USING MITOCHONDRIAL DNA GENE TREES AND NUCLEAR RFLPS TO PREDICT GENEALOGICAL PATTERNS AT NUCLEAR LOCI: EXAMPLES FROM THE AMERICAN OYSTER

Matthew P. Hare
Organismic and Evolutionary Biology Department
Harvard University
Biological Laboratory #166
16 Divinity Ave.
Cambridge, MA 02138
mhare@oeb.harvard.edu

Abstract

It is now widely appreciated that the stochastic sorting of neutral ancestral polymorphisms can lead to gene trees at independent loci that disagree with one another and with the organismal pattern of descent. To empirically address this problem in studies of closely related species or in phylogenies with short internodes, increased efforts have been made to collect data on DNA variation at multiple nuclear loci. DNA-level data on population structure in the American oyster (C. virginica) has been reported for mtDNA RFLPs, single copy nuclear RFLPs, and nucleotide sequences from several nuclear loci. These data are reviewed here with an examination of how a mtDNA gene tree and scnRFLP population frequencies can each provide predictions about the degree of genealogical concordance across neutral nuclear loci. A re-analysis of the published mtDNA data for C. Virginica shows both Atlantic and Gulf of Mexico populations to be monophyletic, but also predicts the polyphyletic pattern observed for Atlantic and Gulf alleles at three independent nuclear loci. At the CV-32 locus in the oyster, RFLP data motivate alternative hypotheses of ancestral polymorphism, gene flow and restriction site homoplasy to explain the clinal population frequencies, and an allelic CV-32 genealogy provides evidence against the latter two. Criteria are examined for distinguishing among these hypotheses with RFLP data alone. These heuristic tools are offered with the goal of moving from single gene-tree analyses or RFLP level data to testable predictions about genome-wide patterns of genealogical descent.
Introduction

Assays of molecular variation have become indispensable for the inference of population histories and demography (Avise, 1994). Analysis of DNA variants, particularly mtDNA haplotypes, provides a hypothesis of the intraspecific pattern of genealogical descent from which demographic parameters, such as gene flow, can be estimated. The remarkable power of molecular assays has been demonstrated most profoundly in those species that are either morphologically invariant (e.g. horseshoe crabs) or environmentally plastic (e.g. oysters). The American oyster (*C. virginica*) is an intriguing example because molecular markers have revealed population structure between Atlantic and Gulf of Mexico ('Gulf hereafter) populations in spite of a life history that appears to promote panmixture (Reeb and Avise, 1990; Karl and Avise, 1992). American oysters spawn synchronously, have external fertilization, and have an extended larval stage of development in which long distance dispersal is possible (Galtsoff, 1964). In addition, *C. virginica* inhabits estuarine environments along the east coast of North America where barriers to dispersal are not readily apparent.

Given the fundamental assumption of nearly neutral evolution required to use molecular markers for inferences of historical demography and population structure, it is expected that patterns of differentiation will generally agree across independent loci, with any among-locus variation in pattern due to sampling error or the inherent variance of stochastic processes (mutation, lineage sorting, etc.). In this paper I will examine theoretical expectations for agreement in genealogical patterns across independent neutrally-evolving loci. Empirical DNA-level research on oyster population history has progressed in a predictable fashion from assays of mtDNA to nuclear RFLP frequencies and ultimately nuclear sequence variation. In each progression there is a question of how well the existing data predict a higher resolution assay or data from multiple independent loci. Thus, I will use the accumulated data from *C. virginica* to illustrate a predictive framework in which data from RFLPs, mtDNA sequences, and nuclear sequences are progressively predictive about genome-wide (neutral) patterns as well as informative about organismal historical demography. This goal is built upon the assumption that inferences of historical demography are more accurate when based on patterns observed across unlinked loci rather than from a single gene tree. The reasons for this assumption have been thoroughly described elsewhere (Neigel and Avise, 1986; Pamilo and Nei, 1988; Wu, 1991; Maddison 1997).
The oyster has been the subject of two controversial observations of discordance across molecular data sets: (1) geographically uniform allozyme allele frequencies (Buroker, 1983) compared to differentiation of both mtDNA and single copy nuclear (scn) RFLP variants in the Atlantic and Gulf of Mexico (Karl and Avise, 1992), and (2) the same contrast between markers in two studies of scnRFLPs (Karl and Avise, 1992; McDonald et al., 1996). Although these observations continue to be enigmatic and are certainly relevant to predictions of genome-wide patterns of genealogical structure, they will not be considered further here.

