

Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep

STEVEN T. KALINOWSKI*† & PHILIP W. HEDRICK‡

†Department of Biology, Arizona State University, Tempe, AZ 85287, USA and ‡Conservation Biology Division, National Marine Fisheries Service, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

The great expansion of population genetic data using molecular techniques now allows examination of the extent of linkage disequilibrium for many pairs of loci, each locus often with multiple alleles. The expectation-maximization (EM) algorithm for generating maximum likelihood estimates of gametic frequencies from multiallelic genotypic data is described and applied. The EM algorithm is used in desert bighorn sheep where the population size, and consequently the sample size, is often small. We calculated haplotype frequencies for all pairwise combinations of five major histocompatibility loci and three microsatellite loci in 14 populations; the performance of the algorithm is discussed. Disequilibrium values are calculated and tested for statistical significance. High levels of disequilibrium are found between all pairs of major histocompatibility complex (MHC) loci and between MHC and a linked microsatellite locus.

Keywords: EM algorithm, major histocompatibility complex, microsatellites, *Ovis canadensis*.

Introduction

The recent accumulation of multilocus data has led to increasing interest in describing associations between alleles at different loci. Such associations may be useful to infer population history (Reich *et al.*, 2001), map the location of genetic diseases (Kruglyak, 1999), address forensic applications (Ayres & Balding 2001, and document associations in different parts of the genome (Huttley *et al.*, 1999). Quantifying the amount of statistical correlation, linkage (or gametic) disequilibrium, however, between alleles at different loci from molecular data is difficult when individuals are heterozygous at more than one locus. Gametic frequencies cannot be obtained directly from genotypes because the haplotypic phase of multilocus heterozygotes cannot be determined. Pedigree analysis (Eaves *et al.*, 2000) or direct molecular characterization of gametes (Taillon-Miller *et al.*, 2000) can solve this problem, but it is not generally possible for autosomal loci in wild populations.

Hill (1974) extended the gene counting method of Ceppellini *et al.* (1955) to generate maximum likeli-

hood estimates of gametic frequencies in randomly mating populations. This work has been expanded and generalized by others (Hill, 1975; Dempster *et al.*, 1977; Excoffier & Slatkin, 1995; Long *et al.*, 1995; Slatkin & Excoffier, 1996; Weir, 1996) and is now referred to as the expectation-maximization (EM) algorithm.

In this paper, we describe the EM algorithm that is used to estimate haplotype frequencies in a sample. We then use the EM algorithm to estimate the amount of gametic disequilibrium in the bighorn sheep data of Boyce *et al.* (1997). In this rare species, the sample sizes are often not large, even though a large proportion of the population has been sampled. For highly variable loci having multiple alleles, the small sample size appears to accentuate the difficulty in estimating the extent of disequilibrium. Finally, we evaluate some aspects of the EM algorithm.

Bighorn sheep data

Boyce *et al.* (1997) examined genetic variation at five ovine class II major histocompatibility complex (MHC) loci and three microsatellite loci in 14 populations of bighorn sheep, *Ovis canadensis*. Samples were collected

*Correspondence and present address. E-mail: philip.hedrick@asu.edu

from six adjacent populations of bighorn sheep in the Peninsular Mountains of southern California (Carrizo Canyon, Palm Canyon, Vallecito Mountains, Coyote Canyon, Santa Rosa Mountains, and San Jacinto Mountains), five scattered populations in the Mojave Desert of California and Nevada (San Bernadino Mountains, Eagle Mountains, Orocopia Mountains, Old Dad Mountains, and Muddy Mountains), two populations in the Chihuahuan Desert of New Mexico (San Andres Mountains and Red Rock Refuge), and one population of Rocky Mountain sheep transplanted to New Mexico from Alberta, Canada. Sample sizes ranged from nine to 29. Although these samples sizes are not large, they represent more than 20% of the individuals of this difficult to capture species in most populations (Boyce *et al.*, 1997).

MHC class II variation was evaluated with hybridization probes that identified two DQB1-like (DQB1-1 and DQB1-2) and three DRB3-like loci (DRB3-1, DRB3-2, and DRB3-3). Two to four restriction fragment length polymorphism (RFLP) alleles were identified at each of these loci. Microsatellite typing was performed for markers D5S2 (Steffen *et al.*, 1993), OARFCB11 (designated here as FCB11, Buchanan *et al.*, 1993), and DRB3 (designated here as MDRB3, Ellegren *et al.*, 1993). The D5S2 and FCB11 loci are composed of simple GT dinucleotide repeats and are on different chromosomes, unlinked to the MHC. The MDRB3 locus is a collection of GT and GA repeats located in the second intron of DRB genes and appears to be conserved in mammalian genomes (e.g. Andersson *et al.*, 1987; Reiss *et al.*, 1990; Ellegren *et al.*, 1993).

Statistical background

Consider two loci: *A* with *m* alleles, A_1, A_2, \dots, A_m and *B* with *n* alleles, B_1, B_2, \dots, B_n with respective frequencies p_1, p_2, \dots, p_m and q_1, q_2, \dots, q_n . Let f_{ij} represent the frequency of gamete $A_i B_j$ and *N* the number of individuals in the sample. The amount of linkage (or gametic) disequilibrium between alleles A_i and B_j can be expressed as the difference between the observed and expected gametic frequencies

$$D_{ij} = f_{ij} - p_i q_j. \quad (1)$$

Observed gametic frequencies are estimated with the EM algorithm as described below.

Application of the EM algorithm requires two data sets: multilocus genotypic data, and an initial estimate of gametic frequencies. Let n_{hijk} represent the number of individuals in a sample having the genotype $A_h A_i B_j B_k$.

