Exact calculation of the joint allele frequency spectrum for generalized isolation with migration models

An Investigation Submitted to the Population and Evolutionary Genetics Section of *Genetics*

Andrew D. Kern* and Jody Hey†

July 20, 2016

*Department of Genetics, Rutgers University, Piscataway, NJ 08554, USA

†Department of Biology, Temple University, Philadephia, PA 19122 USA

Running Head: Exact AFS from IM models

Key Words: Isolation with Migration, AFS, Markov chains, composite likelihood

Corresponding Author:

Andrew D. Kern

Department of Genetics
Rutgers University
604 Alison Rd.
Piscataway, NJ 08854
kern@biology.rutgers.edu

Abstract

Population genomic datasets collected over the past decade have spurred interest in developing methods that can utilize massive numbers of loci for inference of demographic and selective histories of populations. The allele frequency spectrum (AFS) provides a convenient framework for such analysis and accordingly much attention has been paid to predicting theoretical expectations of the AFS under a number of different models. However, to date, exact solutions for the joint AFS of two or more populations under models of migration and divergence have not been found. Here we present a novel Markov chain representation of the coalescent on the state space of the joint AFS that allows for rapid, exact calculation of the joint AFS under generalized isolation with migration (IM) models. In turn, we show how our Markov chain method, in the context of composite likelihood estimation, can be used for accurate inference of parameters of the IM model using SNP data. Lastly, we apply our method to recent whole genome datasets from *Drosophila melanogaster*.

INTRODUCTION

The explosion in availability of genome sequence data brings with it the promise that longstanding questions in evolutionary biology might now be answered. In particular, understanding the forces at work when populations begin to diverge from one another is crucial to our understanding of the process of speciation. Population genomic sampling of multiple individuals from closely related populations provides our clearest view of the evolutionary forces at work during divergence, however it remains a challenge as to how best to analyze such massive datasets in a population genetic framework (Sousa and Hey 2013).

A popular model for population divergence is the so-called isolation with migration (IM) model (Wakeley 1996; Nielsen and Wakeley 2001; Hey and Machado 2003), in which a single ancestral population spits into two daughter populations at a given time and the daughter populations then have some degree of geneflow between them. IM models are a convenient framework for statistical estimation of population genetic parameters as the models described by various parameter combinations exist along a continuum between pure isolation after divergence to panmixia among daughter populations. More complex models of divergence, for instance secondary contact after isolation or geneflow that stops after a certain period of time, are also readily modeled in the IM framework. As a result numerous methods are now available from estimation of IM parameters.

Generally there exist two classes of methodology for the estimation of IM model parameters: genealogical samplers which aim to accurately compute the probability of a population sample under the assumption of no recombination within a given locus (e.g. IMa2; Hey and Nielsen 2007; Hey and Nielsen 2004) and methods which make use of the joint allele frequency spectrum (AFS) and assume free recombination between SNPs (e.g. δaδi; Gutenkunst et al. 2009). While genealogical samplers yield maximum likelihood or Bayesian estimates of population parameters, they become somewhat unwieldy for use with genome-scale data, due to the assumption of no recombination. Thus with the enormous increase in population genomic data from both model and non-model systems, much recent effort has been devoted to AFS based approaches that rely upon composite likelihood estimation (Gutenkunst et al. 2009; Naduvilezhath et al. 2011; Lukić and Hey 2012; Excoffier et al. 2013).

Estimation methods based on the joint AFS between populations rely upon calculating the probability of an observed AFS given the vector of parameters that describe the population history. The method for calculation of this expected AFS is thus central, and varies between competing methods. For instance, Gutenkunst et al. (2009) took the approach of numerically solving a diffusion approximation to the population allele frequency spectrum, whereas more recent methods of demographic inference rely upon coalescent simulation to estimate the expected sampled AFS (Naduvilezhath et al. 2011; Excoffier et al. 2013). While both of these approaches have been shown to be reliable for demographic inferences under many parameterizations, both are approximate and may contain error to various degrees across parameter space.

Here, we introduce a method for exact calculation of the joint AFS under generalized two-population IM models. Our method uses a coalescent Markov chain approach that is defined on the state space of the AFS itself. Using this newly defined state space, in combination with the rich mathematical toolbox of Markov chains, we are able to compute the expected AFS of a given IM model for moderate sample sizes. We compare our coalescent Markov chain calculations of the AFS to diffusion approximations and that obtained via simulation. Further, using simulation we show how our approach can be used for accurate inference of demographic parameters. Lastly we apply our software package implementing the method, IM_CLAM, to population genomic data from *Drosophila melanogaster*.

MODEL

Here we present a strategy for exact calculation of the joint AFS under the IM model, and the subsequent inference of its associated parameters, that relies upon both discrete time and continuous time Markov chains (DTMC and CTMC respectively). In outline our approach involves first enumerating the complete state space associated with a given configuration of samples from two populations (i.e. sample sizes), followed by construction of a transition matrix to be used for a DTMC (or the analogous CTMC), and finally through the use of standard Markov chain techniques, the calculation of the implied joint AFS. For reasons that will become clear below, we begin by describing how one would calculate the exact joint AFS from a two population island model, before moving on to the full-blown IM model.

A markov chain on the state space of the joint AFS: The first step in our approach requires the complete enumeration of the state space associated with our Markov chains given a sample configuration. The state space we describe is on the space of the allele frequency spectrum. That is to say that each state of our model is implies a unique contribution to the joint AFS of the model in question. To track the allele frequency contribution implied by each state we will track the number of descendent leaf lineages in each population that each gene copy present is ancestral to. We will need to track this quantity independently for each population to deal with migration. To introduce our state space consider a sample that consists of one allele for population 1 and one allele from population 2, and let n_1 and n_2 be the sample sizes such that $n_1 = n_2 = 1$ (Figure 1). Although this is a trivially small case, it is adequate for accurately describing the form of the state space. Our initial state (i.e. the configuration at the time of sampling), call it A_0, looks like the following

$$A_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

where the left and right matrices represent the state in populations 1 and 2 respectively, and the entry at i, j represents the number of gene copies ancestral to i sampled alleles in population 1 and j sampled alleles in population 2. By convention these state space matrices are zero indexed, and there will never be a non-zero value at the position $(0, 0)$ as the models does not track lineages that are not ancestral to the sample. The initial state A_0 indicates that there is a single allele in population 1 that is ancestral to one of the sampled gene copies from population 2 and a single allele in population 2 that is ancestral to one of the sampled gene copies from population 2. Moving back in time in Figure 1 the first event is a migration event from population 1 to population 2. Thus in state A_1 the matrix representing population 2 now has two alleles, one of which is ancestral to a single allele in population 1 and the other which is ancestral to a single allele in population 2. Further notice that the left hand matrix, representing population 1, is empty. Finally to two alleles coalesce to find the MRCA in population 2, as indicated in state A_3.

