

Inference of historical changes in migration rate from the lengths of migrant tracts

John E. Pool¹ and Rasmus Nielsen^{1,2}

¹Department of Integrative Biology, University of California, Berkeley

²Department of Statistics, University of California, Berkeley

Running title: Inferring changes in migration rate

Keywords: migration, demographic inference, admixture, hybridization, *Mus musculus*

Corresponding author:

John E. Pool

Department of Integrative Biology, University of California, Berkeley

3060 Valley Life Sciences Bldg #3140, Berkeley, CA, 94720-3140, United States

Tel: +1-510-642-1233. Fax: +1-510-643-6264

E-mail: jpool@berkeley.edu

Abstract:

After migrant chromosomes enter a population, they are progressively sliced into smaller pieces by recombination. Therefore, the length distribution of “migrant tracts” (chromosome segments with recent migrant ancestry) contains information about historical patterns of migration. Here we introduce a theoretical framework describing the migrant tract length distribution and propose a likelihood inference method to test demographic hypotheses and estimate parameters related to a historical change in migration rate. Applying this method to data from the hybridizing subspecies *Mus musculus domesticus* and *Mus musculus musculus*, we find evidence for an increase in the rate of hybridization. Our findings could indicate an evolutionary trajectory toward fusion rather than speciation in these taxa.

Introduction:

An accurate understanding of population history is essential for such diverse applications as the search for recent signatures of positive selection in population genetic data (e.g. Jensen et al. 2005), the study of admixed human populations to identify disease-associated genetic variants (e.g. Montana et al. 2004, Patterson et al. 2004), and the definition of management units in conservation (Pearse and Crandall 2004). Patterns of genetic variation contain information about past changes in population size (e.g. Cornuet and Luikart 1996; Marth et al. 2004), the timing of population splitting events (e.g. Nielsen and Wakeley 2001), and levels of migration between populations (e.g. Beerli and Felsenstein 2001).

Since the advent of molecular markers, researchers have sought to gauge the genetic differentiation of populations, and to draw conclusions about the level of migration between them. Wright's F_{ST} (Wright 1952) has served as the classic metric of population differentiation, and under ideal conditions, the population migration rate can be estimated by $N_e m = \frac{1 - F_{ST}}{4F_{ST}}$, where N_e is the effective population size, m is the per-generation probability of being a migrant, and $N_e m$ is thus equal to the number of migrants exchanged each generation. However, this relationship relies on several assumptions that may not be valid for most natural populations (reviewed in Whitlock and MacCauley 1999), including that of a constant rate of migration. A given F_{ST} value between two populations could be produced by a constant level of migration over a long period of time, or by genetic drift following a relatively recent split between the two populations, or by recent admixture between historically isolated populations, or by any number of more complex scenarios. The MDIV/IM inference framework (e.g. Nielsen and Wakeley 2001; Hey 2005) offers a way to differentiate ongoing migration between populations from lineage sorting in isolated populations, while estimating relevant demographic parameters.

As in the case of IM, most population genetic methods that estimate demographic parameters assume that all sites or markers under study are either completely linked (no recombination) or completely unlinked (free recombination), though see Becquet and Przeworski (2007). And correspondingly, most population genetic data has been collected with these criteria in mind. Assuming either full linkage among sites or else independence among loci can greatly simplify the task of modeling the histories of molecular markers. However, the bulk of the genome in most organisms consists of DNA that is subject to recombination, and furthermore, the pattern of recombination events within a sample of chromosomes may hold valuable information concerning population history. For example, we know that haplotype statistics (Depaulis *et al.* 2003) and linkage disequilibrium (Wall *et al.* 2002) across short loci are quite sensitive to the effects of population bottlenecks. The recent availability of genome-scale polymorphism data should facilitate investigation of longer-range linkage patterns, which may shed new light on the recent histories of populations.

Patterns of diversity at partially linked markers may be especially informative concerning the historical pattern of migration between populations. Once a migrant chromosome enters a new population, recombination will break it down into progressively shorter segments. The lengths of these “migrant tracts” – or admixture “chunks” (Falush *et al.* 2003) – therefore contain information about how long ago migration occurred. This logic has been utilized to estimate the timing of recent admixture events (e.g. Patterson *et al.* 2004; Hoggart *et al.* 2004; Koopman *et al.* 2007), but its applicability should extend beyond such cases. We suggest that migrant tract lengths are expected to have a certain equilibrium distribution under a constant migration rate model. An excess of long migrant tracts would indicate a recent increase in migration rate, while the opposite pattern would suggest recently reduced gene flow. We use theoretical predictions

and simulations to explore the migrant tract length distribution under a variety of demographic scenarios, and we assess the potential of this approach for inferring demographic parameters related to migration rate changes.

