TECHNICAL NOTE

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

Steven T Kalinowski

Department of Ecology, 310 Lewis Hall Montana State University, Bozeman MT 59717

Correspondence:

Steven Kalinowski
skalinowski@montana.edu
Phone (406) 994-3232
FAX (406) 994-3190

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 has been 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 extirpation of many of plants and animals (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, especially when hybrid individuals or populations have had only a small genetic contribution from non-native taxa. 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 cutthroat trout (e.g., Shepard et al. 2003). 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 of F1 hybrids to the native taxon 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 conducting the analysis (e.g. Anderson and Thompson 2002, Pritchard et al. 2000), but when the species are evolutionarily well-differentiated, the simplest way to estimate the ancestry of potentially hybridized individuals to use taxon-specific diagnostic markers, and count the proportion of genes 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 be identified in which all individuals in the native taxon have one allele and all the individuals in the non-native taxon have an alternative allele. Finding such diagnostic loci is often not difficult, and the resulting data is 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 difficulty.

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 have may contain ancestry from all three taxa. SNP data from a population of Westslope cutthroat trout in Yellowstone National Park (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 *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 should have a genotype of *CC*, 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) can be used to estimate the genetic composition of individuals and populations in the same manner as it is used to estimate the frequency of *A*, *B*, and *O* blood antigens (Ceppellini et al. 1955; see Weir 1996, Chapter 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 for a population. I will present the method in the context of estimating the ancestry of a single individual

The following notation is useful. Let P_i represent the frequency of the i^{th} taxon's genes in an individual or population ($\sum_i P_i = 1$). Let n_{jk} represent the number of times that allele *k* is observed at locus j within an individual. Let the indicator variable X_{ijk} equal 1 if all individuals in taxon *i* have allele *k* at locus j, and equal 0 if all individuals in taxon *i* have an alternative allele. Let N_{Loci} denote the number of co-dominant diploid loci genotypes that have been genotyped. Let N_{Sample} represent the number of genes sampled for the individual (if there is no missing data, $N_{Sample} = 2N_{Loci}$). Lastly, let $N_{Alleles(j)}$ represent the number of alleles at locus j. For most applications with SNPs and indels, this will equal 2, but there is no restriction on the total number of alleles (provided all individuals in the taxa have the same allele).

The EM algorithm uses iteration to find maximum-likelihood estimates of taxon-specific allele frequencies. Given an estimate of the allele frequencies in a taxon, P_i, a better estimate, P_i', can be obtained from

$$P_i' = \frac{1}{N_{Sample}} \sum_{j=1}^{N_{Loci}} \sum_{k=1}^{N_{Alleles(j)}} n_{jk} \left(\frac{X_{ijk} P_i}{\sum_{i'}^{N_{Taxa}} X_{ijk} P_{i'}} \right)$$

Once P_i' is obtained, it can used as estimate of P_i to obtain an even better estimate (P_i') (using the above equation). Iteration is continued until estimates converge. In practice, it is convenient to stop iteration when the total sum of the absolute value of changes between iterations is less than 10^{-6}.

The method above is equally useful for estimating the frequency of taxon-specific alleles in a sample. In this application, N_{Sample} in the equation above is the total number of genes in the sample. If there is no missing data, this will equal 2 × N_{Loci} × the number of individuals sampled.

A computer program, *Clarcki*, is available from the author's website (www.montana.edu/kalinowski) for estimating the ancestry of individuals and populations using SNP data. The program runs on the Windows operating system. A user's manual and sample data files are also available.

Acknowledgements

Funding for this work was provided by Montana Fish Wildlife and Parks and the National Science Foundation (DEB 071745).

Table 1. Sample genotypes for nine diagnostic SNP loci in 10 trout of unknown ancestry. The population is within the range of Westslope cutthroat trout. Alleles that 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).

	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	AA	GG
YCT allele	C	G	A	C	C	C	A	GG	TT
RBT allele	T	T	T	C	C	C	G	AA	GG
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

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	

References

Allendorf, F. W., Leary, R. F., Spruell, P. & Wenburg, J. K. 2001 The problems with hybrids: setting conservation guidelines. Trends Ecol. Evol. 16, 613–622.

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

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.

Finger AJ , Stephens MR, Clipperton NW, May B (2009) Six diagnostic single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trouts. Molecular Ecology Resources 9, 759-763.

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

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

Leary, R. F., W. R. Gould, and G. K. Sage. 1996. Success of basibranchial teeth in indicating pure populations of rainbow trout and failure to indicate pure populations of westslope cutthroat trout. North American Journal of Fisheries Management 16:210–213.

Kalinowski ST, ML Taper (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics* 7:991-995.

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

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, NW Clipperton, B May (2009) Subspecies-informative SNP assays for evaluating introgression between native golden trout and introduced rainbow trout. Molecular Ecology Resources 9:339-343.

Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. (1997) Human domination of Earth’s ecosystems. Science 277, 494–499.