Determining the pathways of marine bioinvasion: genetical and statistical approaches.

Neil Davies and George K. Roderick

Center for Conservation Research & Training, University of Hawaii, Honolulu, Hawaii.
and
Department of Environmental Science, Policy and Management, University of California, Berkeley.

Running header: Marine Bioinvasion Genetics

Key Words: multilocus, genetic, DNA, population, source

Corresponding author: Neil Davies, Center for Conservation Research & Training, 3050 Maile Way, Gilmore Hall 409, University of Hawaii, Honolulu, HI 96822, USA. Email: ndavies@Hawaii.edu; Fax: 1-808-956-9608; Phone: 1-808-956-3744

(after 1 August, Division of Insect Biology, Dept. of Environmental Sciences, Policy and Management, 201 Wellman Hall, University of California, Berkeley, CA 94720-3112 USA. Email: ndavies@nature.berkeley.edu; tel: (510) 642-3327, fax: (510) 642-7428)
**Abstract.** Advances in genetic technology have enabled biologists to reconstruct the history of populations, their evolutionary relationships and geographical origins. Such information is essential in understanding the biology of invasions and in designing successful management responses. Unfortunately, the wholesale transfer of ‘traditional’ population genetic methodology to identify the origins of marine bioinvasions is inappropriate. By definition, invading populations are characterized by rapid and recent range expansion. This has two important genetic consequences: i) genetic diversity is often very low due to the bottlenecks in population size associated with the founding of new populations, and ii) evolutionary relationships among genes may bear no relation to the history of populations. These characteristics of invading populations limit our ability to reconstruct their geographic history. The situation is further complicated by the fact that many bioinvasions occur as a dynamic series of sequential or overlapping invasion events, the totality of which can be termed a metainvasion. Here, we evaluate the genetic markers and statistical methods currently being used to determine invasion pathways. Analyses of molecular genetic data fall into two categories: those based on phylogenies, and those based on frequency differences of genetic markers. We describe these two approaches and outline the conditions under which they are appropriate and useful in marine bioinvasions. We also outline recent technical and analytical developments that may assist in the study of marine bioinvasions.
Determining the pathways of marine bioinvasion: genetical and statistical approaches.

Neil Davies and George K. Roderick

Following the discovery of an alien species, nearly all attempts at eradication or control (sterile male release, biological control, the use of transgenic plants etc.) have a higher probability of success when source regions and the demographic parameters of invading and source populations are known. For example, very different management measures are required when invasions involve a species with panmictic population structure, compared to those that involve a species with sharply defined and genetically isolated populations (Carey 1991). Here we review genetic methods of determining the source of a bioinvasion. We begin with a brief description of the problem and evaluate the various genetic tools that are available for reconstructing invasion pathways. Next, we review the statistical developments that are needed to make sense of the new genetic data. Finally, we propose some future developments that might help to elucidate the invasion process.

Mapping a metainvasion

When an invading species is recognized in a particular location, assigning the invading individuals to a source population is an important task (Davies et al. 1999; Roderick et al. 1998). Unfortunately, many bioinvasions are hierarchical, consisting of several sequential or overlapping invasion events, which together constitute a ‘metainvasion’
(Davies et al. in press). In many cases metainvasions are global in nature and any single invasion will often have multiple, genetically similar potential sources, many of which are likely to have been recently established themselves. Species that spread in association with human activity, especially invasive species, are likely to be characterized by newly established populations and these represent a severe obstacle for genetic analysis. First, new populations are often genetically impoverished due to the population bottleneck associated with colonization (Nei et al. 1975), making it hard to find sufficiently variable markers. Second, what little variation is present tends to be ancestral, rendering phylogenetic methods less appropriate or meaningless. These are the principle challenges faced by those seeking to identify the source of marine bioinvasions and to reconstruct their invasion pathways.

Allozyme electrophoresis is a useful technique for rapidly assessing the similarity of various populations, however, only a small amount of the underlying genetic variation at any given locus can be discerned using protein electrophoresis. Consequently, allozymes have provided only very limited resolution in invading populations of species such as the medfly, *Ceratitis capitata* (Huettel 1980; Roderick 1996b). A number of workers have identified heritable markers that reveal genetic variation at the DNA sequence level for reviews see (Geller 1996; Palumbi & Baker 1996; Roderick 1996a), with mitochondrial DNA being the most commonly used genetic marker in population studies (Avise 1994). Unfortunately, the reduction in diversity associated with colonization bottlenecks is exacerbated for mitochondrial genes because they have only a quarter of the effective population size ($N_e$) of nuclear genes (Moore 1995). In the medfly, for example, most New World populations are less than one hundred years old and have a single high
frequency and a single low frequency haplotype. By contrast, ancestral medfly populations in Africa display up to six haplotypes (Gasparich et al. 1997). Fortunately, new markers, such as microsatellites (Queller et al. 1993; Weber & May 1989), introns (Palumbi & Baker 1994), randomly amplified polymorphic DNA (RAPDs) (Welsh & McClelland 1990; Williams et al. 1990), and restriction length polymorphisms (RFLPs) (Aquadro et al. 1992) assay nuclear DNA variation. These markers have revealed high levels of diversity in ancestral and invading populations, even when mtDNA and allozymes are relatively impoverished (Baruffi et al. 1995; Villablanca et al. 1998). The greater variability of these markers can reveal population structure over a much finer scale. For example, genetic analysis of DNA sequence variation at four intron loci revealed significant population structure among previously indistinguishable (Gasparich et al. 1997) medfly populations in California, Central America and eastern South America (Davies et al. in press).

