Molecular Ecology (2008) **17**, 1020–1035 doi: 10.1111/j.1365-294X.2008.03669.x

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

Correspondence: Dr Hugh Maclsaac, Fax: 519-971-3616;
E-mail: hughm@uwindsor.ca

provide information on the history of invasion and human transport (Collins *et al*. 2002; Downie 2002; Cognato *et al*. 2005; Goolsby *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

Table 1 Descriptors used to search the ISI Web of Knowledge database for studies employing mt and cpDNA to trace the origins of NIS. Descriptors were entered in the Title/Keyword/Abstract fields and searches were run in the following combinations: A with C and D, B with C and D

Descriptors			
A	B	C	D
Biological invasions	Alien species	Mitochondrial DNA	Phylogenetics
Invasions	Nonindigenous species	Molecular marker	Phylogeography
Introduction	Introduced species	COI	Invasion genetics
	Invasive species	Cytochrome *b*	Source
		Chloroplast DNA*	
		CpDNA	
		*Rbc*L	

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

than range expansions and invasions in evolutionary time. We excluded studies that used length polymorphisms as opposed to sequence data to ensure that recovered studies had comparable levels of phylogenetic resolution (Wilson *et al*. 1989).

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 PHYLOGR (Díaz-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:

$$c_i = \left(1 - \frac{\min d_{ij}}{\frac{1}{m}\sum_{j=1}^{m} d_{ij}} \right),$$

where $\min d_{ij}$ 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).

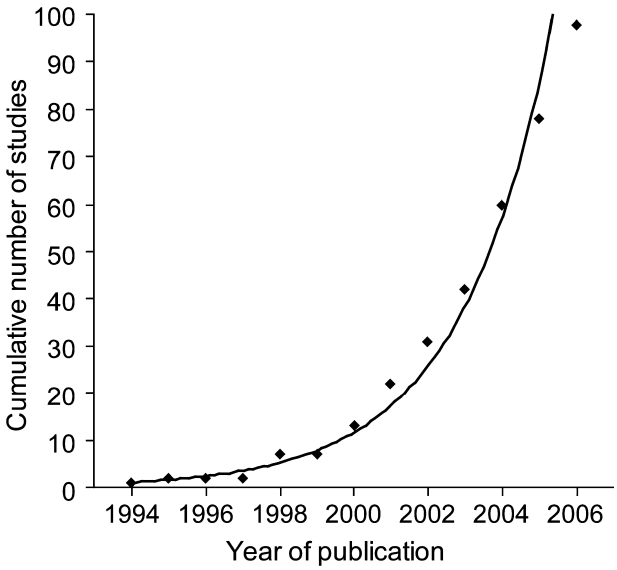

Fig. 1 Cumulative number of studies published between 1994 and 2006 that have used mtDNA and cpDNA sequence variation to locate the source of an introduced species. The line was fit using an exponential function ($y = 0.708e^{0.399x}$; $r^2 = 0.98$).

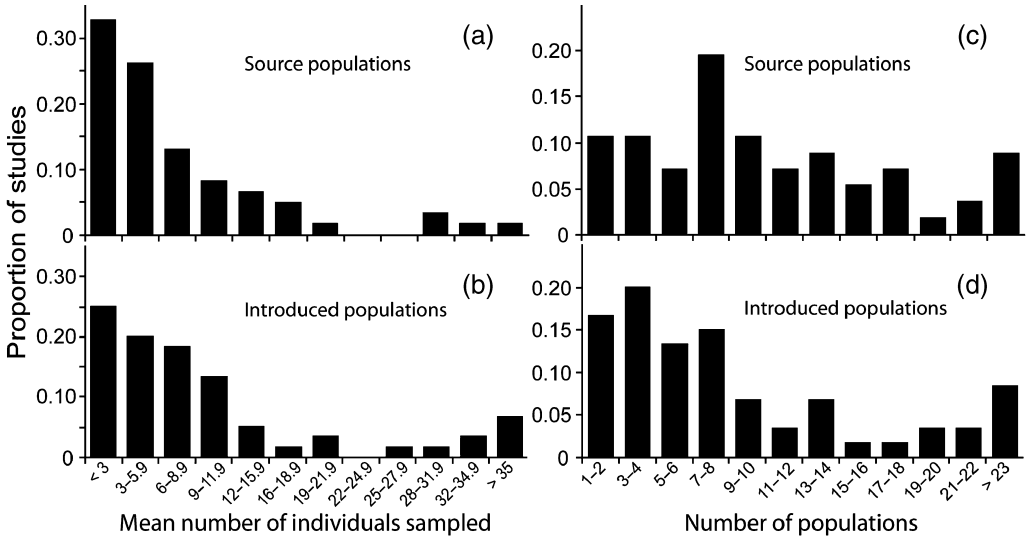

Fig. 2 Mean number of individuals assayed per population (a, b), and number of populations assayed (c, d) for published studies. A and C, source populations; B and D, introduced populations.

