# A Model-Based Method for Identifying Species Hybrids Using Multilocus Genetic Data

**E. C. Anderson**[1] **and E. A. Thompson**

*Department of Statistics, University of Washington, Seattle, Washington 98195*

Manuscript received October 3, 2001
Accepted for publication December 24, 2001

ABSTRACT

We present a statistical method for identifying species hybrids using data on multiple, unlinked markers. The method does not require that allele frequencies be known in the parental species nor that separate, pure samples of the parental species be available. The method is suitable for both markers with fixed allelic differences between the species and markers without fixed differences. The probability model used is one in which parentals and various classes of hybrids ($F_1$'s, $F_2$'s, and various backcrosses) form a mixture from which the sample is drawn. Using the framework of Bayesian model-based clustering allows us to compute, by Markov chain Monte Carlo, the posterior probability that each individual belongs to each of the distinct hybrid classes. We demonstrate the method on allozyme data from two species of hybridizing trout, as well as on two simulated data sets.

HYBRIDIZATION between individuals from genetically distinct populations is of interest across many fields within biology. Hybrid zones, regions where individuals from genetically distinct populations interbreed to form genetically mixed offspring, have been recognized as fertile grounds for evolutionary studies concerning models of speciation, selection, recombination, the maintenance of species boundaries, and the evolution of host-parasite interactions (Hewitt 1988; Boecklen and Spellenberg 1990; Harrison 1990). In conservation biology and resource management, hybridization between endemic species and introduced species (Goodman *et al.* 1999) or between wild and cultured populations (Elo *et al.* 1995; Jansson and Oest 1997) is a topic of great concern.

Mendelian genetic markers (Avise 1994) provide valuable tools for studying species hybridization because they allow the characterization of individuals as purebred individuals or hybrids. Such a characterization is useful, if not crucial, to the research goals in many studies of hybridizing populations. For example, identifying individuals as belonging to pure, $F_1$, $F_2$, or backcrossed classes is important for documenting gene exchange and introgression between species.

Several methods have been advanced for identifying hybrid individuals (Campton and Utter 1985; Nason and Ellstrand 1993; Barton 2000; Miller 2000; Young *et al.* 2001). One family of methods relies on the use of alleles that are unique to each species. Nason and Ellstrand (1993) present a maximum-likelihood method for estimating the proportion of individuals from a sampled population belonging to the six genealogical classes of parentals, $F_1$'s, $F_2$'s, and backcrosses that are the possible first- and second-generation products of all possible matings between two species. To employ their method, individuals in the sample must be classified exclusively to one of the six categories or be excluded from the analysis altogether because they fall into the "ambiguous" category. This requires that parental species can be sampled separately, so that the frequency of different alleles among the parentals may be estimated and then used as if known without error. Additionally, it requires that each parental species be segregating for unique alleles, though even then it is not always possible to unambiguously assign individuals to the different categories. Epifanio and Philipp (1997) note that error rates when classifying individuals to hybrid categories may be quite high if few loci are available. This is so even with diagnostic loci: loci that are fixed for alternate alleles in the different species. Boecklen and Howard (1997) give expressions and recommendations for the number of markers needed to achieve a desired level of classification error, under several restrictive assumptions such as unidirectional backcrossing and diagnostic loci. Miller (2000) provides a similar analysis describing the probability of misclassification using diagnostic dominant markers.

Other methods for identifying hybrids with genetic data do not necessarily require that the different species possess unique alleles. Campton and Utter (1985) derive a statistic that is a simple function of the conditional probability of an individual's genotypes at multiple loci given the parental species' allele frequencies. Once again, this method requires that the parental allele frequencies are already known separately from the sample of hybrids. And further, this method allows only for the

[1]*Corresponding author:* Department of Integrative Biology, University of California, Berkeley, CA 94720-3140.
E-mail: eriq@u.washington.edu

resolution of individuals into pure and hybrid categories and does not indicate directly whether individuals are $F_1$, $F_2$, or backcrossed hybrids. Barton (2000) suggests that, rather than classifying hybrids into genealogical classes, they could be classified by the number of alleles derived from each taxon and the number of heterozygous loci they possess. He suggests a moment-based method for doing such classification using the information implicit in the linkage disequilibrium present in hybridizing populations.

The above methods for classifying hybrids require that the allele frequencies of each species are known or can be estimated separately. When it is not possible to sample the different species separately and hence obtain estimates of the parental allele frequencies, the classification of hybrids is more difficult. Young *et al.* (2001) recently demonstrated the use of principal coordinate analysis (a general multivariate statistical technique) to cluster pure individuals of two species of trout and their hybrids. Individuals intermediate between the two species clusters were assumed to be hybrids. These were removed from the sample before estimating allele frequencies in the parental species. This method of separating hybrids from pure individuals has the drawback that the principal coordinate analysis is not based upon a genetic model, so the clusters are not readily interpretable, and, further, the parental allele frequencies so obtained are made under the assumption that the classification of hybrids by the principal coordinate analysis is correct.

We present a new Bayesian statistical method for identifying hybrids. Rather than assigning individuals to a single hybrid category, our method computes the posterior probability that an individual in the sample belongs to each of the different hybrid categories. This posterior probability reflects the level of certainty that an individual belongs to a hybrid category. This is an improvement over previous methods, which do not explicitly compute the probability of misclassification of particular individuals. It further allows the inference of all model parameters to be made while integrating over the uncertainty in hybrid category classifications, rather than making those inferences conditional on a single classification of individuals to hybrid or pure categories. Our method also has the following attractive features: (1) It is based upon a genetic model, so the results are easily interpreted; (2) it does not require that parental classes be sampled separately, though if they can be, then those samples can be included as prior information in the model; (3) it does not require that loci be diagnostic, or even that the species possess unique alleles—it can make use of the information in frequency differences between alleles that are not fixed in either species; (4) it incorporates the uncertainty due to the fact that allele frequencies are always estimated and are not known without error; and (5) it provides a modeling and computational framework that may be easily adapted to special cases.

Our method is related to the method of Rannala and Mountain (1997), but differs substantially in that it treats all the individuals in a sample simultaneously, rather than on a one-by-one basis. Our method also is similar to Pritchard *et al.*'s (2000) Bayesian method for analyzing structured populations. However, their method focuses on a genetic inheritance model specified in terms of the proportion of an individual's genome originating from each of a set of possible subpopulations. This is a useful heuristic model for populations with structure of unknown origin; however, when populations are known to consist of pure individuals and recent hybrids of two species, a more detailed analysis using an inheritance model defined in terms of genotype frequencies, as pursued here, is possible.

In this article, we assume that we have a sample of individuals drawn from a hybridized population and genotyped at $L$ unlinked loci. We describe the population model and the genetic model for hybridization and then describe the likelihood function that these models imply. We then describe how that likelihood is used in a Bayesian specification of the problem and how Markov chain Monte Carlo (MCMC) is carried out for simulating from the Bayesian posterior distribution. Once the method is developed, we apply it to multilocus genetic data from juvenile steelhead trout (*Oncorhynchus mykiss*), cutthroat trout (*O. clarki clarki*), and hybrids of the two species collected from a coastal stream in Washington state. We then demonstrate the method on two simulated data sets and discuss the results. One data set has many relatively uninformative markers and the other has nearly diagnostic markers. Finally, in the discussion, we note several useful extensions that could be easily handled within the framework described here.

## PROBABILITY MODEL AND COMPUTATIONAL METHODS

**Genotype frequency classes:** We consider a group of individuals in the wild that consists of sympatric populations of two species, $A$ and $B$, and hybrids of the two species that have occurred from $n$ potential generations of interbreeding. We take $n$ to be known or assumed. For example, it may be known that species $A$ was introduced $n$ generations ago to species $B$'s range, or it may be that the hybrids have reduced fitness, so hybrids remaining in the population will be hybridized for no more than $n$ generations. More practically, it is well known that individuals arising from many generations of backcrossing are difficult to distinguish from pure individuals even with many diagnostic markers (Boecklen and Howard 1997); hence the quality of the data limits the extent of the biological inferences one can make. With few markers or with low genetic differentiation between the species it could be impossible to distinguish

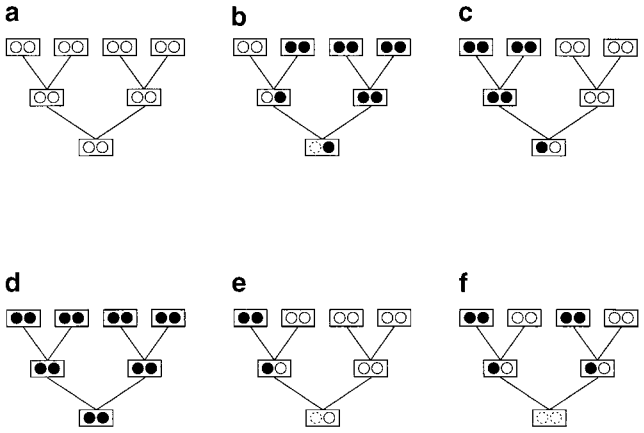


FIGURE 1.—Six arrangements of founders on a pedigree of $n = 2$ generations. Each box represents a locus. The circles within each box represent the two genes possessed by the diploid organism at the locus. The founders are the individuals in the top row of each pedigree. Black gene copies are those originating from the species *A* population, and the white genes are from species *B*. Genes that are not determined to be either black or white by the pedigree and the founders in it are denoted by broken circles. The individual at the bottom of each pedigree belongs to a different hybrid class, determined by the arrangement of species among the founders. a–f represent six distinct *genealogical classes*. a–f also represent six distinct *genotype frequency classes*. There are, however, only five distinct *gene frequency classes*; the individuals at the bottoms of pedigrees c and f are both in the same gene frequency class.

between all the hybrid categories generated by as few as $n = 2$ or $n = 3$ generations of potential interbreeding.