To enumerate the complete state space associated with a given sample configuration (n_1,n_2), we use a recursive approach that considers all possible coalescent and migration moves among present gene copies to exhaustively find all possible states, including MRCA states that will represent the absorbing states of our Markov chain. Note that in this two population island model only two absorbing states are possible– the MRCA could be found in population 1 or it could be found in population 2. In the case of $n_1 = n_2 = 1$ as shown in Figure 1, there are a total of 6 possible states however the number of states grows extremely quickly with increasing sample size (See Appendix). For instance when $n_1 = n_2 = 2$ there are 46 possible states, and $n_1 = n_2 = 3$ there are 268 states. Figure 2 shows how the state space grows in sample size, and while growth is sub-exponential it clearly explodes for larger samples.

Markov chain transition matrix: Having defined the state space we next consider the form of the transition matrix associated with the DTMC. Transitions between states in our coalescent markov chain depend both on parameters of the model (e.g. population sizes, migration rates) and on the combinatoric probability involved in the chain move. For instance let n_i be the number of active lineages in population i within a state of the chain, let x represent the multiplicity of a specific lineage at the current state, and y be the number of gene copies involved in the move to the next state. Further let N_i be the population size of population i, and the coalescent rate be $C_i = n_i(n_i - 1)/4N_i$. Then the probability of a coalescent event in population 1 that moves the chain from state A_i at time t to A_j at time $t+1$ would be

$$Prob(x(t) = A_i | x(t+1) = A_j) = \frac{\prod \binom{x}{y}}{\binom{n_1}{2}} \times \frac{n_i(n_i - 1)}{4N_i}$$

where the product in the numerator of the first term is over each lineage involved in the move (maximum of two different terms). Here the first term represents the combinatorics involved in the move and the second, parameter-dependent term corresponding to the type of move (either coalescent or migration).

In the case of a migration event from population i to j the terms of the transition matrix take the form

$$Prob(x(t) = A_i | x(t+1) = A_j) = \frac{x}{n_i} \times n_i M_i$$

where M_i is the migration rate from population i scaled by effective population size such that $M_i = 4N_i m$ where m is the fraction of the focal population made up of migrant individuals each generation.

Turning our attention back to the case of $n_1 = n_2 = 1$, whose complete state space is given in the appendix, the transition matrix associated with the DTMC, call it P, would be

$$P = \begin{pmatrix} 0 & M_1 & 0 & M_2 & 0 & 0 \\ M_2 & 0 & M_2 & 0 & 0 & \frac{1}{2N_2} \\ 0 & M_1 & 0 & M_2 & 0 & 0 \\ M_1 & 0 & M_1 & 0 & \frac{1}{2N_1} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

where now each matrix entry P_{ij} is scaled such that each row sums to one such that $\sum_j P_{ij} = 1$ The P matrix also implies an analogous CTMC transition matrix, call it Z, whose rows are constrained such that $\sum_j Z_{ij} = 0$. With these transition matrices in hand we now turn attention to computing the SFS of the island (or IM) model.

Calculating the AFS As said above, each state implies an associated contribution to the allele frequency spectrum. Let F represent the joint AFS from a two population sample. F will be matrix valued of size n_1+1 rows and n_2+1 columns, where n_1 and n_2 are the sample sizes from populations 1 and 2 respectively. Entries of F, F_{ij}, will be the number of SNPs sampled with i derived alleles in population 1 and j derived alleles in population 2. To map a given state A_i to its contribution to F, we need only ask how long the system stays in a given state and then add that amount of time to each of the corresponding cells of F from the non-zero entries in both the right and left hand matrices of the A_{ith} state.

We can use the tools of Markov chains to then perform the two calculations needed to exactly calculate the AFS under a given model: 1) calculate the expected number of times each state is visited before absorption (i.e. reaching the MRCA), and 2) calculate the expected length of time the chain is in each state to compute the AFS. The latter calculation is simply the exponentially distributed wait time under the coalescent with migration, which itself is a function of the number of gene copies active in a given state, population sizes, and migrations rates.

Calculating the expected number of visits to each state is move involved. We can rearrange our transition matrix P into what is called "canonical form". We assume that P has r absorbing states and t transient states, such that

$$P = \begin{pmatrix} Q & R \\ 0 & I_r \end{pmatrix}$$

where Q is a $t \times t$ matrix, R is a $t \times r$ matrix, and I_r is the identity matrix of rank r (Kemeny and Snell 1976). Using this factorization we can next compute the fundamental matrix of our Markov chain, N, by using the relationship

$$N = (I_t - Q)^{-1} \tag{1}$$

where the entries of N_{ij} represent the expected number of visits to state j given the chain started at state i, and I_t is a rank t identity matrix. It is important to note that this calculation will thus require the inversion of a potentially very large matrix, thus complicating our implementation. For the calculation of the island model however, we are only interested in one row of N, as the starting state is known with certainty (i.e. the observed sample), so this is readily solved. Also, note that N gives us the expected number of visits to each state by the DTMC until absorption (i.e. the MRCA). For the island model this describes the complete stochastic process as in that case we are dealing with a time homogenous process. For models with changes in population size or populations splitting we will need to consider different "phases" of the demographic history separately, as the transition rates through the system, or indeed even the state space of the system will change moving back in time.

