

**MITOCHONDRIAL GENOMICS AND NORTHWESTERN ATLANTIC
POPULATION GENETICS OF MARINE ANNELIDS**

By

Robert M. Jennings

B.S., University of Michigan, 1998

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

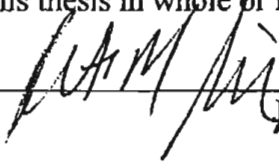
WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2005

© 2005 *Robert M. Jennings*
All rights reserved.

The author hereby grants to MIT and WHOI permission to reproduce paper and electronic copies of this thesis in whole or in part and to distribute them publicly.

Signature of Author



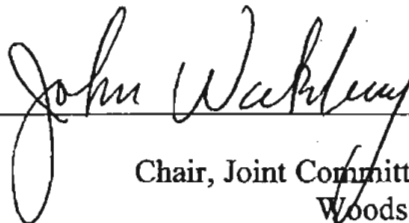
Joint Program in Biological Oceanography
Massachusetts Institute of Technology
and Woods Hole Oceanographic Institution
September 6, 2005

Certified by



Lauren S. Mullineaux
Thesis Supervisor

Accepted by



John Waterbury
Chair, Joint Committee for Biological Oceanography
Woods Hole Oceanographic Institution

Mitochondrial genomics and Northwestern Atlantic population genetics of marine annelids

by
Robert M. Jennings

Submitted in Partial Fulfillment of the Requirements for the degree of
Doctor of Philosophy in Biological Oceanography

ABSTRACT

The overarching goal of this thesis was to investigate marine benthic invertebrate phylogenetics and population genetics, focused on the phylum Annelida. Recent expansions of molecular methods and the increasing diversity of available markers have allowed more complex and fine-scale questions to be asked at a variety of taxonomic levels. At the phylogenetic level, whole mitochondrial genome sequencing of two polychaetes (the deep-sea tubeworm *Riftia pachyptila* and the intertidal bamboo worm *Clymenella torquata*) supports the placement of leeches and oligochaetes within the polychaete radiation, in keeping with molecular evidence and morphological reinvestigations. This re-interpretation, first proposed by others, synonymizes “Annelida” and “Polychaeta”, and lends further support to the inclusion of echiurids, siboglinids (previously called vestimentiferans) within annelids, and sipunculans as close allies. The complete mt-genome of *C. torquata* was then rapidly screened to obtain markers useful in short timescale population genetics. Two quickly evolving mitochondrial markers were sequenced from ten populations of *C. torquata* from the Bay of Fundy to New Jersey to investigate previous hypotheses that the Cape Cod, MA peninsula is a barrier to gene flow in the northwest Atlantic. A barrier to gene flow was found, but displaced south of Cape Cod, between Rhode Island and Long Island, NY. Imposed upon this pattern was a gradient in genetic diversity presumably due to previous glaciation, with northern populations exhibiting greatly reduced diversity relative to southern sites. These trends in *C. torquata*, combined with other recent short time scale population genetic research, highlight the lack of population genetics models relevant to marine benthic invertebrates. To this end, I constructed a model including a typical benthic invertebrate life cycle, and described the patterns of genetic differentiation at the juvenile and adult stages. Model analysis indicates that selection operating at the post-settlement stage may be extremely important in structuring genetic differentiation between populations and life stages. Further, it demonstrates how combined genetic analysis of sub-adult and adult samples can provide more information about population dynamics than either could alone.

Thesis Supervisor: Lauren S. Mullineaux
Title: Professor of Biological Oceanography

TABLE OF CONTENTS

ABSTRACT	3
ACKNOWLEDGEMENTS	7
CHAPTER 1: Introduction.....	9
CHAPTER 2: Mitochondrial Genomes of <i>Clymenella torquata</i> (Maldanidae) and <i>Riftia pachyptila</i> (Siboglinidae): evidence for conserved gene order in annelids.....	21
CHAPTER 3: Assessment of the Cape Cod phylogeographic break using the bamboo worm <i>Clymenella torquata</i> (Annelida: Maldanidae).....	59
CHAPTER 4: Stage-specific selection structures geographic genetic patterns in marine benthic invertebrate populations.....	95
CHAPTER 5: Summary, broader impacts, and future research.....	121
APPENDIX 1: PCR amplification and primer information for Chapter One.....	128
APPENDIX 2: MATLAB model codes used in Chapter Four.....	132

ACKNOWLEDGEMENTS

This thesis belongs as much to the people in my labs, my family, and my friends as it does to me. Without their constant support, guidance, encouragement, and love, I would not have accomplished anything. First, I would like to thank my several advisors for inviting me into their labs and teaching me how to do quality science. I am grateful to Ken Halanych for his willingness to talk at any time, for the enthusiasm with which he engages in scientific discussion, and for the sense of energy and excitement he instills in the people who work in his lab. I thank Lauren Mullineaux for adopting me into her lab, and enabling me to benefit from her perceptive mind, her caring persona, and constant support. Although Tim Shank was not technically an advisor, he welcomed me into his lab as if I were his student, for which I am very thankful. Tim was also always eager to discuss my work, offer insightful advice, and never lose sight of what made science exciting.

I would also like to thank the other members of my thesis committee, Mike Neubert and Dan Rothman. They were always at hand to provide expert guidance and a fresh perspective. Combined, my five committee members brought a tremendous array of knowledge and skill to my thesis work, and the integrative nature of my research reflects their ability to work within and outside of their disciplines with incredible talent. I am grateful for all the scientific and communicative skills I learned from interacting with each of them.

I am especially thankful to have been surrounded by so many wonderful coworkers and friends during my time at MIT and WHOI. My lab mates—Yale, Nan, Annette, Thomas, El, Susan, Lara, Heidi, Diane, Carly, Amy, Stace, Kate, Abby, and Walter have been subjected to countless drafts, practice talks, and often random lines of questioning and have responded each time with aplomb. They have provided great ideas, cheered me on, and made me laugh. To name the larger list of people in labs at both WHOI and MBL whose advice, machines, and reagents I have sought out (and hopefully all returned) would take too long; I think I have gone to just about every person in the Biology Department at some time or other, and I thank them all for your friendliness and help.

My family has suffered more of my personal storms and crises while I worked on this thesis, and responded with calming words and much needed encouragement. Their love and support has been invaluable over these past years. My incredible husband David has been my rock, and occasionally my reluctant-but-amenable lab mate. He has learned far more about worms, DNA, and estuaries than he probably ever intended to, and moreover managed to keep me sane throughout the whole process.

There is the family who was stuck with me (sorry guys), and the family of friends by whom I am blessed to have been chosen. Their names are far too many to list, but Christie, Brad, Ann, Randolph, Gordon, Mimi, Helen, Sandy, Sarah, Jean, Regina, Ami, Rudi, Andy, Astri, and Dicky deserve special mention.

Financial support was provided by an Academic Programs Office fellowship, a CICOR fellowship and research grant, and a National Science Foundation research grant.

Chapter One: INTRODUCTION

In the last 20 years, the field of molecular biology has witnessed an explosion of new techniques and types of markers, as well as considerable increases in throughput. After the allozyme investigations of the 1970's, the advent of PCR made markers available such as DNA sequences, RFLPs, AFLPs, microsatellites, RAPIDs, and many others. These advances can be seen as a sort of “bigger, better, faster” process, where the goals have been to discover and employ high resolution markers for the desired analysis, to analyze high numbers of markers in the same analysis, and to increase the number of individuals in datasets, all while maintaining economic feasibility. As low-cost, high throughput molecular methods have made larger, complex datasets more attainable, new avenues of research have opened up, often crossing the lines that have traditionally separated (for example) phylogenetics and population genetics.

One molecular marker that has this potential is the mitochondrial genome (mt-genome). The mitochondrial DNA (mtDNA) of bilaterian animals is usually on the order of 15 kilobases (kb), does not recombine, and is inherited in a simple fashion (Boore 1999). Further, it contains different types of markers, such as DNA sequences of genes (tRNA, rRNA, and protein-coding), protein sequences of protein-coding genes, and the order of the genes in the molecule (Boore 1999, Blanchette et al 1999). This range of data types enables mt-genomes to bridge phylogenetics and population genetics, because the one molecule provides different suites of characters with resolution at the varying scales relevant to these fields. Mitochondrial genomes, therefore, have the potential to free both fields from their reliance on the commonly used, easily sequenced genes of the

past (e.g. cytochrome oxidase c subunit I (COI, see Folmer et al. 1994) or the large ribosomal subunit (16S, see Palumbi 1996)). In doing so, they allow new questions to be asked, and offers new insights into old questions. Many of these new endeavors are also fueling advances in modeling and analysis as well. My dissertation highlights some of the new avenues in phylogenetics and population genetics, focusing on marine worms of the phylum Annelida.