When hybridization between two species has been potentially occurring for $n$ generations, the possible *genealogical classes* into which an individual may fall can be enumerated and described by considering the possible arrangements of different species among the founders in an $n$-generational pedigree, up to changes in branching order at any node in the binary tree of the pedigree. The individual of interest is taken to be the member at the bottom of the pedigree and is assumed to be noninbred over the last $n$ generations; hence we assume there are no loops in its $n$-generational pedigree. Figure 1 illustrates this for the case of $n = 2$. With data only on unlinked loci, it is not possible to resolve all the genealogical classes for $n \geq 3$. For example, the expected proportions of different multilocus genotypes composed of unlinked markers for the $F_2$ and $F_3$ genealogical classes are identical. Instead, with unlinked marker data, one can only resolve what we refer to as *genotype frequency classes*.

Before describing genotype frequency classes, however, it is convenient to consider a simpler classification of hybrid individuals into *gene frequency classes*. Members of the same gene frequency class have the same expected proportion of all their genes originating from species *A*. This is determined by the number of founders (without regard to their arrangement) from each species at the top of the pedigree. Since there are $2^n$ founders for a pedigree with $n$ generations, there are $2^n + 1$ different gene frequency classes that are determined by the number, $a$, of founders originating from the species *A* population ($a = 0, 1, \ldots, 2^n$). In Figure 1, both c and f belong to the gene frequency class with $a = 2$. The individuals at the bottoms of the other pedigrees belong to the remaining four distinct gene frequency classes.

We use uppercase $Q$ to denote the proportion of an individual's genome derived from the species *A* population. This quantity, which we refer to as the "genetic heritage proportion," is discrete and is determined by the gene frequency class to which an individual belongs, namely $Q = a/2^n$, with $a$ defined as in the previous paragraph. We note that $Q$ is the quantity estimated by BARTON and GALE's (1993) familiar hybrid index $z$ and also that $Q$ is closely related to the latent variable $q_k^{(i)}$ described in the model with admixture developed by PRITCHARD *et al.* (2000) as the proportion of the genome of the $i$th individual originating from population $k$. $Q$ is necessarily discrete in our approach since we perform the analysis conditional upon $n$, the number of generations of potential interbreeding. By contrast $q_k^{(i)}$ is used as a continuous variable by PRITCHARD *et al.* (2000), although they do treat $q_k^{(i)}$ as a discrete variable in their model for detecting immigrants with prior population information. The PRITCHARD *et al.* (2000) model for detecting immigrants is similar to what we propose here for detecting hybrids; however, their model is restricted to the case where the number, $a$, of immigrant founders on a sampled individual's pedigree does not exceed one, and it does not make use of the expected frequencies of single-locus genotypes that we discuss below.

With gene frequency classes and the genetic heritage proportion so defined, it is straightforward to enumerate and define the genotype frequency classes. The members of a genotype frequency class all have the same expected proportion of single-locus genotypes possessing 0, 1, or 2 genes originating from species *A*. For the $g$th genotype frequency class we denote these expected proportions by $\boldsymbol{G}_g = (G_{g,0}, G_{g,1}, G_{g,2})$, respectively. Enumerating these genotype frequency classes and computing the expected proportions of the genotypes follows from Mendel's laws. Since each individual receives one gene copy randomly selected from the two in its mother and another randomly selected from the two in its father, the expected proportions of the genotypes in an individual are determined by the gene frequency classes to which its parents belong. For the $g$th genotype frequency class, we have

$$\begin{aligned} G_{g,0} &= (1 - Q_m)(1 - Q_f) \\ G_{g,1} &= Q_m(1 - Q_f) + Q_f(1 - Q_m) \\ G_{g,2} &= Q_m Q_f, \end{aligned} \tag{1}$$

where $Q_m$ and $Q_f$ are the genetic heritage proportions of the individual's mother and father (and the sexes of the parents are interchangeable). Straightforward algebra verifies that two individuals $i$ and $j$ will belong to the same genotype frequency class if and only if the parents of $j$ belong to the same gene frequency classes as the parents of $i$. Consequently, the number of distinct genotype frequency classes after $n$ generations of possible interbreeding can be calculated as the number of unordered pairs that may be formed from the $2^{n-1} + 1$ gene frequency classes after $n - 1$ generations: $(2^{n-1} + 1)(2^{n-1} + 2)/2$. We denote this quantity by $\mathcal{G}_n$. For $n \geq 2$ there are always more genotype frequency classes than there are gene frequency classes. With data on multiple unlinked loci, it is possible to distinguish between individuals in different genotype frequency classes. This is our primary inference goal and will be pursued in the Bayesian context by computing the posterior probability that each individual belongs to each of the $\mathcal{G}_n$ genotype frequency classes. The following section describes the data and the probability model for making such inference.

**Genetic data and probability model:** We have a sample of $M$ individuals drawn for genetic analysis. For now, we assume that individuals are sampled randomly and independently of whether they are purebred individuals of either species or are hybrids. This sort of sampling would arise if, for example, the two species, or hybrids thereof, were difficult to distinguish on the basis of morphology—so-called "cryptic" hybridization. Each individual in the sample is genotyped at $L$ unlinked loci. Let the $\ell$th locus possess $K_\ell$ alleles detected in the sample. We denote the allele frequencies in species $A$ and $B$, $n$ generations ago, by $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$, respectively. Each of these $\boldsymbol{\Theta}$'s is a collection of vectors, with each vector giving the allele frequencies at a particular locus. For example, for species $A$, $\boldsymbol{\Theta}_A = (\boldsymbol{\theta}_{A,1}, \ldots, \boldsymbol{\theta}_{A,L})$, where $\boldsymbol{\theta}_{A,\ell} = (\theta_{A,\ell,1}, \ldots, \theta_{A,\ell,K_\ell})$ are allele frequencies at the $\ell$th locus. The alleles found in individuals from species $A$ and species $B$ are assumed to be drawn randomly from the allele frequencies $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$, respectively, $n$ generations ago. Likewise, individuals $n$ generations before sampling are assumed to be in Hardy-Weinberg and linkage equilibrium with reference to their contemporaneous conspecifics; thus, linkage disequilibrium and Hardy-Weinberg disequilibrium in the mixed population are assumed to result entirely from the mixing and admixing of the gene pools of species $A$ and species $B$.

Within an individual, the two gene copies carried at any locus are considered to be ordered and indexed by $j = 1$ or 2. The order of the gene copies is arbitrary; for example, it may merely be the order in which the genetic data on that locus in that individual happened to be recorded. We do not know from which species each of an individual's gene copies descended, but we denote that unknown information by the latent variable $\boldsymbol{W}_{i,\ell} = (W_{i,\ell,1}, W_{i,\ell,2})$. $W_{i,\ell,j}$ takes the value 1 if the $j$th gene copy at the $\ell$th locus of the $i$th individual originated from the species $A$ population, and it takes the value 0 if that gene copy originated from species $B$. We use $\mathbf{W}_i = (\boldsymbol{W}_{i,1}, \ldots, \boldsymbol{W}_{i,L})$ to denote all the latent gene origin indicators in the $i$th individual, and $\mathbf{W}$ denotes the latent gene origin indicators in all the individuals.

The allelic types of the two gene copies at locus $\ell$ in individual $i$ are denoted by $\boldsymbol{Y}_{i,\ell} = (Y_{i,\ell,1}, Y_{i,\ell,2})$, with each of $Y_{i,\ell,1}$ and $Y_{i,\ell,2}$ taking an integer value between 1 and $K_\ell$, inclusive, corresponding to the possible allelic types at the $\ell$th locus. The $L$ single-locus genotypes in the $i$th individual are denoted by $\mathbf{Y}_i = (\boldsymbol{Y}_{i,1}, \ldots, \boldsymbol{Y}_{i,L})$, and all of the genetic data over all $M$ individuals in the sample is $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_M)$. We introduce some notation here to avoid awkward subscripting: let $\theta_A\langle i; \ell; j\rangle$ denote the frequency in the species $A$ population of the allele possessed by the $i$th individual at the $j$th ($j = 1, 2$) gene copy of its $\ell$th locus. This is a shorthand for the doubly subscripted $\theta_{A,\ell,Y_{i,\ell,j}}$. We also introduce the latent variable $Z_i$: $Z_i = g$ indicates that individual $i$ in the sample belongs to genotype frequency class $g$.

