The Demographic Histories of the M and S Molecular Forms of *Anopheles gambiae s.s.*

Jacob E. Crawford* and Brian P. Lazzaro

Department of Entomology, Cornell University

***Corresponding author**: E-mail: jc598@cornell.edu.

Associate editor: Matthew Hahn

Abstract

Anopheles gambiae is a primary vector of *Plasmodium falciparum*, a human malaria parasite that causes over a million deaths each year in sub-Saharan Africa. Population genetic tests have been employed to detect natural selection at suspected *A. gambiae* antimalaria genes, but these tests have generally been compromised by the lack of demographically correct null models. Here, we used a coalescent simulation approach within a maximum likelihood framework to fit population growth, bottleneck, and migration models to polymorphism data from Cameroonian *A. gambiae*. The best-fit models for both the "M" and the "S" molecular forms of *A. gambiae* included ancient population growth and a high rate of migration from an unsampled subpopulation. After correcting for differences in effective population size, our models suggest that the molecular forms expanded at different times and both expansions significantly predate the advent of agriculture. We show that correcting null models for demography increases the power to detect natural selection in *A. gambiae*.

Key words: *Anopheles gambiae*, demographic history, population growth, maximum likelihood, population genetics.

Anopheles gambiae sensu stricto (hereafter *A. gambiae*) is a primary vector of *Plasmodium falciparum* (Collins and Paskewitz 1995), a human malaria parasite responsible for the death of an estimated 1 million people each year in sub-Saharan Africa, most of whom are children under the age of 5 years (WHO/UNICEF World Malaria Report 2008). Development of novel technologies for controlling disease transmission, including genetic engineering of *Plasmodium*-resistant transgenic mosquitoes (Alphey et al. 2002), depends on knowledge of the basic biology and evolution of the vector and the parasite. One approach to obtaining such information is to use population genetic data to identify *Anopheles* loci that evolve under pathogen-mediated natural selection, and a number of candidate loci have been tested for selection in *A. gambiae* (e.g., Cohuet et al. 2008; Obbard et al. 2009). Tests for selection in this system tend to rely on the site frequency spectrum (SFS; the frequency distribution of polymorphic mutations in the population) due to the lack of a suitable outgroup for interspecies molecular evolutionary comparisons (Obbard et al. 2007). However, tests of the SFS are also sensitive to demographic processes such as population growth and bottlenecks (Tajima 1989a, 1989b; Fu and Li 1993). One way to improve the power to distinguish patterns generated by selection from those generated by demography is to test selective hypotheses against a null model based on the demographic history of the species (e.g., Haddrill et al. 2005; Stajich and Hahn 2005), but the absence of genome-wide polymorphism data has prevented development of an adequate demographic null for *A. gambiae*. In this work, we use sequence polymorphism data from 109 *A. gambiae* genes recently published by Cohuet et al. (2008) to infer the demographic history of Cameroonian *A. gambiae*.

Several nonequilibrium demographic hypotheses have been previously proposed to describe *A. gambiae*. *Anopheles gambiae* is highly anthropophilic and ecologically dependent on humans and has been hypothesized to have undergone a range and population expansion coincident with agriculture-related shifts in human populations (Coluzzi et al. 2002; Costantini et al. 2009). A study of microsatellite polymorphism from Kenyan *A. gambiae* found evidence for population growth (Donnelly et al. 2001), and the SFS in this system tends to be enriched with low-frequency alleles (e.g., Cohuet et al. 2008; Obbard et al. 2009) consistent with an historical population expansion. Such patterns of polymorphism could also, however, derive from population bottlenecks (Tajima 1989a, 1989b). Additional evidence for a bottleneck stems from transposable element insertion site frequency data that are suggestive of population bottlenecks (e.g., Esnault et al. 2008) possibly related to founding events associated with the formation of incipient species or population fluctuations during the last glacial maximum (Weijers et al. 2007). Migration among subpopulations may also be an important demographic factor in this system. Extant *A. gambiae* are divided into two largely reproductively isolated units referred to as the "M" and "S" molecular forms (della Torre et al. 2001). Geographic and microecological substructure has been identified within both molecular forms as well (e.g., Lehmann et al. 2003; Slotman et al. 2007).