Marine Annelid Phylogenetics

Annelid phylogenetic relationships have remained difficult to determine despite numerous morphological analyses (e.g., Rouse and Fauchald 1995; Rouse and Fauchald 1998, Rouse and Pleijel 2001) and molecular analyses (reviewed by McHugh 2000; Halanych, Dahlgren, and McHugh 2002; and Halanych 2004). Recent incorporations of taxa originally described as separate phyla (i.e. Echiura, McHugh 1997; Vestimentifera, Halanych et al. 1998) have further complicated annelid relationships. Note that the newer name Vestimentifera has reverted to the previously proposed Siboglinidae (McHugh 1997). Newer views of annelid phylogeny have moved away from the traditional view of two main groups, Clitellata (the Oligochaetes and Hirudineans) and Polychaeta, in support of the hypothesis that clitellates, echiurans, and vestimentiferans fall within the polychaete radiation (McHugh 2000; Halanych, Dahlgren, and McHugh 2004; and Halanych 2004). This new view synonymizes Annelida with Polychaeta, and alters the previous, long-accepted phylogeny at a deep taxonomic level. However, the vast potential for annelid morphological adaptation implied by these revisions is perhaps

not surprising, given the enormous diversity of annelid body plans, habitats, and life histories. Because of this wide array of annelid morphologies, understanding the relationships among annelid groups speaks to the more general themes of how organisms speciate, evolve, and adapt to new environments.

Because the Annelida contain a large amount of species diversity, full mt-genome analysis is a particularly appealing tool for resolving phylogenetic relationships. DNA sequence analysis of mitochondrial genes has the potential to resolve fine scale taxonomic relationships, while the more slowly evolving protein sequences should be meaningful at a larger taxonomic scale. Finally, the even lower evolutionary rate of gene order (at least as seen in annelids to date, cf. Boore and Brown 2000) might be best suited to the larger questions of relationships between the Clitellata and Polychaeta. Presently, only two complete and two incomplete annelid genomes are available (the polychaete *Platynereis dumerilii* and the oligochaete *Lumbricus terrestris*, and the polychaete *Galathealinum brachiosum* and leech *Helobdella robusta*, respectively; Boore and Brown 2000). One of the polychaetes differs in gene order from the other three annelids (including the other polychaete); however, since several major annelid clades remain unsampled, it is unclear to what extent this pattern fits with the systematic hypotheses outlined above. By using the deep-sea tubeworm *Riftia pachyptila* (Siboglinidae) and the intertidal bamboo worm *Clymenella torquata* (Maldanidae) in the phylogenetic study, I have included representatives of all the major clades defined by Rouse and Fauchald (1997).

Marine Annelid Phylogeography: *Clymenella torquata*

In addition to providing multilayered data for phylogenetic analysis, whole mt-genome sequencing also allows phylogenetics to be easily extended to the intraspecies scale. Within species, the study of the genetic relationships between geographically separated populations is known as phylogeography (*sensu* Avise and Felley 1979, Avise et al. 1987). Although the entire mt-genome sequence itself has been used as a phylogeographic marker (notably in humans, e.g. Watson et al. 1997; Maca-Meyer et al. 2001), obtaining the complete sequence for a sufficient number of individuals and populations is still not usually feasible for marine annelids (or invertebrates in general). However, obtaining all of the sequence for one individual allows new primers to be designed for rapid screening of typically underused genes (e.g. the small ribosomal subunit (12S) and the ATP and NADH families) across a geographic range of interest; thus a phylogenetic investigation can produce data that are easily adaptable to the population genetic scale. Indeed, whole mt-genome analyses can and have extended across these scales in the same study.

Of the two polychaetes examined in the phylogenetic study, *Clymenella torquata* in particular presents an interesting paradox that whole mt-genome research could help resolve. This intertidal species has a large geographic range in the Atlantic, extending from New Brunswick, Canada to Florida, USA (Mangum 1962). In contrast, however, after the adults reproduce synchronously in late spring, the larvae are only dispersed in the water column for a few days (Newell 1951). The paradox of extremely low dispersal potential coupled with a large geographic range is especially surprising because of

evidence suggesting that Cape Cod, Massachusetts, USA is a barrier to gene flow for many species in the Northwest Atlantic (reviewed in Wares 2002). Although different species appear to respond to the Cape Cod barrier in different ways (or not at all), one would expect a weakly dispersive annelid like *C. torquata* to be extremely sensitive to any restriction to gene flow. This is because larvae that stay in the water column for longer periods of time experience a wider range of water movements (in terms of spatial and temporal variability) and can potentially be advected past a barrier by rare or weak water movements. Further, the wide spacing of sampled populations in some of the studies described by Wares (2002), many of which were not designed to specifically address Cape Cod, only allowed determination of the barrier's location at a relatively coarse scale (i.e., in the vicinity of Cape Cod). In such situations, closely spaced sampling sites and quickly evolving markers provide the best opportunity to pinpoint the nature and location of any barrier(s) to gene flow. Possessing *C. torquata*'s mt-genome, which gives access to the more quickly evolving markers, combined with the species' ubiquity in the region, makes it easier to obtain high-resolution markers with which to further investigate gene flow in the Northwest Atlantic.

Marine Phylogeography: Stage-Structured Gene Flow

The use of more and newer markers in benthic invertebrate population genetics has allowed questions to be asked on a more ecological scale. The information in DNA markers is integrated over a time scale related to its rate of evolution. That is, a slowly evolving gene or marker will retain the genetic signature of previous events (for example

a drastic reduction in genetic diversity) for a longer period of time than would a quickly evolving marker. As population geneticists develop more quickly evolving markers and create genomic scale datasets, they are increasingly using these powerful tools to look at short time scale processes such as single migration events, or the period between settlement of larvae to the bottom and their survival and incorporation into the pool of adults. Many of these processes are also important to fisheries and management groups seeking to understand how best to protect and maintain their resources. This interest has led to increased focus to the sub-adult life stages on which these processes act, and increased understanding of the importance of invertebrate life cycles in shaping the genetics of populations.

Many marine benthic invertebrates (annelids and others) exhibit type III reproductive curves (see Hunt and Scheibling 1997), in which thousands to millions of young are spawned, most of which die before reaching reproductive maturity. Because the term gene flow implies survival to reproduction, the details of life history might not matter to population geneticists only interested in estimating gene flow. However, as fisheries and marine management facilities increasingly turn to genetic estimates of dispersal itself (i.e., movement of larvae between populations regardless of their fate afterward), genetic samples of larvae, juveniles, and recruits become more appealing because they are closer to the actual migration event than the eventual adult survivors.

Although studies involving adult and non-adult genetic samples have already revealed interesting and often contrasting patterns of gene flow (e.g. Johnson and Black 1984, Johnson and Wernham 1999, Moberg and Burton 2000, Drouin et al. 2002,

Crivello et al. 2004), there currently exist no population genetic models that treat stages separately or that consider the possible genetic effects of invertebrate life cycles. Classic population genetics models such as Wright's island model (1943) Kimura and Weiss's stepping stone model (1964), or the phylogeographic method of Avise et al. (1987) still provide the basis for estimating gene flow, but new models should be developed to accommodate the new markers, sub-adult samples, and finer-scale questions becoming more common in population genetics. While the biologically generic formulation of existing population genetic models has made them easily extendable to a wide array of species and environments, the dynamics specific to marine benthic invertebrates should be incorporated explicitly into a model to understand how forces acting at multiple life stages create patterns of genetic differentiation.

Hand in hand with attention to non-adult stages has come a broadening view of the processes important to shaping genetic patterns of populations at all stages. Population genetic theory has long held selective neutrality to be a necessary characteristic of any marker, with the ideal marker being one that passively records (in a genetic sense) the population dynamics important to the species in question. This view can be represented by Kimura's formulation of the neutral theory (1983). Although some population geneticists have constructed models involving selection, this has historically been almost completely in the case of phenotypic traits that vary quantitatively (see for example Chakraborty and Nei 1982), often as related to breeding programs and artificial selection. The neutral and nearly-neutral (Ohta and Kimura 1971) schools have been much more central to the gene flow literature, with selection seen as a confounding factor

that obscures the patterns of interest. In the face of differing genetic patterns observed in (for example) benthic invertebrate recruits and adults, however, selection has been increasingly implicated as the cause (e.g. Hellberg 1996, Planes and Romans 1994). For instance, Moberg and Burton (2000) found geographic structure to the pattern of genetic differentiation in recruits of the sea urchin *Strongylocentrotus franciscani*, but no pattern in adults. They hypothesize that selection acting after settlement (often a time of severe mortality) was the cause of the difference.

The incorporation of stage-structured population genetic models that capture the biological dynamics imposed by marine benthic invertebrate life histories is critical for understanding how populations of these taxa are structured, and how they evolve. As new markers and sub-adult samples allow us to focus on specific processes in marine environments, these models will go far in determining how the patterns of genetic differentiation observed in adult populations are shaped throughout the entire life cycle.

