Efficient Bayesian inference under the multispecies coalescent with migration

Tomáš Flouri, Xiyun Jiao, Jun Huang, Bruce Rannala, and Ziheng Yang

Published October 23, 2023.

Significance

Inference of gene flow using genomic data requires powerful methods as the process of coalescent, migration, and mutation is highly stochastic. However, it is challenging to implement the multispecies coalescent with migration (MSC-M) model in a full likelihood framework correctly and efficiently. We developed Markov chain Monte Carlo algorithms under the MSC-M model and implement them in our Bayesian program BPP to achieve efficient computation. We conduct extensive validations and tests and show that our implementation is reliable and can handle large datasets with thousands of loci. We analyzed genomic data from Anopheles mosquitoes to demonstrate the feasibility of using genomic data to test for gene flow and to estimate the rate of gene flow.

Analyses of genome sequence data have revealed pervasive interspecific gene flow and enriched our understanding of the role of gene flow in speciation and adaptation. Inference of gene flow using genomic data requires powerful statistical methods. Yet current likelihood-based methods involve heavy computation and are feasible for small datasets only. Here, we implement the multispecies-coalescent-with-migration model in the Bayesian program BPP, which can be used to test for gene flow and estimate migration rates, as well as species divergence times and population sizes. We develop Markov chain Monte Carlo algorithms for efficient sampling from the posterior, enabling the analysis of genome-scale datasets with thousands of loci. Implementation of both introgression and migration models in the same program allows us to test whether gene flow occurred continuously over time or in pulses. Analyses of genomic data from Anopheles mosquitoes demonstrate rich information in typical genomic datasets about the mode and rate of gene flow.

One of the most important findings arising from comparative analyses of sequenced genomes during the past two decades is the prevalence of interspecific gene flow. Hybridization has been inferred in both plants [e.g., Arabidopsis (1)] and animals, including Anopheles mosquitoes (2), Panthera cats (3), cichlid fishes (4), and Hominins (5). Hybridization transfers genetic variation across species and by creating new allelic combinations at multiple loci may contribute to ecological adaptation (6, 7). Inference of gene flow can further our understanding of speciation (8, 9), help delineate species boundaries (10, 11), and guide efforts to conserve biodiversity and detect invasive species.

Gene flow is often inferred using simple summaries of genomic data. For example, the D statistic (12) and HyDe (13) test for gene flow using genome-wide site-pattern counts in a species quartet, while SNAQ (14, 15) uses the reconstructed gene tree topologies. These methods have low power and are often unable to identify gene flow between sister lineages, or to infer the direction, timing, and strength of gene flow (16–18).

Likelihood methods based on the multispecies coalescent (MSC) model (19) make full use of information in the data, providing rich inference using genomic datasets (20). Two simple models of gene flow have been developed under the MSC, representing different modes of gene flow (16, 17). The MSC-with-introgression [MSC-I; (21)] model, also known as multispecies network coalescent [MSNC, (22, 23)], assumes that gene flow occurs at a particular time point in the past. The MSC-with-migration (MSC-M) model, also known as the isolation-with-migration (IM) model, assumes that gene flow occurs continuously at a certain rate every generation after species divergence (24, 25). In both models the rate of gene flow should be considered an ‘effective’ rate, reflecting the combined effects of gene flow and natural selection on introgressed alleles, influenced by genetic drift and local recombination rate (9). The two models are simple extremes as in reality the rate of gene flow may be expected to vary over time (26). Here, we focus on MSC-M.

Under the MSC-M model, the gene genealogy at any locus includes not only the tree topology and coalescent times (branch lengths) but also detailed migration history (the number, directions, and timings of migration events). There may be no upper limit to the number of migration events at each gene locus. Likelihood inference has to average over the gene genealogy underlying the sequence data at each locus, including the migration history. There have been two approaches to dealing with the migration history (16). The first relies on a theory developed in the structured coalescent framework, in which the backward-in-time process of coalescence and migration is described using a continuous-time Markov chain, to integrate out the migration history at each locus analytically (27–30). However, the number of states in the Markov chain grows explosively with the increase in the number of species and the number of sequences (30).
Thus, this approach is feasible for very small numbers of species and sequences but can deal with many loci. The maximum likelihood program SES (29, 31) is limited to three species and three sequences per locus. The program MIST (32) is implemented for two species/populations only and can handle eight sequences per locus. Both programs can handle >10,000 loci.

The second approach uses Markov chain Monte Carlo (MCMC) to average over the gene trees including the migration history numerically. Both IMAs (24, 25) and G-PhoCS (33) take this strategy. While G-PhoCS assumes that the MSC-M model is fixed, IMAs also allows the species phylogeny to change during the MCMC. Both programs involve a high computational load: IMAs has been used to analyze data of a few hundred loci while G-PhoCS has been used to analyze data of a few thousand.

Population genetic models of population subdivision and migration (34) have been implemented in the program MIGRATE (35–37). A major difference of these models from MSC-M is that they do not account for the population/species phylogeny or the history of population divergences. They can be considered special cases of MSC-M with divergence times approaching ∞ (see below). The structured coalescent model was also implemented in the program BEAST2 as the MULiTYPE TREE package (38), and approximations were made to improve computation in the MASCOT package (39). These are designed for phylogenetic analysis of viral sequence data, treating geographical locations as subpopulations, with the aim of estimating migration rates, reconstructing transmission histories, and tracing the emergence of outbreaks in a pandemic (38). The so-called “migration” model treats migrations between geographical regions as a continuous-time Markov chain, such as used to model mutations, and assumes that the migration process does not influence the shape of the genealogical tree. The model thus has major deviations from the structured coalescent or MSC-M, leading to unreliable inference of migration rates and high sensitivity of the inferred root location to sampling biases (40). They do not appear to be suitable for analysis of multilocus sequence alignments from different species under the MSC.

Overall, current likelihood methods under the MSC-M model for multilocus sequence data including both IMAs and G-PhoCS involve heavy computation and do not scale well with genomic datasets. The algorithmic challenge is not mainly due to the expanded state space because of migration histories at multiple loci; rather, it lies in the extremely stringent constraints placed by the gene trees on the species tree or on the MSC-M model. For example, if we propose to modify a species divergence time with all gene trees fixed, only tiny changes are permissible. In the context of MSC without gene flow, we have found that smart MCMC moves that make coordinated changes to both the species tree and the gene trees can dramatically improve MCMC mixing (19, 41, 42), making it possible to analyze datasets with >10,000 loci (21, 43).

