The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium

Noah Zaitlen[1], Scott Huntsman[1], Donglei Hu[1], Melissa Spear[1], Celeste Eng[1], Sam S. Oh[1], Marquitta J White[1], Angel Mak[1], Adam Davis[2], Kelly Meade[2], Emerita Brigino-Buenaventura[3], Michael A LeNoir[4], Kirsten Bibbins-Domingo[1], Esteban G Burchard[1], and Eran Halperin[*5]

[1]Department of Medicine, University of California San Francisco, San Francisco, CA
[2]Childrens Hospital and Research Center Oakland, Oakland, CA
[3]Department of Allergy and Immunology, Kaiser PermanenteVallejo Medical Center, Vallejo, CA
[4]Bay Area Pediatrics, Oakland, CA
[5]Department of Computer Science, Tel Aviv University, Tel Aviv, Israel

May 30, 2016

1 Abstract

Statistical models in medical and population genetics typically assume that individuals assort randomly in a population. While this simplifies model complexity, it contradicts an increasing body of evidence of non-random mating in human populations. Specifically, it has been shown that assortative mating is significantly affected by genomic ancestry. In this work we examine the effects of ancestry-assortative mating on the linkage disequilibrium between local ancestry tracks of individuals in an admixed population. To accomplish this, we develop an extension to the Wright-Fisher model that allows for ancestry based assortative mating. We show that ancestry-assortment perturbs the distribution of local ancestry linkage disequilibrium (LAD) and the variance of ancestry in a population as a function of the number of generations since admixture. This assortment effect can induce errors in demographic inference of admixed populations when methods assume random mating. We derive closed form formulae for LAD under an assortative-mating model with

*Address correspondence to: Noah Zaitlen: noah.zaitlen@ucsf.edu or Eran Halperin: heran@post.tau.ac.il

and without migration. We observe that LAD depends on the correlation of global ancestry of couples in each generation, the migration rate of each of the ancestral populations, the initial proportions of ancestral populations, and the number of generations since admixture. We also present the first evidence of ancestry-assortment in African Americans and examine LAD in simulated and real admixed population data of African Americans. We find that demographic inference under the assumption of random mating significantly underestimates the number of generations since admixture, and that accounting for assortative mating using the patterns of LAD results in estimates that more closely agrees with the historical narrative.

2 Introduction

One of the most common assumptions in human population genetics analyses is that of Hardy-Weinberg Equilibrium (HWE). The HWE assumption in turn enforces a set of additional conditions including the absence of selection, infinite population size, and importantly, random mating. We and others have shown that assortative mating is a common phenomenon[1, 2, 3] and many phenotypes including height, education level, and personality traits are correlated between spouses [4]. For Latinos and other admixed populations the African, Native-American, and European proportions of individual's genomes can be correlated between spouses. We recently demonstrated that the genomic ancestry of Latino couples is highly correlated [1], and refer to this as ancestry-assortative mating. Thus, the assumption of random mating and therefore Hardy-Weinberg Equilibrium is not satisfied in practice, and the implication of this observation for population and evolutionary genetic studies remains unclear.

The assumption of random mating is used in many types of population and quantitative genetics analyses. Particularly, random mating is assumed both in analysis of population genetics data and when inferring population parameters such as recombination rates, mutation rates, selection, heritability, and others. Moreover, methods for quality control and data cleaning often make the random mating assumption. For example, methods for haplotype phasing typically compute the likelihood of the genotype as the product of the likelihoods of each of the haplotypes, and this derivation is based on the random mating assumption[5]. Similarly, such likelihood derivations are also common in methods for the inference of identity by descent and inference of ancestry from genomic data[6]. Thus far, the sensitivity of these methods to the assumption of assortative mating has not been evaluated. In principle, realistic violations of the random mating assumption may not be detrimental to existing methods, however this needs to be taken to the test.

In this paper we explore the robustness of specific genetic features and their inference from genetic data to assortative mating. Because assortative mating in Latinos has been shown

to be affected by ancestral proportions, we focused our analysis on the behavior of ancestry linkage disequilibrium under assortative mating. We propose a random generative model for population dynamics under assortative mating which is due to population structure. Our model follows the spirit of the Wright-Fisher model, and makes the assumption that the correlation of ancestry proportions between spouses stays fixed across generations. Particularly, when the correlation of ancestry proportions is zero, our model is equivalent to the Wright-Fisher model.

We develop mathematical theory that describes the decay of local ancestry disequilibrium (LAD) as a function of assortative mating strength, migration rate, recombination rate, and the number of generations since admixture began. Thus, one can use these results in order to infer the demographic history of admixed populations. Several methods for demographic inference in admixed populations exist including ones that use patterns of LD decay [7], local ancestry track length distribution [8], and the distribution of identity by descent segments[9]. However, these methods assume random mating, and under assortative mating LD decay follows a different pattern [10]. Using simulations we demonstrate that our mathematical derivation matches empirical LAD decay. Furthermore, we develop the theory with migration rates from the ancestral populations, and we demonstrate that in the presence of assortative mating, one may erroneously conclude that there has been active migration, and vice-versa.

We applied our analysis to a dataset of 1730 African Americans from the Study of African Americans, Asthma, Genes and Environments (SAGE) study[11]. We first used ANCESTOR [12] to show that the correlation of African ancestry between the spouses in the last generation is approximately 0.32. We then used our analysis to infer the number of generations and migration patterns in the African American population. Under the assumption of no migrations and random mating, an analysis of LAD resulted in an estimated of the number of generations since admixture of 3. Adding assortment and migrations we find that the estimated number of generations since the admixture event is 15. Assuming a generation time of 25 years this places the initial migrations in the mid 17th century which is consistent with the history of African Americans[13].

3 Methods

The model We assume the following alternative to Wright-Fisher. Let N be the number of individuals in each population (effective population size divided by 2). Each individual has two haplotypes, so the total number of haplotypes is $2N$. Also, we assume the population is a recently admixed population with two ancestral populations (referred to as population 1 and population 2), and let θ_i denote the fraction of the genome with population 1 ancestry.

In the next generation, each individual picks two parents from the current generation, such that the correlation between the ancestry of the two parents is a fixed value P. One way of generating such mating in silica is the following. We randomly pick the set of mothers (with or without replacements) from the original distribution. We then randomly choose the set of fathers (with or without replacements). Now, for each of the parents we give a score $score_i = \theta_i + \epsilon_i$, where θ_i is the global ancestry of the parent, and ϵ_i is drawn from a normal distribution $N(0, \sigma^2)$. We then sort the mothers and the fathers based on their score and we let the mother with $i-th$ largest score marry the father with the $i-th$ largest score. We then compute the correlation between $corr(\theta_m, \theta_f)$, where θ_f, θ_m are the ancestries of the mother and the father. We choose σ such that $P = corr(\theta_m, \theta_f)$. We note that our analysis below does not rely on this specific procedure; particularly, the distribution of parents for the new generation can be quite general, and our only assumption is that P is constant across the generations. Note that this assumption may seem restrictive at first, however the case of random mating is far more restrictive, since there one requires that $P = 0$ in all generations.