To distinguish among potential demographic hypotheses describing *A. gambiae*, we performed coalescent simulations under various parameterizations of the above demographic models (supplementary fig. S1, Supplementary Material online) and employed a modified approximate likelihood method (Weiss and von Haeseler 1998) to test the fit of simulated models to synonymous

Table 1. MLE (and 95% CIs) for Model Families and Model Comparisons.

	M-form			S-form		
Model	Growth	Bottleneck	Migration	Growth	Bottleneck	Migration
Generations since growth (N_{curr})[a]	3.52 (2.88–4.12)	3.44 (2.92–4.16)	3.00 (2.54–3.50)	3.08 (2.58–3.50)	3.04 (2.60–3.52)	2.60 (2.18–2.88)
Fold growth ($N_{\mathrm{curr}}/N_{\mathrm{anc}}$)[a]	1,000 (4.65–ND)[b]	10,000 (4.05–ND)[b]	10,000 (13.0–ND)[b]	2,000 (4.65–ND)[b]	1,000 (5.10–ND)[b]	100 (13.00–ND)[b]
Reduction during bottleneck ($N_{\mathrm{prebottle}}/N_{\mathrm{anc}}$)[a]	—	10,000 (NA)[c]	—	—	667 (NA)[c]	—
Duration of bottleneck (T_{bot})[a]	—	0.2 (NA)[c]	—	—	0.2 (NA)[c]	—
Migrants per generation ($4Nm$)[a]	—	—	5 (ND)	—	—	10 (ND)
Size of unsampled subpopulation (relative to sampled)	—	—	0.40 (0.23–0.58)	—	—	0.50 (0.35–0.75)
Log likelihood	−217.1336	−216.9759	−213.6302	−210.4305	−210.3666	−197.9891
AIC (k)	438.2672 (2)	441.9518 (4)	435.2604 (4)	424.861 (2)	428.7332 (4)	403.9782 (4)
Λ_{SN}[d] (P value relative to equilibrium)	−28.7046 (<0.0001)	—	—	−44.7847 (<0.0001)	—	—
Λ_{G}[e] (P value relative to growth)	—	3.6846 (NS)	−3.0068 (0.0019)	—	3.8722 (NS)	−20.8828 (<0.0001)

Note.—AIC = −2(log likelihood − k), where k is the number of free parameters in model; CI, confidence interval, and AIC is the Akaike Information Criterion (Supplementary Material online).

[a] Parameter units.

[b] ND indicates cases where only one boundary of the CI could be determined.

[c] CIs were not estimated for these parameters.

[d] Λ_{SN} indicates comparisons made between the AIC under the MLE and the AIC under the SNE model.

[e] Λ_{G} indicates comparisons made between the AIC under the MLE and the AIC under the growth model.

autosomal polymorphism data for each molecular form independently (Cohuet et al. 2008; Supplementary Material online). The data set of Cohuet et al. (2008) consists of short coding fragments from 72 immune-related and 37 functionally random genes sequenced in M-form (n = 10–16 chromosomes) and S-form (n = 10–18 chromosomes) mosquitoes collected in Cameroon. We treated the demographic models in a hierarchy of increasing parameter number, such that the standard neutral equilibrium (SNE) model was the null hypothesis, population growth was the first alternative, and the bottleneck and migration models were alternatives to the growth model. We found that the population growth model fits the empirical data significantly better than the equilibrium model for both the M- and the S-forms (table 1; $P_M < 10^{-4}$, $P_S < 10^{-4}$). No support was found for a population bottleneck in either molecular form (table 1). However, models that included both population growth and migration fit the data significantly better than the simple growth model for both molecular forms (table 1; $P_M = 0.0019$, $P_S < 10^{-4}$). We confirmed that our best-fit models were able to adequately reproduce the empirical data by showing that the average number of pairwise differences and the number of segregating sites in samples simulated under the best-fit migration models (Supplementary Material online) match those summary statistics from the empirical data very well (supplementary figs. S2 and S3, Supplementary Material online). Furthermore, approximately 10% of simulations were accepted for each model (supplementary table 1, Supplementary Material online), implying a good fit considering we used the 20% threshold approach within the approximate likelihood method, which should reject as high as 80% of simulations even when the model perfectly matches the evolutionary process underlying the data.