Here, we implement the MSC-M model in BPP, a coalescent-based Bayesian MCMC program (44, 45). We develop MCMC algorithms for efficient mixing and conduct extensive simulations to validate our algorithms. We also show that the MSC-M model can be used to analyze classical population genetic models of subdivision and structure such as the finite-island model (46, 47) and the stepping-stone model (48) which are special cases of the MSC-M model. We show that BPP outperforms existing methods in both reliability and scalability. We applied both MSC-M and MSC-I models to genomic data from the Anopheles gambiae group of African mosquitoes (2, 43), to test for gene flow and to estimate the rate of gene flow, as well as other major population parameters such as species divergence times and population sizes.

Having both MSC-I and MSC-M in the same program allows us to examine their differences when both are applied to the same data and to compare their goodness of fit.

Results

Gene-tree Density under the MSC-M Model. Fig. 1A illustrates the MSC-M model, which involves three sets of parameters: species divergence times (τ), population sizes (θ), and migration rates (M), with Θ = {τ, θ, M}. Time is scaled by mutations so that one time unit is the expected time to accumulate one mutation per site. Thus, both τ and θ are measured in expected number of mutations per site. At this time scale, two sequences from a population of size θ coalesce at the rate of θ/2. Migration rate Mij is defined as the expected number of migrants from species i to j per generation, with Mij = Nijmiij, where Nij is the (effective) population size of species j and miij is the proportion of immigrants in population j from population i. Note that we use the real-world view (with time running forward) to define the migration rate parameter. Let X = ⎛X(i)⎞ denote the data, with ⎛X(i)⎞ to be the sequence alignment at locus i. Let G = {G(i)} be the gene trees, where G(i) includes the rooted tree, the coalescent times, and the migration history at the locus (including the number, directions, and timings of migration events). We assume no recombination among sites in the sequence of the same locus and free recombination between loci, so that all sites at the same locus share the same gene tree while gene trees at multiple loci are independent. A recent simulation suggests that inference under the MSC is robust to moderate levels of recombination (49).

We implement MCMC algorithms to sample from the joint conditional distribution of the parameters and the gene trees.

\[ p(\Theta, G | X) \propto p(\Theta)p(G | \Theta)p(X | G). \]

where \( p(\Theta) \) is the joint prior, \( p(G | \Theta) \) is the probability density of the gene trees given the parameters in the MSC-M model, and

![Fig. 1.](https://doi.org/10.1073/pnas.2310708120)

**Fig. 1.** (A) A species tree for three species (A, B, C) with migration between species B and C showing model parameters, \( \Theta = \{\tau_{AB}, \tau_{AC}, \tau_{BC}, \theta_A, \theta_B, \theta_C, M_{AB}, M_{AC}, M_{BC}\} \). There are three sets of parameters in the model: species divergence times (\( \tau_{AB} = \tau_{AC} = \tau_{BC} \)), population sizes (\( \theta_A, \theta_B, \theta_C, \theta_R, \theta_S \)), and migration rates (\( M_{BC}, M_{AC} \)). Both \( \tau \) and \( \theta \) are measured in the expected number of mutations per site. The (population) migration rate is defined as \( M_{ij} = N_{ij}m_{ij} \), the expected number of migrants from species i to j per generation, where \( N_{ij} \) is the (effective) population size of species j and \( m_{ij} \) is the proportion of immigrants in population j from population i. (B) A possible gene tree with the complete history of coalescent and migration events at a locus with two sequences from each of the three species (\( a_1, a_2, b_1, b_2, c_1, c_2 \)) with migration from A to C at time \( t_1 \) and from C to B at time \( t_2 \), with \( t_1 < t_2 < t_3 < t_4 < t_5 < t_6 \).
The probability density of the gene trees under the MSC-M model, \( p(G|\Theta) \) in Eq. 1, is essentially given by the structured coalescent process of coalescent and migration (24, 27, 28, 34, 35) operating within each time interval between species divergences. To accommodate the species phylogeny, we simply reset the process when we reach a species divergence event, with an update to the number of populations, migration rates, and other population parameters. The MSC-M process is also a variable-rate (piecewise constant-rate) Poisson process, in which the coalescent and migration rates change at any coalescent event, migration event, and species divergence event (51).

We break the time period for each species \( j \) into \( K_j \) time segments at locus \( i \), during which the coalescent and migration rates are constant. Let \( t_{ijk} \) be the duration of the \( k \)th time segment, and \( n_{jk} \) be the number of lineages, with \( k = 1, \ldots, K_j \). Let \( w_{ij} \) be the number of migration events from species \( i \) to \( j \) at locus \( i \) (with time running forward). We define the indicator \( I_{ijk} \) to be 1 if migration from \( i \) to \( j \) is possible during time segment \( k \) (i.e., if both species \( i \) and \( j \) exist in time segment \( k \) and are permitted to exchange migrants) and 0 otherwise. The probability density for the gene trees under MSC-M is then a product over species and segments at locus \( i \), with six sequences is

\[
\prod_j \prod_i p(G_i|\Theta) = \prod_j \prod_i \left[ \frac{2}{\theta_{ij}} e^{-\frac{2}{\theta_{ij}} n_i} \prod_i \left( \frac{4M_{ij}}{\theta_{ij}} \right)^2 \times \exp \left\{ -\sum_{k=1}^{K_j} \left( \frac{n_{ijk}(n_{ijk} - 1)}{\theta_{ij}} + \frac{n_{ijk} \cdot 4 \sum_i I_{ijk} M_{ij}}{\theta_{ij}} \right) t_{ijk} \right\} \right].
\]

where \( \theta_{ij} \) is the heredity/ploidy scalar for the \( r \)th locus (e.g., 1 for autosomes, \( \frac{1}{2} \) for X-linked, \( \frac{1}{4} \) for Y-linked or mtDNA loci), and \( c_{ij} \) is the number of coalescent events in species \( j \) at locus \( i \). Here \( \frac{2}{\theta_{ij}} \) is the coalescent rate per time unit for a pair of sequences in species \( j \) at locus \( i \) and \( \frac{4M_{ij}}{\theta_{ij}} \) is the (mutation-scaled) migration rate from species \( i \) to \( j \), with time running forward.