Given the population allele frequencies, the gene origin indicators, and the genotype frequency class to which an individual belongs, it is easy to compute the probability of that individual's single-locus genotype at the $\ell$th locus. For our purposes later, it is more useful to have an expression for the joint probability of the genotype and the gene origin indicators. At the $\ell$th locus in individual $i$ belonging to genotype frequency class $g$, this joint probability is

$$P(\boldsymbol{Y}_{i,\ell}, \boldsymbol{W}_{i,\ell} | Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}) = \begin{cases} \theta_A\langle i; \ell; 1\rangle\theta_A\langle i; \ell; 2\rangle G_{g,2}, & \text{if } W_{i,\ell,1} = W_{i,\ell,2} = 1 \\ \theta_A\langle i; \ell; 1\rangle\theta_B\langle i; \ell; 2\rangle G_{g,1}/2, & \text{if } W_{i,\ell,1} = 1,\ W_{i,\ell,2} = 0 \\ \theta_B\langle i; \ell; 1\rangle\theta_A\langle i; \ell; 2\rangle G_{g,1}/2, & \text{if } W_{i,\ell,1} = 0,\ W_{i,\ell,2} = 1 \\ \theta_B\langle i; \ell; 1\rangle\theta_B\langle i; \ell; 2\rangle G_{g,0}, & \text{if } W_{i,\ell,1} = W_{i,\ell,2} = 0. \end{cases} \quad (2)$$

The product of the two allele frequencies in the above expressions follows from the assumption that each gene copy in the founders ($n$ generations ago) of the $i$th individual is sampled randomly from the alleles present in its population of origin. Then, $G_{g,2}$ is the probability that an individual in genotype frequency class $g$ has both gene copies originating from species $A$, $G_{g,1}/2$ is the probability that the first (second) gene copy originates from species $A$ and the second (first) originates from species $B$, and $G_{g,0}$ is the probability that both gene copies originated from species $B$.

For a given genotype frequency class, the marginal probability of the $i$th individual's genotype at locus $\ell$ is computed by summing (2) over the latent gene origin indicators:

$$P(\boldsymbol{Y}_{i,\ell} | Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}) = \sum_{\substack{0 \leq W_{i,\ell,1} \leq 1 \\ 0 \leq W_{i,\ell,2} \leq 1}} P(\boldsymbol{Y}_{i,\ell}, \boldsymbol{W}_{i,\ell} | Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}). \quad (3)$$

Finally, under the assumption of unlinked markers in Hardy-Weinberg and linkage equilibrium among conspecifics $n$ generations ago, the probability of the $i$th individual's multilocus genotype is just the product over the $L$ single-locus genotype probabilities:

$$P(\mathbf{Y}_i|Z_i = g, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B) = \prod_{\ell=1}^{L} P(\mathbf{Y}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}). \quad (4)$$

This gives us an expression for the probability of the data on a single individual. We now must derive the probability of the data on all $M$ individuals in the sample. We do this by modeling the hybridized population as a mixture with unknown proportions of individuals from the different genotype frequency classes.

As shown earlier, given $n$ generations of potential interbreeding between the species, the members of the genetic sample may fall into $\mathcal{G}_n = (2^{n-1} + 1)(2^{n-1} + 2)/2$ genotype frequency classes. We model the individuals in the sample as being randomly and independently drawn from a mixture of individuals, each belonging to one of the $\mathcal{G}_n$ genotype frequency classes with probability $\pi_g$, $g = 1, \ldots, \mathcal{G}_n$, $\sum_{g=1}^{\mathcal{G}_n} \pi_g = 1$. Using $\boldsymbol{\pi}$ to denote the vector of mixing proportions, $(\pi_1, \ldots, \pi_{\mathcal{G}_n})$, we may now write the probability of all the observed data $\mathbf{Y}$ conditional on $n$, $\Theta_A$, $\Theta_B$, and $\boldsymbol{\pi}$ as the product over the $M$ members of the sample of the probability of each of their multilocus genotypes:

$$P(\mathbf{Y}|\boldsymbol{\Theta}_A,\boldsymbol{\Theta}_B,\boldsymbol{\pi}) = \prod_{i=1}^{M}\left(\sum_{g=1}^{\mathcal{G}_n} \pi_g P(\mathbf{Y}_i|Z_i = g,\boldsymbol{\Theta}_A,\boldsymbol{\Theta}_B)\right). \quad (5)$$

Assigning each genotype frequency class a separate mixing proportion provides a means of accounting for possible differential fitness of the different classes.

**A Bayesian specification:** Equation 5 is the likelihood for $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$, and $\boldsymbol{\pi}$. To pursue Bayesian inference in this problem requires prior distributions $P(\boldsymbol{\Theta}_A)$, $P(\boldsymbol{\Theta}_B)$, and $P(\boldsymbol{\pi})$, so that the posterior distribution may be computed. We wish to make inferences not only about $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$, and $\boldsymbol{\pi}$, but also the latent variables $\mathbf{W}$ and $\mathbf{Z} = (Z_1, \ldots, Z_M)$, so we are concerned with the joint posterior distribution, which is proportional to the joint probability of all the variables, $P(\mathbf{Y}, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}, \mathbf{Z}, \mathbf{W})$. With the latent variables present, this joint density factorizes as

$$P(\mathbf{Y}, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}, \mathbf{Z}, \mathbf{W}) = P(\boldsymbol{\Theta}_A)P(\boldsymbol{\Theta}_B)P(\boldsymbol{\pi}) \times \prod_{i=1}^{M} P(\mathbf{Y}_i|\mathbf{W}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)P(W_i|Z_i)P(Z_i|\boldsymbol{\pi}), \quad (6)$$

which, as we see in the next section, allows straightforward MCMC sampling from the posterior distribution, $P(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}, \mathbf{Z}, \mathbf{W}|\mathbf{Y})$.

It is computationally convenient and biologically reasonable to take the specific form of the prior distributions $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$ to be Dirichlet distributions, independent over the $L$ unlinked loci. That is, $P(\boldsymbol{\theta}_{A,\ell})$ is Dirichlet $(\lambda_{A,\ell,1}, \ldots, \lambda_{A,\ell,K_\ell})$. Since the Dirichlet distribution is the conjugate prior for the multinomial distribution, this choice facilitates simulation from the full conditional distributions for $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$. The Dirichlet distribution is also the multivariate generalization of the beta distribution, which arises theoretically as the equilibrium distribution for gene frequencies in the presence of genetic drift and linear pressure from migration or mutation (Wright 1938, 1952). Specification of the parameters $\boldsymbol{\lambda}_{A,\ell} = (\lambda_{A,\ell,1}, \ldots, \lambda_{A,\ell,K_\ell})$ and $\boldsymbol{\lambda}_{B,\ell} = (\lambda_{B,\ell,1}, \ldots, \lambda_{B,\ell,K_\ell})$ provides a way to incorporate prior information about the allele frequencies among the two species at the $\ell$th locus. If, at locus $\ell$, previous studies have indicated that species $B$ has very low frequency of allele $j$ while species $A$ has high frequency, then $\lambda_{B,\ell,j}$ should be chosen small, relative to the other components of $\boldsymbol{\lambda}_{B,\ell}$, while $\lambda_{A,\ell,j}$ should be chosen large. If, on the other hand, very little prior knowledge is available about allele frequencies in the two species, then a sensible choice of prior $\lambda_{A,\ell,j} = \lambda_{B,\ell,j} = 1/K_\ell$ for $j = 1, \ldots, K_\ell$. This has the form of the Jeffreys prior for a multinomial proportion (see Gelman *et al.* 1996).

The conjugate prior for $\boldsymbol{\pi}$ is also a Dirichlet distribution, so it is helpful to let $P(\boldsymbol{\pi}) \equiv$ Dirichlet $(\zeta_1, \ldots, \zeta_{\mathcal{G}_n})$. As with the allele frequencies, the parameters $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{\mathcal{G}_n})$ could be specified so as to reflect prior knowledge of the biology of the situation. For example, if it was well known that backcrosses between $F_1$ hybrids and species $A$ had low fitness, then that could be reflected in the prior for $\boldsymbol{\pi}$. Additionally, if hybridization and backcrossing were known to be fairly rare, then this prior knowledge could be reflected by having smaller $\zeta_g$'s for those genotype frequency classes that required more episodes of interbreeding within the last $n$ generations. In the absence of prior information on hybridization rates, the prior $\zeta_g = 1/\mathcal{G}_n$, $g = 1, \ldots, \mathcal{G}_n$ is again a suitable choice.