Returning for a moment to the island model then, having calculated N we are ready to compute the expected AFS. As we said before, the expected AFS will simply be the sum of the products of the number of visits to each state and the length of time spent in each state. For the island model in the case where $n_1 = n_2 = 1$ there will be 6 terms in the summation to find F, one for each state.

Isolation with Migration To calculate the AFS for the IM model, we calculate the contributions to the AFS from two sources: that of the island model phase of the model prior to divergence (looking back in time), and the contribution to the AFS from the single, ancestral population (see Figure 3). The contribution to the AFS from the island model portion, call it F_I, can be computed by first calculating the total AFS from the island model from time zero to absorption, F_{tot}, and then subtracting off the portion of the AFS contributed from the population divergence time, T_{div}, until absorption (e.g. Wakeley and Hey 1997). Let the vector $\pi(t)$ be the probability of being in each state of our Markov chain at time t. We need to calculate $\pi(T_{div})$ both to find F_I and to figure out where our system begins the single population phase of the IM model. We use a CTMC representation of our same transition matrix from the island model (denoted Z) to compute $\pi(T_{div})$ using the matrix exponential such that

$$\pi(T_{div}) = \pi(0)e^{T_{div}Z}. \tag{2}$$

With $\pi(T_{div})$ in hand, we then can use the fundamental matrix of the island model, N, to compute the number of visits to each state conditional on starting in each state at T_{div} with probability $\pi(T_{div})$ as $N_g = \pi(T_{div})N$, where N_g is subscripted g in reference to the fact that these represent "ghost visits," unseen in the actually IM model. F_I then can simply be calculated as $F_I = F_{tot} - F_g$, where F_g is the AFS implied by N_g.

Once we have the contribution to the AFS from the island phase, F_I, there is only one portion remaining– the contribution to to the AFS from the single population, ancestral phase, call it F_A (Figure 3). To compute this we map the state space of the island model onto a reduced state space of a single population model, use that mapping to fold $\pi(T_{div})$ to the state space size, and then compute a new DTMC transition matrix for the single population phase, changing population size as necessary and removing migration. With the new transition matrix we can compute the fundamental matrix for the ancestral phase, N_A, and from that its contribution to the AFS, F_A. Finally the AFS for the complete IM model, F_{IM}, is equal to the combined sums of the AFS contributions from the two phases such that $F_{IM} = F_I + F_A$.

IMPLEMENTATION

Our strategy for computing the AFS from the IM model relies upon taking the inverse of two large, sparse matrices, corresponding to functions of the transition matrix from the DTMC, and exponentiating one matrix. Such calculations are extremely expensive computationally, so in our implementation of this method we have used parallel, scalable algorithms where ever possible. Our software package, IM_CLAM, performs these calculations with help from two open source packages, the CSPARSE library (Davis 2006) and the PETSc package (Balay et al. 1997; Balay et al. 2015a; Balay et al. 2015b). In particular we use PETSc to distribute all sparse matrix calculations across a parallel compute environment that uses MPI. For matrix inversion, we compute row by row of the inverse matrix using a direct solver from CSPARSE and distribute those solves across cores. The matrix exponential is calculated using the Krylov subspace method as implemented in the SLEPc add-on to the PETSc package (Hernandez et al. 2005). IM_CLAM and its associated open source code are available for download from GitHub (`https://github.com/kern-lab/im_clam`).

APPLICATION TO *DROSOPHILA MELANOGASTER* DATA

We apply our method to recent whole genome sequencing projects from *Drosophila melanogaster* in which two population samples, one from North America (North Carolina) (Mackay et al. 2012) and a second from Africa (Zambia) (Lack et al. 2015) have been sequenced to good depth. We obtained aligned datasets from the Drosophila Genome Nexus resource (v1.0; Lack et al. 2015), and subsequently filtered from those alignments regions that showed strong identity-by-decent (IBD) and admixture using scripts provided with the alignments. This yielded sample sizes of $n = 197$ genomes from African lines and $n = 205$ genomes from N. American lines. The joint AFS was then constructed, using alignments to *D. simulans* and *D. yakuba* to determined the derived and ancestral allele at a given SNP. Tri-allelic positions were ignored. In an effort to sample the AFS from regions of the genome that should be less likely to affected by linked selection, we only examined intergenic regions that were at least $5kb$ away from genes, and that did not contain simple repeats, repeat masked regions, annotated transcription factor binding sites, or annotated regulatory elements. This yielded 5530 regions of the genome with a total length of 4.43Mb. From this we constructed a joint AFS that we then downsampled to a smaller size ($n = 6$ African and $n = 6$ N. American alleles) to allow for calculation using IM_clam.

RESULTS

Simulation We first set out to compare the expected AFS calculated with IM_CLAM versus that calculated from coalescent simulations. As our calculations result in the exact AFS, we were interested in comparing the convergence of the simulated AFS to the true AFS as a function of the number of simulations. In Figure 4 we show the mean percentage error of the AFS computed from simulating a given number of independent genealogies, shorthanded on the axis label as number of SNPs. In Figure 4 the AFS was computed using $n_1 = n_2 = 6$, a symmetric migration of rate $m_{12} = m_{21} = 1.0$, and a divergence time of $t_{div} = 0.25$. As the number of simulated genealogies increases the mean percentage error between the simulated AFS and that calculated by IM_CLAM drops quickly. However after 10^6 simulations the amount of Monte Carlo error plateaus at approximately 0.3% and then decays very slowly even after 10^9 simulations. Thus brute force simulation of the AFS seems ill advised for IM models, as it will be computationally quite expensive to converge to the correct distribution of allele frequencies, although approximately correct calculation could be done with considerably fewer simulations.