For example, the density for the gene tree of Fig. 1B for a locus with six sequences is

\[
p(G|\Theta) = \left[ \frac{2}{\theta_{CA}} e^{-\frac{2}{\theta_{CA}} n_i} \right] \times \left[ \frac{2}{\theta_{BC}} e^{-\frac{2}{\theta_{BC}} n_i} \right] \times \left[ \frac{2}{\theta_{CB}} e^{-\frac{2}{\theta_{CB}} n_i} \right] \times \left[ \frac{2}{\theta_{CR}} e^{-\frac{2}{\theta_{CR}} n_i} \right] \times \left[ \frac{2}{\theta_{RC}} e^{-\frac{2}{\theta_{RC}} n_i} \right] \times \left[ \frac{2}{\theta_{RC}} e^{-\frac{2}{\theta_{RC}} n_i} \right].
\]

The five terms in the square brackets correspond to contributions from the five populations: \( A, B, C, S, \) and \( R \), respectively. For species \( A \), the contribution is \( \frac{2}{\theta_{CA}} e^{-\frac{2}{\theta_{CA}} n_i} \), as there is one coalescent event, between \( a_1 \) and \( a_2 \) at time \( t_1 \). In species \( B \), there is no coalescent, so the probability of having no coalescent when there were two sequences during the time periods \( (0, s_1) \) and \( (t_2, s_2) \) is \( e^{-\frac{2}{\theta_{BC}} \left( t_1 + t_2 \right)} \). There is a migration event; hence, the rate \( \frac{4M_{BC}}{\theta_{BC}} \), while the probability that no migration occurs during time periods \( (0, s_1), (s_1, s_2), \) and \( (t_2, s_2) \), when there are 2, 1, and 2 sequences, respectively, is \( \exp \left\{ -\frac{4M_{BC}}{\theta_{BC}} \left[ 2(s_1 + t_2 + s_2) + (t_2 - s_1) \right] \right\} \).

Given the MSC-model and parameters, the gene trees are assumed to be independent among loci, so that the density for all gene trees is a product over the loci. Bayesian estimation of parameters in the model (\( \Theta \)) involves averaging over all possible gene trees for all loci, which is achieved by the MCMC algorithm.

Overview of MCMC Algorithms under the MSC-M Model. We modified the MCMC proposals in ref. 19. We added sliding-window moves to update the migration rates (\( M \)) and the migration times on the gene trees. The subtree-pruning-and-regrafting (SPR) proposal for changing the gene-tree topology (19) was modified to accommodate migration events on the gene tree. We prune off a subtree and regraft it back to the gene-tree backbone by simulating the backward-in-time process of coalescent and migration using the current values of parameters (\( r, \Theta, M, \)). This proposal was originally implemented in MIGRATE (35) and used in G-PHOCS (33). The proposal that updates population sizes (\( \Theta \)) remains unchanged. The mixing move is modified to rescale both migration times and coalescent times on the gene trees, together with the species divergence times (\( r \)).

Note that in this paper, the MSC-M model (including the species phylogeny, the number of migration rate parameters and the direction and populations involved in migration) is fixed. We leave it to future work to implement cross-model algorithms to move in the space of MSC-M models.

We introduce two major changes to the rubber-band algorithm for updating species divergence times (19).

Composite-space algorithm for migration-rate parameters. First, we implement a trans-model MCMC algorithm to deal with the disappearance and reappearance of a migration-rate parameter when species divergence times change in the proposal. This is necessary because migration rate \( M_{ij} \) exists only when populations \( i \) and \( j \) are contemporary and changes in \( r \) may cause contemporary populations to become noncontemporary or vice versa. As a result, a migration rate parameter which exists in the current model may not exist in the proposed model. Consider the species tree (\( (A, B), (C, D) \)) of Fig. 2, with migration rate \( M_{SC} \) from \( S \) to \( C \). This exists only if \( r_S < r_T \), when species \( S \) and \( C \) coexist during the time interval (\( r_S, r_T \)). Otherwise, the two species are not contemporary, and \( M_{SC} \) is not a parameter in the model. The problem also appears when migration occurs in the opposite direction (\( C \to S \)) or in both directions (\( C \leftrightarrow S \)), and in large species trees with migration involving ancestral species.

In effect, the species tree represents two different models, depending on whether \( r_S < r_T \), and the rubber-band algorithm for updating the species divergence time (\( r \)), which is a within-model proposal under MSC with no gene flow, becomes trans-model and trans-dimensional under MSC-M. We combine different strategies of trans-dimensional MCMC algorithms (52–54),
and implement a Carlin-Chib-Green-O’Hagan-metropolized (CCGOm) algorithm. See SI Appendix, Methods and Materials for details.

**Extended rubber-band algorithm for the MSC-M model.** Second, we extend the rubber-band proposal (19) to accommodate migration events on gene trees. In G.PhoCS (33), a simple modification to the algorithm was introduced, whereby the rubber-band proposal is executed ignoring migration events but the proposal is abandoned if the proposed changes to coalescent times are in conflict with the migration events. Such a rejection approach may lead to poor mixing of the MCMC when there are many migration events on the gene trees as may happen at high migration rates or in large datasets with many loci. Here, we instead extend the rubber-band proposal to accommodate migration events explicitly, avoiding rejection. We identify a time interval \((\tau_S, \tau_T)\) affected by the proposal and populations in the time interval affected by migration, and then rescale affected migration times in the time interval according to the rubber-band algorithm (SI Appendix, Fig. S1). See SI Appendix, Methods and Materials for details.

We implemented the rejection algorithm (33) as well. Both the rejection and extended rubber-band algorithms showed good mixing when there are not many migration events on the gene trees across all loci, as in the analysis of small 100-loci datasets from the Anopheles genomes (Fig. 6A). However, in analyses of large datasets of Anopheles chromosomal arms, migration events were common, and the rejection algorithm mixed poorly due to frequent rejections, as seen from large differences among replicate runs (SI Appendix, Fig. S2). The extended rubber-band algorithm had better mixing properties for these data, producing consistent results among runs (SI Appendix, Fig. S3). Trace plots and posterior summaries for the two algorithms are shown for one dataset with 2223 coding loci on chromosome 2L1 (SI Appendix, Figs. S4 and S5).

**Bayesian Simulation to Validate the MCMC Algorithms.** Extensive tests have been conducted to validate our MCMC implementation (51). If no sequence data are used or if the likelihood is always set to 1, the MCMC algorithm should be sampling from the prior distribution, which is either known or analytically tractable. This test was effective in revealing most errors during the debugging stage.