The influence of prior distributions in this sort of hierarchical model is not easy to predict, and determining a good prior distribution to use in the absence of prior information is not a simple task (Kass and Wasserman 1996). A second reasonable choice of prior is the uniform Dirichlet distribution—$\lambda_{A,\ell,j} = \lambda_{B,\ell,j} = 1$ for $j = 1, \ldots, K_\ell$ or $\zeta_g = 1$, $g = 1, \ldots, \mathcal{G}_n$. However, this uniform prior represents a substantial amount of information when the number of components is large (for example, if $K_\ell$ or $\mathcal{G}_n$ is large). For this reason we prefer the use of the previously described Jeffreys-type priors, which reflect the amount of information contained in a single observation, rather than in $K_\ell$ or $\mathcal{G}_n$ observations. Because of the fact that some of the genotype frequency classes will have few if any members in the sample, and because some alleles will be found only in low frequency in either of the species, there is potential that the posterior will be sensitive to the choice of prior. We used various combinations of the uniform and Jeffreys-type priors in analyzing the data sets described later in the article and found that, although there were differences in some of the specific posterior probabilities computed, they were not large enough to

alter the conclusions made from the analysis based on the Jeffreys-type priors.

**MCMC simulation from the posterior distribution:** It is not possible to compute directly the posterior distributions for only the variables that we are interested in. However, simulating from the joint posterior distribution of all the variables by MCMC can be done via Gibbs sampling in a manner similar to that for normal finite mixture models (Diebolt and Robert 1994). Gibbs sampling is a special case of the Hastings (1970) algorithm for constructing an ergodic Markov chain having a unique stationary distribution for which the normalizing constant may be unknown. In this case, the desired stationary distribution is the posterior distribution of all unknown variables in the model and is known up to scale (*i.e.*, without knowing the normalizing constant) by Equation 6. Given initial starting values for all the variables in the model, Gibbs sampling proceeds by successively simulating new values for particular variables in the model from their full conditional distributions (Geman and Geman 1984). After a sufficient period of burn-in, a sample of variables drawn from this joint posterior distribution allows Monte Carlo estimation of the posterior distribution of any subset of variables of interest, either marginally or conditional on the values taken by another subset of variables.

We denote full conditional distributions by $P(\cdot|\cdots)$ and refer to a standard iteration of our MCMC algorithm as a "sweep." A sweep consists of a series of steps in which each of the variables in the probability model (except for the data, $\mathbf{Y}$, which are fixed) is updated once. Here, with the probability distributions given in more detail below, the steps in a single sweep are as follows:

1. For $\ell = 1, \ldots, L$, simulate new values for $\boldsymbol{\Theta}_{A,\ell}$ and $\boldsymbol{\Theta}_{B,\ell}$ from $P(\boldsymbol{\Theta}_A|\cdots)$ and $P(\boldsymbol{\Theta}_B|\cdots)$, respectively.
2. Simulate a new value of $\boldsymbol{\pi}$ from $P(\boldsymbol{\pi}|\cdots)$.
3. For $i = 1, \ldots, M$ and $\ell = 1, \ldots, L$, simulate a new value of $\boldsymbol{W}_{i,\ell}$ from $P(\boldsymbol{W}_{i,\ell}|\cdots)$.
4. For $i = 1, \ldots, M$, simulate a new value of $Z_i$ from $P(Z_i|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi})$.

By sampling the current states of all the variables after each sweep, one acquires a dependent sample suitable for Monte Carlo estimation of most quantities of interest. In particular, a Monte Carlo estimate of the posterior probability that individual $i$ is of the $g$th genotype frequency class is obtained by averaging the values of $P(Z_i = g|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi})$ computed during each sweep.

The full conditional distributions are easily derived. By conjugacy,

$$P(\boldsymbol{\theta}_{A,\ell}|\cdots) \equiv \text{Dirichlet}(\lambda_{A,\ell,1} + r_{A,\ell,1}, \ldots, \lambda_{A,\ell,K_\ell} + r_{A,\ell,K_\ell}), \tag{7}$$

where $r_{A,\ell,j}$ is the number of gene copies of allelic type $j$ at the $\ell$th locus currently allocated to species $A$ (*i.e.*, gene copies of allelic type $j$ for which the corresponding $W_{\cdot,\ell,\cdot} = 1$). An analogous expression exists for $P(\boldsymbol{\theta}_{B,\ell}|\cdots)$. The full conditional distribution for $\boldsymbol{\pi}$ is also easily computed by conjugacy as

$$P(\boldsymbol{\pi}|\cdots) \equiv \text{Dirichlet}(\zeta_1 + s_1, \ldots, \zeta_{\mathcal{G}_n} + s_{\mathcal{G}_n}), \tag{8}$$

where $s_g$ is the number of individuals in the sample currently allocated to genotype frequency class $g$. The full conditional distribution for the pair of gene origin indicators $\boldsymbol{W}_{i,\ell}$ in the $i$th individual currently included in the $g$th genotype frequency class is obtained by Bayes' law as

$$P(\boldsymbol{W}_{i,\ell}|\cdots) = \frac{P(\mathbf{Y}_{i,\ell}, \boldsymbol{W}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})}{P(\mathbf{Y}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})}, \tag{9}$$

where the numerator and denominator are given in (2) and (3), respectively. Finally, the full conditional distribution for $Z_i$ would be $P(Z_i|\mathbf{W}_i, \boldsymbol{\pi})$. However, it is possible to integrate out the $\mathbf{W}_i$ conditional on the remaining variables and hence simulate new values of $Z_i$ from $P(Z_i|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi})$, instead of from $P(Z_i|\mathbf{W}_i, \boldsymbol{\pi})$. This is an example of improving MCMC efficiency through "collapsed Gibbs sampling" (Liu 1994). In our case, this is particularly attractive, since computing $P(Z_i|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi})$ incurs almost no extra cost—the quantities needed to calculate it have already been computed in step 3 of the sweep. By Bayes' law

$$P(Z_i = z|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}) = \frac{\pi_z P(\mathbf{Y}_i|Z_i = z, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B)}{\sum_{g=1}^{\mathcal{G}_n} \pi_g P(\mathbf{Y}_i|Z_i = g, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B)}, \quad z = 1, \ldots, \mathcal{G}_n. \tag{10}$$

Multiple chains are run from different starting values to diagnose mixing problems. In this case, it is easy to assign overdispersed starting values by simulating values of $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$, and $\boldsymbol{\pi}$ from their prior distributions rather than their full conditional distributions in steps 1 and 2 of the first sweep. We used Gelman's (1996) estimated scale reduction potential factor to monitor convergence of the chains to the desired posterior distribution. This quantity is computed for each scalar variable in the posterior distribution. For each such variable, the potential scale reduction is computed as the square root of the ratio of the variance of the variable estimated by using information from all of the multiple chains to the variance of the variable estimated by using the values simulated from just a single one of the chains. Values of the scale reduction potential near 1 indicate that the chains have converged to the target distribution.

In the case of data from cutthroat trout and steelhead trout described in the following section this convergence occurs very rapidly. Burn-in then requires little time. This will typically be the case for genetically well-separated species. Poor mixing of the Markov chain may occur for species that are not genetically well separated. Pritchard *et al.* (2000) discuss strategies for specifying the allele frequency prior distributions to improve mix-

ing in such cases. However, it is likely that with genetic differentiation low enough to have mixing problems with the MCMC, it will be very difficult, if not impossible, to distinguish most genotype frequency classes.

## ANALYSIS OF THREE DATA SETS

We demonstrate our method by analyzing one real and two simulated data sets. The real data consist of 74 juvenile trout from Whiskey Creek, Washington state, typed at 30 polymorphic protein loci, having between two and four alleles, as part of a large genetic survey conducted by the National Marine Fisheries Service and the Washington Department of Fish and Wildlife. The sample was collected under the belief that the fish were juvenile coastal cutthroat trout (*O. clarki clarki*); however, the Hardy-Weinberg and linkage disequilibrium in the sample, and the presence of homozygotes for alleles common in steelhead trout (*O. mykiss*) populations but rare in cutthroat trout populations, suggested that the sample might be a mixture of cutthroat, steelhead, and their hybrids. Hybrids between these two trout species have been documented in several rivers on the West Coast (Campton and Utter 1985; Neillands 1990).

The report of Johnson *et al.* (1999) gives more details about the sampling and the genotyping of the trout. It also summarizes the available literature from the field and the laboratory on hybridization between *O. clarki clarki* and *O. mykiss*. They report that there are no severe developmental abnormalities that occur in hybrids of the two species; hybrid offspring are clearly viable. However, hybrids may possess morphological and behavioral traits that reduce their fitness in natural environments. This accords well with the observation in other studies that hybrid individuals are detected typically among juvenile trout, but adult hybrids are seldom observed, and with the observation that although hybridization may occur each year (in cases where it has been monitored over time it has been found to be ongoing) the two species still remain distinct. Nonetheless, in some studies, the fish sampled and analyzed possess genotypes suggesting they belong to a hybrid class involving more than just one generation of hybridization. Here, we apply our method to the genetic data from Whiskey Creek with particular reference to the task of distinguishing between $F_1$ and later hybrids.

