Identifying the source of species invasions: sampling intensity vs. genetic diversity

JIM R. MUIRHEAD, DEREK K. GRAY, DAVID W. KELLY, SANDRA M. ELLIS, DANIEL D. HEATH and HUGH J. MACISAAC

Great Lakes Institute for Environmental Research, University of Windsor, Windsor, Ontario, Canada N9B 3P4

Abstract

Population geneticists and community ecologists have long recognized the importance of sampling design for uncovering patterns of diversity within and among populations and in communities. Invasion ecologists increasingly have utilized phylogeographical patterns of mitochondrial or chloroplast DNA sequence variation to link introduced populations with putative source populations. However, many studies have ignored lessons from population genetics and community ecology and are vulnerable to sampling errors owing to insufficient field collections. A review of published invasion studies that utilized mitochondrial or chloroplast DNA markers reveals that insufficient sampling could strongly influence results and interpretations. Sixty per cent of studies sampled an average of less than six individuals per source population, vs. only 45% for introduced populations. Typically, far fewer introduced than source populations were surveyed, although they were sampled more intensively. Simulations based on published data forming a comprehensive mtDNA haplotype data set highlight and quantify the impact of the number of individuals surveyed per source population and number of putative source populations surveyed for accurate assignment of introduced individuals. Errors associated with sampling a low number of individuals are most acute when rare source haplotypes are dominant or fixed in the introduced population. Accuracy of assignment of introduced individuals is also directly related to the number of source populations surveyed and to the degree of genetic differentiation among them ($F_{ST}$). Incorrect interpretations resulting from sampling errors can be avoided if sampling design is considered before field collections are made.

Keywords: chloroplast DNA, exotic species, mitochondrial DNA, nonindigenous species, phylogeography, sampling bias

Received 10 June 2007; revision received 26 November 2007; accepted 28 November 2007

Introduction

Population geneticists seeking to identify the complement of haplotypes present in a population, and ecologists attempting to quantify species diversity in a community, have long recognized the importance of sampling strategies (e.g. see Pons & Petit 1995; Davies et al. 1999; Colwell et al. 2004; Dixon 2006; Hortal et al. 2006; Latch & Rhodes 2006). Haplotype or species diversity is usually positively correlated both with the number of samples collected within a habitat and with the number of locations surveyed (e.g. Muirhead et al. 2006). In many but not all cases, these curves are well represented as asymptotic functions of sampling effort. Exceptions to this generality include constrained genetic or species diversity in challenging environments (e.g. Vasquez et al. 2005), and genetic diversity of some invading populations which may be represented by a single haplotype (e.g. Cristescu et al. 2001).

Human-mediated species invasions are increasing worldwide (Mack et al. 2000; Ruiz et al. 2000). Identifying the vectors and sources responsible for these introductions is an important prerequisite for development of effective prevention and management strategies (Kolar & Lodge 2002; Fofonoff et al. 2003; Lodge et al. 2006; Rollins et al. 2006). Moreover, knowledge of the origins of nonindigenous species (NIS) can provide insight into their biology, facilitate the selection of optimal biological control agents, and
provide information on the history of invasion and human transport (Collins et al. 2002; Downie 2002; Cognato et al. 2005; Goosby et al. 2006). Determining the source of NIS is also essential for ecological and evolutionary studies. For example, data collected from the genuine source provides an evolutionary baseline that allows comparisons of phenotypic traits, allele frequencies, and genetic diversity among introduced and native populations (Wares et al. 2005). Moreover, ecological studies designed to explore determinants of invasion success often require information on the source of introduced individuals to facilitate the selection of appropriate contrast groups (Colautti et al. 2004; Hierro et al. 2005).

Two differing approaches are available for identifying the source of NIS invasions. If the introduction vector is known, information on trade routes and patterns can be combined with data on the species’ distribution to identify probable donor regions (Ruiz & Carlton 2003; Drake & Lodge 2004; Tatem & Hay 2007). This approach identifies the dominant invasion pathways, which can then be targeted for management efforts aimed at reducing propagule pressure to vulnerable destinations (Rouget & Richardson 2003; Verling et al. 2005). However, this method provides only coarse resolution, and does not definitively link an introduced population with a specific source. An alternative approach uses genetic markers to assess the relationship between introduced and putative source populations. This approach can be subdivided into methods based upon phylogeography and those based upon population genetics. The former combines genealogical and geographical information to reconstruct the history of an invasion, while the latter uses individual assignment based on allele proportions to assess the likelihood that an individual belongs to a specific population (Avise 2000). Strengths and weaknesses of these approaches were reviewed by Wares et al. (2005).

A variety of markers are available to provide data for phylogeographical and allele frequency analyses, including mitochondrial DNA (mtDNA), chloroplast DNA (cpDNA), microsatellite DNA, introns, amplified fragment length polymorphisms (AFLP) and randomly amplified polymorphic DNA (RAPD). The utility of these markers has been reviewed in several studies (Bossart & Prowell 1998; Cruzan 1998; Davies et al. 1999; Sunnucks 2000; Manel et al. 2005; Bazin et al. 2006; Selkoe & Toonen 2006).

Problems relating to sampling may hinder the successful use of molecular markers to identify the source of NIS populations. These problems relate to the number of individuals sampled per source population and/or the number of putative source populations surveyed, issues well-recognized but rarely quantified in population genetics and community ecology. In this study, we take two approaches to highlight sampling problems in the invasive species literature. First, we review published studies that used either mtDNA or cpDNA markers to identify the source(s) of NIS populations. Second, we conduct simulations with an mtDNA database for the amphipod Gammarus tigrinus, which is native to the Atlantic seaboard of North America but introduced to Europe and to the Great Lakes. These simulations highlight and quantify the impact of the number of individuals sampled in source populations, the number of source populations surveyed, and the degree of genetic differentiation between source populations for correct assignment of individuals to source populations.