Models and Methods:

Constant migration rate

A large set of different population genetic models converge to the same coalescence process as the population size becomes large ($N \rightarrow \infty$; Kingman 1981a,b). In two-island models (Wright 1931), an ancestral process arises (*e.g.* Hudson 1983) which can be described by a Markov pure jump process $\{X(t), t \geq 0\}$ with state space on $\{0, \dots, n_1\} \times \{0, \dots, n_2\} \setminus (0, 0)$, initial state (n_1, n_2) , absorbing states $(0, 1)$, $(1, 0)$, and transition rates

$$q((i, j) \rightarrow (i-1, j)) = \binom{i}{2} \text{ if } i \geq 2$$

$$q((i, j) \rightarrow (i, j-1)) = \binom{j}{2} \frac{N_1}{N_2} \text{ if } j \geq 2 \quad (1)$$

$$q((i, j) \rightarrow (i-1, j+1)) = N_1 m_{21} i \text{ if } i \geq 1$$

$$q((i, j) \rightarrow (i+1, j-1)) = N_2 m_{12} j \text{ if } j \geq 1$$

where n_1 and n_2 are the sample sizes from population 1 and 2, respectively, and N_1 and N_2 are the population sizes. Migration occurs from population 2 to 1, and from 1 to 2, at rates m_{21} and m_{12} , respectively. Time is measured in units of N_1 generations, and $N_j m_{ij}$ can be interpreted as the proportion of individuals in population j that are replaced with individuals from population i in each generation.

Consider the ancestry of a single lineage from population 1. The waiting time in number of generations until the last migration event for this lineage is exponentially distributed with mean $1/m$ (letting $m = m_{21}$ here and in the following to simplify the notation). We now introduce recombination, and measure distances in the genome as genetic distances. By using genetic distances, we may assume that recombination in each generation occurs according to a Poisson process with rate 1 along the chromosome. We assume that migrant tracts do not recombine together, we disallow back-migration events (*i.e.* assume $m_{12} = 0$), and we ignore the effect of the ends of the chromosome (but later we will evaluate violations of these assumptions). Then, after t generations, the distribution of tracts lengths follows an exponential distribution with mean $1/t$:

$$f(x;t) = te^{-tx}. \quad (2)$$

Because we can only reliably infer migrant tracts over a certain length, we will be interested in the distribution of tracts and the expected proportion of a chromosome in tracts larger than a certain threshold, C . The proportion of a migrant chromosome from time t which is in tracts on a size $> C$, p_C , can be found from the convolution of two independent and identically-distributed exponential random variables with parameter t :

$$E[p_C | t] = 1 - \int_0^C te^{-ty}(1 - e^{-t(C-y)})dy = e^{-tC}(1 + Ct). \quad (3)$$

These two variables represent, respectively, the distance to the left and right on the chromosome from the point of inspection to the nearest recombination event. Integrating over t , we find

$$E[p_C] = \int_0^{\infty} me^{-mt}e^{-tC}(1 + Ct)dt = \frac{m(2C + m)}{(C + m)^2}. \quad (4)$$

The expected number of fragments in the population of a migrant chromosome of length L is

$$E[k(t)] = 1 + Lt \quad (5)$$

after t generations, *i.e.* the contribution of migrant tracts from generation t to the population is proportional to $me^{-mt}(1 + Lt)$. Again ignoring recombination among migrant tracts, the density of tract lengths will be formed as a mixture distribution of tracts from different times

$$f(x) = \frac{\int_0^{\infty} te^{-tx}(1 + Lt)me^{-mt} dt}{\int_0^{\infty} (1 + Lt)me^{-mt} dt} = \frac{m^2(2L + m + x)}{(L + m)(m + x)^3}. \quad (6)$$

The conditional tract length distribution of tracts of a length larger than C is then

$$f(x | x > C) = \frac{f(x)}{\int_C^{\infty} f(x) dx} = \frac{(C + m)^2(2L + m + x)}{(C + L + m)(m + x)^3}. \quad (7)$$

These expressions do allow for genetic drift. However, they assume that recombination events between descendents of the same or different migration event contribute to the breakdown of chromosomes into smaller distinguishable tracts. In practice, we cannot distinguish between non-recombinants and recombinants between the same allele. The approximations we derive here are, therefore, expected to break down when t becomes so large compared to N_1 that migrant alleles may have drifted to appreciably high allele frequencies, thereby allowing for

recombination between migrant tracts. However, this is not a fundamental problem as we can only infer relatively large tracts which, with high probability, are descendents of recent migrants. If C is sufficiently large, it is highly probable that only fragments for which t is small have been sampled. The chance that a migrant allele of size $> C$ has drifted to high frequencies is small if $C \gg 1/N$ (since recombination will break down tracts below this threshold before drift can substantially elevate them in frequency). Problems identifying recombinants between migrant alleles are, therefore, avoidable if C is sufficiently large. For the same reason, for large C , inferences based on Equation (7) should be relatively robust to violations of the assumption of no back-migration, *i.e.* $m_{12} = 0$.