For this analysis, we assume there are six genotype frequency classes to which individuals might belong—the six classes arising from $n = 2$ generations of potential interbreeding. Table 1 lists the expected proportions of the different single-locus genotypes in these six classes and also gives the names that we use to refer to them. Though prior information is available on allele frequencies in other steelhead and cutthroat populations, we do this analysis using the prior $\lambda_{\ell,j} = 1/K_\ell$, $j = 1, \ldots, K_\ell$ for allele frequencies and the prior $\zeta_g = 1/6$, $g = 1, \ldots, 6$ for the mixing proportions, $\boldsymbol{\pi}$, of the different genotype frequency classes. A total of 100,000 sweeps of five chains started from overdispersed starting values were run. This required 4.6 hr on a laptop computer with a 266 Mhz G3 (Macintosh) processor.

**TABLE 1**

**Genotype frequency classes assumed for the analyses**

| $g$ | $Q$ | $G_{g,2}$ ($A$, $A$) | $G_{g,1}$ ($A$, $B$) or ($B$, $A$) | $G_{g,0}$ ($B$, $B$) | Name |
|---|---|---|---|---|---|
| 1 | 1.00 | 1.0000 | 0.0000 | 0.0000 | Pure Cutt |
| 2 | 0.00 | 0.0000 | 0.0000 | 1.0000 | Pure St |
| 3 | 0.50 | 0.0000 | 1.0000 | 0.0000 | $F_1$ |
| 4 | 0.50 | 0.2500 | 0.5000 | 0.2500 | $F_2$ |
| 5 | 0.75 | 0.5000 | 0.5000 | 0.0000 | Cutt Bx |
| 6 | 0.25 | 0.0000 | 0.5000 | 0.5000 | St Bx |

These are the six genotype frequency classes that arise from $n = 2$ generations of potential interbreeding. $G_{g,2}$, $G_{g,1}$, and $G_{g,0}$ are the expected frequencies of loci having 2, 1, or 0 genes originating from species $A$, as described in the text. The final column gives names that we use to refer to these genotype frequency classes.

After applying the method to the Whiskey Creek data set, we demonstrate its use on simulated data sets 1 and 2. Data set 1 is a simulated set of steelhead and cutthroat trout data. To simulate data set 1, we started with two species having allele frequencies at 30 unlinked loci that were the posterior mean estimates of allele frequencies in the cutthroat and steelhead populations at the 30 loci used in the analysis of the Whiskey Creek data. We simulated a sample of size 300 individuals with 155 pure cutthroat (Pure Cutt), 100 pure steelhead (Pure St), 25 $F_1$, 6 $F_2$, 14 cutthroat backcross (Cutt Bx), and 0 steelhead backcross (St Bx) individuals. This sort of sample might be encountered from a population in which steelhead and cutthroat hybridize infrequently, and the $F_1$'s tend to mate assortatively, being more likely to mate with other $F_1$'s or with pure cutthroat than with steelhead. To simulate the $i$th individual belonging to the $g$th genotype frequency class, the species origin of each gene copy at a locus was randomly assigned according to $\boldsymbol{G}_g$, and then the allelic type of each gene was randomly selected from its species of origin according to the posterior mean allele frequencies estimated from the Whiskey Creek analysis.

Simulated data set 2 is a sample of 300 individuals with the same number of individuals belonging to the different genotype frequency classes as in data set 1. However, data set 2 consists of 20 nearly diagnostic loci for distinguishing the two simulated species, $A$ and $B$. That is, there are assumed to be 20 diallelic, codominant loci at which the frequency of the first allele is 0.995 in species $A$ and the frequency of the alternate allele is 0.995 in species $B$. This represents considerably more power to distinguish species than is present in the trout data set.

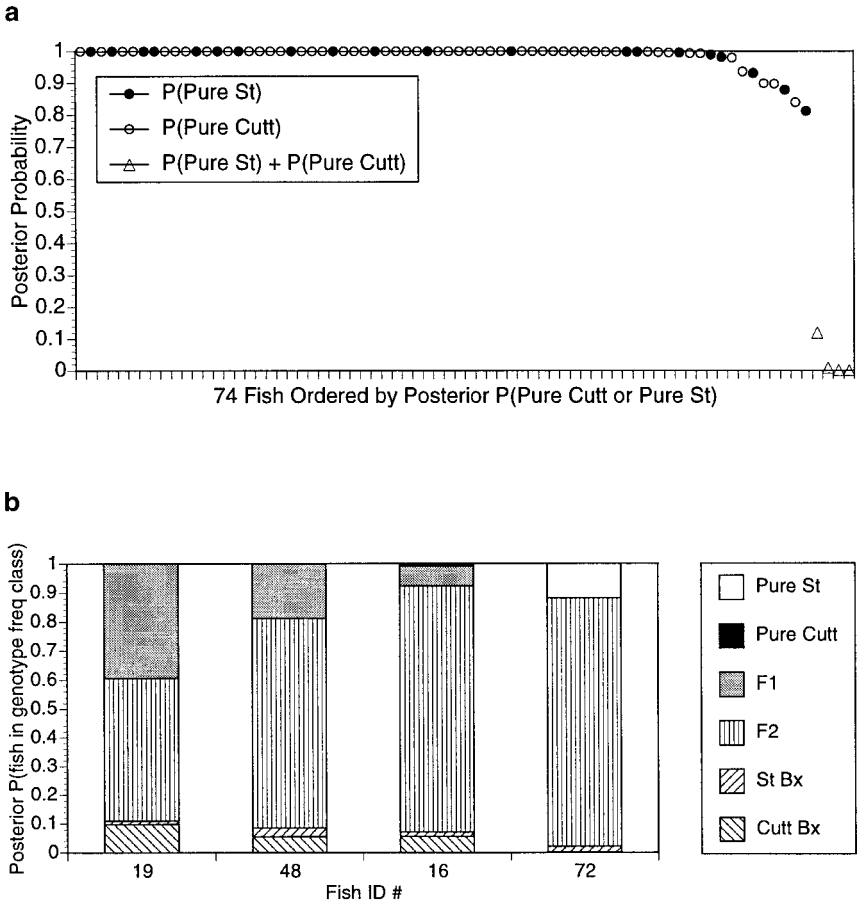


FIGURE 2.—Graphical summary of results for the Whiskey Creek data set. (a) The 20 probable Pure St individuals appear as solid circles while the 50 probable Pure Cutt fish appear as open circles in the graph with the height of the circles determined by the posterior probability of being Pure St or Pure Cutt, respectively. The four remaining fish are plotted on the graph as open triangles at heights given by the posterior probability that they are either Pure St or Pure Cutt. The fish are ordered by their posterior probability of being purebred. (b) Posterior probabilities of genotype frequency class for fish 19, 48, 16, and 72—the four plotted as triangles in a. The height of the different patterns in the column denotes the posterior probability of each fish belonging to each of the six different genotype frequency classes.

Both of the simulated data sets were analyzed assuming $n = 2$ and using the same priors on the allele frequencies and the mixing proportions that were employed in the analysis of the Whiskey Creek data. Five chains, started from overdispersed starting points, were simulated for 45,000 sweeps each. This required 8 hr for data set 1 and 5.3 hr for data set 2 on the same 266 Mhz G3 processor.

## RESULTS

**Whiskey Creek data:** Using the multilocus genotype data on the 74 fish in the data set, our method successfully estimated allele frequencies $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$ for the two putative species contributing to the sample. Inspecting these allele frequencies, it was clear that one set of frequencies corresponded to the steelhead group and the other to the cutthroat group, so we were able to label them as such. In this analysis, each fish is assigned a posterior probability of belonging to one of the six different genotype classes. Two of those classes are purebred categories (Pure Cutt and Pure St). Twenty of the fish in the sample have posterior probability >0.8 of being in the Pure St category, and for 15 of those fish, that posterior probability is >0.99. Fifty fish have posterior probability >0.8 of being Pure Cutt, with 45 of those having posterior probability >0.99 of being Pure Cutt. All of the fish with posterior probability >0.8 of being Pure St have negligible posterior probability of being Pure Cutt and vice versa. Of the remaining 4 fish, 3 of them, nos. 19, 48, and 16, have posterior probability <0.01 of being either Pure Cutt or Pure St. This means that, given the data and the assumptions of the model and the priors used, those fish have probability >0.99 of being hybrids of some sort. The fourth fish (no. 72) has posterior probability near 0.12 of being either Pure Cutt or Pure St; hence, posterior probability near 0.88 of being a hybrid of some sort. Figure 2 graphically represents this.