An initial estimate of gamete frequencies, \hat{f}'_{ij} can be obtained by assuming all haplotypes are in gametic equilibrium so that

$$\hat{f}'_{ij} = p_i q_j. \quad (2a)$$

An improved estimate of the gametic frequencies, \hat{f}''_{ij} can be obtained if the population is assumed to be in Hardy-Weinberg proportions as

$$\hat{f}''_{ij} = \frac{1}{2N} \left[2n_{iijj} + \sum_{k \neq j} n_{iijk} + \sum_{h \neq i} n_{hiij} + \sum \sum n_{hijk} \frac{\hat{f}_{ij} \hat{f}_{hk}}{\hat{f}_{ij} \hat{f}_{hk} + \hat{f}_{hj} \hat{f}_{ik}} \right] \quad (2b)$$

(Hill, 1974).

The EM algorithm is iterative: estimates of gamete frequencies obtained from equation (2b) can be further improved by inserting the current estimate into the right-hand side of the equation. This iteration produces gamete frequency estimates that are more likely to produce the observed data (assuming random mating, Dempster *et al.*, 1977). Simply, the likelihood, *L*, of the observed data, given the gametic frequency estimates, increases with each iteration. If mating is random and sampling multinomial, then

$$L = a \prod_{hijk} (p_{hijk})^{n_{hijk}} \quad (3)$$

where the product is taken over all genotypes, *a* equals the multinomial coefficient, and p_{hijk} represents the probability of two randomly selected gametes producing the genotype $A_h A_i B_j B_k$. The probability equals $f_{ij} f_{hk}$ if $h=i$ and $j=k$, $2f_{ij} f_{ik}$ if $h=i$ and $j \neq k$, $2f_{ij} f_{jk}$ if $h \neq i$ and $j=k$, and $2f_{ij} f_{hk} + 2f_{ik} f_{hj}$ if $h \neq i$ and $j \neq k$. Iteration continues until the gametic frequencies converge on a set of maximum likelihood gametic frequencies. Iteration may, however, lead to different estimates of gametic frequencies when different initial gametic frequencies are used to begin the iteration (Weir & Cockerham, 1979; Excoffier & Slatkin, 1995).

The path and destination of EM iteration is dependent on the likelihoods of sets of gametic frequencies producing the observed data. Figure 1 compares the likelihoods of alternative gametic compositions of double heterozygotes. When the EM algorithm is applied to the data sets in Fig. 1(a), f_{11} will rapidly converge on the maximum likelihood frequency of 0.625 (corresponding to the top alternative in the figure) from all initial frequencies and will never converge on the less likely

Observed two locus genotypes			Composition of potential double heterozygotes	L'	Frequencies of gametes in sample			
B_1B_1	B_1B_2	B_2B_2			f_{11}	f_{12}	f_{21}	f_{22}
a)								
A_1A_1	2	1		1.000	0.625	0.250	-	0.125
A_1A_2	1			0.556	0.500	0.375	0.125	-
A_2A_2								
b)								
A_1A_1	1	1		1.000	0.500	0.333	-	0.167
A_1A_2	1			1.000	0.333	0.500	0.167	-
A_2A_2								
c)								
A_1A_1	2	2		1.000	0.500	0.333	-	0.167
A_1A_2	2			0.818	0.417	0.417	0.083	0.083
A_2A_2				1.000	0.333	0.500	0.167	-
d)								
A_1A_1	2	2		0.780	0.300	0.200	0.200	0.300
A_1A_2	2			1.000	0.250	0.250	0.250	0.250
A_2A_2	2			0.780	0.200	0.300	0.300	0.200

Fig. 1 Relative likelihoods of alternative gametic compositions of double heterozygotes. The first column lists two-locus genotypes in four hypothetical samples and the second column depicts the possible gametic composition of the double heterozygote(s). The relative likelihood, of each gametic composition. This is obtained by calculating the likelihood for the gametic frequencies of each alternative and dividing each likelihood by the largest likelihood.

lower alternative of $f_{11} = 0.500$. However, iteration of the data in Fig. 1(b,c) will converge on alternative gametic frequencies, depending on the initial starting frequencies. Such situations can appear easily as a component of highly polymorphic data sets. Figure 1(c,d) displays different ways of partitioning the gametic composition of double heterozygotes when the possible constituent gametic pairs have equal frequencies in the homozygotes and single heterozygotes.

Figure 2 depicts the relative likelihoods of the entire possible range of gametic frequencies of two samples: a data set from the Eagle Mountains and a data set similar to an observed data set from the San Jacinto Mountains. These examples have been chosen to illustrate contrasting likelihood surfaces. Although these

samples contain up to six gametes, only two gametic frequencies, e.g. f_{11} and f_{12} , are independent when the allele frequencies are specified, these frequencies are shown. Figure 2(a) gives an example of a two-locus data set with a small range of plausible gametic frequencies. Iteration from all starting points rapidly converges on the gametic frequencies, $f_{11} = 0.29$ and $f_{12} = 0.06$. Gametic frequencies for the case of gametic equilibrium for the two-locus data table in Fig. 2(a) are indicated on the figure ($f_{11} = 0.23$ and $f_{12} = 0.14$). Although the gametic frequencies at gametic equilibrium have a relatively low likelihood of 0.04, they are not statistically different from the maximum likelihood gametic frequencies using the bootstrap method described below. Iteration of gametic frequencies from random initial values for the

