Conservation Genet Resour
DOI 10.1007/s12686-009-9146-0

TECHNICAL NOTE

How to use SNPs and other diagnostic diallelic genetic markers to identify the species composition of multi-species hybrids

Steven T. Kalinowski

Received: 15 July 2009 / Accepted: 9 November 2009

Abstract Hybridization with non-native species is a threat to many taxa, but hybrids can be difficult to identify based on morphology. Genetic data is useful for estimating the ancestry of admixed populations, and diallelic markers such as single nucleotide polymorphisms are popular for such applications. When taxa are evolutionarily well diverged, loci frequently become fixed for different alleles in each taxa, and the degree of genetic admixture between two taxa can be estimated by counting diagnostic alleles for each taxa. However, when there is hybridization between more than two taxa, and loci have only two alleles, the origin of each allele cannot be assigned ambiguously to a taxon. In this note, I show how the expectation–maximization algorithm can be used to solve this problem. A computer program for implementing this approach is available at www.montana.edu/kalinowski.

Keywords Hybridization · Estimation · SNP · Diagnostic · EM algorithm

Invasive species are one of the greatest threats to global biodiversity (Vitousek et al. 1997). Of the many negative effects that non-native species can have on native taxa, hybridization and genetic introgression is one of the most pernicious (Rhymer and Simberloff 1996). Genetic introgression and outbreeding depression have contributed to the extinction of many plant and animal species (Allendorf et al. 2001), and even small amounts of genetic admixture can substantially lower fitness in the wild (e.g., Muhlfeld et al. 2009).

One of the challenges to managing species that interbreed in the wild is accurate identification of hybrids and admixed populations (Allendorf et al. 2001). When species are morphologically similar, this can be difficult. For example, cutthroat trout (*Oncorhynchus clarki*) and rainbow trout (*Oncorhynchus mykiss*) readily interbreed in the wild (Benke 2002), and this introgression presents a serious threat to the persistence of many subspecies of cutthroat trout (e.g., Shepard et al. 2003; Muhlfeld et al. 2009). However, identifying rainbow/cutthroat hybrids using morphology is difficult—especially when only a small proportion of the ancestry of a hybrid cutthroat trout is from rainbow trout (Leary et al. 1996).

Molecular markers offer a useful tool for accurately estimating the ancestry of hybrid individuals and populations. When F1-hybrids are fertile, and backcrosses are common, multiple loci must be used to estimate the ancestry of fish and populations. There are several types of molecular markers that can be used to this, and a variety of statistical methods available for analyzing the data (e.g., Anderson and Thompson 2002; Pritchard et al. 2000), but when taxa are evolutionarily well-differentiated, the simplest way to estimate the ancestry of potentially hybridized individuals is to use taxon-specific diagnostic alleles, and count the proportion of alleles in an individual or population that are non-native. Single nucleotide polymorphisms (SNPs) (Finger et al. 2009; Stephens et al. 2009) and insertion/deletions (Ostberg and Rodriguez 2004) are popular for such applications, because diagnostic loci can often be identified in which all individuals in the native taxon have one allele and all individuals in the non-native taxon have an alternative allele. Finding such diagnostic loci is often not difficult, and the resulting data is

S. T. Kalinowski (✉)
Department of Ecology, 310 Lewis Hall Montana State
University, Bozeman, MT 59717, USA
e-mail: skalinowski@montana.edu

Published online: 27 November 2009

unambiguous when *two* taxa are compared. However, when hybridization may have occurred between *three* or more taxa, diallelic loci can be difficult to interpret. An example illustrates the problem.

Westslope cutthroat trout (*Oncorhynchus clarki lewisi)* are native to the Rocky Mountains of the northern United States. Yellowstone cutthroat trout (*Oncorhynchus clarki bouvieri*) and rainbow trout have been extensively introduced throughout the range of westslope cutthroat trout, so that some populations in the historic range of Westslope cutthroat trout may contain ancestry from all three taxa. A population of westslope cutthroat trout in Yellowstone National Park that has been genotyped for nine SNP loci (S. Kalinowski unpublished) contains such a mixture (Table 1). The ten individuals in the sample clearly show low levels of genetic introgression from Yellowstone cutthroat and rainbow trout. For example, Trout #1 has a Yellowstone cutthroat trout allele at *Locus9*, and Trout #2 has rainbow trout alleles at *Locus2* and *Locus3*. The possibility of admixture among all three species leads to ambiguity in estimating the degree of hybridization among individuals. Trout #9 exemplifies the problem. This fish has Yellowstone cutthroat ancestry at *Locus8* and *Locus9* and rainbow trout ancestry at *Locus2*. Given this complex ancestry, the genotype of Trout #9 at *Locus1* (*CC*) is ambiguous. Both westslope and Yellowstone cutthroat trout have a genotype of *CC* at *Locus1*, so the ancestry of this fish cannot be estimated by simple gene counting. This problem extends to the sample as a whole. Given the ambiguity present in the diallelic loci, the frequency of westslope, Yellowstone, and rainbow alleles cannot be estimated by simply counting the number of alleles from each taxon.