We may look more closely at the four probable hybrid fish. Figure 2b shows the posterior probabilities that fish 19, 48, 16, and 72 belong to each of the six different genotype frequency classes. Note that they all have highest posterior probability of belonging to the $F_2$ class. This is unexpected because one would suspect there would be, in general, more $F_1$'s in a population than $F_2$'s since the $F_2$'s would have to be formed by mating between $F_1$'s. In fact, the posterior probabilities for all these fish are such that no classification or assignment of any of them could be made with great certainty to a single one of the genotype frequency classes. As already noted, with posterior probability 0.12, fish 72 may belong to the Pure St category. Further, it cannot be determined with great certainty that fish 19 and 48 are not

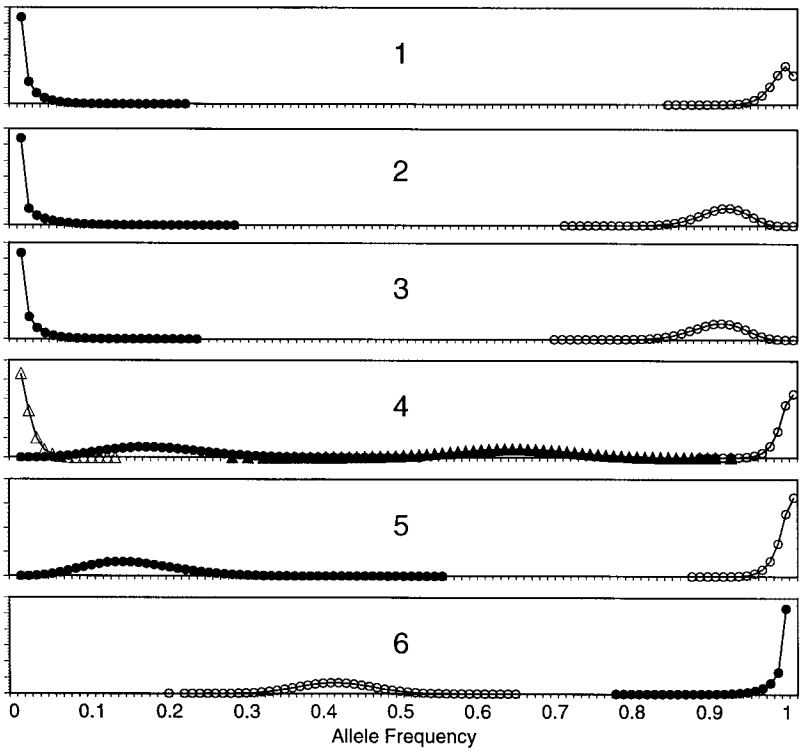


FIGURE 3.—Histograms showing the posterior distribution of allele frequencies estimated for the steelhead and cutthroat populations in Whiskey Creek. The solid symbols are for alleles in the steelhead population, and the open symbols denote the cutthroat allele frequencies. The vertical axis in each graph is proportional to posterior probability density. The numbers 1–6 refer to different loci ranked in order or how informative they are for distinguishing between the species. Locus 4 has three alleles, and the others have two alleles. Note that locus 1 is the only one that is close to being diagnostic. For a contrast, consider locus 6: The steelhead population has high posterior probability of being fixed for a single allele, but that allele is at a frequency around 0.4 in the cutthroat population.

$F_1$ hybrids. And even fish 16 has a posterior probability of almost 0.07 that it is an $F_1$ hybrid. In other words, no conclusions may be made with great confidence about the presence of fish in the sample that are hybrids belonging to categories beyond $F_1$. This is due to the lack of clear separation between the two species in this data set. Only about eight of the loci are very informative, and none of them are strictly diagnostic when informative priors for the allele frequencies are not used.

A statistically reasonable way to measure the degree to which a locus is useful in separating the species is by the Kullback-Leibler divergence (KULLBACK and LEIBLER 1951) between the two species at that locus. Briefly, at locus $\ell$ the Kullback-Leibler divergence between species $A$ and $B$ is $\mathcal{T}_\ell(A, B) = I_\ell(A{:}B) + I_\ell(B{:}A)$, where $I_\ell(A{:}B)$ is the expected Kullback-Leibler information for distinguishing the species from which an allele at locus $\ell$ originated given that it came from species $A$. Mathematically, for a given value of the allele frequencies,

$$I_\ell(A{:}B) = \sum_{k=1}^{K_\ell} \theta_{A,\ell,k} \log \frac{\theta_{A,\ell,k}}{\theta_{B,\ell,k}}. \qquad (11)$$

By averaging the values of $\mathcal{T}_\ell(A, B)$ obtained for the different allele frequencies visited by the Markov chain in doing MCMC, it is possible to compute the posterior mean Kullback-Leibler divergence between the species at each locus. We have done this to identify the six most informative loci in the data set, and we have plotted the posterior distribution of the allele frequencies at those loci in Figure 3. From this figure, it is apparent that only the locus labeled "1" comes close to being a diagnostic locus. For all the others, there is almost zero posterior probability that an allele appearing in the steelhead species does not appear in the cutthroat species or vice versa. Having such a small number of loci that are distinctive between the species makes it difficult to make fine-scale inferences about the specific genotype frequency classes to which an individual belongs. This point is made clear again by the analysis of the first simulated data set.

**Simulated data set 1:** This data set was simulated assuming that allele frequencies in the two species were equal to the posterior mean estimates from the Whiskey Creek data set. In this case, since we know the true genotype frequency classes to which each individual belongs, we can better assess how well the method works on this data set. Figure 4 shows how well purebred individuals in the sample could be distinguished from the hybrids. For most of the Pure Cutts and the Pure St the posterior probability or being in the correct genotype frequency class is $>0.99$. There are 32 Pure Cutts with posterior probability of being Pure Cutt $<0.98$. Almost all of the remaining probability for these individuals goes to the Cutt Bx category. There are, in contrast, only seven Pure St individuals with posterior probability of Pure St $<0.98$. This is most likely due to the fact that there are no St Bx individuals in the sample, so the proportion of St Bx individuals in the population is estimated to be small, and hence even the fish that are not well distinguished solely by their genotype between the Pure St and the St Bx classes will tend to have a higher posterior probability of being in the Pure St class.

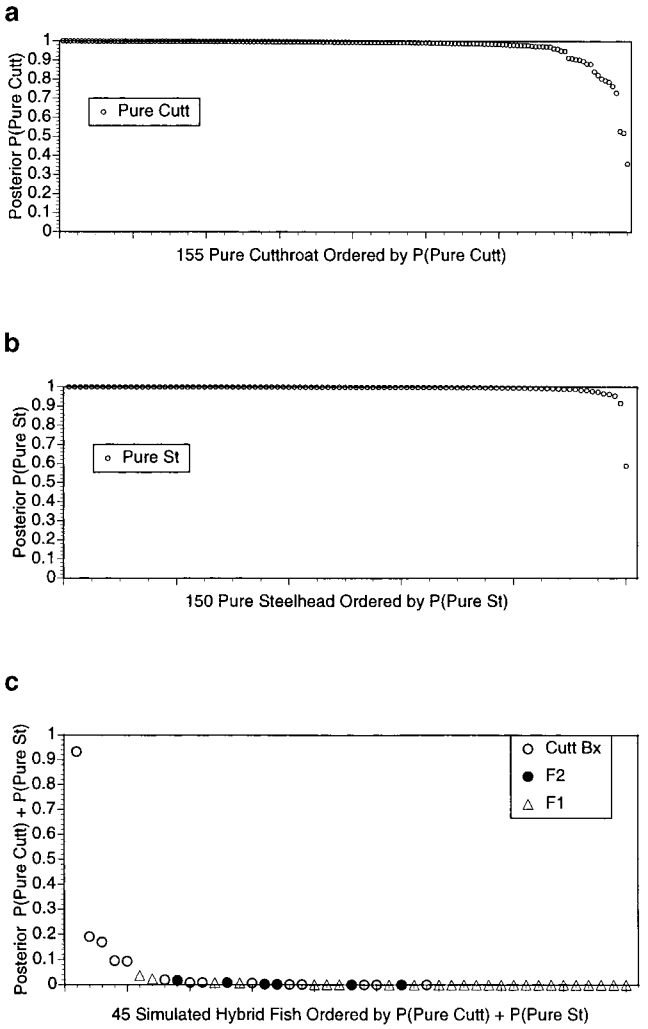


FIGURE 4.—Results for simulated data set 1. (a) Posterior probability of Pure Cutt for 155 simulated cutthroat trout. (b) Posterior probability of *P* (Pure St) for 100 simulated steelhead trout. (c) Posterior probability of either Pure Cutt or Pure St for 45 simulated hybrid trout of genotype frequency classes denoted by the different symbols as given in the inset.