Goals of this study

The goals of my dissertation were to further our understanding of the molecular evolution of marine invertebrates, specifically annelids, at a variety of scales. At the phylogenetic scale, I sought to use whole mt-genomes as multifaceted tools in resolving poorly understood relationships among major annelid groups. In doing so, I also investigated the patterns of mt-genome evolution in annelids. I used this phylogenetic work as a springboard to investigate annelid population genetics, using the common, weakly dispersive bamboo worm *Clymenella torquata* to assess the hypothesized Cape

Cod phylogeographic barrier. Finally, I develop and analyze a stage-structured population genetic model that investigates the ways in which genetics of marine invertebrate populations are shaped by neutral and selective forces acting throughout the life cycle.

References

- Avise, J.C. and J. Felley. 1979. Population structure of freshwater fishes. I. Genetic variation of the bluegill (*Lepomis macrochirus*) populations in manmade reservoirs. *Evolution* **33**: 15-26.
- Avise, J.C., J. Arnold, R.M. Ball, E. Bermingham, T. Lamb, J.E. Neigel, C.A. Reeb and N.C. Saunders. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* **18**: 489-522.
- Blanchette, M., T. Kunisawa and D. Sankoff. 1999. Gene order breakpoint evidence and animal mitochondrial phylogeny. *Journal of Molecular Evolution* **49**: 193-203.
- Boore, J.L. 1999. Animal mitochondrial genomes. *Nucleic Acids Research* **27**: 1767-1780.
- Boore, J.L. and W.M. Brown. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: Sequence and gene rearrangement comparisons indicate the Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Molecular Biology and Evolution* **17**: 87-106.
- Chakraborty, R. and M. Nei. 1982. Genetic differentiation of quantitative characters between populations of species. *Genetical Research* **39**: 303-314.
- Crivello, J.F., D.J. Danila, E. Lorda, M. Keser and E.F. Roseman. 2004. The genetic stock structure of larval and juvenile winter flounder larvae in Connecticut waters of eastern Long Island Sound and estimations of larval entrainment. *Journal of Fish Biology* **65**: 62.
- Drouin, C.A., E. Bourget and R. Tremblay. 2002. Larval transport processes of barnacle larvae in the vicinity of the interface between two genetically different populations of *Semibalanus balanoides*. *Marine Ecology-Progress Series* **229**: 165.
- Folmer, O., M. Black, W. Hoeh, R. Lutz and R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*. **3**: 294-299.
- Halanych, K.M., R.A. Lutz and R.C. Vrijenhoek. 1998. Evolutionary origins and age of vestimentiferan tube-worms. *Cahiers de Biologie Marine* **39**: 355-358.

- Halanych, K.M., T.G. Dahlgren and D. McHugh. 2002. Unsegmented annelids? Possible origins of four lophotrochozoan worm taxa. *Integrative and Comparative Biology* **42**: 678-684.
- Halanych, K.M. 2004. The new view of animal phylogeny. *Annual Review of Ecology and Evolutionary Systems* **35**: 229-256.
- Hellberg, M.E. 1996. Dependence of gene flow on geographic distance in two solitary corals with different larval dispersal capabilities. *Evolution* **50**: 1167.
- Hunt, H.L. and R.E. Scheibling. 1997. Role of early post-settlement mortality in recruitment of benthic marine invertebrates. *Marine Ecology Progress Series* **155**: 269-301.
- Johnson, M.S. and R. Black. 1984. Pattern beneath the chaos: the effect of recruitment on genetic patchiness in an intertidal limpet. *Evolution* **38**: 1371-1383.
- Johnson, M.S. and J. Wernham. 1999. Temporal variation of recruits as a basis of ephemeral genetic heterogeneity in the western rock lobster *Panulirus cygnus*. *Marine Biology* **135**: 133.
- Kimura, M. and G.H. Weiss. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**: 561-576.
- Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Maca-Meyer, N., González, A.M., Larrunga, J.M., Flores, C., and Cabrera, V.M. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* **2**:13.
- Mangum, C.P. 1962. Studies on speciation in the Maldanid polychaetes of the North American Atlantic Coast. I. A taxonomic revision of three species of the subfamily Euclymeninae. *Yale Peabody Museum Postilla No. 65*: 12 p.
- McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proceedings of the National Academy of Sciences USA* **94**: 8006-8009.
- McHugh, D. 2000. Molecular phylogeny of the Annelida. *Canadian Journal of Zoology* **78**: 1873-1884.
- Moberg, P.E. and R.S. Burton. 2000. Genetic heterogeneity among adult and recruit red sea urchins, *Strongylocentrotus franciscanus*. *Marine Biology* **136**: 773.

- Newell, G.E. 1951. The life-history of *Clymenella torquata* (Leidy). Proceedings of the Zoological Society of London **121**: 561-590.
- Ohta, T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution*, **1**: 18-25.
- Palumbi, S.L. 1996. Nucleic acids II: The polymerase chain reaction. Pp. 205-248 in D. M. Hillis, C. Mortiz and B. K. Mable, eds. *Molecular Systematics*. Sinauer Associates, Sunderland, MA.
- Planes, S. and P. Romans. 2004. Evidence of genetic selection for growth in new recruits of a marine fish. *Molecular Ecology* **13**: 2049.
- Rouse, G.W. and K. Fauchald. 1995. The articulation of annelids. *Zoological Scripta* **24**: 269-301.
- Rouse, G.W. and K. Fauchald. 1998. Recent views on the status, delineation and classification of the Annelida. *American Zoologist* **38**: 953-964.
- Rouse, G.W. and F. Pleijel. 2001. *Polychaetes*. Oxford University Press, New York.
- Wares, J.P. 2002. Community genetics in the northwestern Atlantic intertidal. *Molecular Ecology*. **11**: 1131-1144.
- Watson, E., P. Forster, M. Richards and H.J. Bandelt. 1997. Mitochondrial footprints of human expansions in Africa. *American Journal of Human Genetics* **61**: 691.
- Wright, S. 1943. Isolation by distance. *Genetics* **28**: 114-138.

Chapter 2: Mitochondrial Genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for Conserved Gene Order in Annelida

Abstract

Mitochondrial genomes are useful tools for inferring evolutionary history. However, many taxa are poorly represented by available data. Thus, to further understand the phylogenetic potential of complete mitochondrial genome sequence data in Annelida (segmented worms), I examined the complete mitochondrial sequence for *Clymenella torquata* (Maldanidae) and an estimated 80% of the sequence of *Riftia pachyptila* (Siboglinidae). These genomes have remarkably similar gene orders to previously published annelid genomes, suggesting that gene order is conserved across annelids. This result is interesting given the high variation seen in the closely related Mollusca and Brachiopoda. Phylogenetic analyses of DNA sequence, amino acid sequence and gene order all support the recent hypothesis that Sipuncula and Annelida are closely related. Our findings suggest that gene order data is of limited utility in annelids but that sequence data holds promise. Additionally, these genomes show AT bias (~66%) and codon usage biases, but have a typical gene complement for bilaterian mitochondrial genomes.

INTRODUCTION

Sequencing of complete mitochondrial genomes has become a useful tool for inferring animal phylogeny (e.g. Boore and Brown 1998; Lavrov, Brown, and Boore 2004; Helfenbein and Boore 2004). The haploid, non-recombining properties of animal mitochondrial DNA (mtDNA), coupled with its small size, make it a logical choice when considering phylogenetic events. Determination of the entire mitochondrial genome sequence provides several suites of characters for phylogenetic analysis; for example, DNA gene sequences (rRNA, tRNA, and protein-encoding), inferred amino acid sequences of protein-encoding genes, and the arrangement of genes in the genome. However, there is considerable disparity in taxonomic sampling. Chordata accounts for 75% of the published animal mitochondrial genomes and Arthropoda represents the next

12.5%. Thus, there is still much to learn about how mitochondria evolve in many animal lineages.