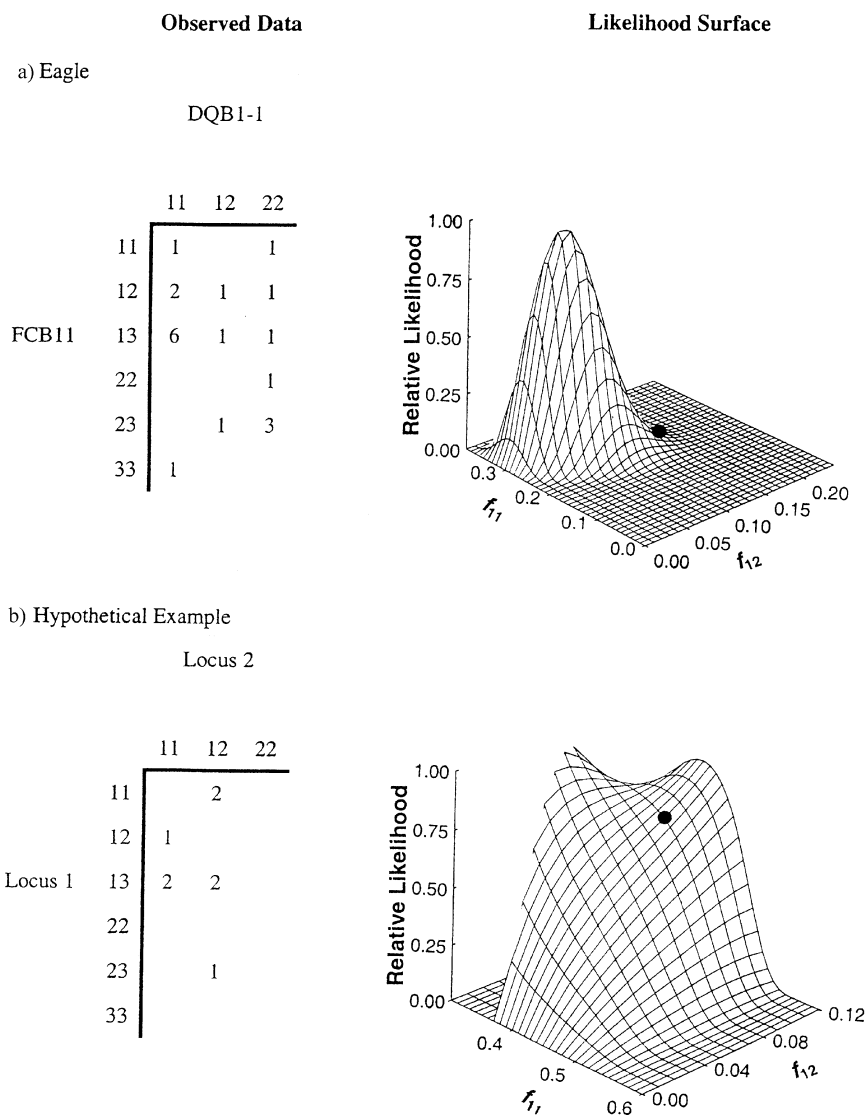


Fig. 2 Two data sets and the relative likelihoods of the range of possible gametic frequencies. The graph in 2b has been rotated to facilitate viewing, and its f_{11} axis is orientated differently than the graph in 2a.

two-locus data table in Fig. 2(b) leads to one of two alternative sets of gametic frequencies: $f_{11}=0.31$ and $f_{12}=0.06$ or $f_{11}=0.40$ and $f_{12}=0.12$ with relative likelihoods, 1.000 and 0.92, for each set of gametic frequencies, respectively. The likelihood surface in Fig. 2(b) suggests that the frequency of gamete A_1B_2 could vary from zero to its maximum value and still produce the given two-locus genotypes with a reasonable probability. The gametic frequencies at gametic equilibrium for this example ($f_{11}=0.39$ and $f_{12}=0.09$) are not significantly different than the maximum likelihood gametic frequencies using the bootstrap method described below.

The likelihood surfaces of gametic frequencies having more than two alleles at one locus and three at another cannot easily be depicted graphically. These surfaces, however, seem to become increasingly complex as the number of alleles in the sample increases because of the

increasing number of possible maximum likelihood solutions. For example, a sample with three alleles at one locus and two at another can possibly have six maximum likelihood solutions if there is no redundancy in the sample (i.e. the gametes potentially present in the double heterozygotes have equal frequencies among the individuals not heterozygous for both loci). Iteration of samples with three alleles at one locus and four at another can possibly lead to convergence for 24 different sets of gametic frequencies. There does not appear to be a simple way to predict the possible number of maximum likelihood solutions based on the number of alleles at the two loci. The likelihood of solutions must be calculated and compared to distinguish local maximum likelihood solutions from the global maximum likelihood solutions.

The complexity and unpredictability of likelihood surfaces makes thorough explanation of the multidimensional gametic frequency space necessary (Excoffier &

Slatkin, 1995). When there are n gametes, this gametic frequency space is bounded by 1 on n axes. Obtaining uniformly distributed numbers in this n dimensional space can be obtained by using the spacings of $n - 1$ uniformly distributed random numbers on the interval $[0,1]$ (Devroye, 1986).