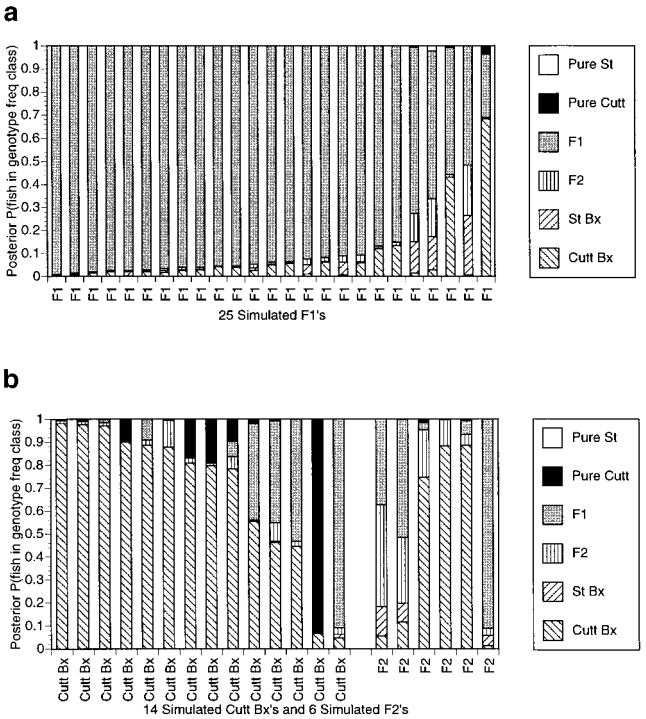


FIGURE 5.—Posterior probabilities of specific genotype frequency categories for the hybrid individuals in simulated data set 1. (a) Twenty-five simulated $F_1$'s. (b) Six simulated $F_2$'s and 14 simulated Cutt Bx's, as indicated.

This is a clear example of how the method gives weight to the proportion of individuals from different genotype frequency classes found in the sample. These results also suggest that, with data similar to those from Whiskey Creek and with many purebred individuals of each species sampled, it is unlikely that any purebred individual will receive high posterior probability of being in a non-purebred genotype frequency category.

All but 1 of the 45 simulated hybrid fish have posterior probability $<0.20$ of being either Pure St or Pure Cutt, and 37 of them have posterior probability $<0.02$ of being purebred. As is apparent in Figure 4c, it is most difficult to distinguish the Cutt Bx's from the purebred categories. As expected, the $F_1$'s are the easiest to distinguish from the purebred individuals.

While the method works well in distinguishing between purebred and hybrid categories, with the allele frequencies used in the simulations it is much more difficult to make clear distinctions between the different hybrid genotype frequency classes. Figure 5a shows the posterior probabilities of inclusion in the six genotype frequency classes for the 25 $F_1$ hybrids in descending order of the posterior probability that they are $F_1$'s. While many of them have high posterior probability of being $F_1$, there is also one with posterior probability $>0.5$ of being in the Cutt Bx category. For the non-$F_1$ hybrids, the situation is even less promising. Of the six $F_2$ individuals (the last six individuals in Figure 5b), not one of them has posterior probability $>0.5$ of being in the $F_2$ category. Finally, for the Cutt Bx genotype frequency class, as the first 14 columns in Figure 5b reveal, there is a great deal of variability among the individuals in the posterior probability that they are Cutt Bx. This results from having few loci in this simulated data set that are strongly distinctive between the species, and it argues that with genetic data of this sort, assignment of individuals, if desired, to specific genotype frequency classes should be made with caution and only in the presence of very strong posterior support (for example, $\approx 0.98$) of that assignment.

**Simulated data set 2:** The analysis with 20 nearly diagnostic loci demonstrates how well the method performs when the populations are genetically well separated by the markers. The 155 species *A* individuals, the 100 species *B* individuals, and the 25 $F_1$'s all had posterior

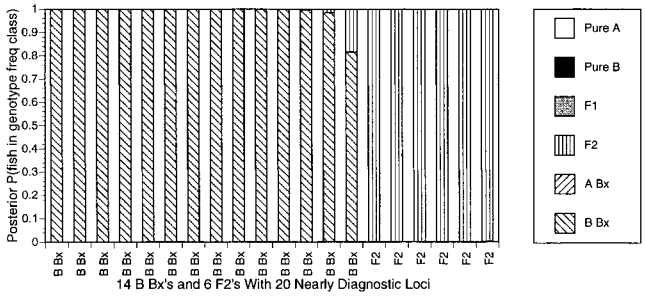


FIGURE 6.—Posterior probability of genotype frequency classes for the 14 *B* Bx individuals and 6 $F_2$ individuals in simulated data set 2.

probabilities >0.9996 of belonging to the correct genotype frequency class. Likewise, the six $F_2$'s each had posterior probability >0.997 of belonging to the $F_2$ genotype frequency class, and all but two of the simulated *B* Bx individuals had posterior probability >0.997 of belonging to the *B* Bx class. As Figure 6 shows, of the remaining two *B* Bx individuals, one has posterior probability 0.983 of being *B* Bx; this individual possessed, by random chance, many more loci heterozygous for the two alleles than expected of a *B* Bx. The remaining *B* Bx individual happened to be one of the rare carriers of a locus homozygous for the species *A* common allele. This can occur because the loci were designed in the simulation to not be perfectly diagnostic. Our method is able to handle such a situation because it does not require that loci are fixed for alternate alleles. The result is just that the posterior probability of belonging to the correct genotype frequency class is reduced for that individual.

## DISCUSSION

We have developed a model-based statistical method for identifying species hybrids using multilocus genetic data. Using Markov chain Monte Carlo in a Bayesian setting, we compute the posterior probability that an individual belongs to each of a set of possible genotype frequency classes. This allows us to utilize all the data simultaneously, while integrating over the uncertainty in individual assignments to genotype frequency classes and in the model parameters—the proportion of individuals from the different genotype frequency classes and the allele frequencies of each of the species—which are seldom known without error. This method also has the advantage that it can perform a mixture deconvolution without *a priori* knowledge of the allele frequencies in the separate species. In other words, one need not be able to separate the two species on the basis of morphology, nor must one be able to sample from the pure species separately, to separate the two species and their hybrids present in a mixture. Finally, though strong genetic differentiation between the species is helpful, this method does not require that each species possess unique alleles.

We have demonstrated the method by analyzing data from cutthroat and steelhead trout. Though the data set includes 30 loci polymorphic in the sample, many of the loci are not very informative. Given only the 74 individuals in the sample, none of the loci appear to be purely diagnostic. While information from samples drawn from other steelhead and cutthroat populations could be used as prior information, we used a vague prior on allele frequencies, reflecting little prior knowledge of allele frequencies, so as to demonstrate the method's performance with data having no apparently diagnostic loci. Despite such restricted data, the method was able to identify four individuals having high posterior probability of being species hybrids.

The analysis of data simulated to mimic the trout data provides insight into the method's ability to distinguish between the different genotype frequency classes that the hybrids belong to. Despite the low level of genetic differentiation between the two trout species, the posterior probability of being purebred was very low for most of the simulated hybrids. However, it was clear that it is much harder to distinguish the specific (*i.e.*, $F_2$ *vs.* backcrossed category) genotype frequency classes of hybrid individuals without many loci showing extreme allele frequency differences between the species. Our analysis of a simulated data set with 20 nearly diagnostic loci demonstrates the method's ability to identify the specific genotype frequency of individuals with great certainty using highly informative genetic data.

A number of assumptions are made to derive the likelihood in this problem. One of the advantages of the model-based approach over a more general multivariate approach (like principal component analysis) is that the assumptions here are explicit. One important assumption is that the loci used in the analysis are unlinked. In studies with a limited number of loci on organisms with many chromosomes, this assumption is not likely violated. However, as the number of loci increases, so does the probability that some of them will be linked. Under the assumption of no linkage, each locus is treated as an independent unit of information; however, the information carried by linked loci will not be independent. Therefore, analyzing data on linked loci with this method will cause one to overestimate one's certainty in identifying species hybrids. If the recombination fractions between the markers were known, it would be possible to account for linked loci. However, modeling the dependence between loci in a manner faithful to the underlying process could incur a heavy computational cost, and the use of an approximation, for example, a Markov approximation to the genotype frequency class process along the chromosome as taken by MCKEIGUE (1998) in a related problem, would be computationally preferable.

In addition to the assumption that the loci are un-

linked, this analysis assumes that the markers used are not tightly linked to any loci that are under selection. Especially with large numbers of markers, this assumption will be violated to an extent. For example, Rieseberg and Linder (1999), reporting on hybrids between two sunflower species of known pedigree, find that selection leads to significant departures from the expected proportions of some marker alleles in the hybrids. We note that with some modification the statistical framework presented here may be useful for identifying loci influenced by selection in naturally hybridizing populations.

We also assume that there is no linkage or Hardy-Weinberg disequilibrium in the parental species $n$ generations before the sampling event. This assumption allows multilocus genotype probabilities to be expressed in terms of a few allele frequency parameters and the mixing proportions $\boldsymbol{\pi}$, and it allows the disequilibrium in the sample to be used to identify two separate species' gene pools: Any disequilibrium in the mixed population is assumed to arise from the mixture of the two species and their hybrids. This assumption could be relaxed only if samples of the pure species' populations were available and the preexisting disequilibrium observed therein could be accounted for by using a parameterization in terms of genotype frequencies for the loci in disequilibrium, rather than simply a parameterization in terms of allele frequencies. The fact that the linkage and Hardy-Weinberg disequilibrium in the sample are a source of information that can be used to estimate the separate species' allele frequencies underscores the necessity of having recently or incompletely hybridized populations. Unless some prior information about species' allele frequencies were available, or if a sample of known, purebred individuals were available, it would not be possible to identify hybrids among species that have been entirely panmictic with one another for enough generations that the linkage disequilibrium at all markers had decayed to low levels.