Although similar in structure, the most likely migration models for the M and S molecular forms differed in their timing of expansion. From profile likelihood curves, we obtained maximum likelihood estimates (MLEs) and approximate 95% confidence regions for the growth parameters (figs. 1 and 2). To evaluate the potential impact selection may have on our demographic inference, we reanalyzed the likelihood surface after removing the six loci with the most extreme Tajima's D values and found that the migration models remained the best-fit models, and MLE parameter values were essentially unchanged from those inferred using whole data sets. From this, we conclude that it is unlikely that any natural selection in the history of the empirical data is biasing our inference process. We estimate that both molecular forms underwent at least 13-fold growth (table 1) but that the M molecular form expanded more recently than the S molecular form (49,000–490,000 years before present [YBP] for M-form vs. 63,000–630,000 YBP for S-form, assuming 10 generations per year and a reasonable mutation rate; table 2). Our estimated growth times likely predate the extant division between the two molecular forms (e.g., Mukabayire et al. 2001). One potential explanation for our estimate of differing times of expansion for the two forms is that the ancestral, premolecular form population underwent an expansion, and then the derived M molecular form underwent a second more recent expansion, which may

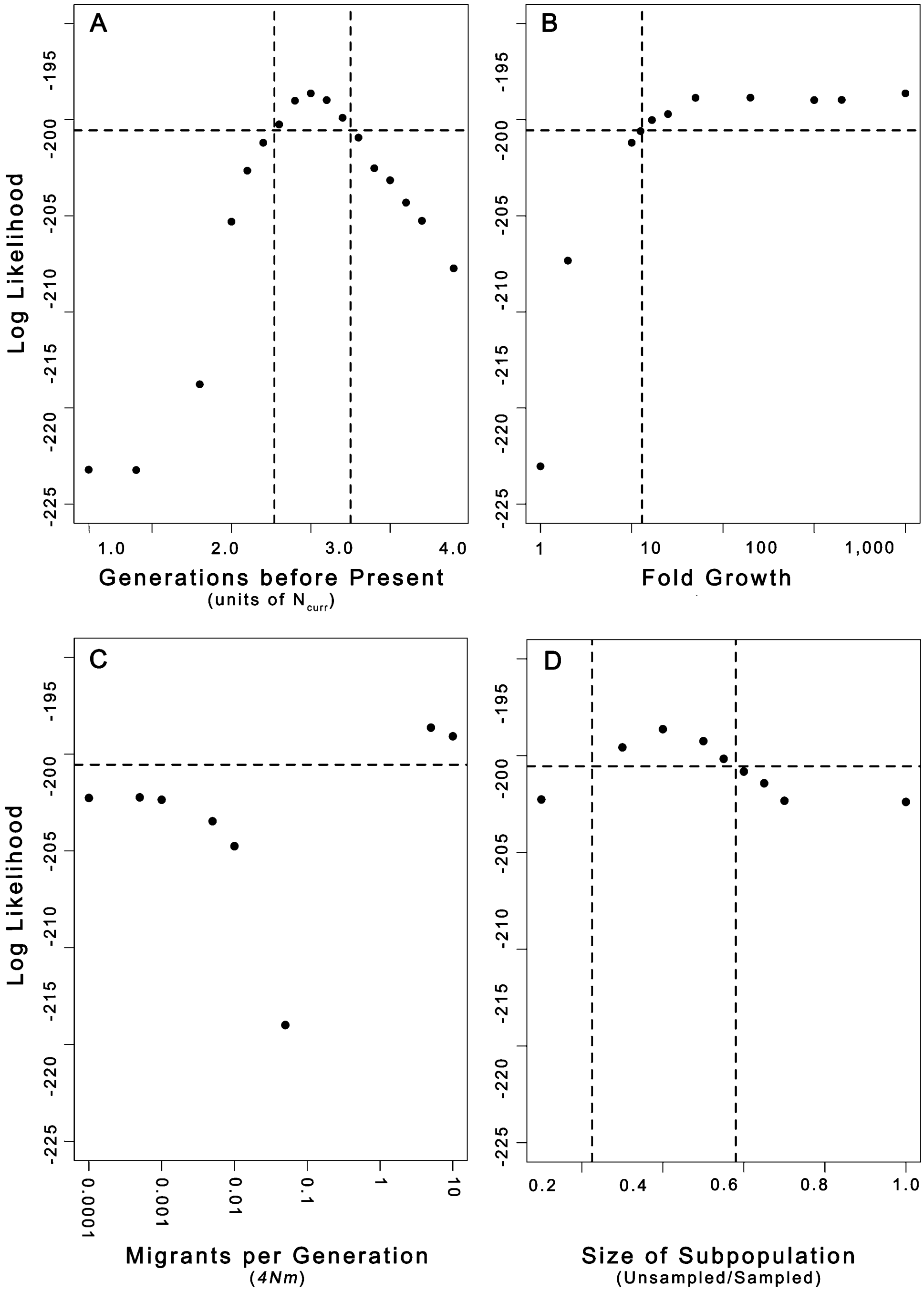

Fig. 1. M molecular form genome likelihood values relative to migration model parameters (*A*) time of expansion in units of N_{curr} generations, (*B*) size of population growth (N_{curr}/N_{anc}), (*C*) rate of migration (4*Nm*) per generation, and (*D*) size of the unsampled subpopulation relative to the sampled subpopulation. For each parameter value, the highest genome likelihood value from all models within the migration model family is plotted. Note log scale in panels (*B*) and (*C*). Horizontal dashed line indicates 95% threshold, such that all genome likelihood values below this threshold are significantly different from the maximum likelihood value. Vertical dashed lines indicate approximate boundaries of the 95% confidence region of the model parameter. For panels (*A*) and (*D*), all parameter values outside the vertical dashed lines are significantly different from the MLE value. For panel (*B*), all parameter values to the left of the vertical line are significantly different from the MLE value. The shape of the curve in panel (*C*) did not allow determination of confidence region.

have been associated with postspeciation niche specialization (Costantini et al. 2009), such that the M-form genome bears a mixed demographic signal from the two expansions.

Models that include migrational exchange with an unspecified second population fit the data best for both molecular forms, but these models should be interpreted with caution. For both molecular forms, the profile likelihood curve for the rate of migration (4*Nm*) is bimodal with local maxima near 4*Nm* of 0 and 10 (figs. 1 and 2). Both these maxima suggest near-panmixia, with little or no real migration component. We therefore next modeled each molecular form under the growth model but manually adjusted the modeled effective population size to be larger (i.e.,