Local Ancestry disequilibrium Denote by γ_1^t the probability of having an allele from ancestry 1 at a given position at generation t. Furthermore, for a pair of positions, let γ_{11}^t denote the probability of having an allele from ancestry 1 at the two positions. We define a new statistic, termed local-ancestry linkage disequilibrium, denoted by LAD. We define $LAD = \gamma_{11} - \gamma_1^2$. We are interested in the expected value of LAD^t (LAD at generation t) as a function of the recombination rate r, the number of generations t, and the original local ancestry linkage disequilibrium LAD^0.

For the following derivation, we will assume that the population size is infinite. We will later show that empirically this assumption does not have a substantial effect for realistic values of N. We will first assume that there is no migration and we will relax this assumption in the next section.

Since there is no migration and the population size is infinite, the mean of θ is fixed across the generations (remember that the marginal distribution of the mothers and the fathers is the same and is simply a random draw from the current generation)[14]. We denote $\mu = E[\theta]$. Let $V_t = Var(\theta_t)$ be the variance of θ in generation t, and let $\rho_t = PV_t$ be the covariance $\rho_t = cov(\theta_m, \theta_f)$. For $t > 1$ we have:

$$\begin{aligned} V_{t+1} &= E[\theta_{t+1}^2] - \mu^2 \\ &= E[(\theta_m^t + \theta_f^t)(\theta_m^t + \theta_f^t)/4] - \mu^2 \\ &= \frac{1}{4}\left(2E[\theta_t^2] + 2E[\theta_m^t \theta_f^t]\right) - \mu^2 \\ &= \frac{1}{2}\left(\mu^2 + V_t + \rho_t + \mu^2\right) - \mu^2 \\ &= \frac{V_t(1+P)}{2} \end{aligned}$$

This demonstrates that the variance of genome-wide ancestry is larger when there is assortative mating. Now, we know

$$\rho_{t+1} = PV_{t+1} = \frac{PV_t(1+P)}{2} = \frac{1+P}{2}\rho_t \tag{1}$$

Note that for $t = 0$, $\rho_0 = V_0$ since there was no assortative mating prior to the admixture event, and therefore for $t = 1$ the above calculation gives $V_1 = V_0$, and $\rho_1 = PV_0 = P\rho_0$. In order to simplify the notation, we change the indices, so that generation $t = -1$ corresponds to the time of encounter of the two population and $t = 0$ is the first generation after admixture. Therefore, we have that Equation 1 holds for every $t \geq 1$.

We now find a recursion formula for LAD^t. Let r be the probability for an odd number of recombinations between the two positions in a given meiosis. Hence,

$$\begin{aligned} LAD^{t+1} &= \gamma_{11}^{t+1} - \mu^2 \\ &= (1-r)\gamma_{11}^t + rE[\theta_m^t \theta_f^t] - \mu^2 \\ &= (1-r)LAD^t + r(E[\theta_m^t \theta_f^t] - \mu^2) \\ &= (1-r)LAD^t + r\rho_t \end{aligned}$$

We are now ready to describe our main result:

Lemma 3.1

$$LAD^t = (1-r)^t LAD^0 + r\rho_0 \frac{(1+P)^t - (1-r)^t 2^t}{2^{t-1}(P+2r-1)}$$

Proof We show this is true by induction. It is easy to verify that since $LAD^1 = (1-r)LAD^0 + r\rho_0$, the base case $t = 1$ holds. Assume the lemma holds for t and we will prove

it for $t+1$.

$$\begin{aligned} LAD^{t+1} &= (1-r)LAD^t + r\rho_t \\ &= (1-r)^{t+1}LAD^0 + (1-r)r\rho_0\frac{(1+P)^t-(1-r)^t2^t}{2^{t-1}(P+2r-1)} + r\rho_t \\ &= (1-r)^{t+1}LAD^0 + r\rho_0\left((1-r)\frac{(1+P)^t-(1-r)^t2^t}{2^{t-1}(P+2r-1)} + \frac{(1+P)^t}{2^t}\right) \\ &= (1-r)^{t+1}LAD^0 + r\rho_0\frac{(1+P)^{t+1}-2^{t+1}(1-r)^{t+1}}{2^t(P+2r-1)} \end{aligned}$$

∎

LAD under migration. We now assume that in each generation a fraction m_1 of the population is replaced by individuals from the first population ($\theta = 1$), and a fraction m_0 of the population is replaced by individuals from the population $\theta = 0$. We denote by $m = m_1 + m_0$, and $\alpha = \frac{m_1}{m}$. Since there is migration, the mean global ancestry is changing over time, and we let $\mu_t = E[\theta_t]$ the average values of θ when an individual is randomly sampled from the population. For simplicity of notation, we denote $x_t = \mu_t - \alpha$, and we note that x_t is exponentially decreasing. Since $\mu_{t+1} = \alpha m + (1-m)\mu_t$, we have that $x_{t+1} = (1-m)x_t$ and therefore $x_t = x_0(1-m)^t$.

We now show the following lemma:

Lemma 3.2 *If there is a sequence $y_0, y_1, \ldots,$ satisfying the recursion equation $y_{t+1} = (1-m)q_1y_t + a_3x_t^2 + a_2q_2^tx_t + a_1x_t + a_0$, then*

$$y_t = b_4x_t^2 + b_3q_1^tx_t + b_2q_2^tx_t + b_1x_t + b_0$$

Proof To prove the base of the induction we need to satisfy $y_0 = b_4x_0^2 + (b_1+b_2+b_3)x_0 + b_0$, which is a simple linear equation. We will show that the induction step adds two more linear equations. Assume the lemma holds for t, and consider y_{t+1}:

$$\begin{aligned} y_{t+1} &= (1-m)q_1y_t + a_3x_t^2 + a_2q_2^tx_t + a_1x_t + a_0 \\ &= (1-m)q_1(b_4x_t^2 + b_3q_1^tx_t + b_2q_2^tx_t + b_1x_t + b_0) + a_3x_t^2 + a_2q_2^tx_t + a_1x_t + a_0 \end{aligned}$$

Now, note that $x_{t+1} = (1-m)x_t$. Therefore:

$$\begin{aligned} y_{t+1} &= \left(\frac{q_1b_4(1-m)+a_3}{(1-m)^2}\right)x_{t+1}^2 + b_3q_1^{t+1}x_{t+1} + \left(\frac{b_2q_1(1-m)+a_2}{q_2(1-m)}\right)q_2^{t+1}x_{t+1} \\ &\quad + \left(\frac{q_1(1-m)b_1+a_1}{1-m}\right)x_{t+1} + ((1-m)q_1b_0 + a_0) \end{aligned}$$