**Predictions of Nuclear Genealogical Patterns from a mtDNA Gene Tree**

Under what conditions does a mitochondrial gene tree predict genealogical patterns at independently evolving neutral nuclear loci? In many cases, for example where a mitochondrial result is surprising to a researcher, it is desirable to confirm the result with data from one or more independent nuclear loci. The mitochondrial pattern might not be representative of the organismal history if a species has sex-specific patterns of dispersal, or because of selection on the mitochondrial genome. Because historical demography is expected to affect all loci similarly (in contrast to selection, but see caveats below), corroboration from an unlinked locus can bolster the inference of organismal history from genealogical patterns. Coalescent theory suggests that among nuclear autosomal genes that have shared the same demographic history there will be stochastic variation in the patterns of lineage sorting from a polymorphic ancestor. This will affect genealogical patterns of descent in two cases; (1) at internal nodes in a phylogeny, when the time between branching events was too short to allow complete lineage sorting, and (2) at the "tips" of a phylogeny, where short divergence times separate sister taxa.

The process of lineage sorting perhaps is most easily depicted in the second circumstance described above. Incipient species will initially share many alleles that were segregating in the ancestral species. Over time, in each species independently, some allelic lineages will be lost and others become fixed in a stochastic lineage sampling (= sorting) process governed by genetic drift. As with rates of change by genetic drift, lineage sorting proceeds faster in smaller populations, ultimately leading to alleles within a species that all share a most recent common ancestor (MRCA) that postdates the speciation event. All other allelic lineages have gone extinct, including those that were shared with the sister taxon. At such time that this "coalescent" event or common allelic ancestor postdates the speciation
event, the alleles in that species are all each others closest relatives, i.e. monophyletic. This lineage sorting is an inevitable consequence of chance variation in reproductive success-- whether mutations have occurred enabling us to empirically distinguish allelic lineages is another matter. The rate of lineage sorting is determined by the evolutionary effective population size of the gene, $2N_e$ for nuclear autosomes, where $N_e$ is the effective number of breeding individuals (Hudson, 1990). On average, with no hybridization, lineage sorting is expected to produce exclusive species clades (reciprocal monophyly in two daughter species) after $4N_e$ generations since speciation (Neigel and Avise, 1986).

In contrast, effective population size for mtDNA is $N_e(f)$, the effective number of breeding females. The reduction is by a factor of two because this subgenome is haploid, and by another factor of two when males and females reproduce equally because only females transmit mtDNA through the ovum. Thus, there should be a four-fold faster rate of mtDNA lineage sorting in species with a sex ratio of one (exceptions to these generalities are discussed in Birky, 1991 and Hoelzer, 1997). The influential argument has been made, based on this theory, that a mtDNA gene tree is preferable to nuclear gene trees for inferring organismal relationships because it has a higher probability of reflecting the species phylogeny (Moore, 1995). Although this logic is sound, it overlooks an important conclusion from coalescence theory that no single gene tree should give us confidence in our organismal inferences made under a neutral model. The possibility of selection influencing variation at any one locus reinforces this message.

However, for many practical reasons mtDNA provides a good starting point for molecular evolutionary studies. Given the above assumption about relative effective population sizes of mtDNA and autosomal genes, coalescent theory provides criteria, the "3x rule", by which a mtDNA gene tree can predict the degree of genealogical concordance at nuclear loci (Palumbi et al., submitted). Few empirical studies have compared genealogical patterns at mtDNA and multiple independent nuclear genes, so tests of these predictions and their underlying theory are needed before single gene trees can be appropriately interpreted in terms of organismal relationships.

The 3x Rule

A predictive framework for empirical tests of the theory described above was established by Palumbi et al. (submitted). They reasoned that the relative branch lengths in reciprocally monophyletic mtDNA gene trees (already available for many taxa) foretell the degree of genealogical
concordance across nuclear loci (Palumbi et al., submitted). Their "3x rule", conceived with reference to the evolution of genealogical monophyly within sister species, operates as follows.