We next turned our attention to comparing our exact AFS to that computed by the popular software package $\partial a \partial i$ (Gutenkunst et al. 2009). $\partial a \partial i$ uses diffusion approximations to model the joint AFS among two populations and thus itself may be susceptible to a certain amount of error for given parameterizations. We compared our exact AFS to that generated from $\partial a \partial i$ under a range of migration rates, $m = \{0, 1, 5, 10\}$, and having fixed population sizes to 1.0 and $t_{div} = 0.5$. Figure 5 shows the element wise percentage error for the $\partial a \partial i$

approximation of this comparison. $\partial a \partial i$ harbors an appreciable amount of error under these parameters, particular at the corners of the matrix, that represent fixed differences among populations. Thus while $\partial a \partial i$ has been shown to be accurate for use in inference, we can see here that the expected AFS produced using the diffusion approximation still strays from the true value.

As a result of this discrepancy we set out to compare the accuracy of inference using IM_CLAM in comparison to $\partial a \partial i$. Our goal here is not to perform an exhaustive comparison between methods, as IM_CLAM is much more limited in scope than $\partial a \partial i$, however we wish to show that our method has utility for parameter inference as well. For this we generated 100 replicate simulated AFS draws using coalescent simulations in a manner as to simulate a large number of independent SNPs. Again we set $n_1 = n_2 = 6$, a symmetric migration of rate $m_{12} = m_{21} = 1.0$, but here we used a divergence time of $t_{div} = 0.1$. We set a low per locus θ, $\theta = 0.001$, and generated 10^6 genealogies. This yielded approximately 3.54×10^5 SNPs per simulated AFS sample. With these simulated datasets we then set out to infer the parameters of the IM model. Figure 6 shows boxplot plots of parameter estimates for both IM_CLAM and $\partial a \partial i$. In general both methods are accurate for this parameterization however it can be seen that a minority of optimizations using $\partial a \partial i$ yielded outlier parameter estimates (note y-axis on Figure 6). It is worth considering that both methods are using the BFGS (Broyden-Fletcher-Goldfarb-Shanno; (Press 1985)) algorithm for optimization, set with the same stopping criterion and bounds on the parameter space explored, thus failed optimization alone seems an unlikely explanation. Indeed similar behavior for $\partial a \partial i$ was observed in an earlier report (Naduvilezhath et al. 2011) although the range of parameters considered in that paper was quite wide.

Application to *Drosophila melanogaster* data The demographic history of *Drosophila melanogaster* in many ways mirrors that of human populations. *Drosophila melanogaster* is commonly thought to have had its origins in sub-Saharan Africa, and have spread out of

log likelihood	N_{NA}	N_{AF}	N_{Anc}	t_{div}	$mig_{N \to A}$	$mig_{A \to N}$
-250022.2608	2.04E+05	1.38E+06	4.28E+05	16980.29048	1.318418	0.876434

Table 1: Point estimates of demographic parameters from *D. melanogaster* populations. The divergence time t_{div} is given in years assuming 15 generations per year.

Africa approximately between 10,000-20,000 years ago (Lachaise et al. 1988; David and Capy 1988; Begun and Aquadro 1993; Li and Stephan 2006). *D. melanogaster* seems to have first migrated to Europe and Asia via the middle east, presumably as a human commensal, and then only much later did it arrive in North America (Lachaise et al. 1988). Here we model the demography of North American and African populations, thus the divergence time we will capture will be from the initial split between African and out-of-African lineages. Using a downsampled joint AFS from these population samples we estimated IM model parameters using IM_CLAM. Point estimates of the population sizes, migration rates, divergence time, and the optimized model likelihood are given in table 1. These estimates are scaled in the number of individuals for population sizes and the number of years for divergence time by assuming a mutation rate per base per generation of $u = 5.49 \times 10^{-9}$ and 15 years per generation. Multiple runs of IM_CLAM on this dataset from different starting points yielded similar point estimates and likelihoods.

Our estimates show that while current day Zambian effective population size has grown $\sim 3.2x$ larger then the ancestral population size, the North American population is quite a bit smaller, undoubtedly due to the strong and potentially repeated bottlenecks it has experienced in its history. Moreover our estimates indicate a good deal of continued gene flow between African and North America in both directions. Finally our estimate of the date of divergence between the lineage leading to North America and that leading to Africa was 16,980 years ago, well in line with previous estimates. Figure 7 shows the close correspondence between the inferred SFS and that predicted from our estimation.

DISCUSSION

Population genetic inference of demographic history has become an increasingly important goal for modern genomics, as the impacts of demography on patterns of genetic variation is now appreciated to directly impair our ability to identify causative disease variation via linkage (e.g. Rogers 2014) as well as shape the genetic architecture of phenotypic variation within populations (Lohmueller 2014; Simons et al. 2014). Moreover, our understanding of human prehistory has been revolutionized in recent years through demographic inference using population genetic data (e.g Botigué et al. 2013; Ralph and Coop 2013; Raghavan et al. 2015; Poznik et al. 2016). While that is so, methods that efficiently utilize whole genome information for inferring rich demographic histories, particularly multiple population histories, still lag behind the huge availability of data (Sousa and Hey 2013). Accordingly, much recent effort has focused on using the joint allele frequency spectrum of samples drawn from multiple populations as a way to summarize genome-wide data for demographic inference (Gutenkunst et al. 2009; Naduvilezhath et al. 2011; Lukić et al. 2011; Lukić and Hey 2012; Excoffier et al. 2013; Kamm et al. 2015).

In this study we present a novel method for calculating the exact joint site frequency spectrum expected from two population Isolation with Migration models. Our method relies upon a Markov chain representation of the coalescent, in which the state space of the chain is the joint AFS at a given point in time. Through the use of this state space, in conjunction with standard Markov chain techniques, we are able to numerically calculate our expected AFS. Our method stands in contrast to other popular techniques that either use diffusion approximations (Gutenkunst et al. 2009; Lukić et al. 2011) or direct Monte carlo simulation (Excoffier et al. 2013) to estimate the expected AFS under a given parameterization. Indeed, as we have shown, estimation of the AFS via diffusion or Monte carlo simulation can lead to persistent error and in some cases numerical instability (see Kamm et al. 2015). While we here use a Markov chain approach to calculate the exact AFS under generalized IM models, a recent, elegant paper by Kamm *et al.* (2015) presented analytic solutions and associated algorithms for computing the exact AFS for multiple population models with arbitrary population size histories but without continuous migration.