We now set:

$$\begin{aligned} b_0 &= \frac{a_0}{1-(1-m)q_1} \\ b_1 &= \frac{a_1}{(1-m)(1-q_1)} \\ b_2 &= \frac{a_2}{(1-m)(q_2-q_1)} \\ b_4 &= \frac{a_3}{(1-m)(1-m-q_1)} \\ b_3 &= \frac{y_0 - b_4 x_0^2 - (b_1+b_2)x_0 - b_0}{x_0} \end{aligned}$$

∎

Next, we observe:

$$\begin{aligned} V_{t+1} &= E[\theta_{t+1}^2] - \mu_{t+1}^2 \\ &= \alpha m + (1-m)E[(\theta_m^t + \theta_f^t)(\theta_m^t + \theta_f^t)/4] - (x_{t+1}+\alpha)^2 \\ &= \alpha m + \frac{1-m}{4}\left(2E[\theta_t^2] + 2E[\theta_m^t \theta_f^t]\right) - ((1-m)x_t + \alpha)^2 \\ &= \alpha m + \frac{1-m}{2}\left(\mu_t^2 + V_t + \rho_t + \mu_t^2\right) - ((1-m)x_t + \alpha)^2 \\ &= \alpha m + (1-m)(x_t+\alpha)^2 - ((1-m)x_t+\alpha)^2 + V_t\frac{(1-m)(1+P)}{2} \\ &= m(1-m)x_t^2 + \alpha m(1-\alpha) + V_t\frac{(1-m)(1+P)}{2} \end{aligned}$$

By Lemma 3.2, we have $V_t = b_4 x_t^2 + b_3 x_t \frac{(1+P)^t}{2^t} + b_0$, for b_4, b_3, b_0 specified in the lemma. Note that based on the lemma's proof, $b_1 = b_2 = 0$. Now,

$$\begin{aligned} LAD_{t+1} &= \gamma_{11}^{t+1} - \mu_{t+1}^2 \\ &= \alpha m + (1-m)\left((1-r)\gamma_{11}^t + rE[\theta_m^t \theta_f^t]\right) - \mu_{t+1}^2 \\ &= \alpha m + (1-m)\left((1-r)\gamma_{11}^t + r(\rho_t + \mu_t^2)\right) - \mu_{t+1}^2 \\ &= (1-m)(1-r)LAD_t + (1-m)(1-r)\mu_t^2 + (1-m)r\mu_t^2 - \mu_{t+1}^2 + \alpha m + \\ &\quad (1-m)r\rho_t \\ &= (1-m)(1-r)LAD_t + (1-m)\mu_t^2 - \mu_{t+1}^2 + \alpha m + (1-m)r\rho_t \end{aligned}$$

Therefore, noting that $\mu_{t+1} = (1-m)x_t + \alpha$, we have

$$\begin{aligned} LAD_{t+1} &= (1-m)(1-r)LAD_t + (1-m)\mu_t^2 - \mu_{t+1}^2 + \alpha m + (1-m)r\rho_t \\ &= (1-m)(1-r)LAD_t + (1-m)(x_t+\alpha)^2 - (\alpha + x_t(1-m))^2 + \alpha m + (1-m)r\rho_t \\ &= (1-m)(1-r)LAD_t + x_t^2 m(1-m) + m\alpha(1-\alpha) + (1-m)r\rho_t \end{aligned}$$

Now, recall $\rho_t = Pb_4x_t^2 + Pb_3x_t\frac{(1+P)^t}{2^t} + Pb_0$. Therefore, we have the form $LAD_{t+1} = (1-m)q_1LAD_t + a_3x_t^2 + a_2q_2^tx_t + a_1x_t + a_0$ satisfying Lemma 3.2 with the following values:

$$\begin{aligned} q_1 &= 1-r \\ q_2 &= \frac{1+P}{2} \\ a_3 &= (1-m)(m+rPb_4) \\ a_2 &= (1-m)rPb_3 \\ a_1 &= 0 \\ a_0 &= \alpha m(1-\alpha) + (1-m)rPb_0 \end{aligned}$$

Thus, for c_0, c_1, c_2, c_3, c_4 taken from Lemma 3.2 we have

$$LAD_t = c_4x_t^2 + c_3q_1^tx_t + c_2q_2^tx_t + c_1x_t + c_0.$$

Plugging in the values of q_1, q_2, and the fact that $x_t = x_0(1-m)^t$, we get

$$LAD_t = c_4x_0^2(1-m)^{2t} + x_0(1-m)^t(c_3(1-r)^t + \frac{c_2(1+P)^t}{2^t} + c_1) + c_0 \qquad (2)$$

4 Results

When applied to the genome, we can estimate the value of LAD for known values of r by averaging the observed LAD across the genome. We can now fit the values of m, t, and P based on the distribution of the LAD as a function of r in the current generation. It is therefore important to understand the dependency of the distribution of LAD for varying values of r as a function of t, P, and m. In what follows, we explore the behavior of LAD under different settings.

We first consider the case where $m_1 = m_2 = 0$, i.e., there is no migration, and $P = 0.6$. In Figure 1 we observe that there is a clear separation between the different curves for the different numbers of generations since admixture, and it should therefore be easy to estimate the time of admixture event under the assumption of no migration and $P = 0.6$.

Next, we study the effect of P on the LAD distribution. In Figure 2 we plot the LAD distribution under no migration, after 10 generations of admixture, with varying values of P. Evidently, strong assortative mating with large values of P results in a substantially different levels of LAD. However, we observe that low values of P are harder to distinguish, and therefore we expect that random mating is a robust assumption for any statistic that uses LAD or its derivatives, as long as assortative mating is weak (e.g., $P < 0.5$).

Since typical analysis of genetic data assumes random mating, we attempted at understanding the potential risk in making the assumption in the presence of assortative mating. Thus, we consider the case where there is assortative mating, and we try to estimate the time of admixture under the assumption of random mating. For ancient admixture the difference between the estimates under assortative mating and random mating is not substantial (about 10% - data not shown). For recent admixture (10-20 generations), we observe that there is a considerable difference between the true LAD curve compared to the LAD curve under random mating, and moreover, the true LAD curve is similar to LAD curves that assume random mating but that are substantially more recent. Specifically, in Figure 3 the admixture event occurred 10 generations ago under a strong assortative mating ($P = 0.8$), however under random mating, the LAD curve that corresponds to $t = 4$ is the most similar to the true LAD curve. In Figure 4 the admixture event occurred 15 generations ago under a somewhat weaker assortative mating ($P = 0.6$), while the estimated number of generations would be 11 under random mating.

Next, we explore the effect of migration on the LAD function. We consider both the case where the two populations migrate at the same rate ($m_1 = m_2$) as shown in Figure 5, as well as the case in which $m_1 = 0$, as shown in Figure 6. Evidently, the theoretical calculations capture the empirical well in the sense that they allow for a clear distinction between different migration rates.