Methods

**Sampling characteristics of mtDNA and cpDNA studies**

We reviewed studies of mtDNA and cpDNA to highlight sampling problems, as they are among the most commonly used markers for determining the source of NIS introductions (e.g. Soltis & Soltis 1998; Roderick 2004). Our findings, however, should be applicable to studies that utilized other markers as our focus is on sampling issues rather than on specific markers. The popularity of these markers for identifying sources of NIS derives from several characteristics that make them suited to this task. First, the rapid sequence divergence of mtDNA compared with nuclear DNA provides the resolution required to distinguish among source populations despite relatively short histories of isolation (Avise 2000). Although sequence divergence is not as rapid in cpDNA, informative levels of intraspecific variation have been found in many species (Hamilton 1999; McIvor et al. 2001). Second, maternal inheritance of cpDNA in angiosperms and mtDNA in animals helps to retain the genetic structure that originated during the introduction, as it is subject to neither introgression nor recombination, as are nuclear markers (Avise et al. 1987; Gaskin et al. 2005). Third, universal primers are available for certain mtDNA and cpDNA genes, allowing for rapid data collection without dedicating effort to novel primer development (Simon et al. 1994; Soltis & Soltis 1998).

Mitochondrial and cpDNA-based studies for identifying sources of NIS are typically performed within a phylogeographical framework, as this type of sequence data represents one set of rigidly linked loci, limiting the power of frequency-based assignment methods (Epifanio et al. 1995; Avise 2000). Within this framework, several sources of variation can hinder the ability to identify the native source population(s). First, sequence variation in the marker used for the study must be high enough to resolve relationships below the species level. Certain markers, such as the rbcL gene in plant chloroplasts, usually fail to provide the required resolution (Soltis & Soltis 1998). Fortunately, particular sequences – such as intragenic chloroplast regions and the mitochondrial control region – evolve rapidly at the nucleotide level, potentially allowing for
discrimination among source populations (Avise et al. 1987; Hamilton 1999; but see Davies et al. 1999). Second, sufficient phylogeographical structure must be present to discriminate among sources. Avise (2000) categorized the outcomes of phylogeographical studies along a continuum from highly structured (i.e. deep gene tree, major allopatric lineages) to those showing little structure (i.e. shallow gene tree, sympatric lineages). In the latter case, it may be more difficult to assign introduced individuals to native populations that are genetically homogeneous or nondifferentiated (Wares et al. 2005). Third, the number of individuals sampled per source population and the number of source populations sampled must also be considered, and both must be sufficient to uncover true phylogeographical patterns (Neigel & Avise 1993; Templeton et al. 1995; Hedin & Wood 2002; Morando et al. 2003; Templeton 2004). The number of source populations surveyed must be high enough to provide a high probability of including the genuine source population.

Phylogeographical methods can use haplotype matching or genetic distance methods to source individuals to geographically structured native populations; however, both approaches generally do not provide statistical confidence estimates (e.g. Knowles & Maddison 2002). The source of an individual can be inferred by matching of population-specific haplotypes from the native range, or from the introduced range in cases of secondary invasions (Cristescu et al. 2001; Castilla et al. 2002). In the event that a population-specific haplotype is not identified, the general source region for the introduced individual can be inferred from haplotype similarity (i.e. minimum genetic distance) to genetically structured source populations (Slade & Moritz 1998; Collins et al. 2002). The ability to recover shared, population-specific haplotypes, or haplotypes closely related to those of an introduced population, can be affected by the number of individuals sampled per source population and the number of source populations. An ideal sampling strategy would involve collecting a large number of individuals from each population in a nonindigenous species’ native and introduced range. However, limited resources or access often precludes this strategy, resulting in sampling of fewer individuals per source population or of a reduced number of source populations (Downie 2002). If the aim of a study is to identify a larger source region, then increasing the number of sources sampled and decreasing the number of surveyed individuals per population would maximize genetic diversity recovered in the study and increase the likelihood of identifying haplotypes closely related to those of introduced individuals (Lynch & Crease 1990; Pons & Petit 1995). However, if the goal is to identify the actual source population, it might be necessary to increase the number of individuals per population that are surveyed, as the probability of recovering specific haplotypes is directly related to the number of individuals sampled and to haplotype frequencies (Ott 1992; B-Rao 2001). Small sample sizes and the presence of rare haplotypes can render difficult the recovery of all haplotypes in a population, decreasing the likelihood of finding haplotypes that match those of introduced individuals (Crossa 1989; B-Rao 2001). Given such trade-offs, and the variety of species and systems assessed, researchers may take a variety of approaches when conducting studies to identify the source population of NIS.

We conducted a search of the ISI database using pairs of keywords (Table 1) to identify studies that utilized either of our markers to determine the source of NIS invasions. Publication dates ranged from 1994 through 2006 (Appendix 1). To gauge the number of individuals sampled per source population and the number of source populations analysed from previous studies, we investigated only those species with clearly defined native and introduced distributions. Limiting the data set in this manner allowed us to readily identify the number of individuals sampled per source population and the number of source populations surveyed for all populations. Recovered studies were chiefly concerned with tracing human-mediated introductions rather
Sampling effort may depend on the conspicuousness, perceived importance, or ease of collection and transport of different taxa. To explore this, we conducted nonparametric analyses of variance tests (Kruskal–Wallis) to determine whether the number of individuals sampled or number of populations surveyed varied across taxa (aquatic invertebrates, insects, fishes, plants, mammals, reptiles, and amphibians) in each of native and introduced ranges. Linear regression was utilized to determine whether mean number of haplotypes recovered was related to the mean individuals sampled per source population or to the number of source populations studied. We also used linear regression to assess whether mean number of individuals sampled per population [log(x + 1)] changed through time.