Despite the importance of Annelida (segmented worms) with over 12,000 described species and its dominance as the most abundant macrofaunal group in the deep sea (69% of the planet), only two complete annelid mitochondria have been sequenced (the nereid *Platynereis dumerilii* and the oligochaete *Lumbricus terrestris*). These genomes differ only slightly in gene order. In addition, partial genomes of the siboglinid *Galathealinum brachiosum* and the leech *Helobdella robusta* (Boore and Brown 2000), match the *L. terrestris* gene order exactly. [Note that Siboglinidae was previously referred to as Pogonophora and Vestimentifera (McHugh 1997; Rouse and Fauchald 1997; Halanych et al. 2001).] Some mtDNA genome data is available for allied Lophotrochozoan taxa; most relevant are mollusks (e.g., Hoffman, Boore, and Brown 1992; Boore and Brown 1994; Hatzoglou, Rodakis, and Lecanidou 1995; Terrett, Miles, and Thomas 1996; Wilding, Mill, and Grahame 1999; Kurabayashi and Ueshima 2000; Grande et al. 2002; Tomita et al. 2002; Serb and Lydeard 2003; Boore, Medina, and Rosenberg 2004; Dreyer and Steiner 2004; DeJong, Emery, and Adema 2004), brachiopods (Stechmann and Schlegel 1999; Noguchi et al. 2000; Helfenbein, Brown and Boore 2001), phoronids (Helfenbein and Boore 2004) and sipunculans (Boore and Staton 2002). Of these taxa, the sipunculan *Phascolopsis gouldii* is the most similar to the known annelid arrangements with 16 of the 19 sipunculan genes examined in the same order as in *L. terrestris* (but in two separate blocks). For this reason, Boore and Staton (2002) hypothesized a close relationship between annelids and sipunculans. Mollusks are

notable because their mitochondrial genomes appear to have experienced numerous large-scale rearrangements and some taxa have even lost the *atp8* gene. Brachiopods also seem to have undergone numerous rearrangements. Of the three complete genomes currently available, *Laqueus rubellus* and *Terebratalia transversa* share 14 gene boundaries composed in 9 blocks; *L. rubellus* and *Terebratulina retusa* share only 8 gene boundaries in 8 separate blocks (Helfenbein, Brown, and Boore 2001).

Recent views of annelid phylogeny have moved away from the traditional view of two main groups, Clitellata (Oligochaetes and Hirudineans) and Polychaeta. Although morphological cladistic analyses have supported this hypothesis (Rouse and Fauchald 1995), multiple sources of data clearly show that the Clitellata, Echiuridae, and Siboglinidae are within the polychaete radiation (reviewed in McHugh 2000; Halanych, Dahlgren, and McHugh 2002; Halanych 2004). Such potential for morphological adaptation is not surprising given the enormous amount of diversity in annelids' body plans, habitats, and life histories. A comprehensive molecular phylogeny of Annelida is wanting, and currently our best understanding of annelid evolutionary history comes from morphological cladistic analyses (Rouse and Fauchald 1997; Rouse and Pleijel 2001), which suggest Annelids contain three major groups, Scolecida, Aciculata, and Canalipalpata. Unfortunately, the Clitellata are not considered in these treatments.

I report here the complete mitochondrial sequence of a bamboo worm *Clymenella torquata* (Maldanidae) and an estimated 80% of the genome of the deep-sea tubeworm *Riftia pachyptila* (Siboglinidae). *Clymenella torquata* and the other members of Maldanidae are called bamboo worms because the shape of their segments gives them a

bamboo-like appearance. *Clymenella torquata* is common in sandy intertidal/subtidal estuaries of the Atlantic U.S. coast, where it builds tubes from the surrounding sand and ingests sediment and the associated interstitial organisms (Mangum 1964). *Riftia pachyptila* inhabits the hydrothermal vents of the East Pacific Rise, and obtains energy from the chemosynthetic endosymbiotic bacteria in a specialized structure called the trophosome (Southward and Southward 1988). Although annelid phylogeny has not been well resolved, available molecular evidence (Halanych, unpublished) places these two annelids in very distant parts of the annelid tree. By including these two taxa, I provide representatives for all major clades outlined by Rouse and Fauchald (1997). Our goals in presenting and analyzing these new genomes are 1) to further characterize the evolution of mitochondrial genome structure among annelids and 2) to explore the potential of mitochondrial genomes in resolving annelid phylogeny.

METHODS

Organisms

Clymenella torquata and *Riftia pachyptila* were chosen to obtain better representation of annelid diversity than is currently available for mitochondrial genomes. *C. torquata* is in Maldanidae within Scolecida and *R. pachyptila* is in Siboglinidae within Canalipalpata. When combined with the available annelid genomes from GenBank (Table 1.1), all of the major clades of Annelida are represented (see McHugh 2000; Rouse and Fauchald 1997; Rouse and Pleijel 2001). All of the genome of *C. torquata* and two-thirds of the *R. pachyptila* genome presented here were sequenced from total DNA extractions of a single

Species	Clade	Nucleotides	GenBank Number
<i>Clymenella torquata</i>	Annelida, Scolecida, Maldanidae	15,538 complete	submitted
<i>Riftia pachyptila</i>	Annelida, Canalipalpata, Siboglinidae	12,016 partial	submitted
<i>Galathealinum brachiosum</i>	Annelida, Canalipalpata, Siboglinidae	7,576 partial	AF178679
<i>Platynereis dumerilii</i>	Annelida, Aciculata, Nereididae	15,619 complete	NC_000931
<i>Lumbricus terrestris</i>	Annelida, Oligochaeta, Lumbricidae	14,998 complete	NC_001673
<i>Helobdella robusta</i>	Annelida, Hirudinea, Glossiphoniidae	7,553 partial	AF178680
<i>Phascolopsis gouldii</i>	Sipuncula	7,470 partial	AF374337
<i>Katharina tunicata</i>	Mollusca, Polyplacophora	15,532 complete	NC_001636
<i>Loligo bleekeri</i>	Mollusca, Cephalopoda	17,211 complete	NC_002507
<i>Albinaria caerulea</i>	Mollusca, Gastropoda	14,130 complete	NC_001761
<i>Cepaea nemoralis</i>	Mollusca, Gastropoda	14,100 complete	NC_001816
<i>Terebratulina retusa</i>	Brachiopoda, Articulata	15,451 complete	NC_000941
<i>Laqueus rubellus</i>	Brachiopoda, Articulata	14,017 complete	NC_002322
<i>Terebratalia transversa</i>	Brachiopoda, Articulata	14,291 complete	NC_003086

Table 1.1 Taxa used in phylogenetic analysis

individual of each species; the remaining *R. pachyptila* sequence reported herein came from a second individual. *C. torquata* was collected in 2002 from Hyannisport, MA (N 41°37'57.9" W 70°19'18.3"). The two *R. pachyptila* were collected in 2000 at 2500m depth near the Tica vent at 9°N on the East Pacific Rise (N 9°50'26.8", W 104°17'29.6"). All organisms were frozen at -80°C after collection.

DNA Extraction and mtDNA sequencing

Total genomic DNA was extracted from approximately 25mm³ tissue using the DNEasy kit (Promega) according to manufacturer's protocols. Throughout this paper, gene nomenclature and abbreviations follow Boore and Brown (2000): *cox1-3* refer to cytochrome oxidase c subunits 1-3, *nad1-6* (incl. *4L*) refer to NADH dehydrogenase subunits 1-6, *atp6* and *atp8* refer to ATPase F0 subunits 6 and 8, and *cob* refers to the

cytochrome oxidase b apoenzyme. tRNA genes are designated *trnX*, where *X* is the single-letter amino acid code. Contrary to Boore and Brown (2000), the large and small ribosomal subunits are here referred to as *mLSU* (mitochondrial large subunit) and *mSSU* (mitochondrial small subunit) respectively.

Clymenella torquata

All mtDNA amplifications of *C. torquata* employed 1µL EXL Polymerase (Stratagene), as well as 5µL EXL buffer, 25pmol dNTPs, 200ng each primer, 1µL stabilizing solution and approximately 10ng genomic DNA per 50µL reaction. The sections *mLSU-cox1* (using primers 16Sar-L/HCO2198), *cox1-cox3* (LCO1498/COIIIr), *cox3-cob* (COIIIr/CytbR), and *cob-mLSU* (CytbF/16Sbr-H) all generated single-banded products. The *mLSU* primers are from Palumbi (1996); *cob* and *cox3* primers are from Boore and Brown (2000), and the COI primers are from Folmer et al. (1994). PCR protocols for these fragments are found in Appendix 1, Table A1.1. Products were verified on an agarose gel, purified using the QiaQuick kit (Qiagen), eluted in 40µL water, and sheared separately in a HydroShear DNA shearer (GeneMachines) to generate random fragments of 1-2kb in length. The sticky ends were polished with the Klenow fragment, and were A-tailed using *Taq* polymerase, an excess of dATP, and incubation at 72°C for 10 min. DNA was then repurified with the QiaQuick kit, and cloned into pGEM-T Easy (Promega). Sequencing reactions were performed using Big Dye (versions 2 and 3) chemistry on an ABI 377 (Applied Biosystems). Fifteen *mLSU-cox1* clones (average coverage 5.3X), 9 *cox3-cob* clones (average 2.9X) and 7 *cob-mLSU*

clones (average 2X) were sequenced in both directions using T7 and SP6 and then assembled to generate contigs. Combined, the assemblies contained ~90% of the sequence of *C. torquata*'s mt-genome. Three clones could not be entirely sequenced using plasmid primers. To complete sequencing on these clones, 19 walking-primers were designed (see Appendix 1, Table A1.2). No clones were recovered containing the largest non-coding region (i.e. the control region or *UNK*) or the approximately 3kb surrounding it (roughly including regions of the *atp6* and *nad4L* genes, and all of *nad5*, *trnW*, *-H*, *-F*, *-E*, *-P*, and *-T*). This region was sequenced by amplification with flanking primers (Ctatp6f2 and Ctnad4r2) and direct sequencing using the walking primers.