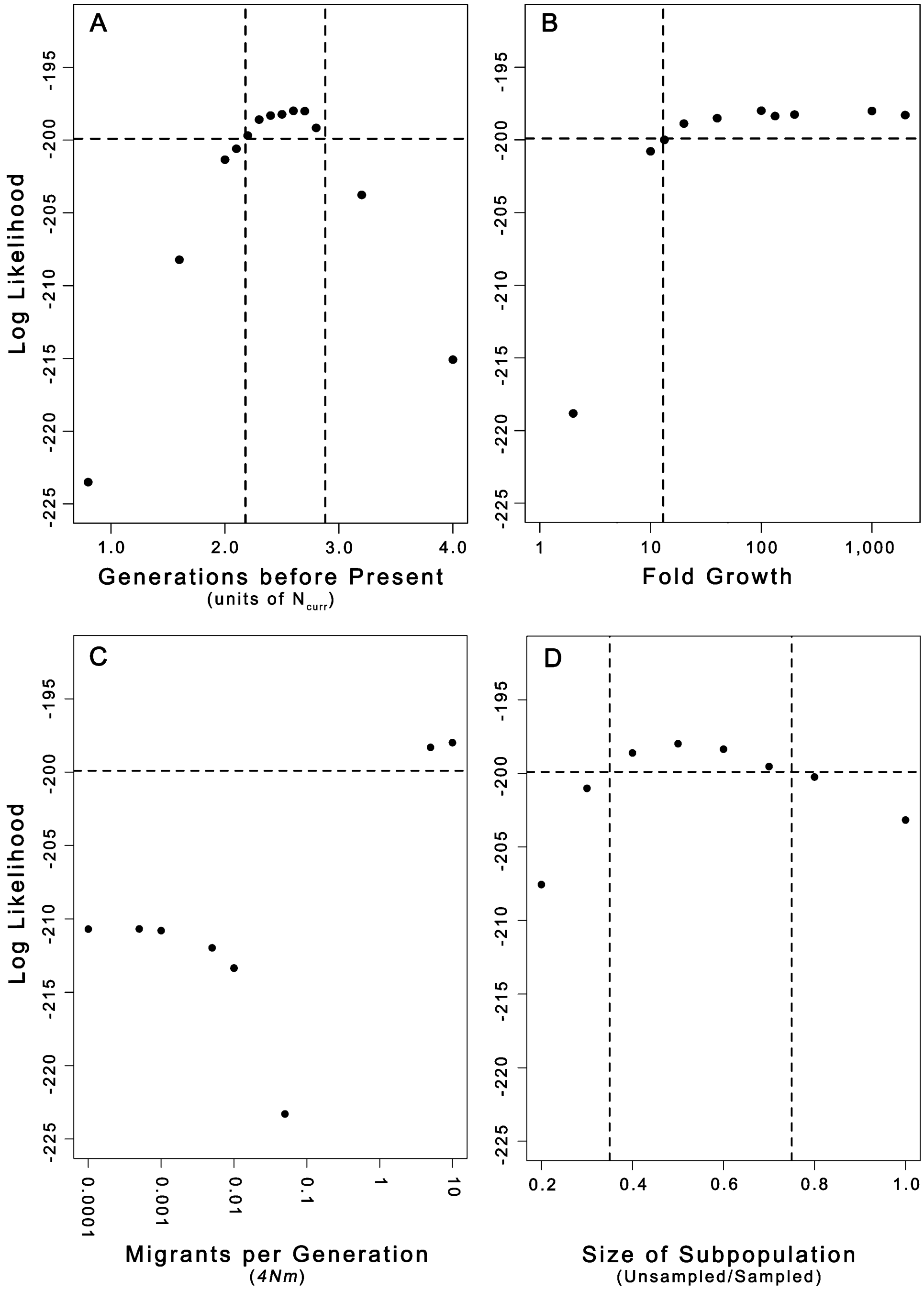

Fig. 2. S molecular form genome likelihood values relative to migration model parameters (*A*) time of expansion in units of N_{curr} generations, (*B*) size of population growth (N_{curr}/N_{anc}), (*C*) rate of migration (4*Nm*) per generation, and (*D*) size of the unsampled subpopulation relative to the sampled subpopulation. For each parameter value, the highest genome likelihood value from all models within the migration model family is plotted. Note log scale in panels (*B*) and (*C*). Horizontal dashed line indicates 95% threshold, such that all genome likelihood values below this threshold are significantly different from the maximum likelihood value. Vertical dashed lines indicate approximate boundaries of the 95% confidence region of the model parameter. For panels (*A*) and (*D*), all parameter values outside the vertical dashed lines are significantly different from the MLE value. For panel (*B*), all parameter values to the left of the vertical line are significantly different from the MLE value. The shape of the curve in panel (*C*) did not allow determination of confidence region.

pooled the sampled and hypothetical unsampled "populations" into a single panmictic unit). We found that both the MLE growth and the MLE migration models fit the data significantly better than the N_e-adjusted