With respect to two daughter species diverging from the latest speciation event, the following predictions can be made from existing mtDNA gene trees under a standard neutral model. Obviously, if mtDNA does not show exclusive species clades then we predict the same for nuclear loci, because our expectation is that nuclear loci will take four times longer to show this pattern. Speciation in this case is too recent for ancestral polymorphisms to have sorted into exclusive clades at either mtDNA or nuclear loci. Imagine that we follow the progress of lineage sorting in the two diverging species through evolutionary time. When mtDNA barely shows exclusive clades (evidenced by all mtDNA alleles in one species being more closely related to each other than to alleles in the other species), then an additional period of time must pass before we expect the majority of nuclear loci to show the same pattern (Fig. 1a). Based upon

Fig. 1. Diagrammatic representation of the 3x rule as applied to (a) a reciprocally monophyletic mtDNA gene tree that has little inter-taxon divergence compared to intra-taxon diversity and (b) to a mtDNA gene tree later in the evolution of the same taxa. In tree (b) the length of time needed to generate intra-taxon diversity is one third the (corrected) divergence time. The intrataxon sequence diversity is represented by a time depth of \(d\). A relative divergence time of 3d as shown in (b), or greater, is necessary at mtDNA before a majority of nuclear loci are also expected to show monophyletic A and B clades (see text).

population genetic theory, that period of time is 3\(d\), where \(d\) is the intraspecific mtDNA diversity (measured as average pair wise sequence divergence). This is true because \(d\) scales with \(N_e\) and provides a lineage-specific estimate of the time required for lineage sorting to be completed (Palumbi et al., submitted). If mtDNA required \(d\) units of time for lineage
sorting to generate exclusive clades, then on average, nuclear loci will require $3d + d = 4d$ units of time. Then, on an mtDNA gene tree in which time is represented by branch lengths, the most recent common ancestor of all mtDNA variants within a species occurs at time depth $d$, and the length of the branch leading from that node back in time must be $3d$ before exclusivity is expected at the majority of nuclear genes (Fig. 1b). Although this genealogical representation provides the logic behind the 3x rule, in practice it is calculated for two sister taxa simultaneously by comparing the average or maximum likelihood distance between the taxa (corrected for within-taxon variation) to the sum of intra-taxon nucleotide diversities (Palumbi et al., submitted).

The 3x rule puts bounds on our expectations for genealogical concordance under a neutral model (within the confidence allowed by a large stochastic variance in lineage sorting across loci). If the 3x rule predictions do not hold for the majority of nuclear loci tested then the assumptions underlying the prediction such as equal effective population sizes of males and females, constant population sizes in the two species, and the absence of hybridization between species must be examined. The reliability of this theory and these assumptions is of fundamental importance for the description of evolutionary significant units in conservation biology (Palumbi and Cipriano, submitted). However, even very basic multi-locus tests of these assumptions and the underlying theory are lacking at all phylogenetic levels, so it is of interest to apply this predictive framework to intraspecific data from the American oyster.

Oyster mtDNA Gene Tree

The available mtDNA data on American oysters consists of RFLP variation in the total molecule revealed by 13 restriction enzymes (Reeb and Avise, 1990). The published tree (Fig. 2a) was generated from an average of 65 restriction fragments per individual, UPGMA clustering of genetic distances calculated from the proportion of shared fragments (Nei and Li, 1979), and midpoint rooting. This 'fragment approach' was used because restriction site differences could not be confidently inferred between Atlantic and Gulf fragment patterns for enzymes that produced many fragments (AvaiI, HincII, HindIII, MspI). Ideally, outgroup rooting and maximum likelihood estimation of branch lengths would be used to assess genealogical structure of mtDNA and make nuclear predictions, but here the UPGMA results will suffice for an initial application of the 3x rule at the intraspecific level.