Simulations with an mtDNA database

We conducted Monte Carlo simulations using an original data set to explore how sampling errors could influence identification of sources of introduced species (see Kelly et al. 2006a). We chose to base our simulations on an extensive, published mtDNA sequence data set to provide assurance that the baseline sequence data were realistic; however, we re-sampled and re-assigned those data to generate quantitative simulations of specific invasion and source population scenarios not present in the original data. Specifically, we sought to quantify both the effect of the number of individuals surveyed per source population and the number of source populations surveyed on the ability to correctly identify the source of an NIS. We used a comprehensive data set of mtDNA sequences for the amphipod *Gammarus tigrinus*, a species that is native to tidal estuaries of eastern North America but introduced to fresh and brackish waters throughout Europe as well as the Laurentian Great Lakes (Kelly et al. 2006a).

In our first set of simulations, we sought to determine the importance of the number of individuals surveyed per native (source) population. The model was created in R (R Development Core Team 2007) with additional packages nnet (Venables & Ripley 2002), ape (Paradis et al. 2005), matrix (Bates & Maechler 2007), and phylot (Diaz-Uriarte & Garland 2007). The data set for the source range consisted of sequences for 177 individuals from nine populations (Appendix 2). The Delaware Estuary was selected as the source for an introduced population, based upon the results of Kelly et al. (2006a). Our survey of this population revealed nine haplotypes consisting of 12 individuals of the most common haplotype, 7, 4 and 2 individuals of the next most common forms, and one individual each of the five remaining haplotypes. For each of 200 bootstrap iterations, 10 individuals from the Delaware Estuary were randomly selected to act as colonists for a simulated introduced population. Nine individuals from each source population were then randomly chosen without replacement. We chose nine individuals per source population for the upper limit as a higher value would have required that we drop native populations with fewer than 10 individuals (i.e. St. John estuary; Appendix 2). Pairwise minimum genetic distances based on the Kimura 2-parameter model (Kimura 1980) were calculated between individuals from the introduced and native populations. We chose this model since it is one of the more common ones used in invasion literature. This process was repeated, decreasing the number of individuals surveyed per native population progressively to two. During each bootstrap iteration, every introduced individual was matched to a native population based upon exact haplotype matching of individuals or, in the absence of a match, by minimum pairwise genetic distance. In the event where minimum genetic distance was tied among individuals from different native populations, the introduced individual was randomly matched to an individual from one of these populations. This form of matching is relatively unbiased as resampling randomizes haplotype matches with each bootstrap iteration. In addition, the number of iterations should be sufficient to track rare misclassification events. To determine if the source population had been correctly identified (classification accuracy), we determined the proportion of introduced individuals that were matched to the known source, the Delaware Estuary.

We estimated assignment confidence for each of the 10 simulated introduced individuals, $c_i$, by:

$$c_i = \begin{cases} 1 - \min_{i,j} d_{ij} & \\
\frac{1}{m} \sum_{j=1}^{m} d_{ij} & 
\end{cases}$$

where $\min_{i,j} d_{ij}$ is the assigned (i.e. minimum) genetic distance between introduced individual $i$ and source individual, $j$. Here, the denominator is simply the average genetic distance between an introduced individual and all source individuals. Assignment confidence ranges from 0 to 1, with 1 indicating an exact match.

Additional analyses were conducted to determine the influence of haplotype frequencies in the introduced population on classification accuracy for differing numbers of individuals per native population. Many studies have revealed that haplotypic composition is a subset of that found in native populations, although some studies have revealed fixation of rare native (e.g. Cristescu et al. 2001; Hänfling et al. 2002; Kelly et al. 2006a) or common native...
(Grapputo et al. 2005; Lindholm et al. 2005) haplotypes in introduced populations. Here, we took either the most common (12/30 individuals) or rare (1/30 individuals) haplotype from the Delaware population and allowed its frequency to vary between 30% and 100% of individuals in the introduced population. Remaining individuals, if any, in the introduced population were selected at random from the remaining haplotypes in the Delaware population. Analyses were repeated 200 times for all values of the number of individuals surveyed per population (i.e. two to nine individuals in each source).

To assess the influence of the number of source populations surveyed on classification accuracy, we conducted simulations that randomly removed source populations. At each bootstrap iteration, 10 individuals for an introduced population were randomly selected from the Delaware population, and between zero and seven source populations were randomly excluded. Two hundred bootstrap iterations were run for each value of the number of populations surveyed (i.e. two to nine source populations). To examine the interaction between the number of individuals per population and the number of populations surveyed, we repeated these simulations using 2, 5, and 9 individuals per source population.

Wright’s $F_{ST}$ provides a measure of interpopulation genetic differentiation. We expect the existence and extent of spatial population genetic structure to influence classification accuracy of invasion tracking studies. To investigate this, we generated simulations using data sets representing a gradient of global $F_{ST}$ values among source populations based on the original sequence data for *G. tigrinus* (Kelly et al. 2006a). The $F_{ST}$ value for our initial data set (0.667) reflects unusually high spatial genetic structuring, as many haplotypes were unique to particular estuaries (Kelly et al. 2006b). The global $F_{ST}$ value for each new data set was lowered (i.e. source populations were sequentially homogenized) through successive exchanges of the two most common Delaware haplotypes with each of the other eight estuarine populations. To maintain original population sample sizes, and to ensure that $F_{ST}$ was lower for each new data set, we replaced the exchanged Delaware haplotypes with haplotypes that were unique but common to the recipient population. This procedure was repeated for haplotypes taken from two other source populations (i.e. St. Lawrence, Hudson Rivers) to further decrease interpopulation differentiation. In this simulation, the introduced population was composed of 10 randomly selected individuals from the Delaware population. We then assessed the effect of interpopulation genetic differentiation and the number of individuals surveyed per population on classification accuracy using 200 bootstrap iterations with two to nine individuals per source population. For each model iteration, the introduced population was composed of 10 individuals resampled from the Delaware. Matches were considered accurate only when individuals from the introduced population were matched back to the Delaware.

