 2001, 2002; della Torre et al. 2005). Genotyping of molecular markers is much less ambiguous than determining inversion karyotypes as the actual number of hybrid karyotypes is difficult to resolve (Touré et al. 1998). Our discovery of a second differentiated region on chromosome 2L, which is in perfect association with the X-chromosome genotype, makes it clear that the 2L and X SNPs are true indicators of a taxonomic split within *A. gambiae*. Because *M* and *S* form individuals with the same karyotype can be compared in Cameroon, this country is the preferred location for future work on speciation: discoveries made in this country can then be investigated in the more complicated populations in Mali, as was done here. Regardless of which population is studied, it is clear that *A. gambiae* is a very interesting model organism for the study of speciation with gene flow.

Supplementary Material

Supplementary table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). Sequence data from this article have been deposited with the GenBank data library under accession numbers: DQ425111–DQ425208, DQ436826–DQ436911, and EF426141-EF426244.

Acknowledgments

We are grateful to G. C. Lanzaro and F. Tripet for providing DNA from karyotyped mosquitoes for this work and to S. V. Nuzhdin, D. J. Begun, and many others for advice and support. This work was supported by National Institutes of Health grant RO1 GM61773–01 and by National Science Foundation grant DEB-0316513 (to S. V. Nuzhdin), by National Institutes of Health grant AI40308 (to G. C. Lanzaro), by the Center for Population Biology at the University of California, Davis (to T.L.T.), and by National Science Foundation grant MCB-0528465 (to M.W.H.).

Literature Cited

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2:e286.

Barton NH, Gale KS. 1993. Genetic analysis of hybrid zones. In: Harrison RG, editor. Hybrid zones and the evolutionary process. New York: Oxford University Press. p. 13–45.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature. 356:519–520.

Begun DJ, Whitley P. 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc Natl Acad Sci USA. 97:5960–5965.

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

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.

della Torre A, Fanello C, Akogbeto M, Dossou-yovo J, Favia G, Petrarca V, Coluzzi M. 2001. Molecular evidence of incipient speciation within *Anopheles gambiae s.s.* in West Africa. Insect Mol Biol. 10:9–18.

della Torre A, Tu Z, Petrarca V. 2005. On the distribution and genetic differentiation of *Anopheles gambiae s.s.* molecular forms. Insect Biochem Mol Biol. 35:755–769.

Emelianov I, Marec F, Mallet J. 2004. Genomic evidence for divergence with gene flow in host races of the larch budmoth. Proc R Soc Lond B. 271:97–105.

Fanello C, Carneiro I, Ilboudo-Sanogo E, Cuzin-Ouattara N, Badalo A, Curtis CF. 2003. Comparative evaluation of carbosulfan- and permethrin-impregnated curtains for preventing house-entry by the malaria vector *Anopheles gambiae* in Burkina Faso. Med Vet Entomol. 17:333–338.

Favia G, Lanfrancotti A, Spanos L, Siden-Kiamos I, Louis C. 2001. Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae s.s.* Insect Mol Biol. 10:19–23.

Ford MJ, Aquadro CF. 1996. Selection on X-linked genes during speciation in the *Drosophila athabasca* complex. Genetics. 144:689–703.

Gentile G, della Torre A, Maegga B, Powell JR, Caccone A. 2002. Genetic differentiation in the African malaria vector, *Anopheles gambiae s.s.*, and the problem of taxonomic status. Genetics. 161:1561–1578.

Gentile G, Slotman M, Ketmaier V, Powell JR, Caccone A. 2001. Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae s.s*. Insect Mol Biol. 10:25–32.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics. 165:1269–1278.

Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics. 116:153–159.

Hunt RH. 1973. A cytological technique for the study of *Anopheles gambiae* complex. Parassitologia. 15:137–139.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics. 123:887–899.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics. 156: 1913–1931.

Kopp A, Frank AK. 2005. Speciation in progress? A continuum of reproductive isolation in *Drosophila bipectinata*. Genetica. 125:55–68.

Machado CA, Kliman RM, Markert JA, Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol Biol Evol. 19:472–488.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. Genet Res. 23:23–35.

McDonald JH, Kreitman MK. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 351:652–654.

Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. 17:481–485.

Ortiz-Barrientos D, Counterman BA, Noor MAF. 2004. The genetics of speciation by reinforcement. PLoS Biol. 2:e416.

Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution. 58:2064–2078.

Post RJ, Flook PK, Millest AL. 1993. Methods for the preservation of insects for DNA studies. Biochem Syst Ecol. 21:85–92.

Reed LK, Markow TA. 2004. Early events in speciation: polymorphism for hybrid male sterility in *Drosophila*. Proc Natl Acad Sci USA. 101:9009–9012.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.

Scott JA, Brogdon WG, Collins FH. 1993. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. Am J Trop Med Hyg. 49:520–529.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol Biol Evol. 22:63–73.

Stephan W, Langley CH. 1998. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics. 150:1585–1593.

Stump AD, Fitzpatrick MC, Lobo NF, Traoré SF, Sagnon NF, Constantini C, Collins FH, Besansky NJ. 2005. Centromere-proximal differentiation and speciation in *Anopheles gambiae*. Proc Natl Acad Sci USA. 102:15930–15935.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays ssp. mays* L.). Proc Natl Acad Sci USA. 98:9161–9166.

Touré YT, Petrarca V, Traoré SF, Coulibaly A, Maiga HM, Sankaré O, Sow M, Di Deco MA, Coluzzi M. 1998. The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. Parassitologia. 40:477–511.

Tripet F, Toure YT, Taylor CE, Norris DE, Dolo G, Lanzaro GC. 2001. DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of *Anopheles gambiae*. Mol Ecol. 10:1725–1732.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. 3:e285.

Wang R, Zheng LB, Toure YT, Dandekar T, Kafatos FC. 2001. When genetic distance matters: measuring genetic differentiation at microsatellite loci in whole-genome scans of recent and incipient mosquito species. Proc Natl Acad Sci USA. 98:10769–10774.

Wright S. 1951. The genetical structure of populations. Ann Eugen. 15:323–354.

Wu CI, Ting CT. 2004. Genes and speciation. Nat Rev Genet. 5:114–122.

Michael Nachman, Associate Editor

Accepted July 12, 2007