We have implemented our approach in a software package called IM_CLAM that allows for inference of generalized IM models using genome-wide joint AFS data by computing the exact AFS. As we have shown above with simulated data, IM_CLAM is quite accurate in its inference of population parameters. Application of IM_CLAM to population genomic data from *Drosophila melanogaster* sampled from North America and Africa recovers point estimates of population sizes and divergence time that are well in line with earlier estimates based on much smaller datasets (Li and Stephan 2006); we show that the North American populations are smaller than both current day African populations and the ancestral population from which both lineages are drawn, with African effective population size roughly 6.7x larger than North American effective population size. Moreover we find that the African population has experienced considerable growth since the divergence of the two populations, as it is now 3.2x larger than the ancestral size, while the North American population has yet to recover from the bottlenecking associated with its establishment. Indeed, larger sample sizes might show evidence of stronger population growth that what we have found here. Our estimated divergence time of 16,980 years ago is again consistent with earlier estimates based on small numbers of loci (Li and Stephan 2006; Duchen et al. 2013). Finally we estimate that there is considerable geneflow in both directions between African and North American populations, with the rate of migration to Africa being approximately 1.5x higher than in the other direction. Approximate Bayesian model selection by Duchen *et al.* (2013) supported a model of admixture between European and African populations in the founding of the North American population over symmetric migration models. While this is so, there is also considerable evidence for strong geneflow from North America to African populations (Pool et al. 2012). Taken on balance, although we are not modeling an admixture event directly here, our migration rate estimates seem reasonable and probably reflect time averaged geneflow between these populations in accordance with their complex history.

While the ability to compute the exact AFS under generalized IM models using our Markov chain approach is an advance, there are many short comings to our methodology. Perhaps most challenging is the fact that the state space of our Markov chain grows nearly exponentially in sample size (fig. 2). This means that our approach is only computationally feasible for smaller sample sizes, as in the current state space the transition matrix associated with larger sample sizes will be too large to represent in memory, even when sparse matrix representations are used as we have done here. While this is so the state space of the Markov chain could potentially be reduced in size if by exploiting lumpability among states (*cf.* Andersen et al. 2014). Even at moderate sizes the computational costs of the matrix inversion and exponentiation needed by our method are still high, thus IM_CLAM needs tens or hundreds of CPUs for optimization runs to complete within hours rather than days. Currently our implementation is quite limited in that it only handles the two population IM case with constant population sizes, and from that model produces only point estimates. Estimation of confidence intervals via the Godambe information matrix which has been utilized recently to provide appropriate interval estimation under composite likelihood (Coffman et al. 2016) should also be possible in this setting.

Despite the computational difficulties associated with the Markov chain approach described here, our method has opened a new avenue in calculating the likelihoods associated with AFS data and might be amenable to other population genetic problems. For instance, in the model presented above we consider the two dimensions of the state space matrices to represent different populations. It is simple to conceive of this dimension as instead two separate loci with recombination acting to make transitions among the numbers of alleles that are ancestral at one or both loci. In this way we have been able to write down a Markov chain that enables calculation of the two-locus allele frequency spectrum that itself might be useful for estimation of demographic parameters and recombination rates.

ACKNOWLEDGMENTS

ADK and JH were supported by NIH R01GM078204. We thank Dan Schrider for suggestion the name IM_CLAM, and Yun Song for helpful conversations about this effort.

LITERATURE CITED

Andersen, L. N., T. Mailund, and A. Hobolth (2014). Efficient computation in the im model. *Journal of mathematical biology 68*(6), 1423–1451.

Balay, S., S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, K. Rupp, B. F. Smith, S. Zampini, and H. Zhang (2015a). PETSc users manual. Technical Report ANL-95/11 - Revision 3.6, Argonne National Laboratory.

Balay, S., S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, K. Rupp, B. F. Smith, S. Zampini, and H. Zhang (2015b). PETSc Web page. `http://www.mcs.anl.gov/petsc`.

Balay, S., W. D. Gropp, L. C. McInnes, and B. F. Smith (1997). Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen (Eds.), *Modern Software Tools in Scientific Computing*, pp. 163–202. Birkhäuser Press.

Begun, D. J. and C. F. Aquadro (1993). African and north american populations of drosophila melanogaster are very different at the dna level.

Botigué, L. R., B. M. Henn, S. Gravel, B. K. Maples, C. R. Gignoux, E. Corona, G. Atzmon, E. Burns, H. Ostrer, C. Flores, et al. (2013). Gene flow from north africa contributes to differential human genetic diversity in southern europe. *Proceedings of the National Academy of Sciences 110*(29), 11791–11796.

Coffman, A. J., P. H. Hsieh, S. Gravel, and R. N. Gutenkunst (2016). Computationally efficient composite likelihood statistics for demographic inference. *Molecular biology and evolution 33*(2), 591–593.

David, J. R. and P. Capy (1988, Apr). Genetic variation of drosophila melanogaster natural populations. *Trends Genet 4*(4), 106–11.

Davis, T. A. (2006). *Direct methods for sparse linear systems*, Volume 2. Siam.

Duchen, P., D. Živković, S. Hutter, W. Stephan, and S. Laurent (2013). Demographic inference reveals african and european admixture in the north american drosophila melanogaster population. *Genetics 193*(1), 291–301.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll (2013). Robust demographic inference from genomic and snp data. *PLoS Genet 9*(10), e1003905.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet 5*(10), e1000695.

Hernandez, V., J. E. Roman, and V. Vidal (2005). SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Trans. Math. Software 31*(3), 351–362.

Hey, J. and C. A. Machado (2003). The study of structured populationsnew hope for a difficult and divided science. *Nature Reviews Genetics 4*(7), 535–543.

Hey, J. and R. Nielsen (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics 167*(2), 747–760.

Hey, J. and R. Nielsen (2007). Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proceedings of the National Academy of Sciences 104*(8), 2785–2790.