Bayesian simulation was used to validate the algorithm more rigorously. Parameters are sampled from the prior and used to simulate each replicate dataset under the likelihood model, which is then analyzed using the same prior to generate the posterior. The average posterior over replicate datasets should then match the prior (55). For the test to be most effective, the datasets should be sufficiently large to allow the posterior for each dataset to be influenced by both the prior and the likelihood (data) and thus to differ among datasets, but small enough to allow inexpensive computation. Bayesian simulation tests both the simulation and inference components of the program.

The first MSC-M model used in our Bayesian simulation assumes one migration event from \(C \rightarrow S\) on the species tree for three species of SI Appendix, Fig. S6A. Each dataset consists of \(L = 250\) loci, with \(S = 4\) sequences per species per locus, and \(N = 500\) sites in the sequence. We included G.PhoCS in the test. For BPP, we observed a close match between the prior and the average posterior for all eight parameters (SI Appendix, Fig. S6B). While the posterior differed among replicate datasets (SI Appendix, Fig. S6C), as the datasets were generated by using different parameter values and influenced by random sampling errors due to the finite data size, the average posterior over replicate datasets matched the priors, as expected.

G.PhoCS is an extension of an earlier version of BPP (19, 56) and uses a different parametrization of the MSC model, so that the two programs implement the same likelihood model, but use different priors. We sampled parameter values from the priors used in G.PhoCS to simulate replicate datasets, and analyzed them using G.PhoCS. See SI Appendix, Methods and Materials. There was a close match between the prior and the average posterior for all parameters (SI Appendix, Fig. S6 D and E).

The second MSC-M model used was a saturated model for three species with eight migration rates (SI Appendix, Fig. S7A). This is parameter-rich, and the rates for ancestral migration \((M_{SC} \text{ and } M_{CS})\) are particularly challenging to estimate. We used \(L = 1,000\) loci, with \(S = 4\) sequences per species per locus and \(N = 500\) sites in the sequence. For BPP, we observed a close match between the prior and the average posterior for all 15 parameters (SI Appendix, Fig. S7).

There was great disparity in information content among the parameters in the model: species divergence times \((\tau)\) and population sizes \((\theta)\) were very well estimated with sharp posteriors, whereas the posterior of migration rates was diffuse (SI Appendix, Fig. S7C). The six migration rates involving extant species were more precisely estimated than the two migration rates involving the ancestral species \(S\). Also \(M_{CS}\) was better estimated than \(M_{SC}\), with sharper posteriors, apparently because twice as many sequences (from \(A\) and \(B\)) reach node \(S\) in the species tree as sequences from \(C\) reaching \(T\) in lineage \(C\), when we trace the genealogy of the sampled sequences backward in time. While the simulation validates the program, datasets with 1,000 loci were not informative enough to estimate all eight migration rates with high precision. We did not use G.PhoCS in this test under the saturated model due to the computational cost, but instead analyzed one dataset of 2,000 loci for comparison with BPP (see below).

**Simulation under a Saturated MSC-M Model for Three Species.** To examine the statistical performance of our method, we simulated replicate datasets under the saturated MSC-M model of figure 3a using increasing numbers of loci \((L = 250, 1,000\) and 4,000 loci) with \(S = 4\) sequences per species per locus. We have two objectives: i) to confirm the correctness of our implementation—indicated by the convergence of Bayesian estimates of parameters to their true values with increasing

---

Fig. 2. A balanced tree for four species with migration from \(S \rightarrow C\) at the rate \(M_{SC}\). The migration rate parameter \(M_{SC}\) exists when \((A) \tau_S < \tau_T\) but disappears when \((B) \tau_S \geq \tau_T\).

MSC simulates each replicate dataset under the likelihood model, which is then analyzed using the same prior to generate the posterior. The average posterior over replicate datasets should then match the prior (55). For the test to be most effective, the datasets should be sufficiently large to allow the posterior for each dataset to be influenced by both the prior and the likelihood (data) and thus to differ among datasets, but small enough to allow inexpensive computation. Bayesian simulation tests both the simulation and inference components of the program.

The first MSC-M model used in our Bayesian simulation assumes one migration event from \(C \rightarrow S\) on the species tree for three species of SI Appendix, Fig. S6A. Each dataset consists of \(L = 250\) loci, with \(S = 4\) sequences per species per locus, and \(N = 500\) sites in the sequence. We included G.PhoCS in the test. For BPP, we observed a close match between the prior and the average posterior for all eight parameters (SI Appendix, Fig. S6B). While the posterior differed among replicate datasets (SI Appendix, Fig. S6C), as the datasets were generated by using different parameter values and influenced by random sampling errors due to the finite data size, the average posterior over replicate datasets matched the priors, as expected.

G.PhoCS is an extension of an earlier version of BPP (19, 56) and uses a different parametrization of the MSC model, so that the two programs implement the same likelihood model, but use different priors. We sampled parameter values from the priors used in G.PhoCS to simulate replicate datasets, and analyzed them using G.PhoCS. See SI Appendix, Methods and Materials. There was a close match between the prior and the average posterior for all parameters (SI Appendix, Fig. S6 D and E).

The second MSC-M model used was a saturated model for three species with eight migration rates (SI Appendix, Fig. S7A). This is parameter-rich, and the rates for ancestral migration \((M_{SC} \text{ and } M_{CS})\) are particularly challenging to estimate. We used \(L = 1,000\) loci, with \(S = 4\) sequences per species per locus and \(N = 500\) sites in the sequence. For BPP, we observed a close match between the prior and the average posterior for all 15 parameters (SI Appendix, Fig. S7).

There was great disparity in information content among the parameters in the model: species divergence times \((\tau)\) and population sizes \((\theta)\) were very well estimated with sharp posteriors, whereas the posterior of migration rates was diffuse (SI Appendix, Fig. S7C). The six migration rates involving extant species were more precisely estimated than the two migration rates involving the ancestral species \(S\). Also \(M_{CS}\) was better estimated than \(M_{SC}\), with sharper posteriors, apparently because twice as many sequences (from \(A\) and \(B\)) reach node \(S\) in the species tree as sequences from \(C\) reaching \(T\) in lineage \(C\), when we trace the genealogy of the sampled sequences backward in time. While the simulation validates the program, datasets with 1,000 loci were not informative enough to estimate all eight migration rates with high precision. We did not use G.PhoCS in this test under the saturated model due to the computational cost, but instead analyzed one dataset of 2,000 loci for comparison with BPP (see below).