Fortunately, there is a straightforward statistical solution to this problem. The expectation–maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1996) can be used to calculate a maximum likelihood estimate of the genetic composition of the sample—in a manner analogous to estimating the frequency of *A*, *B*, and *O* blood antigens (Ceppellini et al. 1955; see Weir 1996, Chap. 2, for a review) and the frequency of null alleles at microsatellite loci (Kalinowski and Taper 2006). The EM algorithm produces maximum-likelihood estimates of the frequency of alleles from each species, under the assumption that the frequency is the same for all loci. The analysis is identical for estimating the ancestry of a single individual or for a sample of individuals from a population. I will present the method in the context of estimating the contribution of multiple taxa to a sample from a population.

The following notation is useful. Let us assume that N_{Taxa} may have hybridized and contributed genes to a population. Let P_i represent the frequency of the *i*th taxon's genes to the sample ($\sum_i^{N_{Taxa}} P_i = 1$) and let $\hat{P}_i$ represent an estimate of P_i. Let n_{jk} represent the number of times that allele *k* is observed at locus *j* in the sample. Let the indicator variable X_{ijk} equal 1 if all non-hybridized individuals in taxon *i* have allele *k* at locus *j*, and equal 0 if all individuals in taxon *i* have an alternative allele. In other words, X_{ijk} will equal 1 if taxon *i* is fixed for allele *k* at locus *j*. Let N_{Loci} denote the number of co-dominant diploid loci that have been genotyped. Let N_{Sample} represent the number of genes sampled (if there is no missing data, $N_{Sample} = 2 \times N_{Loci} \times N_{Inds}$ where N_{Inds} is the number of individuals sampled). Lastly, let $N_{Alleles(j)}$ represent the number of alleles at locus *j*. For applications with SNPs and

Table 1 Sample genotypes for nine diagnostic SNP loci in 10 trout of unknown ancestry

	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7	Locus 8	Locus 9
WCT allele	C	G	A	A	T	T	G	A	G
YCT allele	C	G	A	C	C	C	A	G	T
RBT allele	T	T	T	C	C	C	G	A	G
Trout #1	CC	GG	AA	AA	TT	**C**T	GG	AA	G**T**
Trout #2	CC	G**T**	A**T**	AA	**C**T	TT	GG	AA	GG
Trout #3	CC	GG	A**T**	AA	TT	TT	GG	AA	GG
Trout #4	CC	GG	AA	AA	TT	TT	GG	AA	GG
Trout #5	C**T**	GG	A**T**	AA	TT	TT	GG	AA	GG
Trout #6	CC	GG	AA	**C**A	**C**T	TT	GG	AA	GG
Trout #7	C**T**	GG	AA	AA	**C**T	**C**T	GG	AA	GG
Trout #8	CC	GG	A**T**	AA	TT	TT	G**A**	AA	GG
Trout #9	CC	G**T**	AA	**C**A	**C**T	**C**T	GG	**G**A	G**T**
Trout #10	CC	GG	AA	AA	TT	**C**T	GG	**G**A	GG

The population is within the range of Westslope cutthroat trout. Alleles that are known to be non-native are identified underlined and shown in bold. Loci 1–3 have alleles that are unique in rainbow trout (RBT). Loci 4–6 have alleles that are unique in westslope cutthroat trout (WCT). Loci 7–9 have alleles that are unique to Yellowstone cutthroat trout (YCT)

Table 2 Estimates of species composition for the 10 trout whose genotypes are shown in Table 1

	Proportion		
	WCT	YCT	RBT
Trout #1	0.83	0.17	
Trout #2	0.75		0.25
Trout #3	0.92		0.08
Trout #4	1		
Trout #5	0.83		0.17
Trout #6	0.82	0.09	0.09
Trout #7	0.75		0.25
Trout #8	0.84	0.08	0.08
Trout #9	0.5	0.33	0.17
Trout #10	0.83	0.17	

indels, $N_{Alleles(j)}$ will usually be equal to 2, but there is no restriction on the total number of alleles (provided each taxon is fixed for an allele). If we assume that the amount of hybridization is the same for all loci, the likelihood, L, of the allele counts for the sample is

$$L = \prod_{j=1}^{N_{Loci}} \prod_{k=1}^{N_{Alleles(j)}} \left[\left(\sum_{i}^{N_{Taxa}} X_{ijk} P_i \right)^{n_{jk}} \right]$$

The equation does not assume that population is in gametic equilibrium.