Kamm, J. A., J. Terhorst, and Y. S. Song (2015). Efficient computation of the joint sample frequency spectra for multiple populations. *arXiv preprint arXiv:1503.01133*.

Kemeny, J. G. and J. L. Snell (1976). Finite markov chains. undergraduate texts in mathematics.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner (1988). Historical biogeography of the drosophila melanogaster species subgroup. In *Evolutionary biology*, pp. 159–225. Springer.

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool (2015). The drosophila genome nexus: a population genomic resource of 623 drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics 199*(4), 1229–1241.

Li, H. and W. Stephan (2006). Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS Genet 2*(10), e166.

Lohmueller, K. E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet 10*(5), e1004379.

Lukić, S. and J. Hey (2012). Demographic inference using spectral methods on snp data, with an analysis of the human out-of-africa expansion. *Genetics 192*(2), 619–639.

Lukić, S., J. Hey, and K. Chen (2011, Jun). Non-equilibrium allele frequency spectra via spectral methods. *Theor Popul Biol 79*(4), 203–19.

Mackay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, et al. (2012). The drosophila melanogaster genetic reference panel. *Nature 482*(7384), 173–178.

Naduvilezhath, L., L. E. Rose, and D. Metzler (2011). Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Molecular Ecology 20*(13), 2709–2723.

Nielsen, R. and J. Wakeley (2001). Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics 158*(2), 885–896.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau,

P. Duchen, J. Emerson, P. Saelao, D. J. Begun, et al. (2012). Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. *PLoS Genet 8*(12), e1003080.

Poznik, G. D., Y. Xue, F. L. Mendez, T. F. Willems, A. Massaia, M. A. W. Sayres, Q. Ayub, S. A. McCarthy, A. Narechania, S. Kashin, et al. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide y-chromosome sequences. *Nature genetics*.

Press, W. H. (1985). *Numerical Recipes: The Art of Scientific Computing. Example Diskette (Pascal).* Cambridge University Press.

Raghavan, M., M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M. C. Ávila-Arcos, A.-S. Malaspinas, et al. (2015). Genomic evidence for the pleistocene and recent population history of native americans. *Science 349*(6250), aab3884.

Ralph, P. and G. Coop (2013). The geography of recent genetic ancestry across europe. *PLoS Biol 11*(5), e1001555.

Rogers, A. R. (2014). How population growth affects linkage disequilibrium. *Genetics 197*(4), 1329–1341.

Simons, Y. B., M. C. Turchin, J. K. Pritchard, and G. Sella (2014). The deleterious mutation load is insensitive to recent population history. *Nature genetics 46*(3), 220.

Sousa, V. and J. Hey (2013, Jun). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet 14*(6), 404–14.

Wakeley, J. (1996). Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical population biology 49*(3), 369–386.

Wakeley, J. and J. Hey (1997). Estimating ancestral population parameters. *Genetics 145*(3), 847–855.

APPENDIX

The complete state space for a sample of configuration $n_1 = n_2 = 1$ is given below. The ordering of states shown is arbitrary but identical to the one used in the example markov chain transition matrix in the Model section of the paper.

$$A_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$A_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$A_5 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

FIGURES

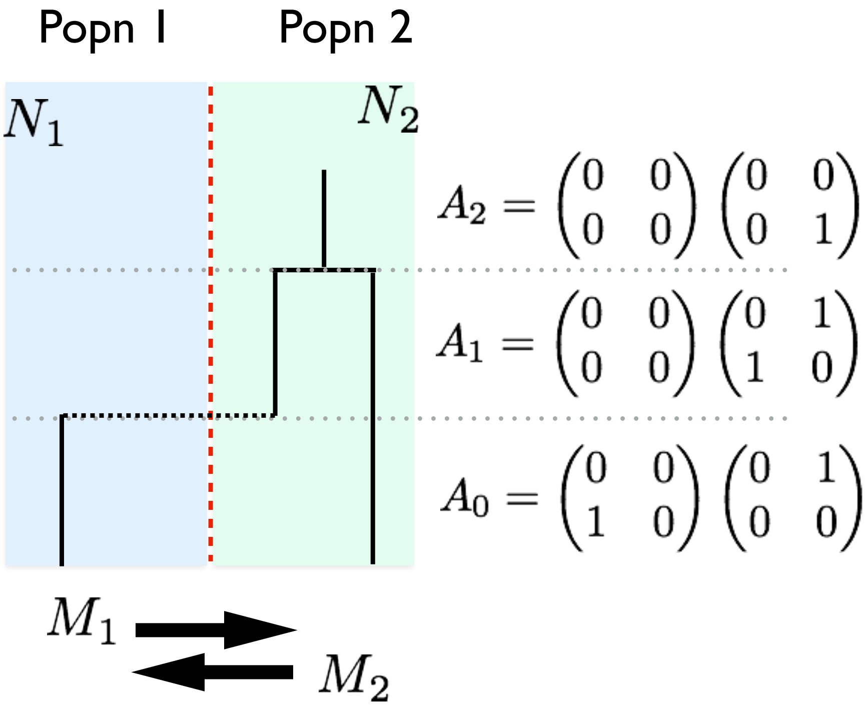
Popn 1
Popn 2
N_1
N_2
$A_2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$
$A_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
$A_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$
M_1
M_2

Figure 1: **Representative State Space for a two population Island Model.** Shown here is an example of a two population island model with sample sizes $n_1 = 1$ and $n_2 = 2$. The model has two population sizes, N_1 and N_2, and two migration rates, m_1 and m_2. The representative state space at each phase in the coalescent tree is shown to the right. package.