We note that migration and assortative mating can result in similar LAD decay. We estimated the LAD curve using the formula of Lemma 3.1 under random mating with migration, as well as under assortative mating with different values of migration. Since the parameter space (m_1, m_2, P) is large, there are triplets of values with very similar LAD curves, thus in practice the model parameters will not necessarily be identifiable. In Figure 7 we present an example where identifiability requires the comparison of LAD decay over dozens of megabases.

Results on real data To examine the properties of our model in real data we used genetic data from 1730 African-American individuals from the the Study of African Americans, Asthma, Genes and Environments (SAGE). The individuals in the SAGE data were genotyped at 800,000 SNPs on the Affymetrix Axiom Genome-Wide LAT 1 Array, and genotype calling and QC were performed as previously described [15].

To compute LAD we first called local ancestry using the LAMP-LD software package[16] and genome-wide ancestry was inferred from mean value of local ancestry for each individual. We measured the LAD decay in 164 **10Mb** overlapping windows with a **1Mb** overlap. We calculated the mean LAD decay across all windows as well as the squared distance of each window to the mean. Regions that are under selection or in which the estimates of recombination rates are inaccurate will result in a different LAD decay. We therefore performed additional QC by removing windows with a LAD decay greater than two standard deviations from the mean. We repeated this process until convergence leaving 96 windows.

We measured the assortative mating over the last generation by applying the method ANCESTOR [12] to the data. ANCESTOR takes as input local and global ancestry and determines the ancestral proportions of the mother and the father of each individual. The Pearson correlation coefficient between the parental ancestries was $P = 0.32$ estimated across all individuals. This establishes that there was strong ancestry based assortment in African Americans in the last generation. If this ancestry-based assortative mating exists in previous generations our theory shows that LAD decay will be affected. Under the assumption that this correlation was stable throughout history, one can use this estimate to constrain the potential demographic histories of African Americans inferred via LAD.

We fitted the migration and assortative mating parameters using a grid search over the entire range of parameters. The best fit resulted in an estimate of $t = 13$ generations, with migration rates $m_1 = 0.01, m_2 = 0.05$, and assortative mating $P = 0.46$ (Figure 8a). Next, we made the assumption of no migration by searching the grid but with the constraint $m_1 = m_2 = 0$, but we allowed for assortative mating. In this case, the number of generations was dramatically shortened to 8 generations, and the assortative mating value increased dramatically to $P = 0.6$ (Figure 8b). Similarly, we search the grid with the constraint $P = 0$ in order to study the case of random mating with migration. In this case the number of generations was 16, and the migration values slightly increased to $m_1 = 0.02, m_2 = 0.05$ (Figure 8c). Finally, under random mating and no migration the estimated number of generation is $t = 3$, which is clearly a vast underestimate of the true number based on the known history of African Americans (Figure 8d). Notably, there is no good fit under random mating and no-migration, and the best fit is obtained in the presence of both migration and assortative mating.

Clearly, the LAD decay is only one summary statistic that depends on the parameters m_1, m_2, t, P, and other statistics may give somewhat different results. For example, it may be possible to examine the distribution of IBD [9], local ancestry [8], and LD [7] under an assortative mating model. Moreover, the LAD decay is not identifiable since different sets of parameters often lead to similar LAD decay. Particularly, in the case of the African Americans in SAGE, the best fit was followed by a few different sets of parameters. Particularly, under the assumption that $P = 0.32$ is fixed across the generations, the best fit

was with $t = 15$ generations, and the migration rates were $m_1 = 0.08, m_2 = 0.01$. Due the computational complexity of the grid search used to estimate model parameters it was not feasible to estimate confidence intervals. However, as was the case in simulations, migration rates and generation times could be altered to accommodate removing assortative mating form the model.

All genetic data are available via dbGAP with the accession number phs000355.v1.p1.

5 Discussion

We presented an adaption of the Wright-Fisher model, which incorporates ancestry-assortative mating in admixed populations. We demonstrated that under this model the linkage disequilibrium of local ancestry (LAD) between markers is a function of their recombination rate, the ancestral population migration rates, and the strength of ancestry based assortment. Assortative mating is likely impacting other estimates of population and medical genetic parameters both within admixed and continental populations including identity by descent distributions, estimates of heritability, joint site frequency spectra, runs of homozygosity, and the distribution of local ancestry track lengths.

While the focus of this work is the definition and presentation the ancestry-assortative model and its properties, we also estimated the parameters of the model in a real African-American data set. Our estimate of 15 generations since admixture in African Americans is larger than previous estimates[8, 9] and it fits considerably better the known history of African Americans[13]. This suggests that taking assortative mating into account may in some cases be critical in order to obtain the correct demographic history or other population parameters.

The approach we presented for estimating the number of generations since admixture using LAD has its limitations. First, this approach involves a very inefficient grid search, resulting in an inability to provide errors around estimates via bootstrap. Second, in some cases both migration and assortative mating can give rise to similar LAD distributions, and therefore in those cases one can mistakenly believe that the migration is higher and assortative mating is lower or vice versa. The latter, however, raises an interesting question: In previous attempts for learning demographic histories of humans and other species, is it the case that the migration coefficients were inflated, or number of generations since admixture deflated due to assortative mating?

Going forward it will be interesting to determine if assortative mating has biased other recent estimates of demographic events such as the introgression of Neanderthals [17] or the domestication of dogs and pigs[18, 19]. We will also explore extensions to multi-way admixed populations and the use of MCMC to provide confidence intervals for parameter

estimates. In addition to altering the distribution of LAD, we showed that assortative mating increases the variance of global ancestry. Under certain polygenic models this will induce a concomitant increase in phenotypic variance, which may have implications for selection and evolution.

Our method makes several strong assumptions, which are likely incorrect, such as constant ancestry-assortment strength and migration rates. However, these are a relaxation of previous methods, since for example under the standard Wright-Fisher model, both random mating and no migration are assumed, and thus both migration rates and ancestry-assortative strengths are fixed across the generations in this case (fixed with value 0). While assortative mating has been well studied, to the best our knowledge this is the first attempt to include ancestry-assortment in the estimation of demographic histories. We also reported, for the first time, the strength of ancestry-assortment in African Americans in the previous generation. In future work we intend to examine the effect of ancestry assortment on other genetic features as well as the resulting impact in population and medical genetics.