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

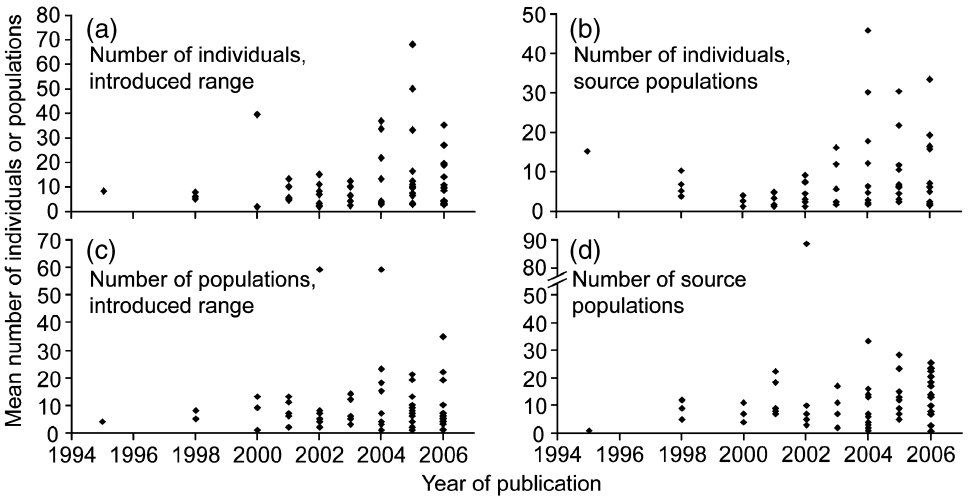

Fig. 3 Mean number of individuals surveyed per population (a, b) and total number of populations surveyed (c, d) by year of studies publication. A and C, introduced populations; B and D, source populations. Linear regressions of number of individuals surveyed per population and number of populations surveyed as a function of study date failed to reveal any significant relationships ($P > 0.10$). In panel B, six and seven studies were published during 2005 and 2006, respectively, with sample sizes of = 6 individuals per source population.

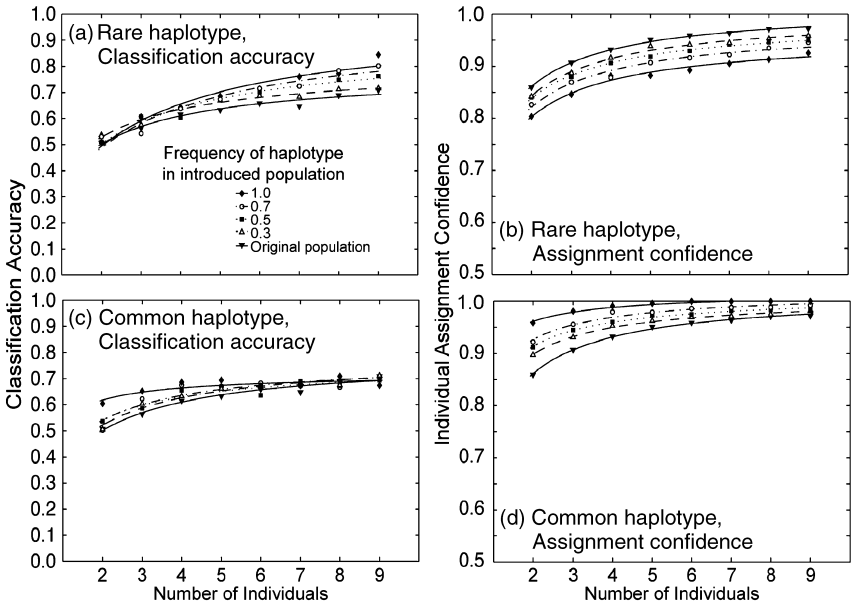

Fig. 4 Relationship between classification accuracy (a,c) and confidence of individual assignment (b,d) of an introduced individual in relation to the number of individuals surveyed per source population (two to nine individuals). Classification accuracy based upon field collections are indicated by original population curves. Analyses were conducted for both rare (a,b) and common (c,d) haplotypes from the source population (Delaware estuary). The frequency of these haplotypes in the introduced population was allowed to vary from 30 to 100%. Curves were fit using the Michaelis–Menten function.

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

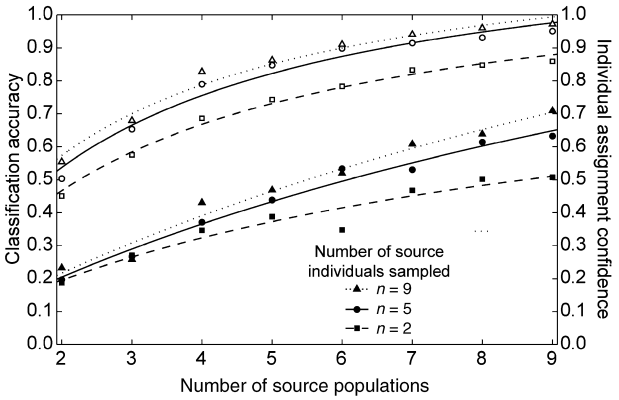

Fig. 5 Classification accuracy of a simulated introduced population and confidence of individual assignment with different values for the number of individuals surveyed per population and the number of populations surveyed. Solid symbols indicated classification accuracy of the simulated population, and outlined symbols indicate confidence of individual assignment. Curves fit as per Fig. 4.