To compare interpopulation mtDNA divergence to intrapopulation diversity, the average genetic distance between Atlantic and Gulf
individuals was corrected for average intrapopulation diversity ($p_{AG} = 0.024$) and then compared to the sum of average sequence divergence in the Atlantic ($p_A = 0.0014$) and Gulf ($p_G = 0.0025$) populations:

$$p_{AG} + (p_A + p_G) = 6.15.$$  

Thus, by the logic of the 3x rule, this mtDNA data indicates that there has been 6x more time than necessary on average for mtDNA coalescence in these populations, suggesting that most nuclear autosomal loci should also show intra-population coalescence in both the Atlantic and Gulf. Because

Although underestimates of $p$ are expected from the fragment approach when $p \geq 0.1$ (Nei, 1987 p. 107), the possibility remains that the large genetic divergence between Atlantic and Gulf haplotypes might be an artifact of the fragment analysis. This could occur if size polymorphisms had a disproportionate affect on the scoring of complicated fragment patterns involving smaller fragments. In the absence of sequence data from oyster mtDNA, I attempted to 'correct' for this potential artifact by re-analyzing data from a subset of enzymes for which restriction site changes can be inferred. Using gel profiles from nine of the 13 restriction enzymes (generously provided by J. Avise), 43 sites were inferred of which 27 were variable and 15 were phylogenetically informative. This is not a random subset of the original data. It excludes the variation most susceptible to artifacts in a fragment analysis, but it also removes that portion of the data that appears to have 'captured' interpopulation differences. However, because one restriction enzyme is no more likely to assay interspecific
divergence than another, this re-analysis is of interest. The resulting site
data, used to calculate nucleotide sequence divergence (Nei and Li, 1979),
produced the UPGMA tree in Fig. 2b. This subset of RFLP data strongly
reinforces the fundamental conclusion of Reeb and Avise (1992) that C.
virginica mtDNA is of two kinds, and these haplotype classes occur in the
Atlantic and Gulf of Mexico, respectively. The corrected interpopulation
divergence in this tree (0.018) is only 1.7 times longer than the sum of the
average intraclade sequence difference (0.0047 Atlantic + 0.0062 Gulf =
0.0109). Accordingly, few nuclear loci are expected to be monophyletic in
the Atlantic or Gulf of Mexico if the difference in $N_e$ between mtDNA and
nuclear genes is four-fold.

[Note: One of the assumptions underlying this analysis that may not hold
for oysters is the four-fold difference in mtDNA and nuclear autosomal
$N_e$. If we take into account the fact that C. virginica is a protandrous
hermaphrodite, breeding as a male early in life and subsequently as a
female, and assume that most individuals survive to breed as females, then
the $N_e$ of mtDNA and autosomal loci might differ only by a factor of two.
In this case both the RFLP fragment analysis and site analysis result in
mtDNA branch lengths that predict monophyly of Atlantic and Gulf of
Mexico populations at the majority of nuclear loci.]

**Inferences from Multiple Nuclear Loci**

Nuclear Gene Genealogies

In an attempt to test the hypothesis that reciprocal monophyly
observed for mtDNA will also characterize genealogies at some nuclear
loci, Hare and Avise (1998) sequenced multiple alleles at three putatively
non-coding loci in Atlantic and Gulf oysters. In all three cases the most
parsimonious genealogy reconstructed from the data showed a polyphyletic
pattern of allelic relationship wherein some Atlantic alleles were more
closely related to Gulf alleles than other Atlantic alleles, and vice versa.
Unfortunately, few phylogenetically informative sites were sampled from
each locus, so the polyphyletic pattern is not statistically supported over a
reciprocally monophyletic alternative in non-parametric tests (Templeton,
1983). However, for the best sampled locus, CV-32, the unconstrained
shortest (polyphyletic) tree was two steps shorter than the shortest tree
possible under a constraint to be reciprocally monophyletic.

Can the polyphyletic pattern result from sampling error? If
additional sequence data were obtained from the CV-32 locus, three or
more diagnostic characters would need to be found among the Atlantic and
Gulf alleles to make either population monophyletic in a maximum
parsimony tree (this was determined by adding to the data matrix progressively more binary "dummy" characters that were diagnostic for Atlantic or Gulf origin and then searching for the shortest trees).

Population sampling error is unlikely to be relevant because no conceivable allele could now be sampled that would alter the polyphyly conclusion. Thus, an initial attempt to examine genealogical patterns across independent nuclear loci in the oyster has not provided evidence for allelic monophyly in either population, as predicted by mtDNA in the re-analyzed RFLP gene tree under the 3x rule. However, the agreement of this result with theoretical predictions will be uncertain until sequence data are obtained for oyster mtDNA and more is known about the variance in reproductive success among protandrous hermaphrodites.