Finally, the analysis we describe is made conditional on the assumed value $n$, the number of generations over which interbreeding has been potentially occurring. In practice, $n$ will typically have to be chosen small, because as $n$ increases, $\mathcal{G}_n$, the number of genotype frequency classes, increases exponentially, and the $\boldsymbol{G}_g$'s of each new class are typically close to those for another class, so distinguishing between them would require an enormous amount of data. It is recommended that the number of genotype frequency classes be kept small in any analysis. One way to do this, while allowing for many generations of potential interbreeding, is to not consider all $\mathcal{G}_n$ possible products resulting from $n$ generations of mating between the two species. For example, if hybridization were rare, it would be sensible to consider only the $F_1$ class and then the simplest backcross categories, omitting the categories involving more than one $F_1$ individual in the pedigree.

We adopted a simple sampling model: Individuals are assumed drawn at random from a population that is a mixture in the proportions $\boldsymbol{\pi}$ of individuals from the different genotype frequency classes. This was probably violated in the case of the cutthroat trout data, because, when the biologists collected the specimens, they were trying to obtain pure cutthroat and hence were throwing back those individuals that looked like steelhead or hybrids. However, it is sometimes difficult to distinguish cutthroat juveniles from steelhead or hybrid juveniles on the basis of morphological characters. To estimate accurately the proportion of hybrids in a locale, or even to estimate accurately the posterior probability that an individual is a hybrid, it would be wise to design the study with those goals in mind. Having an explicit model, like the one described in this article, that includes the sampling of the organisms is an asset, since the model may be tailored to particular sampling schemes. For example, it would be possible to model stratified sampling in which sampled organisms were first put into "possibly hybrid" and "probably purebred" categories on the basis of their morphological traits, and then a random subset of individuals from each of those categories was genetically typed. Or the sampling model could be modified similarly for sampling at several locations along a transect intersecting a hybrid zone.

Throughout, we have been interested primarily in making inference about the genotype frequency class to which individuals in the sample belong. It should be noted, however, that the output for the MCMC sampler could be used to estimate many other quantities of interest. Barton (2000) recently presented a method for estimating multilocus genotype frequencies and multilocus linkage disequilibrium in hybrid populations. He notes that one could try to achieve the same end by describing a population as a mixture of parentals, $F_1$'s, backcrosses, and $F_2$'s, but dismisses such an approach because the mixture is not uniquely determined by the genotype frequencies. Though the mixture is not uniquely determined by the genotype frequencies, the data can be used to determine a posterior distribution for the mixing proportions and the allele frequencies in the two species. This is precisely what we have done here. And further, the mixing proportions $\boldsymbol{\pi}$ and the allele frequencies $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$ determine the multilocus genotype frequencies expected under the model. Since our MCMC method provides values of $\boldsymbol{\pi}$ and the allele frequencies $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$, sampled in proportion to their joint posterior probability, it would be straightforward to also estimate the posterior distribution of the frequency of any multilocus genotype or any linkage disequilibrium measure of interest from the MCMC sample.

We also note that the general framework here could be modified to handle null alleles or dominant markers, like amplified fragment length polymorphisms. For example, if locus $\ell$ included a null allele or was a dominant

marker, then it could be handled by modifying the probability model connecting the latent variables $\boldsymbol{W}_{i,\ell}$ to the observable genetic characteristics of individual $i$ at locus $\ell$. The other portions of the model would remain unchanged. This would allow appropriate weighting of the information from different types of marker systems, used simultaneously, to identify hybrids.

In conclusion, the method we have presented provides a way to use genetic data to identify species hybrids. The method is applicable not only to loci with fixed differences between species, but also to loci without fixed differences. Though prior knowledge may be incorporated into the model, the method is able to cluster individuals in a mixed population without any *a priori* genetic knowledge of the species. The model-based approach is extendable to special sampling scenarios and different types of genetic markers. Software implementing the algorithm described in this article may be obtained for free from the corresponding author.

We thank Matthew Stephens, Joe Felsenstein, and Robin Waples for helpful discussions on this problem; Stuart Baird for extensive comments on the manuscript; and two anonymous referees for numerous insightful and helpful comments. The Whiskey Creek data set was collected by Eric Iwamoto and kindly provided by David Teel, both of the National Marine Fisheries Service. Both authors were supported by National Science Foundation grant BIR-9807747 to E.A.T.

## LITERATURE CITED

Avise, J. C., 1994 *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Barton, N. H., 2000 Estimating multilocus linkage disequilibria. Heredity **84:** 373–389.

Barton, N. H., and K. S. Gale, 1993 Genetic analysis of hybrid zones, pp. 13–45 in *Hybrid Zones and the Evolutionary Process*, edited by R. G. Harrison. Oxford University Press, Oxford.

Boecklen, W. J., and D. J. Howard, 1997 Genetic analysis of hybrid zones: numbers of markers and power of resolution. Ecology **78:** 2611–2616.

Boecklen, W. J., and R. Spellenberg, 1990 Structure of herbivore communities in two oak (*Quercus* spp.) hybrid zones. Oecologia **85:** 92–100.

Campton, D. E., and F. M. Utter, 1985 Natural hybridization between steelhead trout (*Salmo gairdneri*) and coastal cutthroat trout (*Salmo clarki clarki*) in two Puget Sound streams. Can. J. Fish. Aquat. Sci. **42:** 110–119.

Diebolt, J., and C. P. Robert, 1994 Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. Ser. B **56:** 363–375.

Elo, K., J. Erkinaro, J. A. Vuorinen and E. M. Niemelae, 1995 Hybridization between Atlantic salmon (*Salmo salar*) and brown trout (*S. trutta*) in the Teno and Naeaetaemoe River systems, northernmost Europe. Nord. J. Freshwater Res. **70:** 56–61.

Epifanio, J. M., and D. P. Philipp, 1997 Sources for misclassifying genealogical origins in mixed hybrid populations. J. Hered. **88:** 62–65.

Gelman, A., 1996 Inference and monitoring convergence, pp. 131–143 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, New York.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1996 *Bayesian Data Analysis*. Chapman & Hall, New York.

Geman, S., and D. Geman, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6:** 721–741.

Goodman, S. J., N. H. Barton, G. Swanson, K. Abernethy and J. M. Pemberton, 1999 Introgression through rare hybridization: a genetic study of a hybrid zone between red and sika deer (genus Cervus) in Argyll, Scotland. Genetics **152:** 355–371.

Harrison, R. G., 1990 Hybrid zones: windows on the evolutionary process. Oxf. Surv. Evol. Biol. **7:** 69–128.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hewitt, G. M., 1988 Hybrid zones—natural laboratories for evolutionary studies. Trends Ecol. Evol. **3:** 158–167.

Jansson, H., and T. Oest, 1997 Hybridization between Atlantic salmon (*Salmo salar*) and brown trout (*S. trutta*) in a restored section of the River Dalaelven, Sweden. Can. J. Fish. Aquat. Sci. **54:** 2033–2039.

Johnson, O. W., M. H. Ruckelshaus, W. S. Grant, F. W. Waknitz, A. M. Garrett *et al.*, 1999 Status review of coastal cutthroat trout from Washington, Oregon, and California. NOAA Technical Memorandum NMFS-NWFSC-37, National Marine Fisheries Service.

Kass, R. E., and L. Wasserman, 1996 The selection of prior distributions by formal rules. J. Am. Stat. Assoc. **91:** 1343–1370.

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. Ann. Math. Stat. **22:** 79–86.

Liu, J. S., 1994 The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. **89:** 958–966.

McKeigue, P., 1998 Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations by conditioning on parental admixture. Am. J. Hum. Genet. **63:** 241–251.

Miller, L. M., 2000 Classifying genealogical origins in hybrid populations using dominant markers. J. Hered. **91:** 46–49.

Nason, J. D., and N. C. Ellstrand, 1993 Estimating the frequencies of genetically distinct classes of individuals in hybridized populations. J. Hered. **84:** 1–12.

Neillands, W. G., 1990 Natural hybridization between coastal cutthroat trout (*Oncorhynchus clarki*) and steelhead trout (*Oncorhynchus mykiss*) within Redwood Creek, California. Master's thesis, Humboldt State University, Arcata, CA.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA **94:** 9197–9201.

Rieseberg, L. H., and C. R. Linder, 1999 Hybrid classification: insights from genetic map-based studies of experimental hybrids. Ecology **80:** 361–370.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA **24:** 253–259.

Wright, S., 1952 The theoretical variance within and among subdivisions of a population that is in a steady state. Genetics **37:** 313–321.

Young, W. P., C. O. Ostberg, P. Keim and G. H. Thorgaard, 2001 Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Oncorhynchus mykiss*) and coastal cutthroat trout (*O. clarki clarki*). Mol. Ecol. **10:** 921–930.

Communicating editor: M. K. Uyenoyama