6 Acknowledgments

The authors acknowledge the patients, families, recruiters, health care providers and community clinics for their participation. In particular, the authors thank Sandra Salazar for her support as the SAGE II study coordinator. This work was supported in part by the Sandler Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, and the National Institutes of Health (ES015794 and MD006902). NAZ was supported by an NIH career development award from the NHLBI (K25HL121295). EH was supported by the Israel Science Foundation (Grant 1425/13), United States-Israel Binational Science Foundation (Grant 2012304), German-Israeli Foundation (Grant 1094-33.2/2010) and by the National Science Foundation (Grant III-1217615). The SAGE study was supported by the Sandler Family, American Asthma Foundation, NIH/NIHMD- 1P60 MD006902, NIH/NHLBI - 1R01HL117004-01, NIH/NIEHS - R21ES24844-01, NIH/NIHMD - U54MD009523, and TRDRP 24RT-0025.

References

[1] J. Y. Zou, D. S. Park, E. G. Burchard, D. G. Torgerson, M. Pino-Yanes, Y. S. Song, S. Sankararaman, E. Halperin, and N. Zaitlen. Genetic and socioeconomic study of mate choice in latinos reveals novel assortment patterns. *Proc Natl Acad Sci U S A*, 112(44):13621–6, 2015.

[2] N. Risch, S. Choudhry, M. Via, A. Basu, R. Sebro, C. Eng, K. Beckman, S. Thyne, R. Chapela, J. R. Rodriguez-Santana, W. Rodriguez-Cintron, P. C. Avila, E. Ziv, and E. Gonzalez Burchard. Ancestry-related assortative mating in latino populations. *Genome Biol*, 10(11):R132, 2009.

[3] C. A. Mathews and V. I. Reus. Assortative mating in the affective disorders: a systematic review and meta-analysis. *Compr Psychiatry*, 42(4):257–62, 2001.

[4] K. R. Merikangas. Assortative mating for psychiatric disorders and psychological traits. *Arch Gen Psychiatry*, 39(10):1173–80, 1982.

[5] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, P. Donnelly, and Consortium International HapMap. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78(3):437–50, 2006.

[6] S. R. Browning and B. L. Browning. Identity-by-descent-based heritability analysis in the northern finland birth cohort. *Hum Genet*, 132(2):129–38, 2013.

[7] P. R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–54, 2013.

[8] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

[9] S. Gravel, F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio, J. L. Rodriguez-Flores, E. E. Kenny, C. R. Gignoux, B. K. Maples, W. Guiblet, J. Dutil, M. Via, K. Sandoval, G. Bedoya, Project Genomes, T. K. Oleksyk, A. Ruiz-Linares, E. G. Burchard, J. C. Martinez-Cruzado, and C. D. Bustamante. Reconstructing native american migrations from whole-genome and whole-exome data. *PLoS Genet*, 9(12):e1004023, 2013.

[10] B. W. Lambert, J. D. Terwilliger, and K. M. Weiss. Forsim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics*, 24(16):1821–2, 2008.

[11] L. N. Borrell, E. A. Nguyen, L. A. Roth, S. S. Oh, H. Tcheurekdjian, S. Sen, A. Davis, H. J. Farber, P. C. Avila, E. Brigino-Buenaventura, M. A. Lenoir, F. Lurmann, K. Meade, D. Serebrisky, W. Rodriguez-Cintron, R. Kumar, J. R. Rodriguez-Santana, S. M. Thyne, and E. G. Burchard. Childhood obesity and asthma control in the gala ii and sage ii studies. *Am J Respir Crit Care Med*, 187(7):697–702, 2013.

[12] J. Y. Zou, E. Halperin, E. Burchard, and S. Sankararaman. Inferring parental genomic ancestries using pooled semi-markov processes. *Bioinformatics*, 31(12):i190–6, 2015.

[13] H. Schroeder, M. C. Avila-Arcos, A. S. Malaspinas, G. D. Poznik, M. Sandoval-Velasco, M. L. Carpenter, J. V. Moreno-Mayar, M. Sikora, P. L. Johnson, M. E. Allentoft, J. A. Samaniego, J. B. Haviser, M. W. Dee, Jr. Stafford, T. W., A. Salas, L. Orlando, E. Willerslev, C. D. Bustamante, and M. T. Gilbert. Genome-wide ancestry of 17th-century enslaved africans from the caribbean. *Proc Natl Acad Sci U S A*, 112(12):3669–73, 2015.

[14] R. Chakraborty and K. M. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci U S A*, 85(23):9119–23, 1988.

[15] D. G. Torgerson, D. Capurso, E. J. Ampleford, X. Li, W. C. Moore, C. R. Gignoux, D. Hu, C. Eng, R. A. Mathias, W. W. Busse, M. Castro, S. C. Erzurum, A. M. Fitzpatrick, B. Gaston, E. Israel, N. N. Jarjour, W. G. Teague, S. E. Wenzel, J. R. Rodriguez-Santana, W. Rodriguez-Cintron, P. C. Avila, J. G. Ford, K. C. Barnes, E. G. Burchard, T. D. Howard, E. R. Bleecker, D. A. Meyers, N. J. Cox, C. Ober, and D. L. Nicolae. Genome-wide ancestry association testing identifies a common european variant on 6q14.1 as a risk factor for asthma in african american subjects. *J Allergy Clin Immunol*, 130(3):622–629 e9, 2012.

[16] B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, N. Zaitlen, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. G. Ford, P. C. Avila, J. Rodriguez-Santana, G. K. Chen, L. Le Marchand, B. Henderson, D. Reich, C. A. Haiman, E. Gonzalez Burchard, and E. Halperin. Analysis of latino populations from gala and mec studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, 29(11):1407–1415, 2013.

[17] S. Sankararaman, S. Mallick, M. Dannemann, K. Prufer, J. Kelso, S. Paabo, N. Patterson, and D. Reich. The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–7, 2014.

[18] A. H. Freedman, I. Gronau, R. M. Schweizer, D. Ortega-Del Vecchyo, E. Han, P. M. Silva, M. Galaverni, Z. Fan, P. Marx, B. Lorente-Galdos, H. Beale, O. Ramirez, F. Hormozdiari, C. Alkan, C. Vila, K. Squire, E. Geffen, J. Kusak, A. R. Boyko, H. G. Parker, C. Lee, V. Tadigotla, A. Wilton, A. Siepel, C. D. Bustamante, T. T. Harkins, S. F. Nelson, E. A. Ostrander, T. Marques-Bonet, R. K. Wayne, and J. Novembre. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*, 10(1):e1004016, 2014.

[19] L. A. Frantz, J. G. Schraiber, O. Madsen, H. J. Megens, A. Cagan, M. Bosse, Y. Paudel, R. P. Crooijmans, G. Larson, and M. A. Groenen. Evidence of long-term gene flow and selection during domestication from analyses of eurasian wild and domestic pig genomes. *Nat Genet*, 47(10):1141–8, 2015.