Agreement of nuclear genealogical patterns with predictions from the 3x rule will indicate that the underlying assumptions are being met: neutrality, the MRCA of the two taxa was polymorphic, smaller effective population size at mtDNA versus autosome loci, no sex-biased dispersal or variance in reproductive success and constant population size (or parallel changes in Ne) since the MRCA. A test of the 3x rule in baleen whales with actin intron sequences also found agreement of nuclear patterns with predictions based on mtDNA (Palumbi et al., submitted). As nuclear genealogical data (or surrogates thereof, see below) accumulates for other systems, it will be particularly interesting to address two questions: First, in what systems do the data suggest that mtDNA and autosomal loci have similar demographies, and/or that polymorphism in the MRCA was eliminated at the time of 'speciation'? Second, how well does the variance in lineage sorting patterns observed across loci reflect the prediction for a stochastic process among independently evolving units?

RFLP Analysis
Attempts to examine patterns of DNA-level variation across multiple independent nuclear loci most often have been made with RFLP assays. In its simplest form this procedure might be used to assay one polymorphic restriction site per locus (e.g., Karl and Avise, 1992; McDonald et al., 1996; Harrison and Bogdanowicz, 1997), a few closely linked sites per locus (e.g., Karl and Avise, 1993), or multiple sites within larger regions per locus (e.g., Pogson et al., 1995). Studies of this kind (outside of humans and Drosophila) have rarely assayed enough variation to make genealogical analyses seem appropriate at any one locus, and any such endeavor must deal with some ambiguous haplotypes in double heterozygotes (Karl and Avise, 1993). Instead, this data is generally used to follow population allele frequencies in a manner analogous to allozyme data. I would like to argue that there is crude genealogical information in
this relatively low-resolution assay, and these nuclear RFLP studies therefore provide the most comprehensive genome-wide surveys of genealogical structure available.

In the simplest case, when a single restriction site per locus has been surveyed across populations/species, what are the conditions under which the geographic pattern of variation will reflect the underlying genealogy? Another way of asking the same question is if we have a genealogy for a locus, what is the probability that any one variable nucleotide site will portray the deep structure of the tree in a non-homoplasious fashion? This probability could be estimated by simulation with certain substitution rates, but for our purposes it is important that RFLP studies typically screen variation at a locus and analyze or report the most "informative" or diagnostic restriction site. This is explicit in studies for which species-specific markers are sought (Harrison and Bogdanowicz, 1997), and may be implicit in other applications where low frequency variants are relatively uninformative. Because the pattern of interest is usually population structure, and an observation of even one diagnostic variant is unexpected by chance under the null hypothesis of panmixia (Hey, 1991), this screening of RFLP variants seems justified and essentially amounts to a focus on variants at the base of the underlying gene genealogy. Thus, there is good reason to expect that RFLP variants provide useful information about the underlying genealogy, and the question of accuracy reduces to the probability of homoplasy at restriction sites. Without knowledge of the nucleotide change causing restriction site variation this is impossible to estimate, but among closely related taxa it is expected to be small (Dowling et al., 1996).

The American oysters serve to illustrate my point. Out of eight oyster nuclear loci described with RFLP variants (Karl and Avise, 1992; McDonald et al., 1996), none showed Atlantic and Gulf populations to be fixed for different alleles. Several of the loci showed dramatic allele frequency differentiation (Karl and Avise, 1992), and one in particular (CV-32) showed nearly fixed patterns of variation in the Atlantic and Gulf (Karl and Avise, 1992; Hare and Avise, 1996). Alternate (but not mutually exclusive) hypotheses for the existence of 'Atlantic' CV-32 alleles at low frequency in the Gulf, and vice versa, were (1) gene flow between formerly monophyletic populations, (2) incomplete sorting of ancestral polymorphisms (i.e. genealogical polyphyly accurately reflected by the restriction site) or (3) homoplasy at the assayed Nsi I site (i.e. inaccurate representation of an underlying reciprocally monophyletic tree). Subsequent examination of CV-32 sequences for Atlantic and Gulf alleles (discussed above) suggest that the genealogy at this locus is polyphyletic,
rejecting the third hypothesis (Hare and Avise, 1998). Also, in agreement
with independent evidence for contemporary barriers to Atlantic-Gulf gene
flow (Hare and Avise, 1996), none of the alleles sampled in the Gulf were
identical to alleles sampled in the Atlantic despite the sharing of alleles
across long distances within each region. These observations support
ancestral polymorphism over gene flow as an explanation for the RFLP
patterns at CV-32 and reinforce the argument that restriction site variation
is a useful first estimate of genealogical patterns at multiple loci.