**Simulation under a Saturated MSC-M Model for Three Species.** To examine the statistical performance of our method, we simulated replicate datasets under the saturated MSC-M model of figure 3a using increasing numbers of loci \((L = 250, 1,000\) and 4,000 loci) with \(S = 4\) sequences per species per locus. We have two objectives: i) to confirm the correctness of our implementation—indicated by the convergence of Bayesian estimates of parameters to their true values with increasing
numbers of loci, and ii) to address the practical question of whether typical genomic datasets contain enough information to allow reliable estimation of the eight migration rates (and other parameters in the model).

The 95% highest probability density (HPD) credible intervals (CIs) for parameters included the true values and became narrower with the increase of data size (Fig. 3B), as expected from the consistency of Bayesian estimation. The divergence times ($\tau$) and population sizes ($\Theta$) were very well estimated, but the migration rates (especially $M_{SC}$ and $M_{CS}$) involved large uncertainties even with $L = 4,000$ loci. For further confirmation, we simulated a large dataset of $16,000$ loci and found that the estimates of migration rates became much more precise (SI Appendix, Table S1). Overall, the results (Fig. 3B and SI Appendix, Table S1) suggest the correctness of our MCMC implementation. They also suggest that genomic datasets with $>10^4$ loci may contain sufficient information to allow precise and accurate estimation of all migration rates in the saturated model (as well as species divergence times and population sizes).

We analyzed a dataset of $2,000$ loci simulated under the saturated model extensively using BPP and G-PHOCS (SI Appendix, Table S1). Estimates of population sizes were similar between the two programs, but large differences existed in estimates of species divergence times and migration rates. Relative to BPP estimates and to true parameter values, G-PHOCS estimates of $\tau_R$ and $\tau_S$, and of $M_{SC}$ and $M_{CS}$, were too small while those of $\theta_C$ were too large (SI Appendix, Table S1). The differences do not appear to be due to the minor differences between the priors used by the two programs or to mixing issues of the MCMC algorithms in G-PHOCS, and instead suggest that implementation of the MSC-M model in G-PHOCS was not correct.

Simulation under an MSC-M Model with Three Species and Two Migration Rates: Comparison with iMa3. We compared BPP with iMa3 by analyzing datasets simulated under the MSC-M models of Fig. 4 A–C on a species tree for three species. As iMa3 assumes bidirectional migration, each model has a pair of migration rates. We generated 10 replicate datasets under the JC mutation model, each consisting of $L = 500$ loci, $S = 4$ sequences per species per locus, and $N = 500$ sites in the sequence. The results are presented in Fig. 4 D–F.

The BPP results are as expected, with most of the 95% CIs bracketing the true values (Fig. 4 D–F). Migration rates between sister lineages are harder to estimate than between nonsister lineages, and those involving ancestral species are harder than those between extant species. Indeed, the estimates in model B ($B$–$C$ migration) had the narrowest CIs while those in model C ($S$–$C$ migration) had the widest (Fig. 4 D–F).

Overall, iMa3 produced similar posterior results to BPP (Fig. 4 D–F). The correlation between the point estimates (posterior means) from the two programs was 0.997, 1.000, and 0.984, for models A, B, and C, respectively. Under models A and B, the estimates were particularly similar, although iMa3 appeared to overestimate $\theta_S$ and underestimate $\tau_S$ slightly. Under the more challenging model C (C–S migration), most $\theta$ parameters were well estimated by iMa3, but $\tau_S$ was underestimated with the CIs excluding the true value, and the migration rates were underestimated. These “biases” do not appear to be due to mixing issues or to reflect the impact of the prior and instead indicate implementation problems. Note that the prior mean for $\tau_R$ was 0.024, larger than the true value 0.02, while the prior means of $M_{CS}$ and $M_{SC}$ were equal to the true values (SI Appendix, Methods and Materials).

We conclude that iMa3 and BPP produced very similar results under simple migration models with migration involving extant species (Fig. 4 D and E), while BPP was more reliable under challenging models with migration involving ancestral species (Fig. 4F). Also BPP had a computational advantage in large datasets. Note that iMa3 includes cross-model moves that change the species phylogeny (57), whereas the model is fixed in BPP and G-PHOCS.

Simulation under the Stepping-stone and Island Models: Comparison with MIGRATE. Our BPP implementation of the MSC-M model may also be used to perform inference under classical population genetic models of subdivision with migration. Models of population subdivision are typically applied to different populations of the same species and do not incorporate a phylogeny for the populations. They may be viewed as a special case of the MSC-M model with ancient species divergences (e.g., with $\tau \to \infty$). In other words, if the probability is essentially 100% that all sequences sampled from the extant species have coalesced or reached their most recent common ancestor (MRCA) before the time of the most recent species divergence (with time running backward), the two classes of models will be equivalent. The MSC-M model is thus an extension of the structured coalescent model to incorporate a species/population phylogeny.

We conducted a simulation under the stepping-stone and island models (Fig. 5 A and B) and analyzed the data using both BPP and MIGRATE (58). Very large divergence times were used to simulate data so that sequences sampled from all species at any locus coalesce with near certainty before reaching any species divergence event. Under these conditions, we expect BPP estimates
of species divergence times to be very large (and uncertain) and the posterior distribution of ancestral population sizes to match the prior, while migration rates and populations sizes for extant species should be reliably estimated.

Parameter estimates from BPP under the stepping-stone model are summarized in Fig. 5C and SI Appendix, Table S2. As expected, the posteriors for $\theta$ for ancestral populations were nearly the same as the prior. There appeared to be a minimal amount of information about species divergence times ($\tau$) as the CIs became narrower with more loci. The estimates were smaller than the true values, due to the influence of the priors. Our interest is in estimation of population sizes for extant species ($\theta_A, \theta_B, \theta_C, \theta_D$) and migration rates ($M$). These were well estimated, with the posterior means fluctuating around the true values (Fig. 5C) and with the CI becoming narrower with an increase in the number of loci ($L$). The 95% HPD CIs matched the large-sample expectation that quadrupling the number of loci halves the CI width (SI Appendix, Table S2). We note that estimates of migration rates involved considerable uncertainty even in large datasets of $L = 4,000$ loci.

The results under the island model (Fig. 5D and SI Appendix, Table S2) similarly suggest little information in the data concerning the species divergence times ($\tau_A, \tau_S, \tau_T$) and ancestral population sizes ($\theta_A, \theta_C, \theta_T$), but population sizes for extant species ($\theta_A, \theta_B, \theta_C, \theta_D$) were well estimated, as were the migration rates ($M_{AB}, M_{AC}, M_{AD}$).