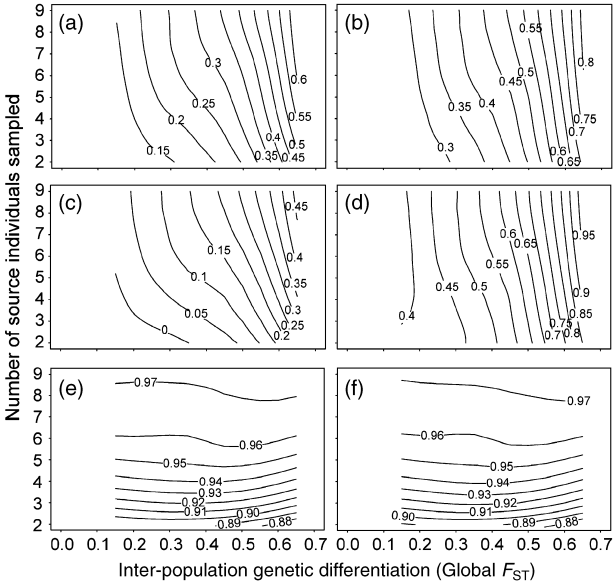

Fig. 6 The effect of the number of individuals surveyed per populations and interpopulation genetic variation (Global F_{ST}) of putative source populations on classification accuracy for introduced individuals. Panels A–C show mean, mean plus and mean minus 1 SD classification accuracy for simulations where the introduced population is matched back to specific source populations. Panel D is the mean classification accuracy of simulations where the introduced population is matched at the regional level. Panels E and F show the confidence index for classifying individuals (eqn 1) for the local and regional level of haplotype matching, respectively. Contours were fit using LOESS smoothing splines.

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 inter-population 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 (Ruiz & 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 difficulty 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 Melania 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

Austin, JW, Szalanski, AL, Scheffrahn, RH *et al*. (2006) Genetic evidence for two introductions of the Formosan Subterranean Termite, *Coptotermes formosanus* (Isoptera: Rhinotermitidae), to the United States. *Florida Entomologist*, **89**, 183–193.

Avise JC (2000) *Phylogeography: the History and Formation of Species*. Harvard University Press, Massachusetts.

Avise JC, Arnold J, Ball RM *et al*. (1987) Intraspecific phylogeography – the mitochondrial-DNA bridge between population-genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Azzurro, E, Golani, D, Bucciarelli, G & Bernardi, G (2006) Genetics of the early stages of invasion of Lessepsian rabbitfish *Siganus luridus*. *Journal of Experimental Marine Biology and Ecology*, **333**, 190–201.

Bachelet G, Simon-Bouhet B, Desclaux C, *et al*. (2004) Invasion of the eastern Bay of Biscay by the nassariid gastropod *Cyclope neritea*: origin and effects on resident fauna. *Marine Ecology Progress Series*, **276**, 147–159.

Bastrop R, Jurss K & Sturmbauer C (1998) Cryptic species in a marine polychaete and their independent introduction from North America to Europe. *Molecular Biology and Evolution*, **15**, 97–103.

Bates D, Maechler M (2007) *Matrix: A Matrix package for R. Pages R Package version 0.999375–3*. URL: http://cran.r-project.org/src/contrib/Descriptions/Matrix.html

Bazin E, Glémin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**, 570–572.

Bossart JL, Prowell DP (1998) Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology & Evolution*, **13**, 202–206.

B-Rao C (2001) Sample size considerations in genetic polymorphism studies. *Human Heredity*, **52**, 191–200.

Castilla JC, Collins AG, Meyer CP, Guinez R, Lindberg DR (2002) Recent introduction of the dominant tunicate, *Pyura praeputialis* (Urochordata, Pyuridae) to Antofagasta, Chile. *Molecular Ecology*, **11**, 1579–1584.

Chu D, Zhang YJ, Brown JK *et al*. (2006) The introduction of the exotic Q biotype of *Bemisia tabaci* from the Mediterranean region into China on ornamental crops. *Florida Entomologist*, **89**, 168–174.

Cognato AI, Sun JH, Anducho-Reyes MA, Owen DR (2005) Genetic variation and origin of red turpentine beetle (*Dendroctonus valens* LeConte) introduced to the People's Republic of China. *Agricultural and Forest Entomology*, **7**, 87–94.

Colautti RI, Ricciardi A, Grigorovich IA, MacIsaac HJ (2004) Is invasion success explained by the enemy release hypothesis? *Ecology Letters*, **7**, 721–733.

Colautti RI, Grigorovich IA, MacIsaac HJ (2006) Propagule pressure: a null hypothesis for biological invasions. *Biological Invasions*, **8**, 1023–1037.

Collins TM, Trexler JC, Nico LG, Rawlings TA (2002) Genetic diversity in a morphologically conservative invasive taxon: multiple introductions of swamp eels to the southeastern United States. *Conservation Biology*, **16**, 1024–1035.

Colwell RK, Mao CX, Chang J (2004) Interpolative, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2717–2727.

Cristescu MEA, Hebert PDN, Witt JDS, MacIsaac HJ, Grigrovich IA (2001) An invasion history for *Cercopagis pengoi* based on mitochondrial gene sequences. *Limnology and Oceanography*, **46**, 224–229.

Cristescu MEA, Witt JDS, Grigorovich IA, Hebert PDN & MacIsaac HJ (2004) Dispersal of the Ponto-Caspian amphipod *Echinogammarus ischnus*: invasion waves from the Pleistocene to the present. *Heredity*, **92**, 197–203.

Crossa J (1989) Methodologies for estimating the sample-size required for genetic conservation of outbreeding crops. *Theoretical and Applied Genetics*, **77**, 153–161.

Cruzan MB (1998) Genetic markers in plant evolutionary ecology. *Ecology*, **79**, 400–412.

Davies N, Villablanca FX, Roderick GK (1999) Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. *Trends in Ecology & Evolution*, **14**, 17–21.