Distinguishing gene flow from ancestral polymorphism is a
formidable challenge (Wakeley, 1996), and I do not wish to imply that
RFLP data, or even multiple gene genealogies will necessarily do the job
conclusively. Introggression is, of course, going to be a common
observation in many systems, even when the two taxa are currently
allopatric. However, in peripatric or partially sympatric taxa there are
signals based on population RFLP frequencies that might distinguish
contemporary or recent gene flow from ancestral polymorphism. Whereas
ancestral polymorphism is not expected to show any geographic pattern
within allopatric populations, gene flow should introgress alleles more
frequently in sympatric populations than in allopatric populations. The
American oysters provide an example of a particularly simple contact zone
along a more or less linear coastline. Along the eastern Florida coast, on
either side of a genetic contact zone at Cape Canaveral (Hare and Avise,
1996), the frequency of alternate CV-32 alleles diminishes in populations
further from Cape Canaveral. Haplotypes from these transitional
populations were not sampled for the CV-32 genealogy, but the clinal
RFLP allele frequency pattern is consistent with introgression (Barton and
Hewitt, 1985). Comparison of allele frequency distributions among
distantly allopatric populations in the Atlantic and the Gulf provides a
different picture. At the RFLP level alternate alleles still appear to be
introgressed, but in these populations they are not clinal, as predicted by
the hypothesis of ancestral polymorphism.

Several historical and demographic factors influence the degree to
which these allele frequency patterns will be useful to distinguish gene flow
from ancestral polymorphism. First, in a species such as the oyster that is
capable of long distance dispersal, there is perhaps only a short period of
time after introgression begins before the affect of gene flow will be
pervasive. Second, the clinal pattern expected from introgression depends
on variation in dispersal distances. Rare episodes of long distance dispersal
can blur the simple distinction that I have described for allele frequency
distributions under the two models (Nichols and Hewitt, 1994). Third, for
taxa that have changed their geographic distributions relatively recently,
historically introgressed alleles are likely to mimic ancestral polymorphisms. Thus, these principles will be most useful for the interpretation of RFLP patterns in low dispersal species. In cases where historical introgression is likely, inferences solely from RFLP patterns will be tenuous. Nonetheless, explicit genealogical inferences from RFLP patterns are a useful intermediate step toward describing and understanding the genome wide distribution of genealogical patterns (Maddison, 1997).

Conclusion

The predictive and analytical tools described here are mostly heuristic, but they serve the goal of moving from single gene-tree analyses or RFLP level data to multi-locus descriptions of genealogical pattern. There is an increasing appreciation that loci which are free to evolve independently, either due to location on different linkage groups or meiotic recombination, can trace fundamentally different genealogical paths through one and the same organismal phylogeny (Moore, 1996; Avise and Wollenberg, 1997; Hoelzer, 1997; Maddison, 1997; Moore, 1997; Ruvolo, 1997). Because neutrality of noncoding sequences is an assumption that applies across the nuclear genome with some uncertainty (Begun and Aquadro, 1992; Charlesworth et al., 1993), it is important to use the data that are already in hand to make testable predictions about genome-wide patterns under a neutral model. Mitochondrial gene trees and nuclear RFLP studies abound for this purpose.

Acknowledgments

I want to thank John Avise for providing the oyster mtDNA RFLP data and for his support and encouragement during my research on C. virginica.

Literature Cited

Birky, C.W., Jr. 1991. Evolution and population genetics of organelle genes: mechanisms and models. In: Evolution at the molecular level,


Palumbi, S.R., F. Cipriano and M.P. Hare. submitted. Predicting nuclear gene coalescence from mitochondrial data: The three-times rule.