7 Figures

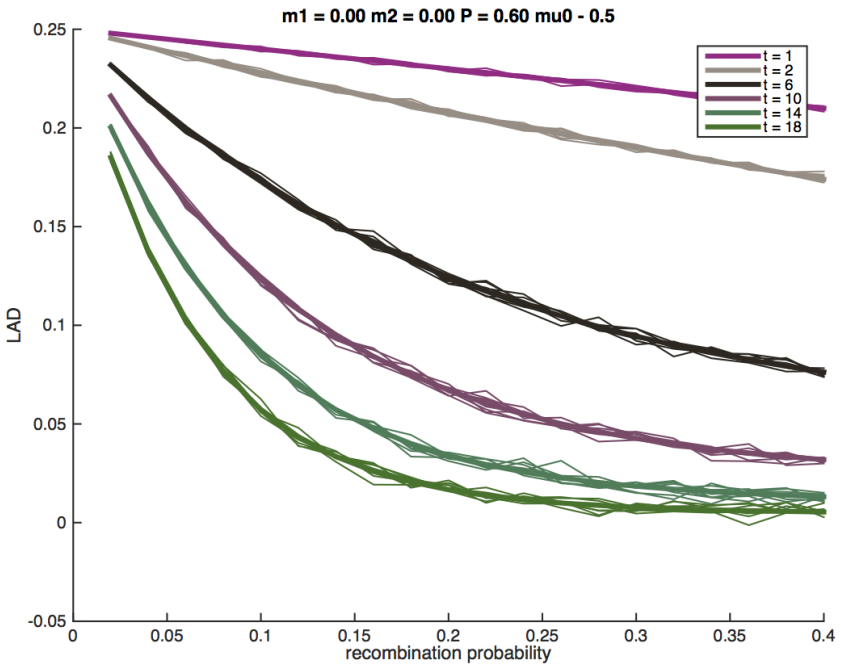

Figure 1: The distribution of LAD for different values of t with no migration (and $P = 0.6$). The thick lines correspond to the expected LAD based on Lemma 3.1, and the thin lines correspond to simulation runs of a single locus in the genome.

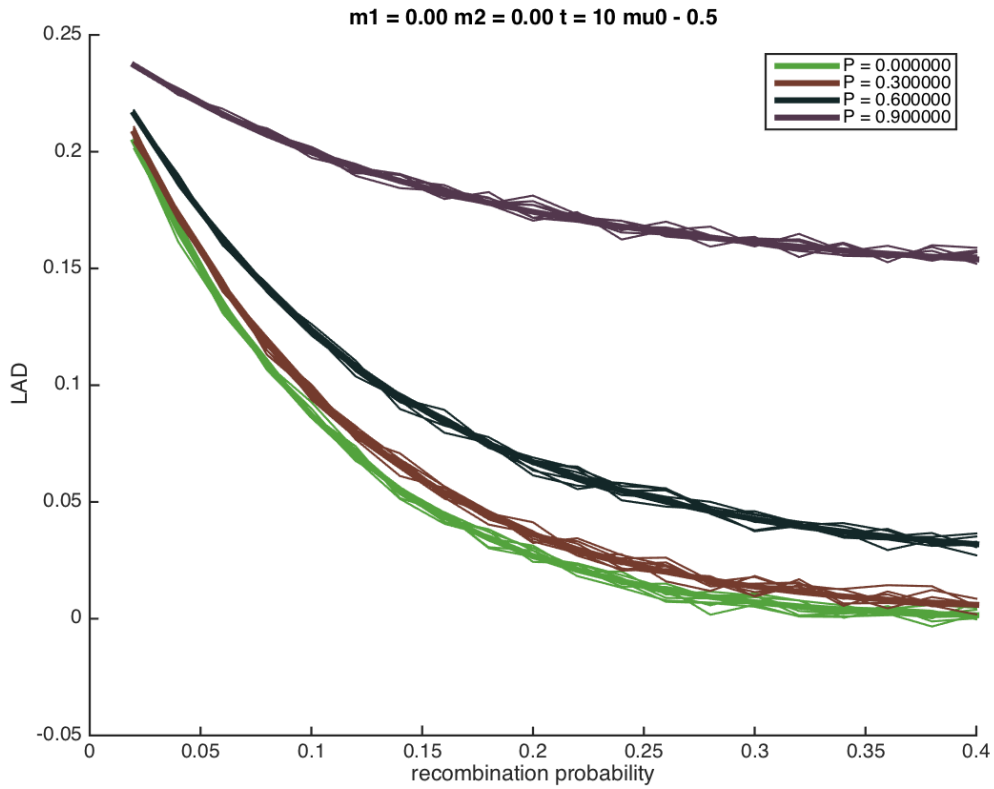

Figure 2: The distribution of LAD for different values of P with no migration. The thick lines correspond to the expected LAD based on Lemma 3.1, and the thin lines correspond to simulation runs of a single locus in the genome.

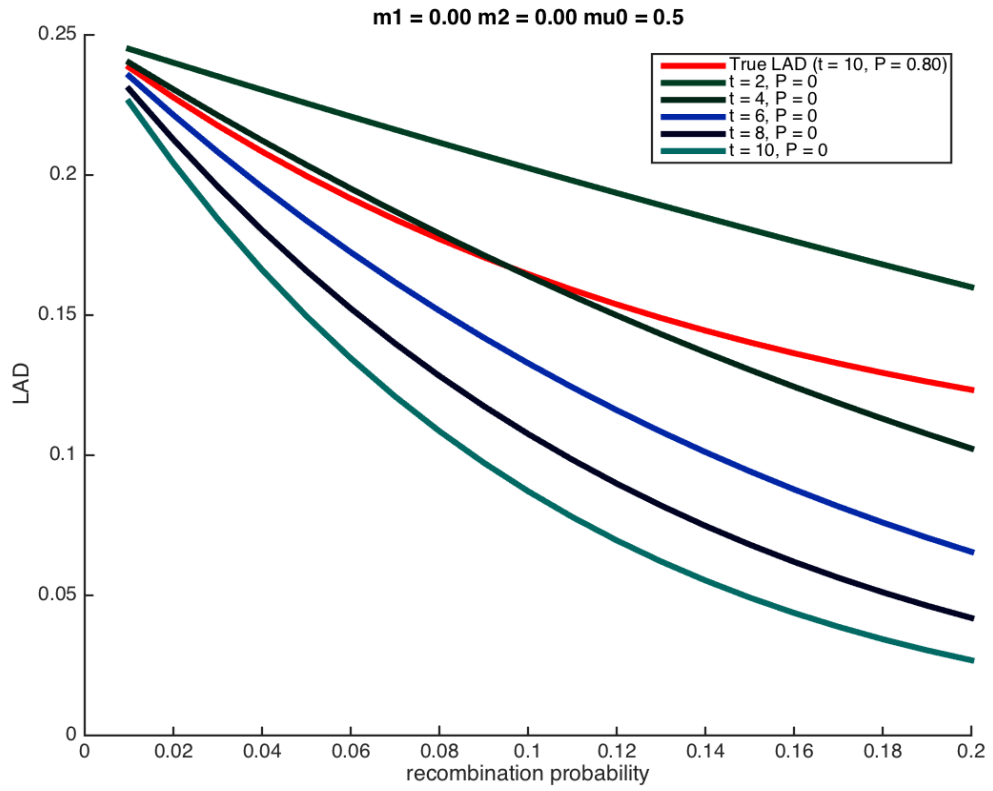

Figure 3: The distribution of LAD for different values of P with no migration. The thick lines correspond to the expected LAD based on Lemma 3.1, and the thin lines correspond to simulation runs of a single locus in the genome.