Changes in the migration rate

We will now extend these results to the case where there has been a discrete change in the rate of migration. Again, we only consider migration into population 1, and assume that the current migration rate is m_1 , and that it before T generations ago was m_2 . We then have

$$\begin{aligned}
 E_2[p_C] &= \int_0^T m_1 e^{-m_1 t} e^{-tC} (1 + Ct) dt + e^{-m_1 T} \int_0^\infty m_2 e^{-m_2 t} e^{-(t+T)C} (1 + C(t+T)) dt \\
 &= \frac{m_1(2C + m_1)}{(C + m_1)^2} - \frac{C^2 e^{-(C+m_1)T} (m_1 - m_2)(2C + m_1 + m_2 + (C + m_1)(C + m_2)T)}{(C + m_1)^2 (C + m_2)^2}. \quad (8)
 \end{aligned}$$

Likewise, setting

$$f_2(x) = \frac{\int_0^T te^{-tx}(1+Lt)m_1e^{-m_1t} dt + e^{-m_1T} \int_T^\infty te^{-tx}(1+Lt)m_2e^{-m_2(t-T)} dt}{\int_0^T (1+Lt)m_1e^{-m_1t} dt + e^{-m_1T} \int_T^\infty (1+Lt)m_2e^{-m_2(t-T)} dt} \quad (9)$$

and conditioning like in Equation (7), we find

$$f_2(x | x > C) = \frac{f_2(x)}{\int_C^\infty f_2(x) dx} = e^{T(C-x)}(C+m_1)^2(C+m_2)^2 \times \frac{a-b}{c} \quad (10)$$

where:

$$a = \frac{m_2(m_2+x+T(m_2+x))^2 + L(2+T(m_2+x)(2+T(m_2+x)))}{(m_2+x)^3}$$

$$b = \frac{m_1((m_1+x)(1-e^{T(m_1+x)}+T(m_1+x)) + L(2-2e^{T(m_1+x)}+T(m_1+x)(2+T(m_1+x))))}{(m_1+x)^3}$$

$$c = e^{(C+m_1)T} m_1(C+L+m_1)(C+m_2)^2 - (m_1-m_2)(-Lm_1m_2 + C^3(1+LT) + Cm_1m_2(1+LT) + C^2(L+m_1+m_2+L(m_1+m_2)T)).$$

Inference

We wish to estimate the parameters, m_1 , m_2 , and T from an observed tract length distribution. As only large tracts can be easily identified, we have to base inferences on equations (8) and (10) and not on Equation (9). We define a composite likelihood function by taking the product of Equation (10) among all tracts in the data above a pre-specified threshold (C). The reason why we consider this a composite likelihood function and not a true likelihood function is that the same tract can be counted twice. However, for real data with C large, this will rarely happen and the estimation function is essentially a true likelihood function.

Equation (10) contains only very little information about the overall amount of population subdivision, because we only look at the relative abundance of tracts with length greater than C . However, much of the information regarding the overall level of population subdivision is captured by our estimate of p_C (Equation 8). We therefore do a constrained optimization of the likelihood function subject to the constraint

$$E_2[p_C] = \hat{p}_C, \tag{11}$$

where \hat{p}_C is the observed proportion of the genome in tracts larger than C . Specifically, we rearrange Equation (8) to express T as a function of C , m_1 , m_2 and $E_2[p_C]$, and we then substitute \hat{p}_C for $E_2[p_C]$. We then perform a two-dimensional optimization for m_1 and m_2 while constraining T to take on the value given by the aforementioned equation. This approach reduces the number of parameters from Equation (10) to be estimated (from three to two) and adds information concerning the total proportion of migrant DNA observed (from \hat{p}_C). Constrained models with one of the two migration rates set to zero are evaluated similarly, via a one-dimensional optimization of the other migration rate. For the constant migration rate model, m can be estimated simply by setting $m_1 = m_2$ in Equation (8), thus using \hat{p}_C to solve for m .

Comparison of likelihood scores from different models allows the testing of demographic hypotheses. Test 1 compares the maximum log likelihood score from the migration rate change model (with m_1 and m_2 allowed to vary) against the null hypothesis of a single, constant migration rate (with m inferred from \hat{p}_C). Test 2 compares the maximum log likelihood score of the migration rate change model against a model where either (A) m_1 is constrained to be zero, or (B) m_2 is constrained to be zero. Generally, Test 2A is performed when $m_1 < m_2$, and Test 2B is

performed if $m_1 > m_2$. Because the distributions of likelihood ratios are not well modeled by standard asymptotic theory for any of these tests, critical values are obtained using data simulated under the null hypothesis. For computational reasons, we obtain critical values using $N_e m = 0.1$ for Test 1, the true values of m_2 and T for Test 2A, and true m_1 and T for Test 2B (rather than using the estimated parameter values for each simulated replicate). In the analysis of empirical data, we use the estimated null model parameter values instead.