Riftia pachyptila

mtDNA amplification for *R. pachyptila* was adapted from the procedure of Boore and Brown (2000). Standard primers were used to amplify short sections of *cox1* (LCO1490 and HCO2198; Folmer *et al.* 1994), and *cob* (CytbF and CytbR; Boore and Brown 2000) with *Taq* polymerase (Promega) in standard 25 μ L PCRs. Products were purified using the QiaQuick Gel Extraction Kit (Qiagen) and sequenced on an ABI 377 automated sequencer. These sequences were used to design *Riftia*-specific primers for long PCR. In *cox1*, the primers Rp1536 and Rp2161 were designed, and in *cob*, CytBRp. Information for all primers can be found in supplementary information.

These primers were then used to amplify long segments of the mt-genome in conjunction with the primers mentioned above: 16Sar-L and Rp1536 amplified the region spanning *mLSU-cox1*, Rp2161 and COIIIr amplified *cox1-cox3*, and COIIIr and CytBRp

amplified *cox3-cob*. PCR conditions are listed in Appendix 1, Table 1.3. These long PCR reactions consisted of 5 μ L 10X *rTth* buffer, approximately 10ng template DNA, 25pmol dNTPs, 30pmol each primer, 0.4 μ L (1U) *rTth* polymerase, and 1 μ L of *Vent* polymerase diluted 1:100 (0.02U) per 50 μ L. Both polymerases are from Applied Biosystems. PCR products were verified, and when necessary size selected, using 1% agarose gels. Single-banded products were purified and single A-overhangs added as above. A-tailed fragments were cloned into the pGEM-T Easy vector (Promega). Initial clone sequencing used the plasmid primers T7 and SP6; complete bidirectional sequencing was accomplished by primer walking, resulting in an average sequencing coverage of 7.8X.

Amplification of the *cob-mLSU* region in *Riftia*, which presumably contains UNK, was difficult. Part of this remaining region was sequenced by designing degenerate primers to *nad4* sequences obtained from the complete genomes of *Lumbricus terrestris*, *Platynereis dumerilii*, and *Katharina tunicata*. These primers (*nad4f*, TGR GGN TAT CAR CCN GAR CG and *nad4r*, GCY TCN ACR TGN GCY TTN GG) amplified a short region of *nad4*, and allowed the design of primers specific to *R. pachyptila* (*Rpnad4bf* and *Rpnad4br*). Using EXL polymerase (Stratagene), the primer combination *Rpnad4bf*/16Sbr-H (Palumbi 1996) amplified the region spanning *nad4-mLSU*, but the region between *cob* and *nad4*, which again was presumed to contain UNK, was still difficult to amplify and could not be cloned successfully after amplification. Three clones containing spliced PCR amplicons for this fragment (see Results) were partially sequenced and provided the remainder of *cob* as well as complete *trnW* and *atp6*

genes. For simplicity, the *R. pachyptila* fragment will henceforth be referred to as the *R. pachyptila* genome.

Genomic Assembly

Assembled sequences were checked by BLAST (Altschul et al. 1990) searches against GenBank. Those sequences that returned strong BLAST hits to mitochondrial protein-encoding genes were translated into amino acids using the *Drosophila* mitochondrial code and aligned in CLUSTAL X (Thompson et al. 1997) with other available lophotrochozoan genome sequences (Table 1.1) obtained from GenBank to ensure correct identification. The full genomes were assembled by resolving ambiguous sequence reads in AutoAssembler (Applied Biosystems), checking against the amino acid alignments, and concatenating the individual alignments to make the complete genome alignment in MacClade 4.03 (Maddison and Maddison 2000).

Candidate tRNA genes were found using the tRNAScan-SE web server (<http://www.genetics.wustl.edu/eddy/tRNAScan-SE>); this identified all but four tRNAs in *C. torquata* and one in *R. pachyptila*. Stretches of mtDNA that did not code for protein genes and were in a similar position to tRNAs in previously published annelid genomes were scanned by eye for potential tRNA secondary structure and the presence of the anticipated anticodon sequence. The tRNA structures reported here are proposed based on the tRNAScan-SE foldings, keeping in mind the general forms suggested by Dirheimer et al. (1995). rRNA genes were identified by sequence homology with BLAST entries, and 5' and 3' ends were assumed to be directly adjacent to up- and

downstream genes. The boundaries of the *C. torquata UNK* were similarly inferred from the ends of the upstream and downstream tRNAs.

Phylogenetic Analysis

Table 1.1 lists the taxa and their GenBank accession numbers used for phylogenetic inference. Outgroups were chosen based on knowledge of Lophotrochozoan evolutionary history (Halanych 2004). Because I hoped to develop a better understanding of the utility of mtDNA in constructing annelid phylogeny, I chose to subsample available lophotrochozoan mtDNA genomes for use as outgroups. For mollusks, I chose the polyplacophoran *Katharina tunicata* for its basal position, the two pulmonate gastropods *Albinaria caerulea* and *Cepaea nemoralis* because they were more easily aligned than other gastropods, and the cephalopod *Loligo bleekeri* to achieve a broader representation of mollusks. Several other molluscan genomes contained large insertions and deletions in several genes relative to annelids, greatly complicating attempts at alignment. All three available brachiopods (*Terebratalia transversa*, *Terebratulina retusa*, and *Laqueus rubellus*) were included in the analyses. To create the final alignment, DNA from protein-encoding genes was aligned in MacClade 4.03 using CLUSTALX alignments of the corresponding amino acids; rRNA genes were aligned manually using secondary structure as a guide, employing phylogenetic conservation diagrams obtained from the RNA database at the University of Texas's Institute for Cellular and Molecular Biology (<http://www.rna.icmb.utexas.edu/topmenu.html>). tRNAs, *UNK*, and non-coding DNA were not included in the alignments due to high

variability (see below). This produced a single multi-partitioned alignment in MacClade 4.03, which is available at TreeBase (<http://www.treebase.org>) and in the supplementary information.

Two sequence-based datasets and one gene-order dataset were created. One sequence-based dataset contained nucleotide sequences from protein-encoding and rRNA genes, and the second contained only inferred amino acid sequences. Regions that could not be unambiguously aligned, and all third codon positions were removed. The amino acids of three protein-coding genes (*atp6*, *atp8*, *nad6*) exhibited high variation, which made alignment difficult, and thus were excluded from both datasets.

All non-annelid taxa herein were treated as outgroups; however, brachiopods are drawn basally for illustrative purposes. Although mollusks, annelids, brachiopods, and sipunculids are closely related, the relationships between them are not well resolved (Halanych 2004). PAUP*4.0b10 (Swofford 2002) was used for parsimony and maximum likelihood (ML) analyses. For both datasets, gaps were treated as missing data. For the DNA dataset, maximum likelihood models and their parameters were determined with hierarchical likelihood ratio tests (hLRT's) using the program MODELTEST 3.5 (Posada and Crandall 1998). Heuristic searches in PAUP under both parsimony and ML employed random sequence addition (parsimony—100 replicates; likelihood—10 replicates) to obtain starting trees, and TBR swapping. Bootstrapping with character re-sampling was performed with 1000 replicates for parsimony and 500 replicates for ML. Decay indices (also called Bremer support, Bremer 1994) were also calculated for the parsimony trees using constraints in PAUP.

The order of genes in the mitochondria was used as a third dataset for phylogenetic analysis. Although breakpoint analysis (Blanchette, Kunisawa, and Sankoff 1999) has proven useful in many cases, I prefer a newer parsimony framework (described in Boore and Staton 2002), which does not condense the data into pairwise distance measures, and allows partial genomes to be included. Briefly, 74 multistate characters were created (“upstream of gene X” and “downstream of gene X” for each of the 37 genes), and character states were coded as “beginning of gene Y” and “end of gene Y”, for a total of 74 states (though obviously a gene cannot appear up- or downstream of itself). The matrix was then analyzed in PAUP under parsimony as previously outlined. Because the gene orders of four taxa (*P. gouldii*, *G. brachiosum*, *H. robusta*, *R. pachyptila*) are incompletely known, missing and ambiguous characters (52) were removed before searching for trees, leaving 22 characters. The brachiopods were again placed as the basal-most outgroup. For comparative purposes, breakpoint and inversion distances were calculated using GRAPPA 1.6 (Bader, Moret and Yan 2001).