In the event where there is substantial gene flow among source populations, classification accuracy of the introduced population decreases in relation to global $F_{ST}$. We repeated the previous simulation under the conditions where a match is considered accurate at the regional level if the introduced individuals (Delaware) were most closely related to populations in either the Delaware or adjacent Poropotank and Elizabeth estuaries in Chesapeake Bay.

**Results**

*Sampling effort from the literature*

Our literature search returned 97 studies that used mt or cpDNA sequence variation to determine the source of species introductions (Fig. 1). Of these, we were able to discern both the number of individuals surveyed per population and the number of sources surveyed for 63 studies (Appendix 1). Seventy-six per cent of studies used mtDNA or cpDNA sequences to identify the source of the invasion, while the other 24% used these markers in combination with microsatellites, allozymes, RAPDs or nuclear introns. These additional markers were used by the authors to enhance discrimination among populations and to allow for the assessment of genetic diversity with multiple independent markers (e.g. Fonseca et al. 2001; Gopurenko et al. 2003).

![Fig. 1 Cumulative number of studies published between 1994 and 2006 that have used mtDNA and cpDNA sequence variation to locate the source of an introduced species. The line was fitting an exponential function ($y = 0.708e^{0.399x}$; $r^2 = 0.98$).](image)

© 2008 The Authors

Journal compilation © 2008 Blackwell Publishing Ltd
Mean number of individuals sampled per native population ranged from 1 to 46. This value was generally low, with 33% of studies averaging fewer than three individuals sequenced per native population, and 59% with fewer than six (Fig. 2a). The number of individuals analysed for introduced populations ranged from 1 to 100, and was skewed towards slightly higher numbers than for native populations (Fig. 2a, b). Only 45% of introduced populations studied included fewer than six individuals, and 13% included 25 or more (vs. 7% for native populations).

Between 1 and 89 populations were surveyed in the native range, whereas values in the introduced range were typically lower (i.e. 1–59 populations; Fig. 2b, d). Twenty-one per cent of studies sampled relatively low (= 4) numbers of populations in the native range, and only 52% of studies sampled > 9 (Fig. 2b). Thirty-seven per cent of studies sampled = 4 populations in the introduced range, whereas 35% sampled > 9 (Fig. 2d).

Neither the number of individuals sampled per population nor the number of populations sampled varied significantly by taxon studied (aquatic invertebrates, insects, fishes, plants, mammals, reptiles, and amphibians) for either native or introduced ranges (Kruskal–Wallis ANOVAs, P > 0.05 in all cases). Thus, easily sampled taxa like aquatic invertebrates were sampled with the same intensity as more mobile ones (e.g. mammals).

Review of haplotype diversity patterns revealed that increased sampling provides a more comprehensive evaluation of genetic diversity throughout a species’ distribution. For example, the average number of haplotypes recovered per population was directly related to the number of individuals [log(\(x + 1\))] surveyed (\(y = 0.28x + 0.11\)), whereas the total number of unique haplotypes recovered was positively related to the number of source populations sampled within a study (\(y = 0.54x + 0.51\)).

Temporal patterns of mean haplotype diversity per study demonstrate that, for both introduced and source populations, some studies were based upon relatively small numbers of individuals surveyed. Of 37 studies published during 2005 and 2006, we were able to determine the sampling characteristics of 25 (Appendix 1). Eight of these studies used mean sample sizes of < 6 individuals to characterize haplotype diversity of introduced populations (Fig. 3a; Lindholm et al. 2005; Smith 2005; Städler et al. 2005; Austin et al. 2006; Chu et al. 2006; Havill et al. 2006; Scheffer & Lewis 2006; Steiner et al. 2006). The pattern was even more pronounced with source populations, in which 13 studies published during these years used < 6 individuals (Fig. 3b; Hingston et al. 2005; Lindholm et al. 2005; Smith 2005; Städler et al. 2005; Williams et al. 2005; Zardus & Hadfield 2005; Austin et al. 2006; Chu et al. 2006; Goolsby et al. 2006; Havill et al. 2006; Muñoz-Fuentes et al. 2006; Scheffer & Lewis 2006; Steiner et al. 2006). Overall, there has been no significant change in mean number of individuals surveyed through time, even though sequencing cost has dropped in recent years (linear regressions, \(P > 0.10\); Fig. 3a, b). Likewise, there has been no significant shift in the number of introduced or source populations surveyed through time (linear regressions, \(P > 0.10\); Fig. 3c, d).

Mitochondrial DNA simulations

When the 10 individuals in the introduced population were randomly drawn from the source population, classification
accuracy increased from a mean of 52% to 69% as the number of individuals surveyed per population increased from two to nine individuals (Fig. 4a). As the frequency of the rare haplotype was increased from its original value of 3.3% to 100% in the introduced population, the number of individuals surveyed had a progressively stronger effect on classification accuracy. For example, when the rare haplotype was fixed in the introduced population, classification accuracy increased to 85% when the number of individuals surveyed was high (nine individuals per source population), although it declined to 51% when the number of individuals sampled was two (Fig. 4a). These results are attributable to the probability that the rare haplotype will be recovered in the subsampled source population.

Confidence in individual assignment was influenced more by the number of individuals sampled per source population than by the proportion of the rare haplotype in the introduced population (Fig. 4b). Overall, assignment confidence was highest when the frequency of the rare haplotype in the introduced population was the same as that in the original population (i.e. low frequency).

The effect of the number of individuals surveyed per source population was less pronounced when the frequency
of the common haplotype was fixed in the introduced population (Fig. 4c). Curves constructed with a differing number of individuals per population, and with different frequencies of the common haplotype, diverge only at very small source population sizes, as even modest sampling is able to recover the haplotype in the source population (Fig. 4c).