Díaz-Uriarte R, Garland T Jr (2007) PHYLOGR: Functions for phylogenetically based statistical analyses. *R Package Version 1.0.6*. URL: http://cged.genes.nig.ac.jp/RGM2/pkg.php?p=PHYLOGR

Dixon CJ (2006) A means of estimating the completeness of haplotype sampling using the Stirling probability distribution. *Molecular Ecology Notes*, **6**, 650–652.

Dougherty JD, Moore WS & Ram JL (1996) Mitochondrial DNA analysis of round goby (*Neogobius melanostomus*) and tubenose goby (*Proterorhinus marmoratus*) in the Great Lakes basin. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 474–480.

Downie DA (2002) Locating the sources of an invasive pest, grape phylloxera, using a mitochondrial DNA gene genealogy. *Molecular Ecology*, **11**, 2013–2026.

Drake JM, Lodge DM (2004) Global hot spots of biological invasions: evaluating options for ballast water management. *Philosophical Transactions of the Royal Society of London. Series B Biological Sciences*, **271**, 574–580.

Eldridge MDB, Browning TL & Close RL (2001) Provenance of a New Zealand brush-tailed rock-wallaby (*Petrogale penicillata*) population determined by mitochondrial DNA sequence analysis. *Molecular Ecology*, **10**, 2561–2567.

Epifanio JM, Smouse PE, Kobak CJ, Brown BL (1995) Mitochondrial-DNA divergence among populations of American shad (*Alosa sapidissima*) – how much variation is enough for mixed-stock analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**, 1688–1702.

Evans JD, Pettis JS & Shimanuki H (2000) Mitochondrial DNA relationships in an emergent pest of honey bees: *Aethina tumida* (Coleoptera: Nitidulidae) from the United States and Africa. *Annals of the Entomological Society of America*, **93**, 415–420.

Facon B, Pointier JP, Glaubrecht M, Poux C, Jarne P & David P (2003) A molecular phylogeography approach to biological invasions of the New World by parthenogenetic Thiarid snails. *Molecular Ecology*, **12**, 3027–3039.

Fofonoff PW, Ruiz GM, Steves B, Carlton JT (2003) In ships or on ships? Mechanisms of transfer and invasion for non-native species to the coasts of North America. In: *Invasive Species: Vectors and Managements Strategies* (eds Ruiz GM, Carlton JT), pp. 152–182. Island Press, Washington, D.C.

Fonseca DM, Campbell S, Crans WJ *et al*. (2001) *Aedes* (*Finlaya*) *japonicus* (Diptera: Culicidae), a newly recognized mosquito in the United States: analyses of genetic variation in the United States and putative source populations. *Journal of Medical Entomology*, **38**, 135–146.

Forsyth DM, Duncan RP (2001) Propagule size and the relative success of exotic ungulate and bird introductions to New Zealand. *American Naturalist*, **157**, 583–595.

Gaskin JF, Zhang D-Y, Bon M-C (2005) Invasion of *Lepidium draba* (Brassicaceae) in the western United States: distributions and origins of chloroplast DNA haplotypes. *Molecular Ecology*, **14**, 2331–2341.

Genner MJ, Michel E, Erpenbeck D, De Voogd N, Witte F & Pointier JP (2004) Camouflaged invasion of Lake Malawi by an Oriental gastropod. *Molecular Ecology*, **13**, 2135–2141.

Gomi T, Muraji M & Takeda M (2004) Mitochondrial DNA analysis of the introduced fall webworm, showing its shift in lifecycle in Japan. *Entomological Science*, **7**, 183–188.

Goolsby JA, De Barro PJ, Makinson JR *et al*. (2006) Matching the origin of an invasive weed for selection of a herbivore haplotype for a biological control programme. *Molecular Ecology*, **15**, 287–297.

Gopurenko D, Hughes JM, Bellchambers L (2003) Colonisation of the south-west Australian coastline by mud crabs: evidence for a recent range expansion or human-induced translocation? *Marine and Freshwater Research*, **54**, 833–840.

Grapputo A, Boman S, Lindstrom L, Lyytinen A, Mappes J (2005) The voyage of an invasive species across continents: genetic diversity of North American and European Colorado potato beetle populations. *Molecular Ecology*, **14**, 4207–4219.

Gwiazdowski RA, Van Driesche RG, Desnoyers A *et al*. (2006) Possible geographic origin of beech scale, *Cryptococcus fagisuga* (Hemiptera: Eriococcidae), an invasive pest in North America. *Biological Control*, **39**, 9–18.

Hamilton MB (1999) Four primers for the amplification of chloroplast intragenic regions with intraspecific variation. *Molecular Ecology*, **8**, 521–523.

Hänfling B, Carvalho GR, Brandl R (2002) mt-DNA sequences and possible invasion pathways of the Chinese mitten crab. *Marine Ecology Progress Series*, **238**, 307–310.

Havill, NP, Montgomery, ME, Yu, G, Shiyake, S & Caccone, A (2006) Mitochondrial DNA from Hemlock Woolly Adelgid (Hemiptera: Adelgidae) suggests cryptic speciation and pinpoints the source of the introduction to eastern North America. *Annals of the Entomological Society of America*, **99**, 195–203.

Hedin M, Wood DA (2002) Genealogical exclusivity in geographically proximate populations of *Hypochilus thorelli* Marx (Araneae, Hypochilidae) on the Cumberland Plateau of North America. *Molecular Ecology*, **11**, 1975–1988.