RESULTS

Genomic Composition

The complete mitochondrial DNA (mtDNA) of *C. torquata* is 15,538 bp in length, and the *R. pachyptila* fragment is 12,016 bp long. Figure 1.1 shows the gene order for both genomes. The *C. torquata* genome is similar in size (i.e., about 15kb) to other lophotrochozoan mitochondrial genomes, and the portion of the *R. pachyptila* genome is

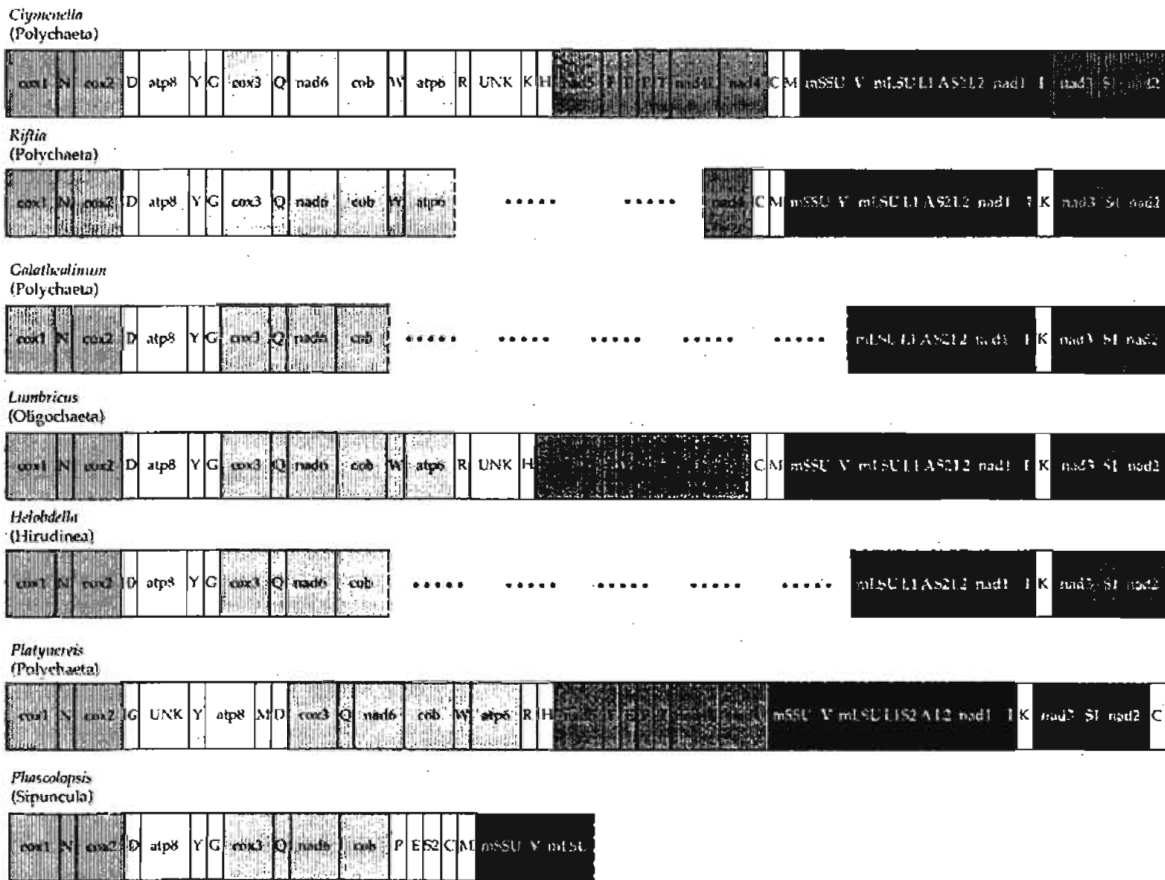


Figure 1.1 Gene orders of annelid and the sipunculan mitochondrial genomes. Abbreviations are as explained in the text. Genomes have been arbitrarily linearized at *cox1* after Boore and Brown (2000). Dashed lines with ellipses in *Riftia*, *Galathealinum*, *Helobdella*, and *Phascolopsis* indicate unsequenced regions whose gene order is unknown. Shaded boxes highlight different sets of gene orders conserved among the taxa shown.

of similar size to the same portions from *C. torquata* and *L. terrestris*. Tables 1.2 and 1.3 show a breakdown of nucleotide composition for *C. torquata* and *R. pachyptila*, respectively. Both genomes show patterns of nucleotide bias and skew¹. The two genomes are AT-rich (~66%), and this bias is consistent across the three main gene types

¹ Herein, “nucleotide bias” refers to unequal nucleotide frequencies (i.e., departures from 25% each) and “codon bias” to unequal frequencies of the codons that code for a single amino acid (e.g., UUA used for leucine more often than UUG). “Skew” will refer specifically to the orientation of hydrogen-bonded pairs in the molecule (e.g. whether the coding strand contains the G of a GC pair or the C).

	Protein Coding				rRNA	tRNA	Whole Genome
	All Positions	1st Positions	2nd Positions	3rd Positions			
A	31.17%	32.19%	18.19%	43.11%	38.99%	36.17%	32.96%
T	35.03%	27.45%	43.81%	33.83%	29.49%	31.74%	34.28%
A+T	66.20%	59.64%	62.00%	76.94%	68.48%	67.91%	67.24%
C	20.66%	19.97%	24.59%	17.45%	17.49%	15.73%	19.46%
G	13.14%	20.40%	13.41%	5.61%	13.99%	16.36%	13.30%
AT-skew	-0.06	0.08	-0.41	0.12	0.14	0.07	-0.02
GC-skew	-0.22	0.01	-0.29	-0.51	-0.11	0.02	-0.19
bp	11146	3716	3715	3715	2116	1424	15538

Table 1.2 Base composition, bias, and skew for *C. torquata*

(those coding for proteins, tRNAs, and rRNAs). T is the most common base, and G the least. Further, the percentage of G's is markedly lower at third codon positions than even the low overall G frequency. In contrast to nucleotide bias, patterns of AT- and GC-skew are not as consistent across gene types. Skew for a given strand is calculated as $(A-T)/(A+T)$ [or $(G-C)/(G+C)$] (Perna and Kocher 1995) and ranges from +1 if the coding strand has A (G) for every AT (GC) pair to -1 if the coding strand always has T (C). On

	Protein Coding				rRNA	tRNA	Whole Genome
	All Positions	1st Positions	2nd Positions	3rd Positions			
A	29.53%	28.66%	17.42%	42.51%	38.04%	34.25%	31.44%
T	36.48%	29.48%	43.39%	36.58%	28.49%	31.70%	34.67%
A+T	66.00%	58.14%	60.80%	79.10%	66.53%	65.95%	66.12%
C	21.90%	21.61%	25.63%	18.45%	19.39%	16.59%	20.99%
G	12.10%	20.25%	13.57%	2.46%	14.08%	17.47%	12.89%
AT-skew	-0.11	-0.01	-0.43	0.07	0.14	0.04	-0.05
GC-skew	-0.29	-0.03	-0.31	-0.76	-0.16	0.03	-0.24
bp	8806	2938	2935	2933	2145	1019	11987

Table 1.3 Base composition, bias, and skew for *R. pachyptila*

the whole AT-skew is slightly negative, and GC-skew is more negative than AT-skew. In both genomes, AT-skew is most positive in 2nd codon positions, and GC-skew is most negative at 3rd codon positions.

The genome of *C. torquata* contains the standard 37 genes found in mtDNAs: 13 protein-coding genes, 2 genes for rRNAs, and 22 genes for tRNAs (Boore 1999). The *R. pachyptila* fragment contains 9 complete protein-coding genes (*atp8*, *cox1*, *cox2*, *cox3*, *cob*, *nad1*, *nad2*, *nad3*, *nad6*) and portions of two others (*atp6*, *nad4*), as well as both rRNA genes (*mLSU*, *mSSU*) and 16 tRNA genes (*trnA*, -C, -D, -G, -I, -K, -L1, -L2, -M, -N, -Q, -S1, -S2, -V, -W, -Y); the remaining genes (*nad4L*, *nad5*, and *trnE*, -F, -H, -P, -R, -T) and the *UNK* are presumably in the unsequenced portion. As seen in all other annelids to date, all genes in both genomes are encoded on a single strand.

Start and stop codon usage also shows patterns of bias. Start codons in protein-coding genes are highly biased towards ATG over ATA; ATG is observed in 12 of 13 coding genes in *C. torquata* (*nad4* uses ATA) and all 10 *R. pachyptila* coding genes for which the 5' end is known. In addition, overlap typically exists between the presumptive stop codon (TAA or TAG) and the 5' end of the next gene. In other words, some stop codon bases appear to be part of the transcript of the down stream gene (illustrated in Supplementary Information). For the purposes of annotation, the stop codon in all such cases is assumed to be incomplete (see Ojala, Montoya and Attardi 1981), and the shared bases assigned to the downstream gene.