Confidence in individual assignment was higher in simulations where the frequency of common haplotypes varied than in simulations where frequencies of the rare haplotype varied (Fig. 4b, d). Confidence ranged from 0.86 to 0.96 when two individuals per source population were sampled to 0.97–1.00 for nine individuals sampled (Fig. 4d).

The ability to correctly match introduced individuals to their source population was strongly related to the number of source populations surveyed (Fig. 5). As the number of source populations increased, classification accuracy increased concomitant with the likelihood of including the Delaware source population. This effect was dependent, however, on the number of individuals surveyed per source population. With only two source populations, there was a negligible (5%) absolute improvement in classification accuracy associated with increasing the number of individuals surveyed per population because it was not as likely that individuals were drawn from the Delaware source. However, when all source populations were considered \((n = 9)\), absolute classification accuracy improved by 20% as the number of individuals surveyed per source population increased from two to nine (Fig. 5). Confidence in individual classification also increased as the number of source populations sampled increased. Confidence index ranged from 0.45 to 0.56 for two source populations sampled to 0.86–0.97 for nine source populations sampled (Fig. 5).

Spatial genetic structure

The ability to accurately classify introduced individuals to a source population was strongly dependent on the degree of genetic differentiation among source populations (Fig. 6). Mean classification accuracy was high only when genetic variation among sources was pronounced. In addition, as genetic differentiation among sources increased (i.e. \(F_{ST}\) increases), the ability to correctly resolve the source of an invasion increased with the number of individuals surveyed per population. At low levels of genetic differentiation among source populations, the number of source individuals sampled had a discernable influence on the classification accuracy. If the source populations exhibit high levels of genetic differentiation, the number of source individuals sampled will have less of an effect on the ability to correctly source individuals from the introduced population. Variation in the classification accuracy is high, reflecting the
genetic structure of both the founding and source populations (Fig. 6b, c). For example, at a global $F_{ST}$ of 0.5, with five individuals sampled per source population, 1 SD in classification accuracy was ±15%. If, by chance, the founding population was composed predominantly of haplotypes shared among source populations, classification accuracy decreased. In our simulations, even at high values of global $F_{ST}$, approximately 10–15% of individuals in the introduced population were matched incorrectly to the two nearest sources.

If only a regional-scale resolution is sufficient, the accuracy in correctly sourcing an introduced population is increased significantly. Mean classification accuracy for a given level of genetic differentiation among source populations was nearly double that from simulations with sources at the population-scale resolution (Fig. 6a, d). Classification accuracy was influenced more by the interpopulation genetic structure than on the number of individuals sampled per source population across the range of global $F_{ST}$ values. Confidence in individual assignment was high across the range of global $F_{ST}$ values for both local and regional-level classification. In both scenarios, confidence was influenced to a greater extent by the number of source individuals sampled per population than by the genetic structure of populations (Fig. 6e, f).

### Discussion

Ecologists use molecular genetic techniques and analyses to address a wide variety of questions, including dispersal (Selkoe & Toonen 2006). Human-mediated introduction and dispersal of species has, in many cases, supplanted natural dispersal. Identifying the source of an introduced population can be difficult given global movement of humans and commodities, although it often can be resolved by analysis of trade patterns or vector movement (Ruíz & Carlton 2003). Molecular tools can provide useful information to link invaded destinations with putative source populations. Results from our literature review illustrate that invasion ecologists have increasingly applied mtDNA or cpDNA surveys to ascertain the source of invading populations (Fig. 1). For example, 37 studies have been published in 2005 and 2006 that used these markers to study patterns of genetic diversity in invading and source populations (Fig. 1). Despite its attractiveness and apparent utility, caution must be applied to studies that utilize genetic markers to assess invasion pathways.

One of the limitations of using phylogeographical methods to identify the source of introduced species is that no simple method exists to estimate the probability of error (Knowles & Maddison 2002). Even in cases of direct haplotype matching with private haplotypes in putative source populations, the possibility exists that the haplotype exists in other unsampled or undersampled source populations.

Nested clade analysis (NCA) is an analytical approach that identifies significant geographical patterns among genetic data (Templeton et al. 1995; Templeton 2004). NCA is of particular value for applications of phylogeography to invasion sourcing, since it provides an objective test for undersampling of individuals at a site, or undersampling of sites within the native range (Templeton 2004). However, such an analysis cannot provide an estimate of confidence in subsequent invaded individual assignment to specific source populations; rather a simulation approach such as that described here must be employed (i.e. ‘statistical phylogeography’; Knowles & Maddison 2002). Furthermore, the results of our simulations also apply to many studies where phylogeographical data and haplotype matching are used to identify common or divergent refugia for subsequent colonized populations.

From our simulations and the results of published studies, the ability to correctly match invading individuals to their source depends on the spatial resolution of accuracy desired and on the genetic structure of the surveyed populations. For example, in order to classify a single invading population to its putative source at the population level of resolution with an accuracy of at least 50%, we recommend at least seven individuals if the global $F_{ST}$ of the source populations is at least 0.6. If classification at a regional level is sufficient, accuracy increases to ~83%. The number of individuals required for accurate classification increases as the frequency of shared haplotypes among source populations increases. Of the 63 papers that reported population sample sizes, only three reported global $F_{ST}$. Of these, two sourced the invasions at a regional level, with a mean classification accuracy of 40% (Grapputo et al. 2005) and 95% (Laffin et al. 2005) according to our simulations. The third study classified the invasion back to a specific source population since the authors had knowledge that the source was a single introduction of captive individuals (Muñoz-Fuentes et al. 2006). Based on the sample size in their study, had they not known the specifics of the introduction, they would only have had ~15% and ~40% chance of matching back to the source population or regional scale, respectively, based on a low global $F_{ST}$.