Hierro JL, Maron JL, Callaway RM (2005) A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *Journal of Animal Ecology*, **93**, 5–15.

Hingston M, Goodman SM, Ganzhorn JU & Sommer S (2005) Reconstruction of the colonization of southern Madagascar by introduced *Rattus rattus*. *Journal of Biogeography*, **32**, 1549–1559.

Holland BS, Dawson MN, Crow GL & Hofmann DK (2004) Global phylogeography of *Cassiopea* (Scyphozoa: Rhizostomeae): molecular evidence for cryptic species and multiple invasions of the Hawaiian Islands. *Marine Biology*, **145**, 1119–1128.

Hortal J, Borges PAV, Gaspar C (2006) Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology*, **75**, 274–287.

Hufbauer RA, Bogdanowicz SM & Harrison RG (2004) The population genetics of a biological control introduction: mitochondrial DNA and microsatellite variation in native and introduced populations of *Aphidus ervi*, a parasitoid wasp. *Molecular Ecology*, **13**, 337–348.

Johnson AJ, Schemerhorn BJ & Shukle RH (2004) A first assessment of mitochondrial DNA variation and geographic distribution of haplotypes in hessian fly (Diptera: Cecidomyiidae). *Annals of the Entomological Society of America*, **97**, 940–948.

Kalinowski ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conservation Genetics*, **5**, 539–543.

Kelly DW, Muirhead JR, Heath DD, MacIsaac HJ (2006a) Contrasting patterns in genetic diversity following multiple invasions of fresh and brackish waters. *Molecular Ecology*, **15**, 3641–3653.

Kelly DW, MacIsaac HJ, Heath DD (2006b) Vicariance and dispersal effects on phylogeographic structure and speciation in a widespread estuarine invertebrate. *Evolution*, **60**, 257–267.

Kim MS, Yang EC, Mansilla A & Boo SM (2004) Recent introduction of *Polysiphonia morrowii* (Ceramiales, Rhodophyta) to Punta Arenas, Chile. *Botanica Marina*, **47**, 389–394.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Kolar CS, Lodge DM (2002) Ecological predictions and risk assessment for alien fishes in North America. *Science*, **298**, 1233–1236.

Kolbe JJ, Glor RE, Schettino LRG *et al*. (2004) Genetic variation increases during biological invasion by a Cuban lizard. *Nature*, **431**, 177–181.

Kurokawa S, Shibaike H, Akiyama H & Yoshimura Y (2004) Molecular and morphological differentiation between the crop and weedy types in velvetleaf (*Abutilon theophrasti* Medik) using a chloroplast DNA marker: seed source of the present invasive velvetleaf in Japan. *Heredity*, **93**, 603–609.

Laffin RD, Dosdall LM & Sperling FAH (2005) Population structure of the cabbage seedpod weevil, *Ceutorhynchus obstrictus* (Marsham) (Coleoptera Curculionidae): origins of North American introductions. *Environmental Entomology*, **34**, 504–510.

Latch EK, Rhodes OE Jr (2006) Evidence for bias in estimates of local genetic structure due to sampling scheme. *Animal Conservation*, **9**, 308–315.

Le Page SL, Livermore RA, Cooper DW & Taylor AC (2000) Genetic analysis of a documented population bottleneck: introduced Bennett's wallabies (*Macropus rufogriseus rufogriseus*) in New Zealand. *Molecular Ecology*, **9**, 753–763.

Libois RM, Michaux JR, Ramalhinho MG, Maurois C & Sara M (2001) On the origin and systematics of the northern African wood mouse (*Apodemus sylvaticus*) populations: a comparative study of mtDNA restriction patterns. *Canadian Journal of Zoology*, **79**, 1503–1511.

Lindholm AK, Breden F, Alexander HJ *et al*. (2005) Invasion success and genetic diversity of introduced populations of guppies *Poecilia reticulata*. Australia. *Molecular Ecology*, **14**, 3671–3682.

Lodge DM, Williams SL, MacIsaac HJ *et al*. (2006) Biological invasions: recommendations for U.S. policy and management. *Ecological Applications*, **16**, 2035–2054.

López-Legentil S, Turon X & Planes S (2006) Genetic structure of the star sea squirt, *Botryllus schlosseri*, introduced in southern European harbours. *Molecular Ecology*, **15**, 3957–3967.

Lynch M, Crease TJ (1990) The analysis of population survey data on DNA-sequence variation. *Molecular Biology and Evolution*, **7**, 377–394.

Mack RN, Simberloff D, Lonsdale WM *et al*. (2000) Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Applications*, **10**, 689–710.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.

Martel C, Viard F, Bourguet D & Garcia-Meunier P (2004) Invasion by the marine gastropod *Ocinebrellus inornatus* in France I. Scenario for the source of introduction. *Journal of Experimental Marine Biology and Ecology*, **305**, 155–170.

May GE, Gelembiuk GW, Panov VE, Orlova MI & Lee CE (2006) Molecular ecology of zebra mussel invasions. *Molecular Ecology*, **15**, 1021–1031.

McIvor L, Maggs CA, Provan J, Stanhope MJ (2001) *rbc*L sequences reveal cryptic introductions of the Japanese red alga *Polysiphonia harveyi*. *Molecular Ecology*, **10**, 911–919.

Meusnier I, Valero M, Destombe C, *et al*. (2002) Polymerase chain reaction-single strand conformation polymorphism analyses of nuclear and chloroplast DNA provide evidence for recombination, multiple introductions and nascent speciation in the *Caulerpa taxifolia* complex. *Molecular Ecology*, **11**, 2317–2325.