Simulation

A forward simulation program was written to allow the generation of migrant tract data. This program simulates each chromosome present in two populations, and models the processes of genetic drift, migration, and recombination under a Wright-Fisher model (Fisher 1930; Wright 1931). It does not generate polymorphism data; instead it directly monitors migrant tract status along chromosomes. When an individual migrates, all previously non-migrant chromosome sections become migrant tracts, and any previous migrant tracts become non-migrant. Tracts are “forgotten” when recombination breaks them down to a size below the threshold length. The program initializes with no migrant tracts present, but goes through a “burn-in” period with migration at rate m_2 . For the analyses shown here, using a threshold tract length of $C = 0.5$ cM, the burn-in time was 2000 generations (results and theory indicated this was more than enough time to reach an equilibrium migrant tract length distribution) and N_e was 10,000. At the end of the burn-in, the migration rate switches to m_1 and the program records all migrant tracts present in each population at a series of time points (T) after this change. An extension to this program allows migrant tracts to be sampled from a specific number of individuals. In testing the performance of the likelihood method, we simulated “genomes” containing 35 chromosomes,

each 100 cM in length (3500 cM is close to the genetic map size of humans and many other mammals; Kong *et al.* 2002), and we sampled 100 haploid individuals from one population.

Application to empirical data

The likelihood method was applied to genome-wide single nucleotide polymorphism (SNP) data from two hybridizing subspecies of the house mouse, *Mus musculus domesticus* and *M. m. musculus*. This data was produced by the Wellcome Trust Center for Human Genetics, and is available at <http://www.well.ox.ac.uk/mouse/INBREDS/>. The strains examined here consist of 7 from *M. m. domesticus* and 8 from *M. m. musculus*, with varying geographic origins (see Harr 2006 for a summary). The data come from wild-derived, inbred mouse strains and are effectively haploid. The few apparently heterozygous sites were recoded as missing data, and invariant SNP's were removed. Since the X chromosome is expected to have a different history, all of the 9,935 SNP's analyzed here were autosomal. The vast majority of these SNP's have inferred genetic map positions (Jensen-Seaman *et al.* 2004), and all analyses were done in terms of genetic distance, rather than physical position. These SNP's had been ascertained in laboratory lines of mixed origin and could be biased in terms of diversity levels and allele frequencies (Boursot 2006), but we do not expect a particular bias for the inference and analysis of migrant tracts.

In general, our likelihood inference method allows the user to decide how migrant tracts should be defined. The sample sizes of the mouse SNP data set seemed too small for published methods for identifying ancestry along recombining chromosomes (e.g. Falush *et al.* 2003). However, the task of tract identification is simplified by the high level of genetic differentiation between the two subspecies, which diverged perhaps 1 million generations ago and show very

high levels of genetic differentiation (Salcedo *et al.* 2007; Baines and Harr 2007). We were therefore able to use a very simple set of criteria for defining migrant tracts in this data. Given the small sample sizes, an individual's SNP allele was deemed to provide evidence for a migrant tract only if it was otherwise absent from the individual's subspecies, but present in the in other subspecies (we call this a "positive SNP"). If an individual's SNP allele is otherwise present in both subspecies, this is a "neutral SNP" neither favoring nor opposing migrant tract status. And if an individual's SNP allele is not present in the other subspecies, it is taken as evidence against migrant tract status (a "negative SNP"). Migrant tracts consisted of two or more positive SNP's with no negative SNP's between them. The minimum tract length was considered to be the genetic distance spanning only the positive SNP's at each end of the tract. The maximum tract length included all sites up to the first negative SNP's flanking the tract.

Given the minimum and maximum length of a migrant tract, we were interested in estimating how far beyond the positive SNP's this tract is expected to extend. To do this we assume that the length of a tract is exponentially distributed with parameter λ . If marker M_i is in a tract, the probability that the next marker, M_{i+1} , is also in the same tract, is $e^{-\lambda D_{i,i+1}}$, where $D_{i,i+1}$ is the genetic distance between markers M_i and M_{i+1} . A log likelihood function for λ is then given by

$$L(\lambda) = \prod_{j: M_j \in Z, M_{j+1} \in Z} e^{-\lambda D_{j,j+1}} \prod_{j: M_j \in Z, M_{j+1} \notin Z} (1 - e^{-\lambda D_{j,j+1}}) \quad (12)$$