A critical assumption behind these studies is that the resolved phylogeographical patterns are accurate and complete representations of the actual genetic diversity in both source and introduced populations. We noted that haplotype diversity was significantly correlated with the number of individuals surveyed per population and with the number of populations sampled in a given study. These patterns are well established in the literature (e.g. Pons & Petit 1995; Kalinowski 2004). We contend that many studies that have utilized mtDNA or cpDNA markers to establish source: destination patterns involving NIS have not utilized precedents in population genetics and community ecology with respect to the importance of sampling design. Many
of the former studies failed to provide a rationale for their choice of sampling design, nor do they discuss the limitations of the sampling strategy employed.

Two sources of sampling error may influence phylogeographical reconstructions of invasions: (i) insufficient numbers of individuals sampled in putative source populations; and (ii) an incomplete list of putative source populations surveyed. Our simulations suggest that two erroneous conclusions could be drawn as a result of these problems. First, investigators could fail to identify the proper source of an invasion. Second, the investigator could overestimate the number of sources required to account for the genetic diversity observed in an introduced population. The first type of error has implications for ecological and evolutionary studies of NIS. Ecological explanations of invasion success (e.g. enemy release hypothesis) require the identification of the correct source population(s) such that appropriate contrasts can be made (Colautti et al. 2004). Similarly, studies pertaining to evolutionary changes during invasion require that contrasts be drawn between the introduced and source populations; the latter group provides the baseline against which changes in traits in the introduced population are assessed (Wares et al. 2005). If incorrect invasion sources are identified, then inappropriate contrast groups may be chosen, leading to false inferences about ecological or evolutionary change resulting from the invasion event. The second type of error — overestimating the number of source populations that contributed to an introduced population — could impact management decisions. Multiple independent invasions inferred from genetic studies suggest that multiple invasion pathways exist and/or that vector strength is high (Forsyth & Duncan 2001; Colautti et al. 2006). Acceptance of these inferences by managers could erroneously inflate the relative importance of particular vectors or pathways over others. Accurate identification of the number of invasion sources and the identity of those sources are critical to management decisions involving attempted extermination of introduced populations. For example, elimination of introduced pests requires information both on the number of individuals being introduced and on their sources (Rollins et al. 2006). Similarly, cost-effective programmes to prevent new pest invasions require correct targeting of the source(s) of pest propagules.

A tell-tale sign of sampling inadequacy is the occurrence of numerous haplotypes in introduced populations that are not observed in any of the censused source populations (e.g. Stepien et al. 2005; Voisin et al. 2005). While individual haplotypes could arise de novo in the introduced population, the occurrence of a series of these seemingly novel forms suggests that source populations were insufficiently sampled. Avoidance of these problems will require that investigators sample a large number of individuals per population in the native range. If the geographical scope of sampling is limited, then care must be taken to ensure that key populations are identified and sampled in order to maximize the likelihood of sampling the genuine source. This objective could be met by considering geographical patterns of functional vectors of NIS dispersal. The completeness with which an individual population has been surveyed cannot be determined until after haplotype richness has been determined in the laboratory. We recommend that investigators utilize Monte Carlo sampling procedures to assess cumulative haplotype diversity as a function of number of individuals assayed. If the resulting relationship appears linear or only mildly asymptotic (i.e. new haplotypes are still being added with the last sample), then further sampling is warranted.

Many studies of genetic diversity are initiated following invasion of a local area by a NIS. Given the ease in sampling locally and the difficulty in collecting and transporting samples from distant sites, our finding that fewer individuals were typically sampled from source populations was not unexpected (Fig. 2a, b). Similarly, it was not surprising that fewer populations were typically sampled in the introduced than native range (Fig. 2c, d), perhaps owing to the time-dependency of the number of introduced populations available for sampling.

We noted that easily sampled taxa were sampled with the same intensity as those more difficult to capture and/or transport. These results were contrary to our expectation that abundant and easily sampled groups — such as aquatic invertebrates and insects — would be sampled more thoroughly than taxa that were more difficult to sample, such as fishes. Also, one could hypothesize that the number of individuals surveyed might reflect the cost of DNA sequencing, rather than being a product of sampling limitations. High cost could potentially limit the total number of individuals sequenced for a particular study, resulting in a compromise between the number of individuals sampled per population and the number of populations assessed (e.g. Slade & Moritz 1998). There is some evidence for this compromise, as we noted an inverse relationship between number of assayed individuals per population and the number of populations assessed (Spearman rank correlations: native range: \( r = -0.46, P = 0.001 \); introduced range: \( r = -0.31, P = 0.029 \)). However, the number of individuals sampled per population and the number of populations surveyed has not increased (Fig. 3) as sequencing cost has dropped over time, suggesting that sample size is not dictated solely by cost and that other factors may influence investigators’ choices.

In summary, while genetic markers provide invasion ecologists with an opportunity to identify the source of an introduced population, care must be taken when designing sampling strategies. Many published studies sample low numbers of individuals in source populations, or sample only a subset of possible source populations. These design limitations may result in erroneous assignment of
introduced individuals to source populations, which, in turn, may result in erroneous ecological or evolutionary comparisons or incorrect management decisions. Our literature analysis and simulations identify the magnitude of sampling issues in invasion studies, and quantify the effect of sampling limitations on potential errors in identification of the true source populations. Furthermore, our results highlight the need for an adaptive analytical approach, where preliminary genetic data should be used to determine the number of individuals sampled per source population and the number of putative source populations sampled.

Acknowledgements

We are grateful to Rob Colautti, Dr Melanie Cristescu, and anonymous reviewers for valuable comments that improved the manuscript. D. Gray, J. Muirhead and D. Kelly were, respectively, supported by NSERC graduate, Ontario Graduate, and GLIER postdoctoral fellowships. This study was supported by NSERC Discovery grants to DDH and HJM, by the NSERC Canadian Aquatic Invasive Species Network (HJM), by a Conservation Genetics Canada Research Chair to DDH and by a DFO Invasive Species Research Chair to HJM.