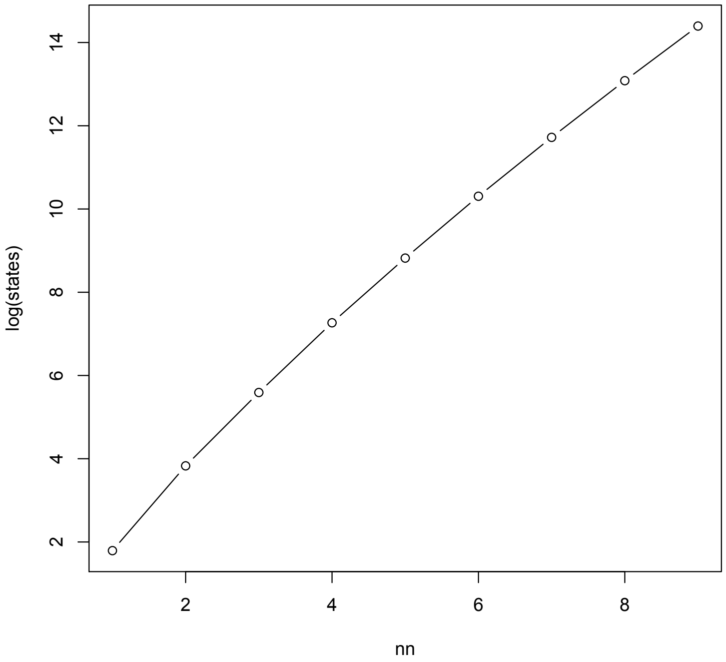

Figure 2: **State space expansion as function of sample size.** Here we show how the state space grows as function of sample size considering symmetric sampling such that $n_1 = n_2$. Note that the y-axis is shown on a low scale. package.

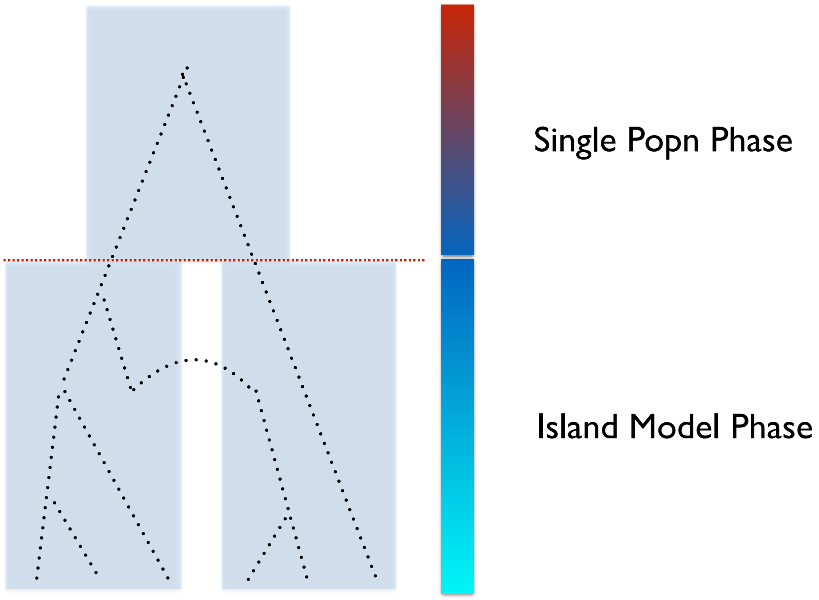

Figure 3: **Two phases of the IM Model.** Here we illustrate the two phases of the IM model, the first of which is an island model phase and the second, the ancestral single population phase. To compute the expected AFS of the IM model

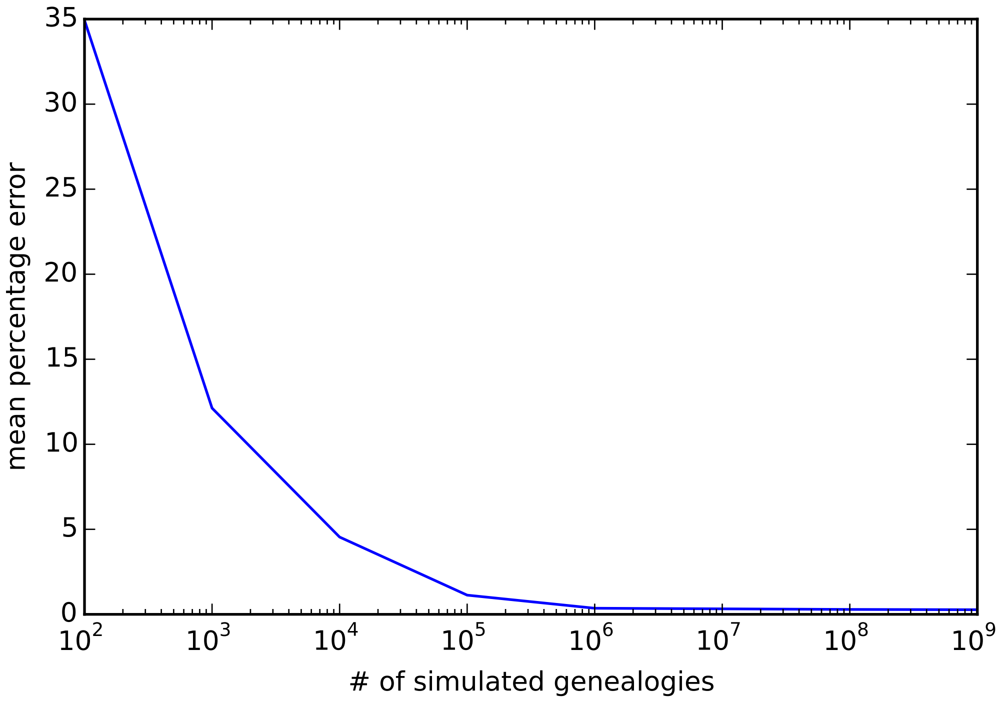

Figure 4: **Monte Carlo error in simulations of the allele frequency spectrum.** Here we show the decline in the mean percentage error in estimates of the joint AFS from simulations where we vary the number of independent coalescent genealogies simulated in comparison to our exact solution.