There is considerable codon usage bias in both genomes as well, with some codons within a group being used more than an order of magnitude more frequently than others (Table 1.4). In codons that exhibit four-fold degeneracy, triplets ending in G tend to be the least used as expected from overall nucleotide frequencies. However, codons ending in A tend to be the most common within a codon group despite the slightly higher prevalence of T's in nucleotide frequency. In 2-fold codon groups, the use of XXG tends to be considerably less than XXA, and use of XXC is somewhat less than XXT. CCG (Pro) and CGG (Arg) were never observed in *R. pachyptila*.

Codon	AA	<i>C. torquata</i>		<i>R. pachyptila</i>		Codon	AA	<i>C. torquata</i>		<i>R. pachyptila</i>	
		N	%	N	%			N	%	N	%
UUU	Phe (F)	211	75	143	61	UCU	Ser (S)	70	33	83	38
UUC	Phe	70	25	92	39	UCC	Ser	37	18	50	23
		281		235		UCA	Ser	99	47	83	38
UUA	Leu (L)	218	95	191	97	UCG	Ser	3	1	3	1
UUG	Leu	12	5	5	3			209		219	
		230		196		CCU	Pro (P)	70	39	79	49
CUU	Leu (L)	108	33	116	40	CCC	Pro	20	11	34	21
CUC	Leu	51	16	45	15	CCA	Pro	77	43	47	29
CUA	Leu	150	46	127	43	CCG	Pro	12	7	0	0
CUG	Leu	14	4	5	2			179		160	
		323		293		ACU	Thr (T)	64	26	75	41
AUU	Ile (I)	238	73	196	77	ACC	Thr	42	17	41	22
AUC	Ile	86	27	59	23	ACA	Thr	132	54	65	36

	324		255		ACG Thr	5	2	2	1
AUA Met (M)	224	90	138	84		243		183	
AUG Met	25	10	26	16	GCU Ala (A)	87	35	69	36
	249		164		GCC Ala	58	23	51	27
GUU Val (V)	46	29	34	26	GCA Ala	96	39	68	36
GUC Val	21	13	19	15	GCG Ala	6	2	2	1
GUA Val	83	52	76	58		247		190	
GUG Val	9	6	2	2	UGU Cys (C)	15	11	18	15
	159		131		UGC Cys	16	12	13	11
UAU Tyr (Y)	76	65	66	74	UGA Trp (W)	80	61	85	73
UAC Tyr	41	35	23	26	UGG Trp	20	15	1	1
	117		89			131		117	
UAA Ter (.)	7	78	3	75	CGU Arg (R)	13	22	10	19
UAG Ter	2	22	1	25	CGC Arg	3	5	5	9
	9		4		CGA Arg	38	66	38	72
CAU His (H)	53	64	49	69	CGG Arg	4	7	0	0
CAC His	30	36	22	31		58		53	
	83		71		AGU Ser (S)	15	15	6	9
CAA Gln (Q)	71	93	56	97	AGC Ser	12	12	6	9
CAG Gln	5	7	2	3	AGA Ser	64	62	53	78
	76		58		AGG Ser	12	12	3	4
AAU Asn (N)	80	54	72	67		103		68	
AAC Asn	67	46	35	33	GGU Gly (G)	36	19	30	19
	147		107		GGC Gly	37	19	19	12
AAA Lys (K)	79	95	62	95	GGA Gly	68	35	101	63

AAG Lys	4	5	3	5	GGG Gly	52	27	10	6
	83		65			193		160	
GAU Asp (D)	33	52	28	53					
GAC Asp	30	48	25	47					
	63		53						
GAA Glu (E)	57	80	59	97					
GAG Glu	14	20	2	3					
	71		61						
TOTAL						3578		2932	

Table 1.4 Codon usage. Stop codons are only listed if complete. AA=Amino Acid, N=number of occurrences in all protein-encoding genes observed.

Putative tRNA structures are depicted for all recovered tRNA genes in Figures 1.2 and 1.3 (*C. torquata* and *R. pachyptila*, respectively). Most possess the common cruciform structure, with an acceptor arm, anticodon arm, TΨC arm, DHU arm, and associated loop regions. In *C. torquata*, *trnS2* and *-V* have shortened TΨC stems, and *trnN* in *R. pachyptila* is missing the TΨC entirely. Additionally, *trnS1* and *-S2* in *R. pachyptila* have no DHU stems. *trnS1* and *-S2* are shown without DHU stems despite the potential for some base pairing; the lack of DHU stems is a widespread feature of mitochondrial tRNA genes (Dirheimer et al. 1995). Also of interest is the single unpaired nucleotide on the 5' side of the acceptor arm of the *trnL2* gene of *R. pachyptila*, confirmed in three independent sequencing reads.

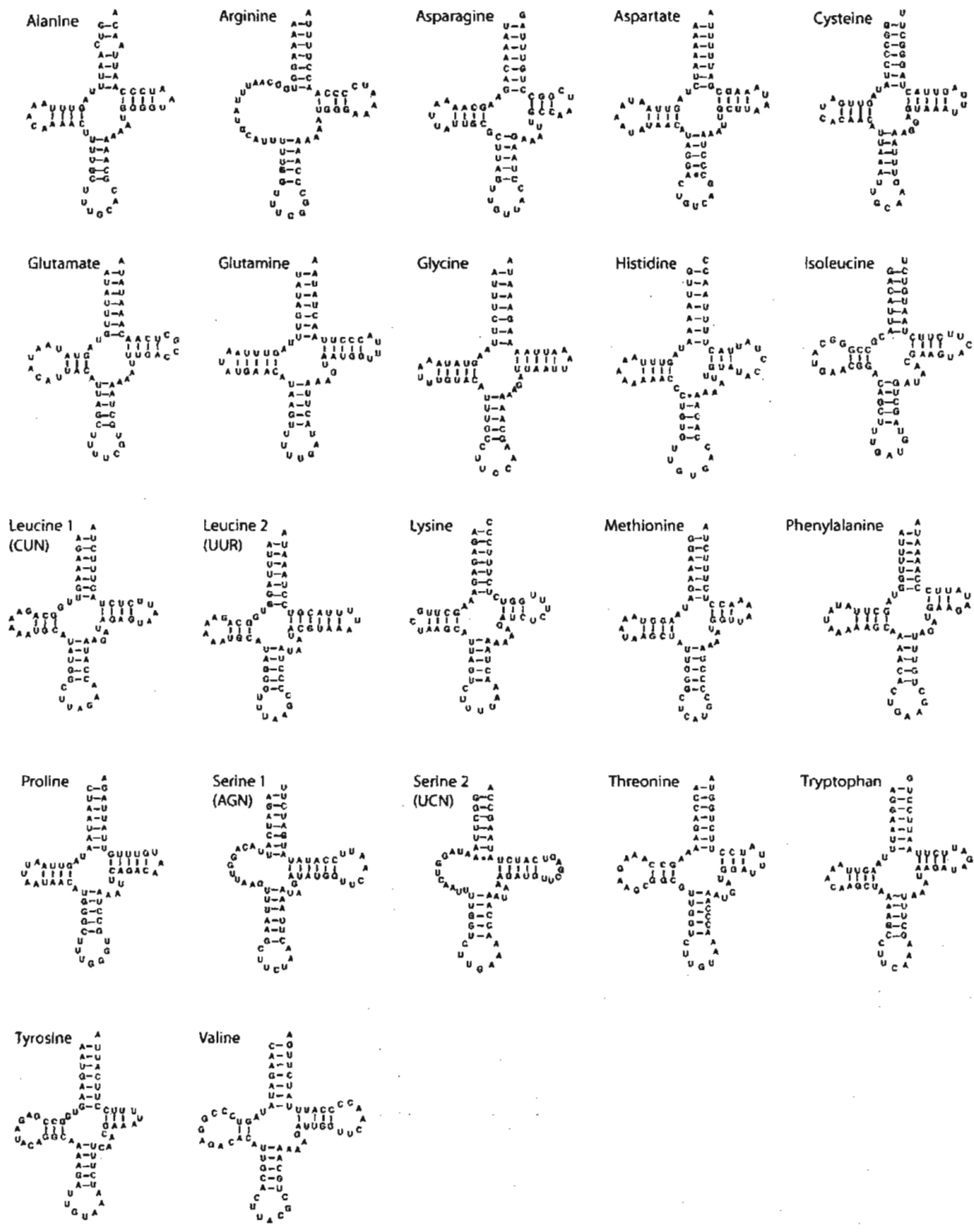


Figure 1.2 *Clymenella torquata* assumed tRNA structures

Phylogenetic analyses

A single shortest tree was recovered under parsimony for both the DNA and AA datasets. The DNA tree is shown in Figure 1.4a (16,680 steps, C.I. =0.549), and the AA tree in Figure 1.4b (12,645 steps, C.I. =0.756). Monophyly of the Annelida was recovered in both trees, as both topologies are consistent with a monophyletic Brachiopoda (100% bootstrap support in both analyses). Also in both trees, *P. gouldii* is sister to Annelida, and the two siboglinids (*R. pachyptila* and *G. brachiosum*) cluster together. There are two main differences between the trees. In the DNA tree, the oligochaete and hirudinean fall outside of the polychaetes, whereas in the AA tree they are inserted among polychaetes. The arrangement of mollusks also differs between the two trees. In the DNA tree, the mollusks are monophyletic with the polyplacophoran basal, the two gastropods together, and the cephalopod in the most derived position. In