References


Jim Muirhead is a postdoctoral fellow in the Department of Biological Sciences at the University of Alberta; he is interested in conceptual and applied models of biological invasion. Derek Gray is a PhD student at Queen’s University, Kingston, ON; he is interested in the ecology and management of biological invasions. David Kelly is a postdoctoral researcher at Landcare Research, New Zealand; he is interested in biological invasion ecology and its study using molecular tools. Sandra Ellis is a researcher with the Department of Fisheries and Oceans Canada; she is interested in prevention and management of ship-mediated invasions. This work was completed while Muirhead (PhD student), Gray (MSc student), Kelly (postdoctoral fellow) and Ellis (MSc student) were resident at the Great Lakes Institute for Environmental Research (GLIER). Dan Heath is professor and Canada Research Chair in Conservation Genetics at GLIER; he is interested in fish reproduction, evolution, and conservation, though he has published extensively on genetics of invading species. Hugh MacIsaac is professor and DFO Invasive Species Research Chair at GLIER; he is Director of the Canadian Aquatic Invasive Species Network, and is interested in vectors of biological invasions.
Appendix 1

Published studies that used mt and cpDNA sequence variation to identify the source of an introduced species, arranged phylogenetically, and including information on additional molecular markers used during the study. DAMD, directed amplification of minisatellite DNA; AFLP, amplified fragment length polymorphisms; RAPD, randomly amplified polymorphic DNA; ISSR, intersimple sequence repeats.

<table>
<thead>
<tr>
<th>Study</th>
<th>Group (Class, Order)</th>
<th>Additional markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turon et al. 2003</td>
<td>Ascidiacea, Enterogona</td>
<td></td>
</tr>
<tr>
<td>Castilla et al. 2002</td>
<td>Ascidiacea, Pleurogona</td>
<td></td>
</tr>
<tr>
<td>López-Legentil et al. 2006</td>
<td>Ascidiacea, Pleurogona</td>
<td></td>
</tr>
<tr>
<td>O Foighil et al. 1998</td>
<td>Bivalvia, Ostreoida</td>
<td></td>
</tr>
<tr>
<td>May et al. 2006</td>
<td>Bivalvia, Veneroida</td>
<td></td>
</tr>
<tr>
<td>Therriault et al. 2002</td>
<td>Branchiopoda, Cladocera</td>
<td>DAMD</td>
</tr>
<tr>
<td>Cristescu et al. 2001</td>
<td>Branchiopoda, Cladocera</td>
<td></td>
</tr>
<tr>
<td>Martel et al. 2004</td>
<td>Gastropoda, Neogastropoda</td>
<td>Allozymes</td>
</tr>
<tr>
<td>Facon et al. 2003</td>
<td>Gastropoda, Neotaenioglossa</td>
<td></td>
</tr>
<tr>
<td>Gennner et al. 2004</td>
<td>Gastropoda, Neotaenioglossa</td>
<td></td>
</tr>
<tr>
<td>Städler et al. 2005</td>
<td>Gastropoda, Neotaenioglossa</td>
<td></td>
</tr>
<tr>
<td>Bachelet et al. 2004</td>
<td>Gastropoda, Sorbecooncha</td>
<td></td>
</tr>
<tr>
<td>Cristescu et al. 2004</td>
<td>Malacostraca, Amphipoda</td>
<td></td>
</tr>
<tr>
<td>Kelly et al. 2006</td>
<td>Malacostraca, Amphipoda</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Gopurenko et al. 2003</td>
<td>Malacostraca, Decapoda</td>
<td></td>
</tr>
<tr>
<td>Hänfling et al. 2002</td>
<td>Malacostraca, Decapoda</td>
<td></td>
</tr>
<tr>
<td>Peterson 2006</td>
<td>Malacostraca, Decapoda</td>
<td></td>
</tr>
<tr>
<td>Zardus &amp; Hadfield 2005</td>
<td>Maxillopoda, Sessilia</td>
<td></td>
</tr>
<tr>
<td>Roy &amp; Sponer 2002</td>
<td>Ophiuroidea, Ophiurida</td>
<td></td>
</tr>
<tr>
<td>Bastrop et al. 1998</td>
<td>Polychaeta, Canaliculipata</td>
<td></td>
</tr>
<tr>
<td>Holland et al. 2004</td>
<td>Scyphozoa, Rhizostomeae</td>
<td></td>
</tr>
</tbody>
</table>