Methods

We sought pairwise disequilibrium values for every combination of polymorphic loci for the five MHC loci and the three microsatellite loci in each of the 14 populations. The EM algorithm was used to estimate gametic frequencies in each two-locus data set. Iteration was started from 50 sets of random gametic frequencies that were generated as described above. Iteration continues until the converging estimates of gametic frequencies were stable at six decimal places. Computationally, this often required the sum of absolute difference between a set of gametic frequency estimates and the previous set of estimates to be less than 10^{-15} . If iteration led to multiple maximum likelihood solutions, then 200 sets of random starting frequencies were used. One thousand sets of starting frequencies were used for the samples between microsatellite loci that led to multiple maximum likelihood solutions, in order to see if all solutions were found with the first 200 trials. Relative likelihoods were calculated for each maximum likelihood solution by dividing the likelihood of each solution by the highest likelihood among the maximum likelihood solutions. The set of gametic frequencies with the highest likelihood was accepted as the observed gametic frequencies in the sample. If iteration led to two sets of gametic frequencies with equally high likelihoods and different disequilibrium values, the sample was not included in further analysis, this occurred in only two cases.

All maximum likelihood solutions were checked to ensure that they were not less than minimum or greater than maximum possible gametic frequencies, as determined from the observed two-locus genotypes

$$f_{ij}(\text{minimum}) = \frac{1}{2N} \left[2n_{iij} + \sum_{k \neq j} n_{iijk} + \sum_{h \neq i} n_{hiij} \right] \quad (4a)$$

$$f_{ij}(\text{maximum}) = \frac{1}{2N} \left[2n_{iij} + \sum_{k \neq j} n_{iijk} + \sum_{h \neq i} n_{hiij} + \sum_{h \neq i} \sum_{j \neq k} n_{hijk} \right] \quad (4b)$$

The number of times iteration led to each maximum likelihood set of gametic frequencies was also recorded.

The effects of sample polymorphism and sample size on the iterative process were examined. First, we sought to determine if the degree of polymorphism affected the probability of the two-locus data set producing multiple maximum likelihood estimates. We used the number of independent gametic frequencies in the sample as a measure of polymorphism. When allelic frequencies are specified, $(m - 1)(n - 1)$ of the mn gametic frequencies are independent. Secondly, we compared the sample size of the two-locus data sets that led to multiple likelihood maximums to the sample sizes of two-locus data sets that produced a single maximum likelihood estimate of gametic frequencies.

Once gametic frequency estimates were obtained, disequilibrium values were calculated as shown above. Because the range of D varies as a function of allelic frequencies, the normalized measure, D' , of Lewontin (1964)

$$D'_{ij} = \frac{D_{ij}}{D_{\max}}, \quad (5)$$

where D_{\max} is the lesser of $p_i q_j$ or $(1 - p_i)(1 - q_j)$ when $D_{ij} < 0$ and is the lesser of $p_i(1 - q_j)$ or $(1 - p_i)q_j$ when $D_{ij} > 0$, was calculated and used in all further analysis. The advantage of this measure of disequilibrium is that it has a range of -1 to 1 , regardless of the allelic frequencies in the sample. The disadvantage of is that extreme values of 1 or -1 are easier to obtain when allelic frequencies are skewed.

Disequilibrium was summarized at three levels: (1) disequilibrium between two loci within one population; (2) disequilibrium between two loci across all populations; (3) disequilibrium across all populations for groups of loci. When a sample has more than two alleles at a locus, the average absolute value of values for the gametes in a population

$$D'_A = \sum_i^m \sum_j^n |D'_{ij}| p_i q_j \quad (6)$$

serves as an effective summary measure of disequilibrium between two loci in a population (Hedrick, 1987). Disequilibrium between two loci across populations was summarized by the sample size weighted average of the values (Nei & Li, 1973). We further classified these disequilibrium values into three categories: disequilibrium between MHC loci (10 sets of values); disequilibrium between microsatellite loci (three sets of values); disequilibrium between MHC and microsatellite loci (15 sets of values). Disequilibrium between MHC and microsatellite loci was further broken down into two

sets: disequilibrium between MHC loci and the linked MDRB3 locus, and disequilibrium between MHC loci and the unlinked FCB11 and D5S2 loci. The disequilibrium values within these categories were summarized by taking the average of these values across populations and by examining their distribution.

Disequilibrium for each of the three levels was tested for statistical significance. To test the significance of observed disequilibrium between two loci in a population we used the test statistic

$$Q = 2N \sum_i^m \sum_j^n \left(\frac{\hat{D}_{ij}^2}{\hat{p}_i \hat{q}_j} \right) \quad (7)$$

$$\text{sum} \quad 2 \sum_{i=1}^k \ln p_i \quad (8)$$

where $2N$ is the number of gametes in the sample. Resampling with replacement from equilibrium gametic frequencies was used to generate an empirical distribution of 1000 Q values. The fraction of these Q values that were larger than the observed Q value was recorded as the probability, P , of obtaining the observed Q value from sampling a population in gametic equilibrium.

The significance of the average disequilibrium values for a pair of two loci across populations was tested by combining the P values for each of the 14 populations. The sum (8) where k equals the number of P values being combined, is distributed with $2k$ degrees of freedom (e.g. Sokal & Rohlf, 1995). Because the natural logarithm of zero is undefined, the probability of 0.001 was used when the estimated P value was 0.000. The P value obtained from this test represents the probability of obtaining the

observed set of k disequilibrium values solely from sampling populations in gametic equilibrium.

The significance of disequilibrium between groups of loci was tested by comparing the observed distribution of D' values with the expected distribution assuming all populations were in gametic equilibrium. The expected distribution was obtained by generating 1000 sets of D' values for all the combinations of loci in each population. The distribution of each set of D' values was found. These 1000 distributions were averaged to obtain the expected distribution. A distribution of 1000 χ^2 values was obtained by comparing each of the sampling distributions with the expected distribution. An observed χ^2 was calculated and compared to the distribution of 1000 sampled χ^2 values.