**Statistical developments in the analysis of bioinvasions**

As the practical difficulties of finding suitable genetic markers have been overcome, severe theoretical problems became apparent. First, new populations violate the assumptions of equilibrium integral to most population genetic theory (e.g. methods of estimating gene flow, Slatkin & Barton 1989). Second, phylogenetic methods that reconstruct the historical biogeography of relatively well established populations (Avise 1994; Roderick 1996a) seem inappropriate in the latter phases of a metainvasion where events are all very recent. For example, in the case of a secondary invasion event, the likely sources may be populations that were themselves only recently established as part
of the primary invasion. Phylogeographic structure (Avise 1989; Avise 1994; Roderick & Gillespie 1998; Roderick & Villablanca 1996) is not expected in such recently founded populations because there has been little time for mutations to occur, and the relationships among alleles mostly reflect evolutionary events in the ancestral range of the species rather than their history in newly occupied areas (Davies et al. 1999; Villablanca et al. 1998). For example, McGuigan et al. (1998) reported significant differences in haplotype frequencies over a fine geographic scale among Australian frog populations, however, a smaller and insignificant F-statistic was obtained when the genetic distance among alleles was considered.

With the traditional approaches being of only limited use for invasion biologists, it is fortunate that a new generation of statistical analyses have been developed based on multilocus genotype data. One of the first applications was in fisheries management, where mixed stock analysis (MSA) was developed to determine fish catch composition, mainly using allozyme data. In MSA, maximum likelihood is used to estimate the combination of potential source populations that best explain the observed allele frequencies in a catch (see Utter & Ryman 1993). Methods, such as MSA, that focus on populations are limited by the need to define those populations a priori and in doing so they risk missing individuals that have an unusual origin. An alternative approach focuses on the most likely origin of an individual, or rather its multilocus genotype, and is known as an assignment test (Paetkau et al. 1995).

Assignment tests use maximum likelihood to assign individual genotypes to potential sources based on the allele frequencies of the source populations. These test are particularly useful for determining the origin of an individual when there are multiple,
genetically similar candidate sources. For example, chinook salmon, *Oncorhynchus tshawytscha*, occur as different temporal populations (runs) that spawn in the same river but at different times of the year. Runs are genetically very similar but have such distinct life histories that some are considered separate (and endangered) species. Microsatellites have been used to assign chinook salmon to particular runs (Banks et al. 1996).

There are two major sources of error associated with the population level data used in an assignment test. First, observed allele frequencies are estimates, so sampling error must be considered. Second, differences in genetic diversity among potential source populations can cause a bias because the likelihood of drawing any genotype is inversely proportional to the diversity of the population from which it is drawn. Rannala and Mountain's (1997) assignment test takes into account the sampling error associated with estimating allele frequencies and the differences in diversity among two potential sources (Davies et al. 1999). Although Rannala and Mountain’s (1997) test applied much needed statistical rigor to source estimation, further modifications of the test are still needed. First, laboratory scoring mistakes must be taken into account. For example, scoring errors occur in allozyme studies with a frequency of about 1% (Lathrop et al. 1983). One can attempt to correct for scoring mistakes prior to analysis, as we did for the errors associated with sequencing cloned PCR products. Alternatively, an error rate factor can be incorporated into the analysis; such an approach was used by Marshall et al. (1998) in their multilocus paternity test. A second source of error that should be considered is the implications of not sampling all the potential sources. Again with a focus on paternity testing, Marshall et al. (1998) presented a simulation method to assess the likelihood that a more probable source remains unsampled. Finally, assignment tests focus on the origin
of single multilocus genotypes, although bioinvasions usually consist of multiple invading individuals, each of which will have their own associated likelihood of being from one source or another. With such multiple assignments, one will be able to plot a distribution of likelihood statistics for the invading population as a whole, adding a new level of complexity to source estimation. For example, a bimodal distribution would imply that the invading population had two sources, but how can one assess the significance of such a conclusion, and how should one correct for multiple comparisons? Such issues are currently being examined in our laboratory using computer simulation (Bohonak et al. in prep.).