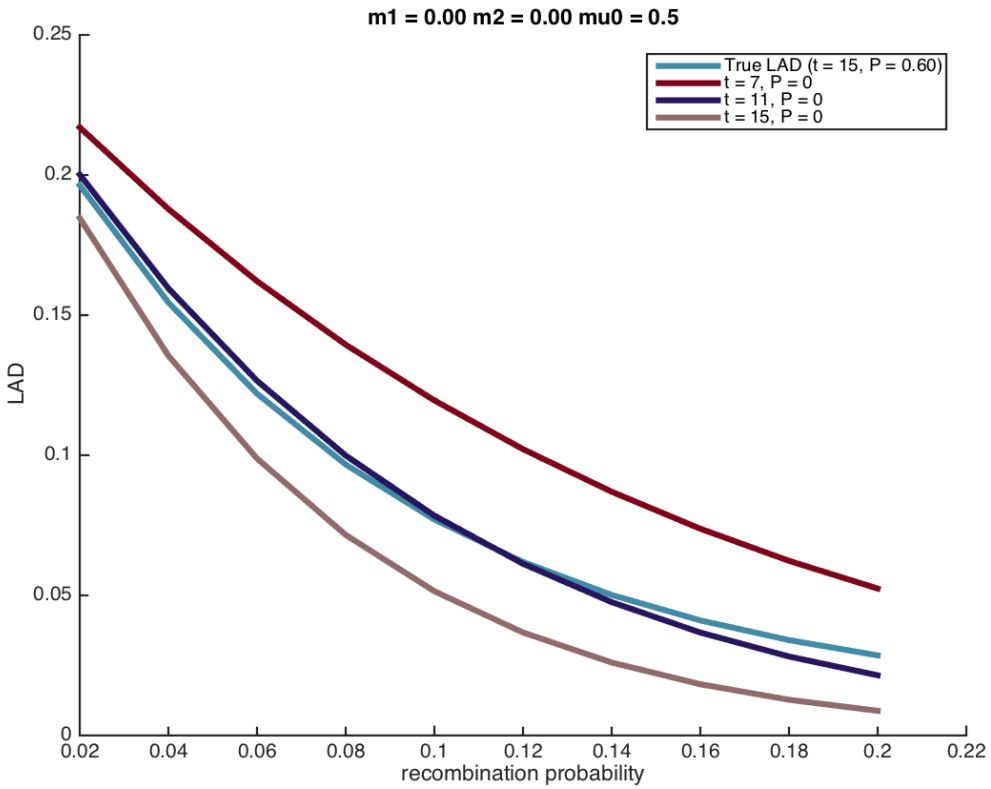

Figure 4: The distribution of LAD for different values of P with no migration. The thick lines correspond to the expected LAD based on Lemma 3.1, and the thin lines correspond to simulation runs of a single locus in the genome.

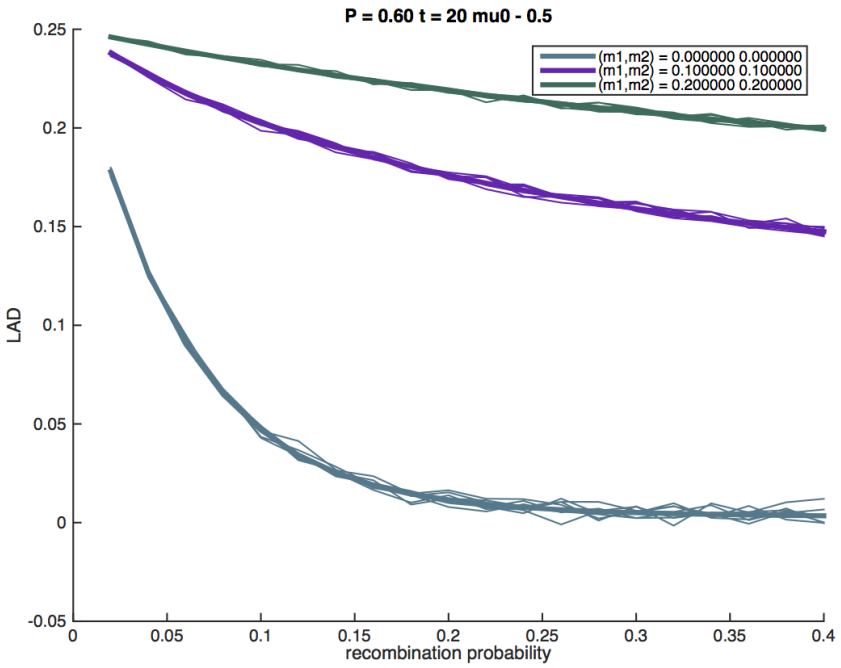

Figure 5: The distribution of LAD for different values of m_1, m_2, with equal migration rates from both populations. The thick lines correspond to the expected LAD based on Equation 2, and the thin lines correspond to simulation runs of a single locus in the genome.

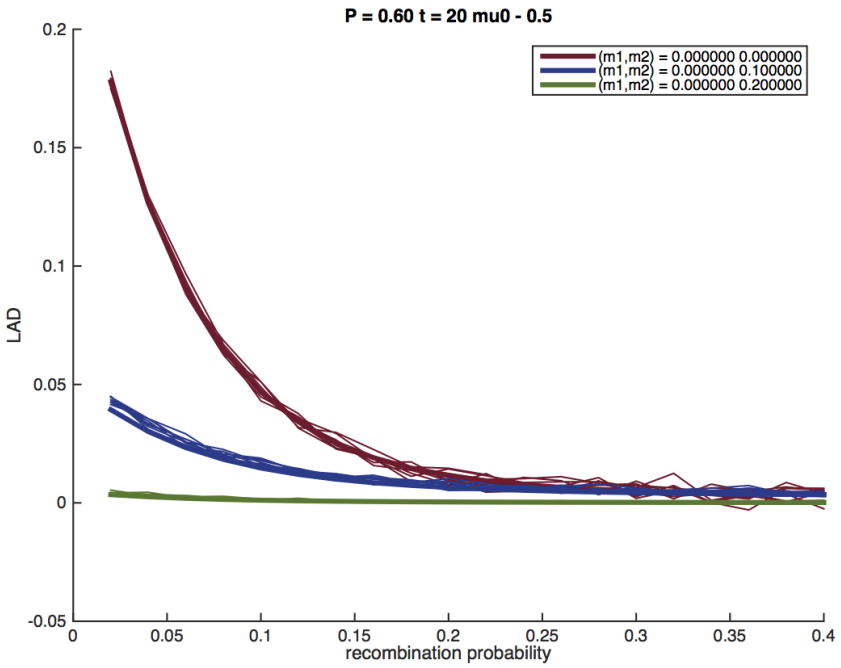

Figure 6: The distribution of LAD for different values of m_1, m_2, with no migration from population 1. The thick lines correspond to the expected LAD based on Equation 2, and the thin lines correspond to simulation runs of a single locus in the genome.