Results

The EM algorithm

Each application of the EM algorithm produced at least one set of maximum likelihood gametic frequency estimates for each of the 300 two-locus data sets polymorphic at both loci. Non-convergence of iteration was generally not a problem, but the sample from the Vallecito DQB1-1 & D5S2 data set required over 4000 iterations for the gametic frequency estimates to converge to a maximum likelihood set of gametic frequencies. All maximum likelihood estimates of gametic frequencies were equal or between the minimum and maximum frequencies defined above.

The EM algorithm was most successful producing estimates of gametic frequencies for the least polymorphic samples, which were usually those between MHC

Table 1 The samples between pairs of microsatellite loci that led to multiple EM algorithm solutions and the relative likelihood of each solution and D'_A (in parentheses except for the locus pair from Eagle with 26 solutions). N is the sample size from each population and n is the number of alleles at a locus in a population

Population (N)	Loci (n)		Relative likelihood (D'_A)							
Santa Rosa (28)	FCB11 (5)	MDRB3 (4)	1.00 (0.74)	0.91 (0.54)						
Muddy (19)	FCB11 (5)	MDRB3 (4)	1.00 (0.35)	0.86 (0.33)						
Eagle (23)	FCB11 (6)	MDRB3 (6)	1.00 0.78	0.73	0.70	0.70	0.55	0.52		
			0.52 0.51	0.43	0.14	0.09	0.09	0.08		
			0.07 0.06	0.05	0.05	0.02	0.02	0.02		
			0.01 0.01	0.01	0.01	0.00				
Eagle (23)	D5S2 (3)	FCB11 (6)	1.00 (0.48)	0.28 (0.51) 0.23 (0.48)						
Orocopia (13)	D5S2 (3)	FCB11 (5)	1.00 (0.44)	0.53 (0.36) 0.16 (0.26)						
Carrizo (22)	D5S2 (3)	FCB11 (5)	1.00 (0.53)	0.08 (0.51)						
Coyote (7)	D5S2 (3)	FCB11 (5)	1.00 (0.39)	1.00 (0.39)						
Muddy (19)	D5S2 (3)	MDRB3 (4)	1.00 (0.26)	0.95 (0.23)						
Eagle (23)	D5S2 (3)	MDRB3 (6)	1.00 (0.53)	0.70 (0.47)	0.55 (0.53)				0.41 (0.52)	
			0.41 (0.73)	0.24 (0.37)	0.22 (0.39)					

loci. To test for gametic equilibrium between each pairwise comparison of five MHC loci in 14 populations required 140 tests. However, 42 of the samples were monomorphic at one of the loci and an additional 14 samples did not have any double heterozygotes. This left 84 samples requiring the EM algorithm to estimate gametic frequencies. The EM algorithm produced a single set of gametic frequency estimates for all of these except the Muddy DQB1-2 & DRB3-1 sample. This sample had a two-locus genotype table similar to that in Fig. 1(b) and the EM algorithm produced a pair of gametic frequency estimates with equal likelihoods.

Many of the more polymorphic data sets yielded multiple maximum likelihood gametic frequencies. The microsatellite loci serve as good examples. The EM algorithm was used to estimate gametic frequencies for 40 pairs of the three microsatellite loci among the 14 populations, and nine of these 40 data sets yielded multiple equilibria. Table 1 lists the relative likelihoods of the nine microsatellite-microsatellite two-locus data sets that led to multiple maximum likelihood solutions. The likelihoods of solutions contrast considerably. The Coyote D5S2 & FCB11 sample produced a pair of gametic frequency estimates with equal likelihoods in a manner analogous to the example in Fig. 1(b). The Muddy FCB11 & MDRB3, Santa Rosa D5S2 & MDRB3, and Muddy D5S2 & MDRB3 samples produced two sets of gametic frequency estimates with nearly equal likelihoods. In contrast, the Carrizo D5S2 & FCB11 sample produced two sets of gametic frequencies with highly unequal likelihoods. The three samples from the most polymorphic population, the Eagle Mountains, each produced three maximum likelihood solutions. The most polymorphic sample, Eagle FCB11 & MDRB3, with six alleles at each locus, produced the most maximum likelihood solutions: 26 sets of gametic frequencies.

Table 1 also shows the average absolute D' values for each of the maximum likelihood solutions for eight data sets between pairs of microsatellite loci that produced multiple maximum likelihood solutions. These disequilibrium values vary less than the likelihoods. Of the nine data sets, only the Coyote D5S2 & FCB11 sample has

equal disequilibrium values for each maximum likelihood solution. Other data sets have similar disequilibrium values, e.g. Muddy D5S2 & MDRB3, Eagle D5S2 & FCB11, and Carrizo D5S2 & FCB11. The 26 Eagle FCB11 & MDRB3 maximum likelihood solutions have the greatest variability of disequilibrium values, 0.373–0.727.

The number of alleles in a two-locus data set appears to affect the probability of the EM algorithm leading to multiple maximum likelihood solutions more than other criteria examined. The samples that produced multiple sets of equilibrium gametic frequencies had more ($P < 0.001$) independent gametic frequencies than a random set of two-locus data sets. Table 2 shows that the probability of a two-locus data set leading to multiple maximum likelihood gametic frequencies increased dramatically with the number of independent gametes in the samples. On the other hand, the sample size did not appear to influence the iterative process. The size of the samples leading to multiple maximum likelihood solutions was 18.3 which is not significantly different than the average size of the data sets leading to only one solution (17.0).