The EM algorithm is a simple iterative method for finding maximum likelihood estimates of P_i. The crux of the EM algorithm is that if an estimate of the allele frequencies in a taxon, $\hat{P}_i$, is available, a better estimate, $\hat{P}'_i$, can be obtained

$$\hat{P}'_i = \frac{1}{N_{Sample}} \sum_{j=1}^{N_{Loci}} \sum_{k=1}^{N_{Alleles(j)}} n_{jk} \left(\frac{X_{ijk}\hat{P}_i}{\sum_i^{N_{Taxa}} X_{ijk}\hat{P}_i} \right)$$

An example helps illuminate the logic behind this equation. Consider *Locus1* in the example provided (Table 1). At this locus, all westslope cutthroat trout have a *C*, all Yellowstone cutthroat trout have a *C*, and all rainbow trout a *T*. Let us assume that 12 *C*'s have been observed in a sample. These alleles are ambiguous because they could come either from a westslope or Yellowstone cutthroat trout. Although we do not know how many of the *C*'s came from westslope cutthroat trout, we can say that *some* fraction came from westslope cutthroat trout. This fraction can be estimated using preliminary estimates of the composition of the population. For example, if we assume that $\hat{P}_{Westslope} = 0.50$, $\hat{P}_{Yellowslope} = 0.25$, and $\hat{P}_{Rainbow} = 0.25$, then we would expect that 0.50/(0.50 + 0.25) = two-thirds of the *C*'s in the sample came from westslope cutthroat trout. This is what the term inside of the parentheses in the equation above tells us. Essentially, this term allocates ambiguous alleles to taxa based on the estimated frequency of the taxa. In the case of the example just described, two-thirds of the 12 *C*'s would be allocated to westslope cutthroat trout. Summation is continued in a similar manner for all loci and alleles to obtain values of $\hat{P}'_i$.

Once $\hat{P}'_i$ is obtained, it can used as an estimate of P_i to obtain an even better estimate (by using the above equation again). Iteration is continued until estimates converge on maximum likelihood estimates of P_i. In practice, it is convenient to stop iteration when the change in likelihood between iterations is less than 10^{-9}. The following estimates are obtained for the sample data in Table 1: $\hat{P}_{Westslope} = 0.81$, $\hat{P}_{Yellowslope} = 0.07$, and $\hat{P}_{Rainbow} = 0.12$. Each step in the iteration is guaranteed to increase the likelihood of the estimates, but estimates may converge on a local optimum instead of the global optimum. I have not encountered examples of this problem, but because it is possible, it is prudent to start iteration for different sets of starting points to verify that the same solution is found. "Broken-stick" random numbers make convenient starting points for iteration because they are uniformly distributed in multidimensional space (Devroye 1986). They can be generated using the spacings of $N_{Taxa} - 1$ uniformly distributed random numbers on the interval [0, 1].

The method above is equally useful for estimating the frequency of taxon-specific alleles in a single individual. In this application, N_{sample} is the total number of genes in the individual's multilocus genotype. If there is no missing data, this will equal 2 × N_{Loci}. Results for the sample data shown in Table 1 are shown in Table 2.

A computer program, *Clarki*, is available from the author's website (www.montana.edu/kalinowski) for estimating the ancestry of individuals and populations using SNP data. The program checks for local maxima and provides a warming if any are found (or if iteration is slow). The program runs on the Windows operating system. A description of the program and sample data files are available on the website.

Acknowledgments Funding for this work was provided by Montana Fish Wildlife and Parks and the National Science Foundation (DEB 071745). I thank two anonymous reviewers for suggestions that improved this paper.

References

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. Trends Ecol Evol 16:613–622

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160:1217–1229

Benke RJ (2002) Trout and salmon of North America. The Free Press, New York

Ceppellini R, Siniscalco M, Smith CAB (1955) The estimation of gene frequencies in a randomly mating population. Ann Hum Genet 20:97–115

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Devroye L (1986) Non-uniform random variate generation. Springer, New York

Finger AJ, Stephens MR, Clipperton NW, May B (2009) Six diagnostic single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trouts. Mol Ecol Resour 9:759–763

Kalinowski ST, Taper ML (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. Conserv Genet 7:991–995

Leary RF, Gould WR, Sage GK (1996) Success of basibranchial teeth in indicating pure populations of rainbow trout and failure to indicate pure populations of westslope cutthroat trout. North Am J Fish Manag 16:210–213

McLachlan G, Krishnan T (1996) The EM algorithm and extensions. John Wiley and Sons, New York

Muhlfeld CC, Kalinowski ST, McMahon TE, Painter S, Leary RF, Taper ML, Allendorf FW (2009) Hybridization reduces fitness of cutthroat trout in the wild. Biol Lett 5:328–331

Ostberg CO, Rodriguez RJ (2004) Bi-parentally inherited species-specific markers identify hybridization between rainbow trout and cutthroat trout subspecies. Mol Ecol Notes 4:26–29

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. Annu Rev Ecol Syst 27:83–109

Shepard BB, May BE, Urie W (2003) Status of westslope cutthroat trout (*Oncorhynchus clarki lewisi*) in the United States: 2002. Westslope Cutthroat Interagency Conservation Team

Stephens MR, Clipperton NW, May B (2009) Subspecies-informative SNP assays for evaluating introgression between native golden trout and introduced rainbow trout. Mol Ecol Resour 9:339–343

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997) Human domination of Earth’s ecosystems. Science 277:494–499

Weir BS (1996) Genetic data analysis II. Sinauer Associates Inc. Sunderland, MA.