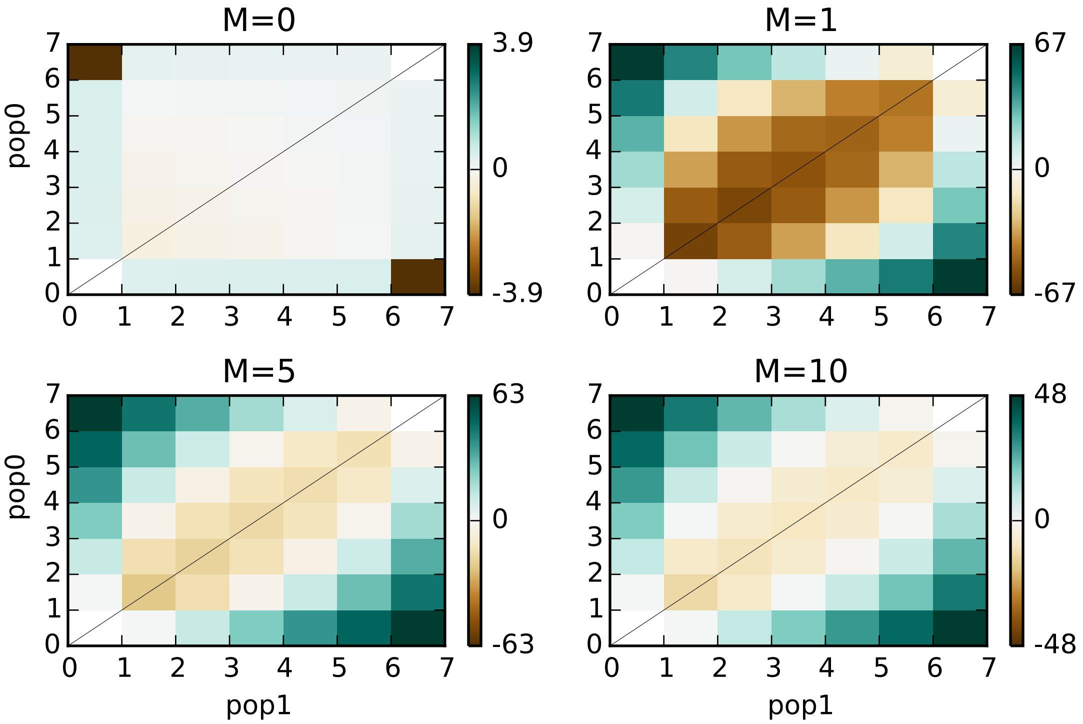

Figure 5: **Percent deviation of expected AFS calculated from $\partial a \partial i$.** Clockwise from top left panel we show the percent deviation of each cell in the expected AFS for four different symmetric migration rates $m = \{0, 1, 5, 10\}$ from $\partial a \partial i$ versus our exact calculation.

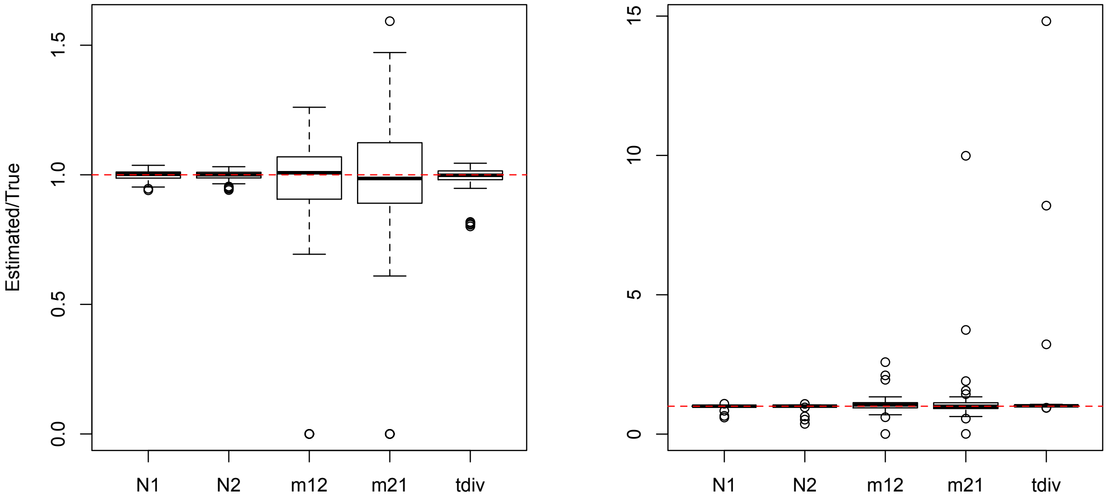

Figure 6: **Accuracy of parameter inference using IM_CLAM and $\partial a \partial i$.** Shown are boxplots of point estimates from 100 replicate simulations with IM_CLAM in the left panel and $\partial a \partial i$ in the right panel. See Methods for simulation parameter values.

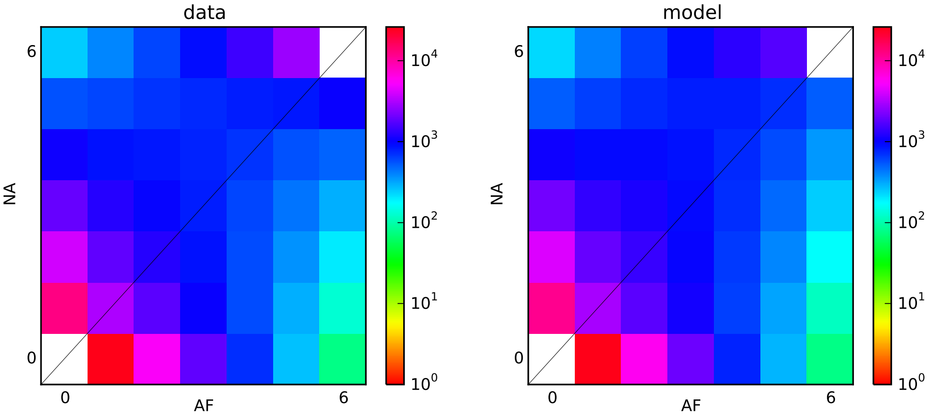

Figure 7: **Comparison of observed joint allele frequency spectrum and that predicted from IM_CLAM from *D. melanogaster* population samples.** The populations are labeled AF for Africa and NA for North America.