where Z is the set of all markers in a migration tract. By entering the lengths of all SNP intervals ($D_{i,i+1}$) where we remain in a migrant tract or leave one, and then maximizing this function, we

obtain a maximum likelihood estimate of λ . Now the expected distance to add to a tract on the right side is

$$E[d_{j,j+1} | M_j \in Z, M_{j+1} \notin Z] = \int_0^{D_{j,j+1}} t\lambda e^{-\lambda t} / (1 - e^{-\lambda D_{j,j+1}}) dt = \frac{D_{j,j+1}}{1 - e^{-\lambda D_{j,j+1}}} + \frac{1}{\lambda} \quad (13)$$

and we similarly add

$$E[d_{j-1,j} | M_j \in Z, M_{j-1} \notin Z] = \frac{D_{j-1,j}}{1 - e^{-\lambda D_{j-1,j}}} + \frac{1}{\lambda} \quad (14)$$

to the left side.

Applying this method to the mouse SNP data, the resulting tract lengths were then used in the likelihood inference method described above. To insure that undetected tracts did not lead to spurious rejection of the null model, migrant tracts from simulated data were subjected to the constraints of the mouse SNP data set. The probability that any given SNP allele is informative concerning migrant ancestry was estimated by replacing each SNP allele in one subspecies with each possible SNP allele from the other subspecies, and monitoring the proportion of transplanted alleles that yielded positive evidence for migrant history under the criteria detailed above. Average SNP informativeness was estimated in this way for each subspecies separately. Tract lengths from constant migration rate simulations were randomly placed on the mouse SNP map, and each tract was detected only if two or more informative SNP's fell within it. This process was repeated until the number of tracts observed in the empirical data was matched.

Results:

Above, we described a theoretical framework for the distribution of migrant tract lengths, and a forward whole-population simulation tool to generate migrant tract data. The simulation program enables several assumptions of the theory to be violated: by allowing back migration, recombinational joining of migrant tracts, and effects of the ends of chromosomes. In all cases examined, including those shown in Figure 1, simulated data closely matched theoretical predictions. Figure 1 depicts the migrant tract length distributions generated by a constant migration rate model, and by admixture beginning 100, 200, or 300 generations ago. The contrasting migrant tract lengths generated by these histories suggested that such data could be informative for demographic inference. But Figure 1 is based on a large number of simulated replicates, and we were interested in testing whether individual data sets would contain enough information for demographic hypothesis testing and parameter inference.

Large, genome-scale data sets were generated for population samples under various demographic histories using the migrant tract simulation method described above. Genomes 3500 cM in size were generated for a sample size of 100, and a minimum tract length of 0.5 cM was used. Likelihood optimization was performed for each simulated data set under the migration rate change model, yielding estimates of m_1 , m_2 , and T . The highest log likelihood value obtained for this model was compared against the log likelihood score for the constant migration rate model, and the significance of likelihood ratios was assessed via comparison with data sets simulated under the constant rate model. Results are presented in Figure 2A and 2B.

The method was found to have high power to reject a constant rate model for a range of histories. The highest power often occurred within the first few hundred generations after a migration rate change—this is not surprising, as only tracts larger than 0.5 cM are considered

here, and recombination will typically break down migrant chromosomes to this size within about 200 generations. In some cases, particularly for strong decreases in migration rate, the method's power lasted well beyond this expectation. Even for the most subtle migration rate changes considered (from $N_e m = 0.1$ to $N_e m = 0.04$ and vice versa), power was fairly high, particularly around the $T = 200$ to $T = 500$ time window.

For histories involving a migration rate decrease, a similar procedure was applied to test whether a model with no current migration ($m_1 = 0$) could be rejected (Test 2A). Here, power was often a bit lower than for Test 1, but generally still quite high (Figure 2C). Conversely, for histories involving a migration rate increase, we tested whether a model with no migration before the rate change ($m_2 = 0$) could be rejected (Test 2B). Power for this test was high for very recent migration rate changes (i.e. 100-200 generations ago), but declined quickly from there (Figure 2D).

Accuracy of parameter estimation under the migration rate change model is shown in Figure 3. For a variety of demographic histories involving isolation, migration rate decreases, migration rate increases, and admixture, estimates of m_1 and m_2 were often quite precise. Although the method can not always distinguish low migration rates from zero, higher migration rates of $1E-5$ ($N_e m$ of 0.1) were estimated quite accurately, often with 95% confidence intervals only extending ~30% above and below the true value. A similar degree of accuracy was observed for T , with confidence intervals spanning a factor of two or considerably less. Parameter estimates for migration rate changes beyond 500 generations ago typically became less precise (data not shown), which makes sense as these data sets become less informative, with few tracts above 0.5 cM having arisen before the migration rate change.