Increasing the number of random starting points for iteration from 200 to 1000 produced few additional maximum likelihood solutions and did not produce a more likely set of gametic frequency estimates. Additional maximum likelihood solutions were found in the two samples already having the most solutions, Eagle FCB11 & MDRB3 and Eagle D5S2 & MDRB3. Two additional solutions were found for the Eagle FCB11 & MDRB3 sample and two additional solutions were found for the Eagle D5S2 and MDRB3 sample. Two sets of maximum likelihood solutions found for the Eagle FCB11 & MDRB3 sample during iteration from 200 random starting points were not found when 1000 random starting points were used, which suggests that other solutions may remain undetected. Six of the seven additional solutions were only produced from one of the 1000 random starting frequencies. All of the new maximum likelihood solutions had small likelihoods compared to the most likely equilibria.

Number of independent gametes	Number of samples	Number producing multiple likelihoods	Proportion of multiple likelihoods
1 or 2	167	4	0.024
3 or 4	82	4	0.049
5 or 6	21	3	0.143
7 or 8	11	3	0.273
9 or 10	9	5	0.556
> 10	8	5	0.625

Table 2 The distribution of independent gametes in the samples and the proportion of them that produced multiple maximum likelihood solutions

Average disequilibria values

D'_A values for each of the two-locus data sets are listed in the Appendix with those with $P < 0.05$ indicated in bold. Many of the data sets were in maximum disequilibrium. For example, all of the data from Carrizo Canyon for pairs of MHC loci had D'_A values of 1.0. Many of the values were significant, but a number of values as high as 1.0 were often not significant. For example, seven of the 10 values of 1.0 for Carrizo Canyon for pairs of MHC loci were not statistically significant. These high, but statistically insignificant, values occur when one or all but one of the alleles in a sample are rare. In contrast, the Eagle Mountains data set for pairs of MHC loci included two (DQB1 & DRB3-3 and DRB3-1 & DRB3-3) values that are close to 0.50 and were significant. Twelve values had a P value of 1.00, indication that all, or virtually all, of the randomly chosen samples from a population in gametic equilibrium would result in the observed amount of disequilibrium.

The 298 D'_A values in Appendix I are summarized in three ways. First, the distribution of all the D'_A values, broken down into three categories based on the types of loci, and compares these distributions to the distributions expected given gametic equilibria. Each of the observed distributions is significantly different from the expected distribution at the 0.001 level. Second, Table 3 shows the average for each pair of loci in the study and indicates the significance of each average. All of the MHC-MHC and MHC-MDRB3 averages are significant at the 0.05 level or greater while none of the MHC-D5S2 or MHC-FCB11 averages were significant. Third, Table 4 gives the mean values of data in Table 3 and the 95% empirical confidence limits. The amount of disequilibrium between pairs of MHC loci was very similar to the amount of disequilibrium between MHC loci and the MDRB3 locus. The average value between pairs of microsatellite loci was much lower, as was the average values between MHC loci and the two unlinked microsatellite loci. The expected disequilibrium generated from sample populations in gametic equilibrium

Table 3 The average D'_A observed across 14 populations between all pairwise comparisons of five MHC loci and three MS loci with the level of statistical significance indicated

		MHC				MS		
		DQB1-2	DRB3-1	DRB3-2	DRB3-3	FCB11	D5S2	MDRB3
MHC	DQB1-1	0.867***	0.879**	0.917***	0.851**	0.509	0.707	0.904***
	DQB1-2		0.877***	0.704*	0.780***	0.464	0.456	0.866***
	DRB3-1			1.000***	0.813***	0.407	0.556	0.827***
	DRB3-2				0.908***	0.540	0.609	0.940***
	DRB3-3					0.626	0.508	0.812***
MS	FCB11						0.416	0.323
	D5S2							0.559

*** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

Table 4 The average D'_A observed in 14 populations and the average expected D'_A from sampling populations in gametic equilibrium for combinations of the five major histocompatibility (MHC) loci and the three microsatellite (MS) loci. The average observed D'_A is the sample size weighted average of the D'_A values of each population for the indicated loci and the average expected D'_A is the average amount of disequilibrium obtained from 1000 random samples from the observed populations under the assumption of gametic equilibrium. The 95% confidence interval was obtained empirically from these 1000 random samples

Loci	Average observed D'_A	Average expected D'_A (95% confidence interval)	Expected/Observed
MHC – MHC	0.854	0.415 (0.344–0.480)	0.49
MHC – MDRB3	0.859	0.380 (0.298–0.449)	0.44
MHC – D5S2 or FCB11	0.530	0.372 (0.320–0.420)	0.70
MS – MS	0.427	0.326 (0.279–0.385)	0.76

represents less than half of the observed MHC-MHC disequilibrium and at least 70% of the disequilibrium observed between the microsatellite loci and unlinked loci.

Discussion

Our examination of linkage disequilibrium in a data set from bighorn sheep shows both the utility of the EM algorithm and a number of its shortcomings. Maximum likelihood estimates of gametic frequencies can be obtained from polymorphic two-locus data sets. The likelihood surfaces, however, of some data sets contain local maxima that prevent iteration from reaching the global maximum likelihood. We found this problem to become increasingly likely as the number of alleles in the sample increases. Excoffier & Slatkin (1995) found that increasing the number of loci examined had the same effect. This makes starting the iteration from many different starting points especially important for highly polymorphic samples. Another related problem is that sometimes the global maximum likelihood is shared by two (or potentially more) sets of gametic frequencies. Perhaps the greatest shortcoming of the EM algorithm is that the probability of a maximum likelihood set of gametic frequency estimates being correct or close to correct is difficult to assess. Other sets of gametic frequencies with nearly equal likelihoods may exist and may be difficult to identify. This risk is present even when iteration from random starting points always leads to the global likelihood maximum. For small data sets, it might be desirable to compute the likelihood of all possible alternative compositions of double heterozygotes to check for other gametic frequency estimates with fairly high likelihoods.