Insects

<table>
<thead>
<tr>
<th>Study</th>
<th>Group (Class, Order)</th>
<th>Additional markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grapputo et al. 2005</td>
<td>Insecta, Coleoptera</td>
<td>AFLPs</td>
</tr>
<tr>
<td>Cognato et al. 2005</td>
<td>Insecta, Coleoptera</td>
<td></td>
</tr>
<tr>
<td>Evans et al. 2000</td>
<td>Insecta, Coleoptera</td>
<td></td>
</tr>
<tr>
<td>Laffin et al. 2005</td>
<td>Insecta, Coleoptera</td>
<td></td>
</tr>
<tr>
<td>Fonseca et al. 2001</td>
<td>Insecta, Diptera</td>
<td>RAPDs</td>
</tr>
<tr>
<td>Johnson et al. 2004</td>
<td>Insecta, Diptera</td>
<td></td>
</tr>
<tr>
<td>Nardi et al. 2005</td>
<td>Insecta, Diptera</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Scheffer 2000</td>
<td>Insecta, Diptera</td>
<td></td>
</tr>
<tr>
<td>Scheffer &amp; Lewis 2006</td>
<td>Insecta, Diptera</td>
<td></td>
</tr>
<tr>
<td>Silva et al. 2003</td>
<td>Insecta, Diptera</td>
<td></td>
</tr>
<tr>
<td>Chu et al. 2006</td>
<td>Insecta, Hemiptera</td>
<td></td>
</tr>
<tr>
<td>Downie 2002</td>
<td>Insecta, Hemiptera</td>
<td></td>
</tr>
<tr>
<td>Gwiazdowski et al. 2006</td>
<td>Insecta, Hemiptera</td>
<td></td>
</tr>
<tr>
<td>Havill et al. 2006</td>
<td>Insecta, Hemiptera</td>
<td></td>
</tr>
<tr>
<td>Smith 2005</td>
<td>Insecta, Homoptera</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Hufbauer et al. 2004</td>
<td>Insecta, Hymenoptera</td>
<td></td>
</tr>
<tr>
<td>Scheffer &amp; Grissell 2003</td>
<td>Insecta, Hymenoptera</td>
<td></td>
</tr>
<tr>
<td>Steiner et al. 2006</td>
<td>Insecta, Hymenoptera</td>
<td></td>
</tr>
<tr>
<td>Tsutsui et al. 2001</td>
<td>Insecta, Hymenoptera</td>
<td></td>
</tr>
<tr>
<td>Austin et al. 2006</td>
<td>Insecta, Isoptera</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Gomi et al. 2004</td>
<td>Insecta, Lepidoptera</td>
<td></td>
</tr>
</tbody>
</table>

Fishes

<table>
<thead>
<tr>
<th>Study</th>
<th>Group (Class, Order)</th>
<th>Additional markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collins et al. 2002</td>
<td>Actinopterygii, Anguilliformes</td>
<td></td>
</tr>
<tr>
<td>Moyer et al. 2005</td>
<td>Actinopterygii, Cypriniformes</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Lindholm et al. 2005</td>
<td>Actinopterygii, Cyprinodontiformes</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Azzurro et al. 2006</td>
<td>Actinopterygii, Perciformes</td>
<td></td>
</tr>
<tr>
<td>Stepiean et al. 1998</td>
<td>Actinopterygii, Perciformes</td>
<td></td>
</tr>
<tr>
<td>Stepiean et al. 2003</td>
<td>Actinopterygii, Perciformes</td>
<td>Nuclear introns</td>
</tr>
<tr>
<td>Dougherty et al. 1996</td>
<td>Actinopterygii, Perciformes</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix 1  Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Group (Class, Order)</th>
<th>Additional markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slade &amp; Moritz 1998</td>
<td>Tetrapods</td>
<td></td>
</tr>
<tr>
<td>Muñoz-Fuentes et al. 2006</td>
<td>Mammalia, Diprotodontia</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Le Page et al. 2000</td>
<td>Mammalia, Primates</td>
<td></td>
</tr>
<tr>
<td>Eldridge et al. 2001</td>
<td>Mammalia, Rodentia</td>
<td></td>
</tr>
<tr>
<td>Modolo et al. 2005</td>
<td>Mammalia, Rodentia</td>
<td></td>
</tr>
<tr>
<td>Hingston et al. 2005</td>
<td>Mammalia, Rodentia</td>
<td></td>
</tr>
<tr>
<td>Libois et al. 2001</td>
<td>Mammalia, Diprotodontia</td>
<td></td>
</tr>
<tr>
<td>Kolbe et al. 2004</td>
<td>Reptilia, Squamata</td>
<td></td>
</tr>
<tr>
<td>Oliverio et al. 2003</td>
<td>Reptilia, Squamata</td>
<td></td>
</tr>
<tr>
<td>Meusnier et al. 2002</td>
<td>Plants and algae</td>
<td></td>
</tr>
<tr>
<td>Goolsby et al. 2006</td>
<td></td>
<td>Nuclear ribosomal DNA</td>
</tr>
<tr>
<td>Kurokawa et al. 2004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Williams et al. 2005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim et al. 2004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plants and algae</td>
<td>Chlorophyceae, Bryopsidales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Filicopsida, Polypodiales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Magnoliopsida, Malvales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Magnoliopsida, Sapindales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rhodophyceae, Ceramiales</td>
<td></td>
</tr>
</tbody>
</table>

### Appendix 2

*Gammarus tigrinus* collection sites in North America and Europe with the number of individuals sequenced per population for the COI gene (see Kelly et al. 2006a)

<table>
<thead>
<tr>
<th>Population type</th>
<th>Location</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native populations</td>
<td>Delaware Estuary, Deemers beach, Delaware</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Elizabeth Estuary, canal locks, Virginia</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Hudson Estuary, New York</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Miramichi Estuary, New Brunswick</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Pawcatuck Estuary, Rhode Island</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Poropotank Estuary, Virginia</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Berry’s Creek, New Hampshire</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>St. John Estuary, New Brunswick</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>St. Lawrence Estuary, Montmagny, Quebec</td>
<td>25</td>
</tr>
<tr>
<td>Native populations with shared haplotypes</td>
<td>Delaware and Poropotank Estuaries</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Delaware and Elizabeth Estuaries</td>
<td>14</td>
</tr>
<tr>
<td>Introduced populations</td>
<td>Hudson and St. Lawrence Estuaries</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>White Sea, Deemers beach, Delaware</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Baltic Sea, Turku, Finland</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Baltic Sea, Anleger Lagoon, Germany</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Baltic Sea, Dierhagen Lagoon, Germany</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Baltic Sea, Vistula Lagoon, Poland</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Salwarpe River, England</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Elbe River, Germany</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mittellandkanal, Germany</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Rhine River, Germany</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Mosel, near Zell, Germany</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Lake Gouwzee, Netherlands</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>R. Waal, Rhine, Netherlands</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Oder River, Brody, Poland</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Oder River, Bytom, Poland</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Poulter River, England</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Barn River, Northern Ireland</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Worra River, Germany</td>
<td>10</td>
</tr>
</tbody>
</table>