growth model for both molecular forms (Supplementary Material online). In principle, the migration models might provide a statistically better fit to the data in the absence of true historical migration if they allow for greater variance in effective population size than the simple growth model does. The signal for ancient growth is strong and clear in both forms regardless of whether historical migration is included in the model. Nonetheless, the best-fitting models for both molecular forms are those in which the focal populations share migrants with an unsampled population that is smaller

Table 2. Calculations of the Approximate Timing of Growth Based on Empirical Parameter Values.

Form (θ_W[a])	μ[b]	N_e[c]	T_1[d]	Generations per year[e]	YBP[f]
M (2.27%)	3.5×10^{-9}	1,623,571	3.0	10	487,071
	3.5×10^{-8}	162,357	3.0	10	48,707
	3.5×10^{-7}	16,236	3.0	10	4,871
	3.5×10^{-9}	1,623,571	3.0	20	243,536
	3.5×10^{-8}	162,357	3.0	20	24,353
	3.5×10^{-7}	16,236	3.0	20	2,435
S (3.4%)	3.5×10^{-9}	2,435,000	2.6	10	633,100
	3.5×10^{-8}	243,500	2.6	10	63,310
	3.5×10^{-7}	24,350	2.6	10	7,500
	3.5×10^{-9}	2,435,000	2.6	20	316,550
	3.5×10^{-8}	243,500	2.6	20	37,499
	3.5×10^{-7}	24,350	2.6	20	3,750

[a] θ_W was estimated from synonymous sites in the data sets of Cohuet et al. (2008) (see Supplementary Material online) and is an estimator of $4N_e\mu$.