Given the generally favorable performance of the likelihood inference method on simulated migrant tract data, we then sought to apply it to empirical data. Because a prerequisite for this method is a set of migrant tracts inferred with reasonable confidence, it is most applicable to populations or taxa that show a high degree of genetic differentiation. One such case is represented by the hybridizing house mouse subspecies *Mus musculus domesticus* and *M. m. musculus* in Europe. We used a simple set of criteria to define migrant (hybrid) tracts in genome-wide SNP data from both subspecies, then applied the likelihood inference method. Due to the limited size of the data set, relatively small numbers of migrant tracts were found: 75 in *M. m. domesticus*, 60 in *M. m. musculus*. However, the length distributions seemed to contain an excess of long tracts relative to equilibrium expectations (Figure 4), and the inference method detected a signal for an increase in the rate of introgression for both subspecies (Table 1).

In spite of having a larger likelihood ratio statistic against the constant rate model than *M. m. musculus*, Test 1 was only marginally significant for *M. m. domesticus*, while being significant for *M. m. musculus* and for a combined analysis of tracts from both subspecies. The weaker result for *M. m. domesticus* is due to a lower level of SNP informativeness in this subspecies: only 18% of *M. m. musculus* alleles would be detected as migrant in *M. m. domesticus*, compared to 38% in the opposite direction. Therefore, smaller tracts more frequently went undetected in the simulations used to assess significance in *M. m. domesticus* (see Models and Methods for details), and likelihood ratios from these simulations were higher. We also confirmed that the combined tract length data set for both subspecies showed the same signal for increased hybridization when each tract was required to have a minimum of three SNP's favoring migrant ancestry, rather than two ($P < 0.01$; results not shown).

M. m. domesticus had an estimate of zero for m_2 , while the estimate for *M. m. musculus* was non-zero (Table 1), but in neither case were the data sufficient to differentiate between no hybridization *versus* low hybridization prior to the inferred rate change. The estimated timing of the rate change was similar in both taxa (202 and 234 generations ago), and in the combined analysis (206). The two subspecies' estimates of m_1 differ by only about a factor of two (Table 1), and both suggest a high contemporary population migration rate between these subspecies.

Discussion

The specific demographic parameter estimates obtained from the house mouse SNP data should be interpreted with caution in light of limitations in the data set. Sample sizes are small - 7 and 8 haploid genomes. Samples originate from various geographic locations (Harr 2006), and our quantitative estimates might depend on the proximity of samples to the hybrid zone. Thus, it would be worthwhile to confirm our conclusions and refine parameter estimates using full genome sequence data from reasonably large population samples of both subspecies.

Still, it is interesting that both taxa yielded estimates of around 200 generations for the time since an increase in hybridization rate. Particularly when using with a minimum tract length as large as 0.5 cM (which is necessitated by the density of SNP's in this data set), the time-scope of our inference method is limited to fairly recent events (Figure 2). Thus, the time of ~200 generations may not represent the first contact between these subspecies in Europe, and indeed, archaeological evidence suggests a more ancient date for this event (reviewed in Boursot *et al.* 1993). However, this timing may still represent an increase in the rate of hybridization. If these mice have approximately two generations per year, the inference method suggests that

hybridization increased about 100 years ago, which seems generally coincident with an increased potential for human-mediated transport in Europe.

The evolutionary trajectories of these hybridizing house mouse subspecies will depend on a variety of factors, but one potential predictor is the current rate of hybridization in terms of $N_e m_1$. The true values of current N_e for European populations of *M. m. domesticus* and *M. m. musculus* are unknown, but long term effective sizes on the order of 1 million have been inferred for ancestral range populations of both subspecies (Baines and Harr 2007). Given the successful relationship of these mice with humans, it seems very plausible that the current N_e is at least this large. If we therefore take 1 million as an estimate for N_e in both taxa, the m_1 estimates obtained here imply that *M. m. domesticus* is currently receiving about 61 immigrants from *M. m. musculus* each generation, while *M. m. musculus* is receiving about 33 immigrants per generation from *M. m. domesticus* (based on the estimate of $N_e m_1$ for each subspecies). Since both of these estimates give $4N_e m_1 \gg 1$, these results could indicate that *M. m. domesticus* and *M. m. musculus* are currently on a path toward fusion rather than speciation. However, the presence of partial incompatibilities between these taxa, particularly on the X chromosome (*e.g.* Good *et al.* 2008), suggests that certain portions of the genome may resist homogenization.