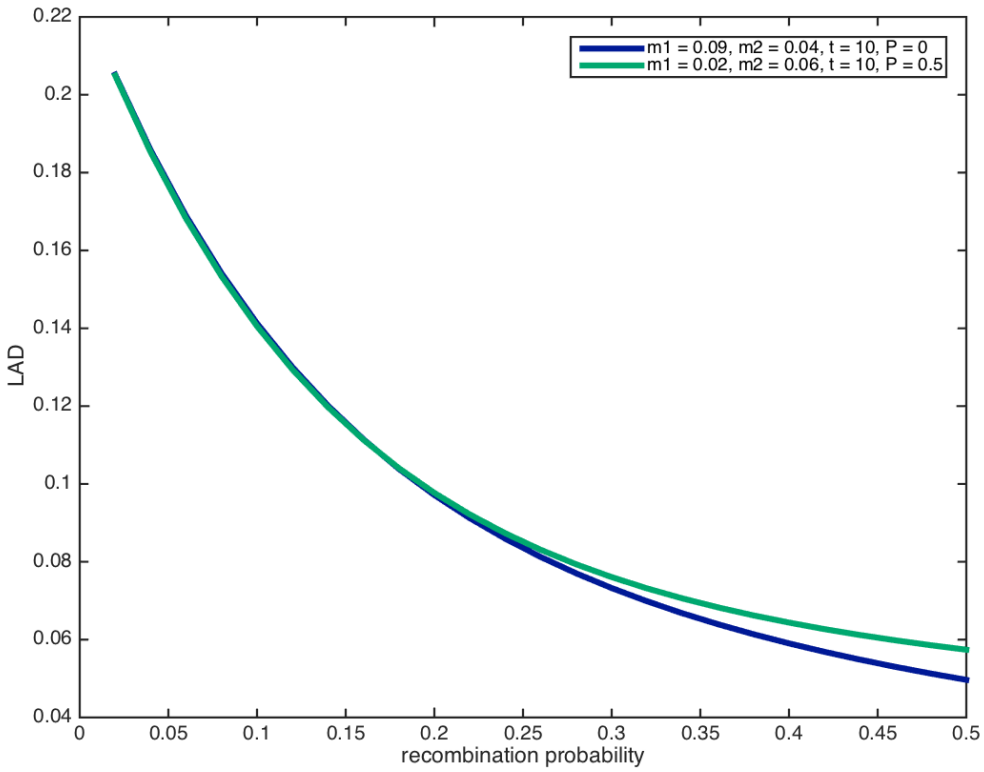

Figure 7: The expected LAD decay under two conditions, one with assortative mating and another with random mating. In the presence of migration the two curves almost overlap, and distinguishing between the two cases will be challenging in practice, particularly if LAD is measured only up to a few dozen centimorgans.

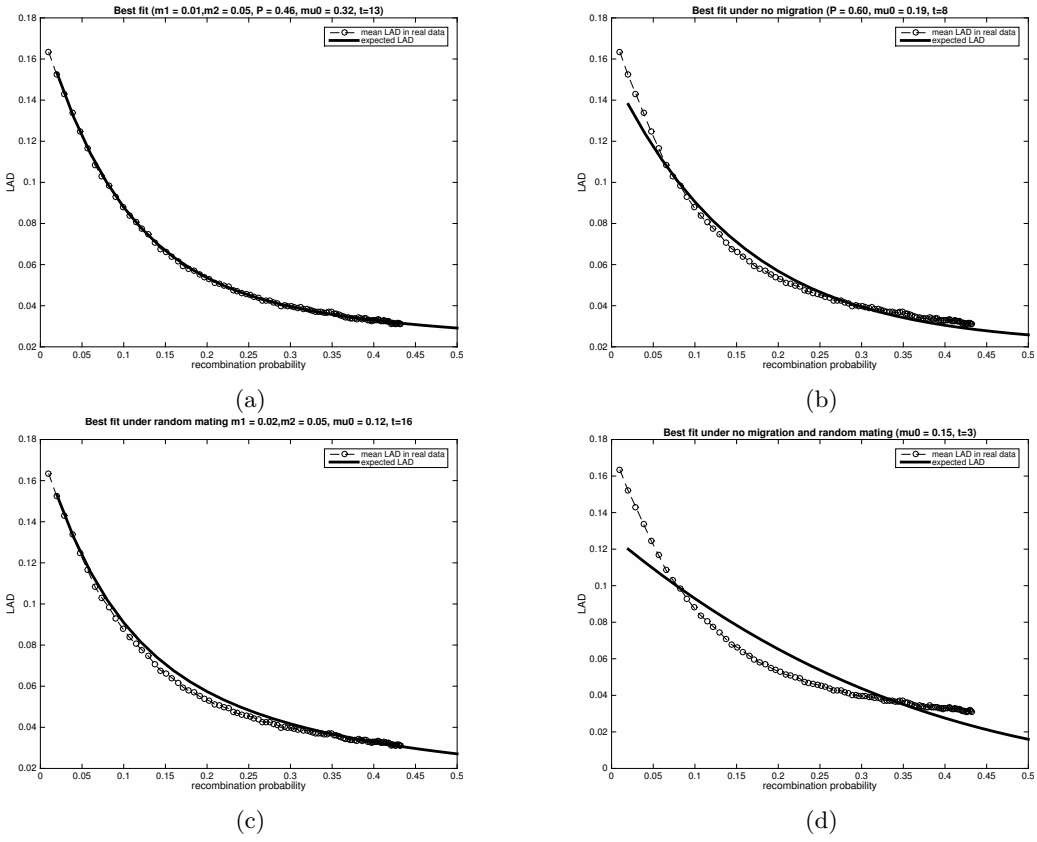

Figure 8: Each of the plots shows the best fit of the parameters to the mean LAD in the African American SAGE dataset: (a) The parameters searched over the entire grid, resulting in the best fit with estimated number of generations 13, migration rates $m_1 = 0.01, m_2 = 0.05$, and correlation $P = 0.46$. (b) The best fit under the assumption of no migration. The number of generations estimated to be 8, and $P = 0.6$. (c) The best fit under the assumption of random mating with migration. The number of generations is estimated as 16. (d) The best fit under the assumption of random mating and no migration - the number of generations is estimated as 3.