Multilocus genotyping is a powerful technique and many different markers, including allozymes, mtDNA, and microsatellites can be analyzed simultaneously. In most cases additional markers will increase the power of these tests, however, some markers may be incompatible with this approach. RAPDs, for example, are very useful in providing high levels of genetic variation and their main advantage over introns and microsatellites is that they can be applied with very little prior genetic knowledge of a species (Williams et al. 1990). Unfortunately, it is not clear how RAPD data can be incorporated into the same statistical framework as introns and other markers, where genotypes can be identified at each locus.

Although we have stated that phylogenetic approaches are less likely to be useful in many bioinvasions (Davies et al. 1999; Roderick et al. 1998; Villablanca et al. 1998), recent data has suggested that phylogenetic data can still be used to help determine sources (Davies et al. in press). Somewhat to our surprise, we found that including an estimate of the genetic distance among alleles revealed a higher level of population
structure among American medflies than estimates based solely on allele frequencies (Davies et al.). This was surprising because medflies only colonized the Americas this century, providing little time for the evolutionary divergence of alleles among American populations. It is possible that phylogeographic structure might reflect multiple colonizations – one of several possibilities we are exploring using computer simulations (Bohonak et al. in prep).

It is possible, therefore, that phylogenetic information might turn out to be useful even in the later stages of a metainvasion. This is especially likely if genetic markers are found with very high evolutionary rates. We have proposed (Davies et al. 1999) that extragenomic markers (EGMs), such as viruses, be used to reconstruct invasion pathways. The RNA genomes of many viruses evolve extremely quickly (Holland 1982) and can be used to reconstruct the geographic history of their hosts (Ho 1993; Yanagihara 1994). Another possible set of markers that provide phylogenetic information and evolve rapidly are transposable element polymorphisms (or TEPs). We are currently working with groups in Greece and Italy to explore the utility of TEPs for the medfly metainvasion. Transposable elements occur widely in nature (Li 1997) and could be applied to marine invasive species should they prove successful in model terrestrial systems such as the medfly.

If EGMs or TEPs do become widely used, approaches that utilize phylogenetic data to detect gene flow (Slatkin 1994) might be applicable. Assignment tests might be modified to consider the relationships among alleles in addition to their frequencies. Currently, assignment tests do not consider the relatedness of alleles because most multilocus markers such as microsatellites (Glenn 1998) do not easily permit the
phylogenetic analysis of alleles. However, sequence data is much more phylogenetically informative, and with introns, EGMs, and TEPs becoming more widely used, restricting analysis to mere allele frequencies may waste useful information. A new test should assess the multilocus likelihood of sampling a given set of alleles from a potential source population based on the distance between alleles as well as their frequencies.

To conclude, the new genetic markers and statistical methods briefly described here will reveal much about the spread and establishment of invasive species. When combined with ecological data, it will eventually be possible to draw up common characteristics of invasive species and to identify the pathways that allow them spread around the globe. Eventually, invasion biology might hope to become a predictive science, identifying species that are likely to invade, and those that will become established should they reach a new area. Such information would finally enable managers to adopt truly proactive policies.

Acknowledgments

We thank A. Bohonak, R. Gillespie, F. Villablanca, S. Palumbi for useful discussions. Our research on metainvasions and non-equilibrium population genetics is supported by California Department of Food and Agriculture, USDA, NSF, NOAA, and University Research Council, University of Hawaii.
References


Queller, D. C., Strassmann, J. E. & Hughes, C. R. 1993 Microsatellites and kinship.  

_Proceedings of the National Academy of Sciences (USA)_ 94, 9197-9201.

Roderick, G. K. 1996a Geographic structure of insect populations: gene flow,  
phylogeography, and their uses. _Annual Review of Entomology_ 41, 263-290.

Roderick, G. K. 1996b Population genetic studies of tephritid flies of economic  
importance. In _Fruit Fly Pests. A World Assessment of Their Biology and  
Management_ (ed. B. A. McPheron & G. J. Steck), pp. 267-271. Delray Beach,  
Florida: St. Lucie Press.

Roderick, G. K., Davies, N., Bohonak, A. J. & Villablanca, F. X. 1998 The interface of  
population genetics and systematics: Invasion genetics of the Mediterranean fruit  
fly (_Ceratitis capitata_). In _Pest Management--Future Challenges. Proceedings of  
the Sixth Australasian Applied Entomological Research Conference_, vol. 1 (ed.  
M. P. Zalucki, R. A. I. Drew & G. G. White), pp. 489-499. Brisbane, Australia:  
University of Queensland.

Roderick, G. K. & Gillespie, R. G. 1998 Speciation and phylogeography of Hawaiian  
terrestrial arthropods. _Molecular Ecology_ 7, 519-531.

Roderick, G. K. & Villablanca, F. X. 1996 Genetic and statistical analysis of  
colonization. In _Fruit Fly Pests. A World Assessment of Their Biology and  
Management_ (ed. B. A. McPheron & G. J. Steck), pp. 281-290. Delray Beach,  
Florida: St. Lucie Press.