Our analysis of simulated data showed that, given the lengths of migrant tracts from a population sample of genomes, the likelihood inference method presented here has high power to detect historical changes in migration (Figure 2), even for rather subtle shifts in migration rate (*i.e.* 2.5-fold changes), and should be useful in testing hypotheses and estimating parameters related to migration rate changes. This approach is conceptually related to methods that estimate the timing of recent admixture events (Patterson *et al.* 2004; Hoggart *et al.* 2004), but it allows for a greater variety of historical scenarios. In terms of its temporal scale, our method falls in

between methods that identify very recent migration events (*e.g.* Rannala and Mountain 1997) and those that estimate long-term migration rates (*e.g.* Beerli and Felsenstein 2001). Although the results presented here suggest that our method is most relevant for detecting migration rate changes within the past 1,000 generations, in many cases it may be possible to use a lower threshold tract length (C) than the 0.5 cM used in this study, and the temporal scope should expand with the inverse of C . The main assumption of the method is that recombination will break down tracts to below the threshold length before genetic drift can lift them to high frequency. Values of C that are less than $1/N_e$ are therefore recommended, but the choice of C will also depend on the level of diversity, the degree of population differentiation, and the density of markers (all of which constrain the inference of short migrant tracts). For *Mus musculus*, a smaller threshold tract length would be a viable option with a denser SNP data set.

Our method does not address the inference of population ancestry along a recombining chromosome and requires that migrant tracts be identified beforehand. Published methods exist for this purpose (*e.g.* Falush *et al.* 2003) and the optimal method may depend on the data set being analyzed. Tract length data obtained from such methods can be used as the input for our analysis, and for methods that allow sampling from a posterior distribution of tract lengths, uncertainty in the tract length inference can be directly incorporated in the likelihood method. Without the use of such methods, the need for confident identification of migrant tracts would make this approach difficult to apply to weakly differentiated populations, but for more strongly differentiated populations or hybridizing subspecies, this method should be very useful in its current form.

To derive the tract length distributions, a number of assumptions were needed. The most troublesome of these, the lack of recombination among migrant tracts, would be very difficult to

relax in the current framework. A full treatment of the problem would require analysis of an ancestral recombination graph in a subdivided population for whole-genome data. Another simplifying assumption is made by ignoring the ends of the chromosome. This assumption is much easier to relax and can be done by considering the conditional distribution in Equation (2). However, as this leads to a considerably less tractable algebraic representation, and since the current approximation performs very well for realistic chromosome lengths, we have chosen not to pursue this further.

The inference method described here may be applicable in a number of biological contexts. As demonstrated by our analysis of the *Mus musculus* SNP data, the migrant tract approach may be especially relevant in testing hypotheses about historical trends of gene flow across hybrid zones, perhaps shedding light on the evolutionary trajectories of hybridizing taxa. The inferences enabled by this method may also find particular relevance in conservation: to test the effect of a new barrier (such as a highway) on the dispersal of an organism with a short generation time, or to infer the rate of migration over relatively recent timescales (rather than over the past $4N_e$ generations) to guide management strategies for species with fragmented habitats. In this context it is important to notice that inferences are done at a time scale more relevant to conservation genetics, and that estimates of time in number of generations are obtained directly and do not rely on inferences of effective population sizes.

For optimal power, this method requires reasonably dense, genome-wide polymorphism data from moderate to large sample sizes. It also requires information about the genetic map position of each marker, which can be estimated by genotyping related individuals such as parent-offspring trios. In light of rapidly improving DNA sequencing technology, we are

optimistic that the inferences described here will be possible for both model and non-model organisms in the near future.

Acknowledgments:

This research was supported by an N.I.H. Kirschstein-N.R.S.A. Postdoctoral Fellowship (F32 HG004182) to JEP, and an N.I.H. research grant (U01HL084706) to R. N.

Literature cited:

- Baines, J. F., and B. Harr. 2007. Reduced X-linked diversity in derived populations of house mice. *Genetics* **175**:1911-1921.
- Becquet, C., and M. Przeworski. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**:1505-1519.
- Berli, P., and J. Felsenstein. 2001. Maximum likelihood estimation of migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci.* **98**:4563-4568.
- Boursot, P., J.-C. Auffray, J. Britton-Davidian, and F. Bonhomme. 1993. The evolution of house mice. *Annu. Rev. Ecol. Syst.* **24**:119-152.
- Boursot, P., and K. Belkhir. 2006. Mouse SNPs for evolutionary biology: Beware of ascertainment biases. *Genome Res.* **16**:1191-1192.
- Cornuet, J. M. and G. Luikart. 1996. Description and power analysis of two tests for detecting population bottlenecks from allele frequency data. *Genetics* **144**:2001-2014.
- Depaulis, F., S. Mousset, and M. Veuille. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *J. Mol. Evol.* **57**:S190-S200.

- Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**:1567-1587.
- Fisher, R. A. 1930. *The genetical theory of natural selection*. Clarendon Press, Oxford.
- Good, J. M., M. D. Dean, and M. W. Nachman. 2008. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* **179**:2213-2228.
- Harr, B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**:730-737.
- Hey, J. 2005. On the number of New World founders: A population genetic portrait of the peopling of the Americas. *PLoS Biology* **3**:e193.
- Hoggart, C. J., M. D. Shriver, R. A. Kittles, D. G. Clayton, and P. M. McKeigue. 2004. Design and analysis of admixture mapping studies. *Am J. Hum. Genet.* **74**:965-978.
- Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**:183-201.
- Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro, and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**:1401-1410.
- Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C.-F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**:528-538.
- Kingman, J. F. C. 1982. The coalescent. *Stochastic Process. Appl.* **13**:235-248.
- Kingman, J. F. C. 1982. On the genealogy of large populations. *J. Appl. Probab.* **19A**:27-43.

- Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, *et al.* 2002. A high-resolution map of the human genome. *Nature* **31**:241-247.
- Koopman, W. J., Y. Li, E. Coart, W. E. Van De Weg, B. Vosman, I. Roldán-Ruiz, and M. J. M. Smulders. 2007. Linked vs. unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression. *Mol. Ecol.* **16**:243-256.
- Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**:351-372.
- Montana, G., and J. K. Pritchard. 2004. Statistical tests for admixture mapping with case-control and cases-only data. *Am. J. Hum. Genet.* **75**:771-789.
- Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics* **158**:885-896.
- Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, *et al.* 2004. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**:979-1000.
- Pearse, D. E., and K. A. Crandall. Beyond F_{ST} : Analysis of population genetic data for conservation. *Conservation Genetics* **5**: 585-602.
- Rannala, B., and J. L. Mountain. 1997. Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci.* **94**:9197-9201.
- Salcedo, T., A. Geraldes, and M. W. Nachman. 2007. Nucleotide variation in wild and inbred mice. *Genetics* **177**:2277-2291.
- Wall, J. D., P. Andolfatto, and M. Przeworski. 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**:203-216.

Whitlock, M. C., and D. E. McCauley. 1999. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity* **82**:117-125.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Wright, S. 1952. The theoretical variance within and among subdivisions of a population that is in a steady state. *Genetics* **37**:312-323.

Table 1. Parameter inference and hypothesis testing for house mouse data

subspecies	# tracts	P	m_1	m_2	T	Test 1	Test 2B
domesticus	75	0.01095	6.08E-05	0	202	$P = 0.08$	n/a
musculus	60	0.00684	3.29E-05	1.00E-06	234	$P = 0.03$	n.s.
combined	135	0.00876	4.71E-05	7.50E-07	206	$P = 0.02$	n.s.

Listed for each subspecies (and for the combined analysis) is the number of tracts greater than 0.5 cM, the proportion of the genome included migrant tracts (P), parameter estimates for m_1 , m_2 , and T , and results of hypothesis tests (n/a indicates that Test 2B is not applicable when the estimate of m_2 is zero; n.s. denotes P values not approaching significance).

Figure Legends

Figure 1. The distribution of migrant tract lengths after the advent of admixture. Models where previously isolated populations begin exchanging migrants at rate $N_e m = 0.1$ either 100, 200, or 300 generations ago are compared against the case in which populations exchange migrants at a constant rate $N_e m = 0.1$ with no prior isolation (the single migration rate, “equilibrium” model). Depicted here are the relative abundance of migrant tracts for 0.01 cM histogram bins between 0.5 cM (the minimum/threshold tract length) and 5 cM. Also shown is the agreement between theoretical predictions (lines) and tracts from 1000 simulated replicates with $N_e = 10,000$ (shapes).

Figure 2. Power to test demographic hypotheses. Shown here first are tests comparing the migration rate change model to the null model of a constant migration rate, for histories involving decreasing (A) or increasing (B) migration rates. For histories involving decreasing migration rates, power to reject a model with m_1 constrained to be zero is shown (C). For histories involving increasing migration rates, power to reject a model with m_2 constrained to be zero is shown (D). Significance was gauged by comparing the difference in log likelihood scores between models to data simulated under the null model. Each data set consisted of 100 simulated haploid genomes, and a threshold tract length of 0.5 cM was used.

Figure 3. Distribution of demographic parameter estimates. Results from the analysis of simulated migrant tract data are shown, including median estimates (diamonds) and 95% confidence intervals (the 2.5 and 97.5 percentiles of the distribution of estimates) for (A) m_1 , (B)

m_2 , and (C) T . The order of parameter sets is the same in each panel (*i.e.* the far left estimates are for true values of $m_1 = 0$, $m_2 = 1 \text{ E } -5$, and $T = 100$).

Figure 4. Migrant tract lengths found in *M. m. domesticus* and *M. m. musculus*, compared to constant migration rate expectations.