For comparison, we used MIGRATE (58) to analyze the small datasets of $L = 250$ loci. For the stepping-stone model, MIGRATE estimates $4 \theta$s and $6 M$ parameters, while BPP estimates $7 \theta$s, $3 \tau$s, and $6 M$ rates. We focus on the shared parameters. While both programs use the same definitions of divergence times ($\tau$) and population sizes ($\Theta$), MIGRATE uses the mutation-scaled migration rate, which is $\hat{M}_{ij} = \frac{4 M_{ij}}{\theta_i}$ in the notation of this paper. We assigned gamma priors on $\theta_j$ similar to priors used in the BPP analysis (SI Appendix, Methods and Materials). For easy comparison with BPP, we then converted the MIGRATE estimates of migration rates into $\hat{M}_{ij} = \hat{\theta}_j \hat{\theta}_i / 4$, using the posterior mean $\hat{\theta}_j$. Estimates of migration rates under the stepping-stone model were very similar between the two programs (Fig. 5E), although the MIGRATE estimates had slightly wider CIs. MIGRATE estimates of $\theta$ for extant species were too small with wide CIs, compared with the BPP estimates and with the true values. For the island model, MIGRATE estimates $4 \theta$s and $3$ migration rates ($M$), while BPP estimates $7 \theta$s, $3 \tau$s, and $3$ migration rates. Estimates of migration rates were similar between the two programs (Fig. 5F). MIGRATE estimates of $\theta_B, \theta_C$, and $\theta_D$ were centered around the true values but had wider CIs than the BPP estimates. MIGRATE estimates of $\theta_A$ were much too small relative to BPP estimates or the true values (Fig. 5F).

The results suggest problems with the MIGRATE implementation of the stepping-stone and island models. MIGRATE does not write the sampled parameter values into a disk file but collects them into predefined bins based on the priors, and the resulting histograms are then smoothed to estimate the posterior probability densities and to calculate the posterior means and HPD intervals. This may cause inaccurate posterior summaries if
Fig. 5. (A) Stepping-stone and (B) island models used for simulating data, for analysis by using BPP and MIGRATE. The population genetic models (Top) are approximately by the MSC-M models with very large divergence times (Bottom). All population sizes are $\theta = 0.002$ except that $\theta_A = 10\theta$ in the island model. The migration rate was $M = 0.15$ migrants per generation. (C and D) Posterior means and 95% HPD CIs of parameters in BPP analyses of 10 replicate datasets (each of $L = 250, 1,000$, or 4,000 loci) simulated under the models. The horizontal lines represent the true values. (E and F) The datasets of 250 loci are also analyzed using MIGRATE (red), in comparison with BPP (blue). MIGRATE uses the mutation-scaled migration rate, $m_{ij} = 4M_{ij}/j$, in the notation here; this is transformed to $\hat{m}_{ij} = \hat{m}_i\hat{m}_j/4$ by using the posterior mean $\hat{\theta}$.

the number of bins is small and if the prior and the posterior are very different. We used a large number of bins (3,000 or 10,000) and found the results to be stable. Similarly, we rule out issues in kernel density smoothing and differences in reparametrizations and the prior as the main reasons for the differences, because the estimated migration rates were similar between the two programs and the large differences were in the population sizes for extant species ($\theta_A - \theta_D$), which should be easy to estimate (59).

**Analysis of Genomic Data from the *A. gambiae* Mosquitoes.** We used the MSC-M model as well as the MSC-I model (21) to analyze the coding and noncoding data from six species of African mosquitoes in the *A. gambiae* species complex. The MSC-I model constructed in ref. 43 includes two introgression events (Fig. 6B), which are replaced by migration to form an MSC-M model (Fig. 6A).

**Bayesian test of gene flow using blocks of 100 loci.** First, we analyzed blocks of 100 loci to test for the presence of gene flow (Fig. 7). We calculated Bayes factors using thermodynamic integration with Gaussian quadrature (42, 60) to compare three models of gene flow:

- $H_0$: MSC with no gene flow;
- $H_1$: MSC + $A \rightarrow GC$ gene flow;
- $H_2$: MSC + $A \rightarrow GC$ and $R \rightarrow Q$ gene flow.

$B_{20}(m)$ and $B_{20}(i)$ are Bayes factors in support of $H_2$ over $H_0$, with migration under the MSC-M and introgression under the MSC-I, respectively (Fig. 7). We considered $B > 100$ or $B < 0.01$ (or $|\log B| > 4.6$) to be “significant”; this is similar to a 1% “significance level.” There is strong evidence for gene flow as $B_{20} > 100$ in every block and under both MSC-M and MSC-I models. In most blocks, $B_{20}(i) > B_{20}(m)$, with the introgression model fitting the data better than the migration model.
Fig. 6. (A) MSC-M and (B) MSC-I models for six species of African mosquitoes in the A. gambiae species complex: A. gambiae (G), A. coluzzii (C), A. arabiensis (A), A. melas (L), A. merus (R), and A. quadriannulatus (Q).

Variable rates of gene flow across blocks of 100 loci. Next, we examine the estimated migration rates. \( M_{A\rightarrow GC} \) and \( M_{R\rightarrow Q} \) under MSC-M varied among the blocks or across the genome, as did the corresponding introgression probabilities under MSC-I (SI Appendix, Fig. S8). There was overall consistency between the coding and noncoding data. \( M_{I\rightarrow GC} \) was high (>0.1) in most blocks except those on 2L1 and 3L1. \( M_{R\rightarrow Q} \) was low for most blocks except those from 3La and 3R. Estimates of \( \varphi_{A\rightarrow GC} \) and \( \varphi_{R\rightarrow Q} \) under MSC-I were similar to those in figure 6 of ref. 21, where inverse-gamma priors were used for \( \tau \) and \( \theta \). Here, we used gamma priors and assumed the same \( \theta \) before and after each introgression event for each branch on the species tree. Both introgression probabilities varied considerably across the genome, with \( \varphi_{A\rightarrow GC} \approx 100\% \) in most blocks.