Modolo L, Salzburger W & Martin RD (2005) Phylogeography of Barbary macaques (*Macaca sylvanus*) and the origin of the Gibraltar colony. *Proceedings of the National Academy of Sciences, USA*, **102**, 7392–7397.

Morando M, Avila LJ, Sites JW Jr (2003) Sampling strategies for delimiting species: genes, individuals, and populations in the *Liolaemus elongatus-kriegi* complex (Squamata: Liolaemidae) in Andean-Patagonian South America. *Systematic Biology*, **52**, 159–185.

Moyer GR, Osborne M & Turner TF (2005) Genetic and ecological dynamics of species replacement in an arid-land river system. *Molecular Ecology*, **14**, 1263–1273.

Muirhead JR, Ejsmont-Karabin J, MacIsaac HJ (2006) Quantifying rotifer species richness in temperate lakes. *Freshwater Biology*, **51**, 1696.

Muñoz-Fuentes, V, Green, AJ, Sorenson, MD, Negro, JJ & Vilà, C (2006) The ruddy duck *Oxyura jamaicensis* in Europe: natural colonization or human introduction? *Molecular Ecology*, **15**, 1441–1453.

Nardi F, Carapelli A, Dallai R, Roderick GK & Frati F (2005) Population structure and colonization history of the olive fly, *Bactrocera oleae* (Diptera, Tephritidae). *Molecular Ecology*, **14**, 2729–2738.

Neigel JE, Avise JC (1993) Application of a random walk model to geographic distributions of animal mitochondrial DNA variation. *Genetics*, **135**, 1209–1220.

O Foighil D, Gaffney PM, Wilbur AE & Hilbish TJ (1998) Mitochondrial cytochrome oxidase I gene sequences support an Asian origin for the Portuguese oyster *Crassostrea angulata*. *Marine Biology*, **131**, 497–503.

Oliverio M, Burke R, Bologna MA, Wirz A & Mariottini P (2001) Molecular characterization of native (Italy) and introduced (USA) *Podarcis sicula* populations (Reptilia, Lacertidae). *Italian Journal of Zoology*, **68**, 121–124.

Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene-mapping. *American Journal of Human Genetics*, **51**, 283–290.

Paradis E, Strimmer K, Claude J *et al*. (2005) *APE: Analyses of Phylogenetics and Evolution*. *R Package version 1.8*. URL: http://cran.r-project.org/src/contrib/Descriptions/ape.html

Peterson, C (2006) Range expansion in the northeast Pacific by an estuary mud crab – a molecular study. *Biological Invasions*, **8**, 565–576.

Pons O, Petit RJ (1995) Estimation, variance and optimal sampling of gene diversity.1. Haploid Locus. *Theoretical and Applied Genetics*, **90**, 462–470.

R Development Core Team (2007) *Royal: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from URL: http://www.R-project.org.

Roderick GK (2004) Tracing the origin of pests and natural enemies: genetic and statistical approaches. In: *Genetics, Evolution, and Biological Control* (eds Ehler LE, Sforza R, Mateille T), pp. 97–112. CABI Publishing, Cambridge, Massachusetts.

Rollins LA, Woolnough AP, Sherwin WB (2006) Population genetic tools for pest management: a review. *Wildlife Research*, **33**, 251–261.

Rouget M, Richardson DM (2003) Inferring process from pattern in plant invasions: a semimechanistic model incorporating propagule pressure and environmental factors. *American Naturalist*, **162**, 713–724.

Roy MS & Sponer R (2002) Evidence of a human-mediated invasion of the tropical western Atlantic by the 'world's most common brittlestar'. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **269**, 1017–1023.

Ruiz GM, Carlton JT (2003) Invasion vectors: a conceptual framework for management. In: *Invasive Species: Vectors and Managements Strategies* (eds Ruiz GM, Carlton JT), pp. 459–504. Island Press, Washington.

Ruiz GM, Carlton JT, Wonham MJ, Anson HH (2000) Invasion of coastal marine communities in North America: Apparent patterns, processes, and biases. *Annual Review of Ecology and Systematics*, **31**, 481–531.

Scheffer SJ (2000) Molecular evidence of cryptic species within the *Liriomyza huidobrensis* (Diptera: Agromyzidae). *Journal of Economic Entomology*, **93**, 1146–1151.

Scheffer SJ & Grissell EE (2003) Tracing the geographical origin of *Megastigmus transvaalensis* (Hymenoptera: Torymidae): an African wasp feeding on a South American plant in North America. *Molecular Ecology*, **12**, 415–421.

Scheffer SJ, Lewis ML (2006) Mitochondrial phylogeography of the vegetable pest *Liriomyza trifolii* (Diptera: Agromyzidae):

diverged clades and invasive populations. *Annals of the Entomological Society of America*, **99**, 991–998.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Silva JG, Meixner MD, McPheron BA, Steck GJ & Sheppard WS (2003) Recent mediterranean fruit fly (Diptera: Tephritidae) infestations in Florida – a genetic perspective. *Journal of Economic Entomology*, **96**, 1711–1718.

Simon C, Frati F, Beckenbach A *et al*. (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America*, **87**, 651–701.

Slade RW, Moritz C (1998) Phylogeography of *Bufo marinus* from its natural and introduced ranges. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 769–777.