[b] Mutation rate per base pair per generation of 3.5×10^{-9} taken from Keightley et al. (2009).

[c] Effective population size calculated from θ_W using the stated mutation rate.

[d] MLE time of growth in units of N_{curr}.

[e] Estimates taken from Lehmann et al. (1998).

[f] YBP calculated as $(T_1 \times N_e)$/generations per year.

than the sampled population. Because the effective population size of the S molecular form is thought to be significantly larger than that of the M molecular form (e.g., Cohuet et al. 2008), this is unlikely to reflect migration between progenitors of extant M- and S-form mosquitoes, at least when the M population is the focal population being modeled.

It has been hypothesized that the advent of agriculture played a major role in the history of *A. gambiae* populations (e.g., Donnelly et al. 2001; Coluzzi et al. 2002), but the empirical sequence data from Cohuet et al. (2008) do not support this hypothesis. Based on the MLE growth parameter values inferred in our study, one would have to assume a per-nucleotide mutation rate of 10^{-7} mutations per generation in order to reconcile the inferred timing of population expansion with the agricultural revolution (<5,000 YBP; Phillipson 2005). Such a mutation rate is orders of magnitude higher than typical per-nucleotide mutation rate estimates for *Drosophila* (e.g., Tamura et al. 2004; Keightley et al. 2009), which provides our best estimate of the *Anopheles* mutation rate. Calculations based on more plausible parameter values (table 2) suggest that earlier anthropogenic events such as the movement out of the ancestral East African range by early humans (ca. 130,000 YBP; Reed and Tishkoff 2006) or subsequent human population expansions (ca. 50,000–70,000 YBP; Rogers and Harpending 1992) may have been key factors allowing mosquito populations to grow.

Genetic substructure in *A. gambiae* has been associated with the incipient speciation between the M- and S-forms (della Torre et al. 2001) as well as with ecological factors and chromosomal inversions (e.g., Slotman et al. 2007), raising the possibility that the demographic signal inferred from any single population may not be universally applicable. With specific respect to our study, the "Forest" M-form population from Cameroon under analysis here is partially differentiated from the "Mopti" M-form populations from West Africa (e.g., Slotman et al. 2007). However, the population size expansion we infer in this study surely predates extant population structure between Forest M and Mopti M, and we are confident that the signature of this M-form demographic history should be shared among extant uninverted M-form autosomes. The same logic can be applied to geographically distinct S-form populations. Sequences within polymorphic chromosomal inversions, particularly on the inversion-rich chromosome II, are likely to bear the signature of more recent demographic and selective events associated with the inversions themselves, which could confound model-based inference of demographic history. As our analysis was based entirely on autosomes with the standard (uninverted) karyotype (Cohuet et al. 2008), thought to be the ancestral form of *A. gambiae* (Ayala and Coluzzi 2005), we believe that our conclusions are insulated from this concern and that they can be taken to provide a baseline ancestral demographic history for the genomes of extant *A. gambiae*.

A primary motivation for establishing correct demographic models in *A. gambiae* and other systems is to accurately identify targets of natural selection. This is especially important in *Anopheles*, where sites of host–pathogen coevolution may serve as targets for malaria-control intervention. To show the effect of including demography in the null population genetic model on the inference of putatively nonneutral patterns of polymorphism, we reevaluated the results from a frequency spectrum–based analysis of *A. gambiae* loci conducted by Obbard et al. (2009). These authors resequenced 16 serine protease inhibitor genes (serpins) and 16 control loci in a West African M-form population from Burkina Faso (BK) and an East African S-form population from Kenya (KY), although we will only consider loci on chromosome III (four serpins and four control loci) to avoid the potentially confounding effects of chromosome II inversions in BK. Of the eight chromosome III loci that had at least four segregating sites (four serpins and four control loci), only control loci BK-5 and BK-6 departed significantly (5% threshold) from a null distribution simulated under the SNE model (Obbard et al. 2009). We compared Tajima's *D* values from all eight loci from BK and KY first to null distributions simulated under SNE and then to null distributions simulated under the MLE migration models we developed here (supplementary table 2, Supplementary Material online). We found that the negative values of *D* observed at control loci 5 and 6 remained significantly inconsistent with neutrality under the MLE migration model (locus BK-5: $P = 0.0160$ and locus BK-6: $P = 0.0076$) and that the positive values of *D* observed at serpins 4C and 6 became significant when compared with the MLE models (KY-4C: $P = 0.0372$ and KY-6: $P = 0.0069$; supplementary table 2, Supplementary Material online). Interestingly, although the mean *D* value under MLE migration models was consistently more negative than those under the SNE, the distributions showed less dispersion around the mean than those simulated under the SNE, resulting in a lower *P* value for control loci BK-5 and BK-6 under the SNE model than under the MLE migration model (supplementary table 2; Supplementary Material online). These results highlight the increased power to detect putative signals of natural selection when using demographically corrected null models. The power gains associated with using correct null models should be even greater when sophisticated genome-scale methods such as the composite likelihood ratio test of Kim and Stephan (2002) are employed.