As the evidence for \( A \rightarrow GC \) gene flow was overwhelming, we focus on the \( R \rightarrow Q \) gene flow under \( H_2 \). There was no good correspondence between the estimated migration rate \( M_{R\rightarrow Q} \) and introgression probability \( \varphi_{R\rightarrow Q} \) (SI Appendix, Fig. S9). The MSC-M model predicts that the probability that any sequence \( Q \) is traced to population \( R \) to be

\[
\varphi_{R,Q} = 1 - e^{-4M_{RQ}R_d/\theta_Q}.
\]

If MSC-M is the true model, Eq. 4 is expected to give an upper bound for the estimate when the data are analyzed under MSC-I: \( \phi < \varphi_0 \) ([26]; figure 1). However, our estimates displayed very poor matches between \( \varphi_{R,Q} \) under MSC-M and \( \varphi_{R,Q} \) under MSC-I (SI Appendix, Fig. S10), even though estimates of \( R_d \) and \( \theta_Q \) were very similar under the two models (SI Appendix, Fig. S11). The reasons for this lack of correspondence are unclear.

Fig. 7. The logarithm of the Bayes factor for testing gene flow obtained from iPP analysis of the 100-loci blocks of the (A) coding and (B) noncoding data from the *Anopheles* mosquitoes (Fig. 6A), calculated using thermodynamic integration with Gaussian quadrature (42). Gene flow is accounted for using either the introgression model (i for MSC-I) or the migration model (m for MSC-M). Model \( H_0 \) is the model with no gene flow. Model \( H_1 \) assumes the \( A \rightarrow GC \) gene flow, with rate \( M_{A\rightarrow GC} \) under MSC-M or \( \varphi_{A\rightarrow GC} \) under MSC-I. Model \( H_2 \) accommodates both gene-flow events, with rates \( M_{A\rightarrow GC} \) and \( M_{R\rightarrow Q} \) under MSC-M or \( \varphi_{A\rightarrow GC} \) and \( \varphi_{R\rightarrow Q} \) under MSC-I. Bayes factor \( B_{21} \) measures the support for \( H_2 \) over \( H_0 \), while \( B_{21} \) measures the support for \( H_2 \) over \( H_1 \). The test is significant when \( |\log B| > 4.6 \) (i.e., if \( B > 100 \) or \( B < 0.01 \)). For example, \( \log B_{21} < -4.6 \) means that the data strongly support the one-rate model with the \( A \rightarrow GC \) migration (with rate \( M_{A\rightarrow GC} \)) over the two-rates model with both the \( A \rightarrow GC \) and the \( R \rightarrow Q \) migrations. Different scales are used for the y-axis over the intervals (−50, −10), (−10, 10), (10, 50), (50, 700).
One possibility is that the MSC-M model (Fig. 6A) is a poor fit to the genomic data: The rate of gene flow might vary over time, but on average, the MSC-I model assuming a pulse of gene flow was closer to reality than the MSC-M model assuming a constant rate over the whole time period. In other words, introgression events detected from the genome data may be largely historical. Estimated frequencies of \( A. gambiae \times A. arabiensis \) F1 hybrids for modern species where the two species are sympatric were low, around 0.15–0.22% ([61] and references therein).

While the MSC-M and MSC-I models make very different assumptions about the mode of gene flow, they produced highly similar estimates of species divergence times \((t_o, t_a, t_e, t_r)\) (SI Appendix, Fig. S11). The results are consistent with the simulation study of ref. 26, which found that species divergence times were well estimated when the mode of gene flow was misspecified. In contrast, the MSC model of no gene flow seriously underestimated divergence times, as found in simulations (62).

**Parameter estimation for chromosomal arms.** Finally, we analyzed all loci for each chromosomal arm as one dataset. The posterior means and 95% HPD CIs for the migration rates are in Table 1, while species divergence times and migration rates are summarized in SI Appendix, Fig. S12 A–D. We have used those large datasets to illustrate mixing properties of the rejection and extended rubber-band algorithms in SI Appendix, Figs. S2–S5.

Similarly to the analyses of the 100-loci blocks, migration rates and introgression probabilities varied considerably among chromosomal arms. \( \hat{M}_{A\rightarrow GC} \) was high for all chromosomal arms, with the smallest being \( \sim 0.2 \) for 3L1 (coding), while \( \hat{\phi}_{A\rightarrow GC} > 0.9 \) for all chromosomal arms except 2La (coding and noncoding) and 2L2 (coding). Note that the \( A\rightarrow GC \) gene flow is so prevalent for the autosomes that the predominant autosomal gene tree has a different topology from the species phylogeny (2, 43). The \( R\rightarrow Q \) gene flow mostly affected 3La and 3L2, while 2L2 and 3R were affected to a lesser extent.

Estimates of species divergence times were highly consistent among the chromosomal arms and between the coding and noncoding data (SI Appendix, Figs. S12 A–D and S13). Most estimates (in particular, those from the noncoding data) had tight CIs because of the large data sizes, although \( t_r \) had wider CI bars as the estimates were affected by the estimated rate of \( A \rightarrow GC \) gene flow.

## Discussion

Models of population subdivision and migration developed in population genetics are special cases of the MSC-M model. The stepping-stone and island models (Fig. 5 A and B) are instances of the MSC-M model with population divergence times approaching \( \infty \). Our results (Fig. 5) suggest that BPP is an efficient and reliable implementation of such population genetic models, allowing them to be fitted to genomic data. We note that other specialized models may also be special cases of the MSC-M model. For example, the isolation-with-initial-migration (IIM) model assumes that migration occurred initially after species divergence but stopped at a certain time point, for example, when reproductive isolation is fully established (63–65). In the secondary contact (SC) model (65), two species initially experienced complete isolation after divergence but came into contact at a certain time point, with subsequent ongoing migration. Both IIM and SC models can be fitted to genomic data as instances of the MSC-M model by including an unsampled ghost species (26).