Smith PT (2005) Mitochondrial DNA variation among populations of the glassy-winged sharpshooter, *Homalodisca coagulata*. *Journal of Insect Science*, **5**, 1–8.

Soltis DE, Soltis PS (1998) Choosing an approach and an appropriate gene for phylogenetic analysis. In: *Molecular Systematics of Plants II, DNA Sequencing* (eds Soltis DE, Soltis PS, Doyle JJ), pp. 1–42. Kluwer Academic Publishers, Boston.

Städler T, Frye M, Neiman M & Lively CM (2005) Mitochondrial haplotypes and the New Zealand origin of clonal European *Potamopyrgus*, an invasive aquatic snail. *Molecular Ecology*, **14**, 2465–2473.

Steiner FM, Schlick-Steiner BC, Trager JC *et al*. (2006) *Tetramorium tsushimae*, a new invasive ant in North America. *Biological Invasions*, **8**, 117–123.

Stepien CA, Dillon AK & Chandler MD (1998) Genetic identity, phylogeography, and systematics of ruffe *Gymnocephalus* in the North American Great lakes and Eurasia. *Journal of Great Lakes Research*, **24**, 361–378.

Stepien CA, Ford AM & Dillon-Klika AK (2003) Risk analysis and genetic identity of the Eurasian source population for the ruffe (*Gymnocephalus cernuus*) invasion in the Great Lakes. In: *Proceedings of Percis III, the 3rd International Symposium on Percid Fishes* (eds Barry TP & Malison JA), pp. 91–92. University of Wisconsin Sea Grant Institute, Madison, Wisconsin.

Stepien CA, Brown JE, Neilson ME *et al*. (2005) Genetic diversity of invasive species in the Great Lakes versus their Eurasian source populations: insights for risk analysis. *Risk Analysis*, **25**, 1043–1060.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology & Evolution*, **15**, 199–203.

Tatem AJ, Hay SI (2007) Climatic similarity and biological exchange in the worldwide transportation network. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1489–1496.

Templeton AR (2004) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, **13**, 789–809.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.

Therriault TW, Grigorovich IA, Cristescu ME *et al*. (2002) Taxonomic resolution of the genus *Bythotrephes* Leydig using molecular markers and re-evaluation of its global distribution. *Diversity and Distributions*, **8**, 67–84.

Tsutsui ND, Suarez AV, Holway DA & Case TJ (2001) Relationships among native and introduced populations of the Argentine ant (*Linepithema humile*) and the source of introduced populations. *Molecular Ecology*, **10**, 2151–61.

Turon X, Tarjuelo I, Duran S & Pascual M (2003) Characterising invasion processes with genetic data: an Atlantic clade of *Clavelina lepadiformis* (Ascidiacea) introduced into Mediterranean harbours. *Hydrobiologia*, **503**, 29–35.

Vasquez EA, Glenn EP, Brown JJ, Guntenspergen GR, Nelson SG (2005) Salt tolerance underlies the cryptic invasion of North American salt marshes by an introduced haplotype of the common reed *Phragmites australis* (Poaceae). *Marine Ecology Progress Series*, **298** (1–8), 2005.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York.

Verling E, Ruiz GM, Smith DL *et al*. (2005) Supply-side invasion ecology: characterizing propagule pressure in coastal ecosystems. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 1249–1256.

Voisin M, Engel CR, Viard F (2005) Differential shuffling of native genetic diversity across introduced regions in a brown alga: aquaculture vs. maritime traffic effects. *Proceedings of the National Academy of Sciences, USA*, **102**, 5432–5437.

Wares JP, Hughes AR, Grosberg RK (2005) Mechanisms that drive evolutionary change: insights from species introductions and invasions. In: *Species Invasions, Insights Into Ecology, Evolution, and Biogeography* (eds Sax DF, Stachowicz JJ, Gaines SD), pp. 229–257. Sinauer & Associates, Sunderland, Massachusetts.

Williams DA, Overholt WA, Cuda JP & Hughes CR (2005) Chloroplast and microsatellite DNA diversities reveal the introduction history of Brazilian peppertree (*Schinus terebinthifolius*) in Florida. *Molecular Ecology*, **14**, 3643–3656.

Wilson ACE, Zimmer EA, Prager EM, Kocher TD (1989) Restriction mapping in the molecular systematics of mammals: a retrospective salute. In: *The Hierarchy of Life* (eds Fernholm B, Bremer K, Jornvall H), pp. 407–419. Elsevier Press, Amsterdam, The Netherlands.

Zardus JD & Hadfield MG (2005) Multiple origins and incursions of the Atlantic barnacle *Chthamalus proteus* in the Pacific. *Molecular Ecology*, **14,** 3719–3733.

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 Maclsaac 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