We have described a method for obtaining random points in the multidimensional gametic space that appears to be different from previous approaches (Excoffier & Slatkin, 1995; Long *et al.*, 1995). Random gametic frequencies can be obtained by dividing n random uniform numbers by their sum, but such numbers have the undesirable property of tending towards $1/n$. Informal trials showed that the method presented here finds maximum likelihood solutions faster, which will help minimize the probability of not finding the global likelihood maximum. We have chosen to randomly select points in the multidimensional gametic frequency space, but the space could be more systematically explored.

It is important to remember that maximum likelihood estimates of gametic frequencies may be influenced by several factors. The EM algorithm assumes random mating in the population, and any process that disturbs a population from Hardy-Weinberg proportions will also alter the likelihood surface climbed in the iterative process

of the algorithm. In addition, Excoffier & Slatkin (1995) have shown that gametic frequency estimates were less likely to be correct when the sample size was small. This suggests the size of our samples may affect the accuracy of our gametic frequency estimates and D' values.

Stephens *et al.* (2001) have recently developed an algorithm to estimate the frequency of multilocus haplotypes from genotypic data collected from a population. Their approach differs from the EM algorithm in its treatment of double heterozygotes that have haplotypes not evident in the other individuals sampled. In this circumstance, they assume that the unsampled haplotypes are likely to be similar to the other haplotypes in the sample. This assumption and a few others helped their phase reconstruction model outperform the EM algorithm when applied to simulated data. This method shows great promise for identifying haplotypes of individuals and estimating haplotype frequencies and could be extended to loci with multiple alleles.

Acknowledgements

We appreciate the advice of Dennis Young on how to explore the multidimensional gametic space and comments of Laurent Excoffier, Bill Hill and Monty Slatkin on an earlier version of the manuscript. S. Kalinowski and P. Hedrick were partially funded for this research on a Heritage Grant from the state of Arizona and the Ullman Professorship, respectively.

References

- ANDERSSON, G., LARHAMMAR, D., WIDMARK, E., SERVENIUS, B., PETERSON, P. A. AND RADK, L. 1987. Class II genes of the human major histocompatibility complex. *J. Biol. Chem.*, **262**, 8748–8758.
- AYRES, K. L. AND BALDING, D. J. 2001. Measuring gametic disequilibrium from multilocus data. *Genetics*, **157**, 413–423.
- BOYCE, W., HEDRICK, P. W., MUGGLI-COCKETT, N. E., KALINOWSKI, S. T., PENEDO, M. C. T. AND RAMEY, R. R. 1997. Genetic variation of major histocompatibility complex and microsatellite loci: a comparison in bighorn sheep. *Genetics*, **145**, 421–433.
- BUCHANAN, F. C., LITTLEJOHN, R. P., GALLOWAY, S. M. AND CRAWFORD, A. M. 1993. Microsatellite and associated repetitive elements in the sheep genome. *Mammal. Genome*, **4**, 258–264.
- CEPPELLINI, R., SINISCALCO, M. AND SMITH, C. A. B. 1955. The estimation of gene frequencies in a random-mating population. *Ann. Hum. Genet.*, **20**, 97–115.
- DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. B.*, **39**, 1–38.
- DEVROYE, L. 1986. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

- EAVES, I. A., MERRIMAN, T. R., BARBER, R. A. *ET AL.* 2000. The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **25**, 320–323.
- ELLEGREN, H., DAVIES, C. J. AND ANDERSSON, L. 1993. Strong association between polymorphisms in an intronic microsatellite and in the coding sequence of the BoLA-DRB3 gene: implications for microsatellite stability and PCR based DRB3 typing. *Anim. Genet.*, **24**, 269–275.
- EXCOFFIER, L. AND SLATKIN, M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
- HEDRICK, P. W. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics*, **117**, 331–341.
- HILL, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **33**, 229–239.
- HILL, W. G. 1975. Tests for association of gene frequencies at several loci in random mating diploid populations. *Biometrics*, **31**, 881–888.
- HUTTLEY, G. A., SMITH, M. W., CARRINGTON, M. AND O'BRIEN, S. J. 1999. A scan for linkage disequilibrium across the human genome. *Genetics*, **152**, 1711–1722.
- KRUGLYAK, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
- LEWONTIN, R. C. 1964. The interaction of selection and linkage. I. General considerations, heterotic models. *Genetics*, **49**, 49–67.
- LONG, J. C., WILLIAMS, R. C. AND URBANEK, M. 1995. An EM algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.*, **56**, 799–810.
- NEI, M. AND LI, W.-H. 1973. Linkage disequilibrium in subdivided populations. *Genetics*, **75**, 213–219.
- REICH, D. E., CARGILL, M., BOLK, S. *ET AL.* 2001. Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
- REISS, O., KAMMERBAUER, C., ROEWER, L. *ET AL.* 1990. Hyper-variability of intronic simple (gt)_n (ga)_m repeats in HLA-DRB genes. *Immunogenetics*, **32**, 110–116.
- SLATKIN, M. AND EXCOFFIER, L. 1996. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity*, **76**, 377–383.
- SOKAL, R. R. AND ROHLF, F. J. 1995. *Biometry*, 3rd edn. W. H. Freeman, San Francisco.
- STEFFEN, P., EGGEN, A., DIETZ, A. B., WOMACK, J. E., STRANZINGER, G. and FRIES, R. 1993. Isolation and mapping of polymorphic microsatellites in cattle. *Anim. Genet.*, **24**, 121–124.
- STEPHENS, J. C., SCHNEIDER, J. A., TANGUAY, D. A. *ET AL.* 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489–493.
- TAILLON-MILLER, P., BAUER-SARDINA, I., SACCONI, N. L. *ET AL.* 2000. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.*, **25**, 324–328.
- WEIR, B. S. AND COCKERHAM, C. C. 1979. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **42**, 105–111.
- WEIR, B. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Appendix D'_A values for 14 populations (listed in the same order as in the text) of bighorn sheep at all pairwise comparisons between five major histocompatibility complex (MHC) and three microsatellite (MS) loci. D'_A values with P values less than or equal to 0.05 are shown in bold