Supplementary Material

Supplementary figures S1–S3 and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to two anonymous reviewers, Laura Harrington, Darren Obbard, Kirk Lohmueller, and members of the Lazzaro Laboratory for helpful discussions and comments on earlier versions of the manuscript. This work was supported by National Institutes of Health grant AI062995.

References

Alphey L, Beard CB, Billingsley P, et al. (23 co-authors). 2002. Malaria control with genetically manipulated insect vectors. *Science* 298:119–121.

Ayala FJ, Coluzzi M. 2005. Chromosome speciation: humans, Drosophila, and mosquitoes. *Proc Natl Acad Sci U S A.* 102(1 Suppl):6535–6542.

Cohuet A, Krishnakumar S, Simard F, Morlais I, Koutsos A, Fontenille D, Mindrinos M, Kafatos FC. 2008. SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system. *BMC genomics.* 9:227.

Collins FH, Paskewitz SM. 1995. Malaria: current and future prospects for control. *Annu Rev Entomol.* 40:195–219.

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.

Costantini C, Ayala D, Guelbeogo WM, et al. (12 co-authors). 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. *BMC Ecol.* 9:16.

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 ss 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 ss molecular forms. *Insect Biochem Mol Biol.* 35:755–769.

Donnelly MJ, Licht MC, Lehmann T. 2001. Evidence for recent population expansion in the evolutionary history of the malaria vectors *Anopheles arabiensis* and *Anopheles gambiae*. *Mol Biol Evol.* 18:1353–1364.

Esnault C, Boulesteix M, Duchemin JB, et al. (12 co-authors). 2008. High genetic differentiation between the M and S molecular forms of *Anopheles gambiae* in Africa. *PloS One.* 3:e1968.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* 15:790–799.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.* 19:1195–1201.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777.

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

Lehmann T, Licht M, Elissa N, Maega BTA, Chimumbwa JM, Watsenga FT, Wondji CS, Simard F, Hawley WA. 2003. Population structure of Anopheles gambiae in Africa. *J Hered.* 94:133–147.

Mukabayire O, Caridi J, Wang X, et al. 2001. Patterns of DNA sequence variation in chromosomally recognized taxa of Anopheles gambiae: evidence from rDNA and single-copy loci. *Insect Mol Biol.* 10:33–46.

Obbard DJ, Linton YM, Jiggins FM, Yan G, Little TJ. 2007. Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection. *Mol Ecol.* 16:3497.

Obbard DJ, Welch JJ, Little TJ. 2009. Inferring selection in the Anopheles gambiae species complex: an example from immune-related serine protease inhibitors. *Malar J.* 8:117.

Phillipson DW. 2005. African archaeology. Cambridge: Cambridge University Press. p. 202.

Reed F, Tishkoff S. 2006. African human diversity, origins and migrations. *Curr Opin Genet Dev.* 16:597–605.

Rogers AR, Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol.* 9:552–569.

Slotman MA, Tripet F, Cornel AJ, et al. (11 co-authors). 2007. Evidence for subdivision within the M molecular form of Anopheles gambiae. *Mol Ecol.* 16:639–649.

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

Tajima F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics* 123:597–601.

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

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol Biol Evol.* 21:36–44.

Weijers JWH, Schefuss E, Schouten S, Damste JSS. 2007. Coupled thermal and hydrological evolution of tropical Africa over the last deglaciation. *Science* 315:1701–1704.

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. *Genetics* 149:1539–1546.

WHO/UNICEF World Malaria Report. 2008. World Malaria Report. Geneva (Switzerland): World Health Organization [cited 2009 Oct]. Available from: http://apps.who.int/malaria/wmr2008/.