Study	Group (Class, Order)	Additional markers
	Aquatic invertebrates	
Turon *et al*. 2003	Ascidiacea, Enterogona	
Castilla *et al*. 2002	Ascidiacea, Pleurogona	
López-Legentil *et al*. 2006	Ascidiacea, Pleurogona	
O Foighil *et al*. 1998	Bivalvia, Ostreoida	
May *et al*. 2006	Bivalvia, Veneroida	
Therriault *et al*. 2002	Branchiopoda, Cladocera	DAMD
Cristescu *et al*. 2001	Branchipoda, Cladocera	
Martel *et al*. 2004	Gastropoda, Neogastropoda	Allozymes
Facon *et al*. 2003	Gastropoda, Neotaenioglossa	
Genner *et al*. 2004	Gastropoda, Neotaenioglossa	
Städler *et al*. 2005	Gastropoda, Neotaenioglossa	
Bachelet *et al*. 2004	Gastropoda, Sorbeoconcha	
Cristescu *et al*. 2004	Malacostraca, Amphipoda	
Kelly *et al*. 2006	Malacostraca, Amphipoda	
Gopurenko *et al*. 2003	Malacostraca, Decapoda	Microsatellites
Hänfling *et al*. 2002	Malacostraca, Decapoda	
Peterson 2006	Malacostraca, Decapoda	
Zardus & Hadfield 2005	Maxillopoda, Sessilia	
Roy & Sponer 2002	Ophiuroidea, Ophiurida	
Bastrop *et al*. 1998	Polychaeta, Canalipalpata	
Holland *et al*. 2004	Scyphozoa, Rhizostomeae	
	Insects	
Grapputo *et al*. 2005	Insecta, Coleoptera	AFLPs
Cognato *et al*. 2005	Insecta, Coleoptera	
Evans *et al*. 2000	Insecta, Coleoptera	
Laffin *et al*. 2005	Insecta, Coleoptera	
Fonseca *et al*. 2001	Insecta, Diptera	RAPDs
Johnson *et al*. 2004	Insecta, Diptera	
Nardi *et al*. 2005	Insecta, Diptera	Microsatellites
Scheffer 2000	Insecta, Diptera	
Scheffer & Lewis 2006	Insecta, Diptera	
Silva *et al*. 2003	Insecta, Diptera	
Chu *et al*. 2006	Insecta, Hemiptera	
Downie 2002	Insecta, Hemiptera	
Gwiazdowski *et al*. 2006	Insecta, Hemiptera	
Havill *et al*. 2006	Insecta, Hemiptera	
Smith 2005	Insecta, Homoptera	
Hufbauer *et al*. 2004	Insecta, Hymenoptera	Microsatellites
Scheffer & Grissell 2003	Insecta, Hymenoptera	
Steiner *et al*. 2006	Insecta, Hymenoptera	
Tsutsui *et al*. 2001	Insecta, Hymenoptera	Microsatellites
Austin *et al*. 2006	Insecta, Isoptera	
Gomi *et al*. 2004	Insecta, Lepidoptera	
	Fishes	
Collins *et al*. 2002	Actinoptergyii, Angulliformes	
Moyer *et al*. 2005	Actinopterygii, Cypriniformes	Microsatellites
Lindholm *et al*. 2005	Actinoptergyii, Cyprinodontiformes	Microsatellites
Azzurro *et al*. 2006	Actinoptergyii, Perciformes	
Stepien *et al*. 1998	Actinoptergyii, Perciformes	
Stepien *et al*. 2003	Actinoptergyii, Perciformes	Nuclear introns
Dougherty *et al*. 1996	Actinopterygii, Perciformes	

Appendix 1 *Continued*

Study	Group (Class, Order)	Additional markers
	Tetrapods	
Slade & Moritz 1998	Amphibia, Anura	
Muñoz-Fuentes *et al*. 2006	Aves, Anseriformes	Microsatellites
Le Page *et al*. 2000	Mammalia, Diprotodontia	Microsatellites
Eldridge *et al*. 2001	Mammalia, Diprotodontia	
Modolo *et al*. 2005	Mammalia, Primates	
Hingston *et al*. 2005	Mammalia, Rodentia	
Libois *et al*. 2001	Mammalia, Rodentia	
Kolbe *et al*. 2004	Reptilia, Squamata	
Oliverio *et al*. 2001	Reptilia, Squamata	
	Plants and algae	
Meusnier *et al*. 2002	Chlorophyceae, Bryopsidales	Nuclear ribosomal DNA
Goolsby *et al*. 2006	Filicopsida, Polypodiales	
Kurokawa *et al*. 2004	Magnoliopsida, Malvales	
Williams *et al*. 2005	Magnoliopsida, Sapindales	Microsatellites
Kim *et al*. 2004	Rhodophyceae, Ceramiales	

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)

Population type	Location	*n*
Native populations	Delaware Estuary, Deemers beach, Delaware	30
	Elizabeth Estuary, canal locks, Virginia	40
	Hudson Estuary, New York	24
	Miramichi Estuary, New Brunswick	12
	Pawcatuck Estuary, Rhode Island	10
	Poropotank Estuary, Virginia	16
	Berry's Creek, New Hampshire	11
	St. John Estuary, New Brunswick	9
	St. Lawrence Estuary, Montmagny, Quebec	25
Native populations with shared haplotypes	Delaware and Poropotank Estuaries	15
	Delaware and Elizabeth Estuaries	14
	Hudson and St. Lawrence Estuaries	44
Introduced populations	Lough Neagh, Northern Ireland	10
	Baltic Sea, Turku, Finland	10
	Baltic Sea, Anleger Lagoon, Germany	10
	Baltic Sea, Dierhagen Lagoon, Germany	10
	Baltic Sea, Vistula Lagoon, Poland	10
	Salwarpe River, England	2
	Elbe River, Germany	4
	Mittlelandkanal, Germany	2
	Rhine River, Germany	2
	Mosel, near Zell, Germany	2
	Lake Gouwzee, Netherlands	10
	R. Waal, Rhine, Netherlands	10
	Oder River, Brody, Poland	10
	Oder River, Bytom, Poland	10
	Poulter River, England	5
	Bann River, Northern Ireland	10
	Werra River, Germany	10