Loc		Ca	Pa	Va	Co	SR	SJ	SB	Ea	Or	OD	Mu	SA	RR	Wh
(MHC)	(MHC)														
DQB1-1	DQB1-2	1.00	1.00	1.00	1.00	1.00	1.00	–	0.72	–	1.00	1.00	1.00	1.00	0.12
DQB1-1	DRB3-1	1.00	1.00	1.00	0.00	1.00	1.00	–	0.47	–	1.00	1.00	1.00	1.00	1.00
DQB1-1	DRB3-2	1.00	1.00	1.00	–	–	–	–	0.71	–	1.00	1.00	–	–	0.83
DQB1-1	DRB3-3	1.00	1.00	1.00	1.00	1.00	0.50	–	0.51	–	1.00	1.00	–	–	0.50
DQB1-2	DRB3-1	1.00	0.83	0.85	1.00	0.64	1.00	1.00	0.78	0.79	1.00	–	1.00	0.77	1.00
DQB1-2	DRB3-2	1.00	1.00	1.00	–	–	–	–	0.88	–	0.18	0.15	–	–	0.52
DQB1-2	DRB3-3	1.00	0.98	0.35	1.00	0.61	0.95	1.00	0.74	0.75	0.76	0.29	–	–	0.76
DRB3-1	DRB3-2	1.00	1.00	–	–	–	–	–	1.00	–	1.00	1.00	–	–	1.00
DRB3-1	DRB3-3	1.00	1.00	1.00	1.00	0.52	1.00	1.00	0.52	1.00	1.00	0.19	–	–	0.90
DRB3-2	DRB3-3	1.00	1.00	1.00	–	–	–	–	0.92	–	1.00	0.32	–	–	1.00
(MS)	(MS)														
FCB11	D5S2	0.53	0.22	0.58	0.39	0.43	0.61	0.35	0.48	0.44	0.32	0.40	1.00	0.42	0.23
FCB11	MDRB3	0.59	0.23	0.29	0.12	0.16	0.56	0.18	0.53	0.42	0.47	0.26	–	0.25	0.19
D5S2	MDRB3	0.48	0.37	0.59	0.64	0.74	0.31	0.48	–	0.79	0.18	0.35	–	0.84	1.00
(MHC)	(MS)														
DQB1-1	FCB11	1.00	0.62	0.29	0.12	0.15	1.00	–	0.36	–	0.34	1.00	–	1.00	0.31
DQB1-2	FCB11	0.33	0.19	0.29	0.12	0.15	0.62	1.00	0.45	0.36	0.95	1.00	–	0.35	0.25

Appendix (Continued)

Loci		Ca	Pa	Va	Co	SR	SJ	SB	Ea	Or	OD	Mu	SA	RR	Wh
DRB3-1	FCB11	0.39	0.32	0.50	0.40	0.29	0.71	0.00	0.41	0.47	0.29	0.20	0	0.05	0.48
DRB3-2	FCB11	1.00	1.00	–	–	–	–	–	0.30	–	0.16	0.19	–	–	0.54
DRB3-3	FCB11	1.00	0.48	1.00	–	0.51	0.64	0.16	0.48	0.25	1.00	1.00	–	–	0.41
DQB1-1	D5S2	0.62	0.64	0.59	0.64	0.56	1.00	–	0.22	–	1.00	–	–	1.00	1.00
DQB1-2	D5S2	0.44	–	0.59	0.64	0.67	0.51	0.12	0.44	0.64	0.18	0.80	–	0.80	0.08
DRB3-1	D5S2	0.47	0.35	0.75	0.75	0.74	0.25	1.00	0.61	–	0.31	0.27	–	0.71	0.52
DRB3-2	D5S2	1.00	1.00	–	–	–	–	–	0.41	–	0.28	0.25	–	–	0.67
DRB3-3	D5S2	1.00	0.28	0.44	–	0.58	0.29	0.43	0.47	0.57	–	0.51	–	–	0.35
DQB1-1	MDRB3	1.00	1.00	1.00	1.00	0.79	1.00	–	0.65	–	1.00	1.00	–	1.00	0.84
DQB1-2	MDRB3	1.00	0.97	1.00	1.00	0.94	1.00	1.00	0.87	0.90	0.76	0.63	–	0.93	0.37
DRB3-1	MDRB3	0.78	0.88	1.00	1.00	0.65	0.77	1.00	0.70	0.79	1.00	0.94	–	0.76	0.68
DRB3-2	MDRB3	1.00	1.00	–	–	–	–	–	–	–	0.83	1.00	–	–	0.87
DRB3-3	MDRB3	1.00	0.90	1.00	–	0.50	0.94	1.00	0.75	0.81	0.89	0.37	–	–	1.00