The MSC-M model extends models of population subdivision to incorporate a population/species phylogeny (16, 66, 67). Besides improving the biological realism of the model, this extension also opens up opportunities for addressing many important questions in evolutionary biology, such as detecting gene flow during and after speciation, delineating species boundaries in the presence of gene flow, inferring historical demographic changes or estimating population sizes for extinct ancestral species, detecting gene flow from extant species that may and may not have extant descendants (16). Likelihood-based implementations of the MSC-M model have involved heavy computation and are impractical for genome-scale data of thousands of loci, although large genomic datasets are routinely generated and are indeed necessary for precise and accurate estimation of the rate of gene flow. Furthermore, it is challenging to implement the model

### Table 1. Bayesian estimates of migration rates (M) and of introgression probabilities (ϕ) from the *Anopheles* genomic data (Fig. 6)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Loci</th>
<th>( \hat{M}_{A\rightarrow GC} )</th>
<th>95% CI</th>
<th>( \hat{\phi}_{A\rightarrow GC} )</th>
<th>95% CI</th>
<th>( \hat{M}_{R\rightarrow Q} )</th>
<th>95% CI</th>
<th>( \hat{\phi}_{R\rightarrow Q} )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>2L1 coding</td>
<td>2,223</td>
<td>0.404 (0.366, 0.443)</td>
<td>0.955 (0.933, 0.975)</td>
<td></td>
<td></td>
<td>0.002 (0.000, 0.003)</td>
<td>0.029 (0.016, 0.043)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2L1 noncoding</td>
<td>4,133</td>
<td>0.311 (0.293, 0.329)</td>
<td>0.963 (0.950, 0.975)</td>
<td></td>
<td></td>
<td>0.000 (0.000, 0.001)</td>
<td>0.016 (0.008, 0.024)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2La coding</td>
<td>2,776</td>
<td>2.451 (2.122, 2.789)</td>
<td>0.791 (0.768, 0.813)</td>
<td></td>
<td></td>
<td>0.005 (0.002, 0.008)</td>
<td>0.038 (0.006, 0.074)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2La noncoding</td>
<td>6,732</td>
<td>2.289 (2.116, 2.466)</td>
<td>0.696 (0.684, 0.708)</td>
<td></td>
<td></td>
<td>0.001 (0.000, 0.001)</td>
<td>0.015 (0.007, 0.022)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2L2 coding</td>
<td>1,362</td>
<td>1.053 (0.874, 1.233)</td>
<td>0.879 (0.847, 0.910)</td>
<td></td>
<td></td>
<td>0.030 (0.020, 0.041)</td>
<td>0.180 (0.134, 0.228)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2L2 noncoding</td>
<td>2,330</td>
<td>0.618 (0.565, 0.672)</td>
<td>0.955 (0.936, 0.974)</td>
<td></td>
<td></td>
<td>0.008 (0.005, 0.012)</td>
<td>0.074 (0.056, 0.091)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2R coding</td>
<td>6,849</td>
<td>0.909 (0.844, 0.977)</td>
<td>0.971 (0.962, 0.979)</td>
<td></td>
<td></td>
<td>0.010 (0.008, 0.013)</td>
<td>0.074 (0.063, 0.085)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2R noncoding</td>
<td>17,027</td>
<td>0.739 (0.712, 0.771)</td>
<td>0.978 (0.974, 0.982)</td>
<td></td>
<td></td>
<td>0.003 (0.002, 0.003)</td>
<td>0.047 (0.042, 0.052)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3L1 coding</td>
<td>983</td>
<td>0.215 (0.189, 0.242)</td>
<td>0.967 (0.948, 0.985)</td>
<td></td>
<td></td>
<td>0.003 (0.001, 0.006)</td>
<td>0.058 (0.034, 0.084)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3L1 noncoding</td>
<td>2,496</td>
<td>0.234 (0.218, 0.249)</td>
<td>0.976 (0.965, 0.987)</td>
<td></td>
<td></td>
<td>0.001 (0.000, 0.002)</td>
<td>0.033 (0.019, 0.047)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3La coding</td>
<td>1,998</td>
<td>1.708 (1.454, 1.971)</td>
<td>0.929 (0.914, 0.945)</td>
<td></td>
<td></td>
<td>0.153 (0.131, 0.176)</td>
<td>0.600 (0.569, 0.631)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3La noncoding</td>
<td>6,208</td>
<td>1.399 (1.299, 1.498)</td>
<td>0.973 (0.968, 0.978)</td>
<td></td>
<td></td>
<td>0.083 (0.077, 0.090)</td>
<td>0.619 (0.604, 0.634)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3L2 coding</td>
<td>764</td>
<td>1.577 (1.261, 1.913)</td>
<td>0.923 (0.896, 0.948)</td>
<td></td>
<td></td>
<td>0.043 (0.029, 0.057)</td>
<td>0.306 (0.234, 0.378)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3L2 noncoding</td>
<td>1,823</td>
<td>2.003 (1.700, 2.300)</td>
<td>0.951 (0.937, 0.964)</td>
<td></td>
<td></td>
<td>0.012 (0.008, 0.017)</td>
<td>0.161 (0.131, 0.192)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3R coding</td>
<td>4,977</td>
<td>0.788 (0.727, 0.853)</td>
<td>0.939 (0.927, 0.952)</td>
<td></td>
<td></td>
<td>0.028 (0.023, 0.034)</td>
<td>0.168 (0.149, 0.188)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3R noncoding</td>
<td>14,323</td>
<td>0.636 (0.612, 0.663)</td>
<td>0.959 (0.953, 0.965)</td>
<td></td>
<td></td>
<td>0.012 (0.011, 0.014)</td>
<td>0.103 (0.095, 0.111)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
correctly: While both G-PHOCS (33) and MIGRATE (58) have undergone extensive testing and validation, our simulation and test suggest errors in implementation.

Here, our stringent tests using Bayesian simulation and conventional methods for evaluating mixing and statistical properties of estimates suggest that the MCMC algorithms in BPP are correctly sampling from the posterior and that our algorithms have improved mixing. While it is computationally demanding, BPP has been applied to datasets of >10,000 loci (Table 1, SI Appendix, Table S1). We suggest that our implementation of the MSC-M model in BPP provides a useful tool for comparative analysis of genomic data to infer gene flow between divergent species or populations and a platform for engineering further algorithmic improvements.

Materials and Methods

Detailed descriptions of algorithms, simulation conditions, and analyses of simulated and Anopheles data are in online SI Appendix, Supplemental Text.


ACKNOWLEDGMENTS. We thank Drs. Ian Gronau, Peter Beerli, and Jody Hey for help with G-PHOCS, MIGRATE, and IMa2, respectively. This study has been supported by Biotechnology and Biological Sciences Research Council grants (BB/R003502/1 and BB/R01356X/1) to Z.Y., a Natural Science Foundation of China (NSFC) grant (12101295) and a Guangdong Natural Science Foundation grant (2022A1515101176) to X.J., an NSFC grant (32200409) to J.H., and an NIH Grant (GM123306) to B.R.

Author affiliations: *Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, United Kingdom; †Department of Statistics and Data Science, China Southern University of Science and Technology, Shenzhen 518055, China; ‡Department of Intelligent Medical Engineering, School of Biomedical Engineering, Capital Medical University, Beijing 100069, China; and ‡Department of Evolution and Ecology, University of California, Davis, CA 95616.