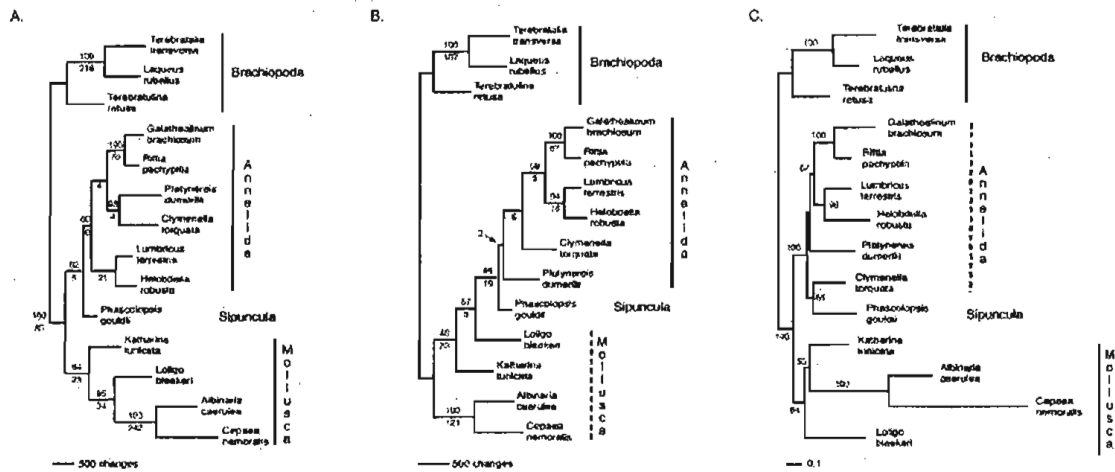


Figure 1.4 Phylogenetic reconstructions. A. The single best DNA sequence parsimony tree (protein-coding genes and rRNA; see text for details). B. The single best amino-acid parsimony tree. Numbers above branches are bootstrap percentages out of 1000 replicates (percentages below 50 not shown). Numbers below branches are Bremer support values (decay indices). C. DNA sequence maximum likelihood tree (model chosen via Modeltest 3.5). Numbers nearest the node indicate bootstrap percentage out of 500 replicates (percentages below 50 not shown).

the AA tree, the cephalopod and polyplacophoran are more closely related to the sipunculan and annelids (bootstrap support 83%) than to the gastropods.

For the ML nucleotide data, Modeltest chose the GTR+I+G model as the best fit to the data (nucleotide frequencies A=0.2557, C=0.1899, G=0.1942, T=0.3602; rates A↔C 1.6203, A↔G 3.4278, A↔T 1.6946, C↔G 2.4572, C↔T 3.6315, G↔T 1.000; proportion of invariable sites 0.1993, gamma shape parameter 0.8916). The single best maximum likelihood tree (-ln likelihood = 67626.63583) obtained with this model bears a strong similarity to both the DNA and AA trees (Figure 1.4c). Bootstrap support of 100% was found for an Annelida+Sipuncula clade with Sipuncula nested within the group as sister to the maldanid *Clymenella torquata*. Limited support (67% bootstrap support) for hirudineans and oligochaetes, the Clitellata, within polychaetes was also found.

The gene-order analysis produced 15 equally parsimonious trees of 112 steps. The strict consensus of these trees (Figure 1.5) contained far less resolution than the trees derived from nucleotide or amino acid sequences with only three supported nodes.

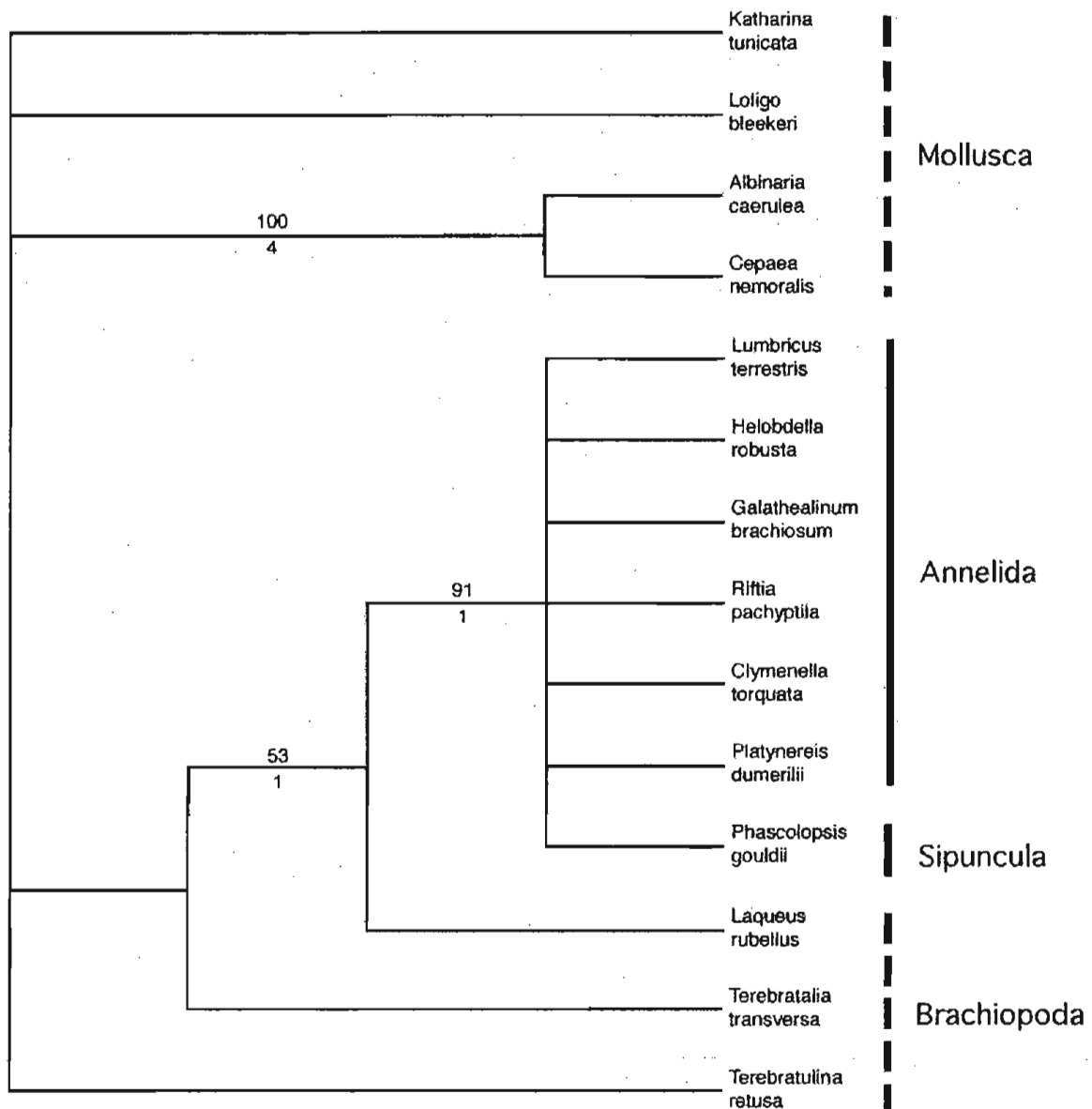


Figure 1.5 Gene Order Topology. The tree produced using the parsimony analysis of gene order from Boore and Staton (2002). Below is the resultant consensus tree with bootstrap support from 1000 replicates shown above the branches and Bremer (1994) values shown on the branches.

Consistent with other analyses, the two gastropods clustered together with 100% bootstrap support. Ninety-one percent support was also recovered for the node containing all annelids and *P. gouldii*. A grouping of this clade as sister to *L. rubellus* had weak support (53%). To determine if this lack of resolution was due to the parsimony method of analyzing gene order or intrinsic to the data, GRAPPA 1.6 breakpoint and inversion distances were also calculated. However, in these trees Brachiopoda and Mollusca interdigitated to a large degree (not shown). Neither algorithm can handle partial genomes; thus, *P. gouldii*, *G. brachiosum*, *H. robusta*, and *R. pachyptila* had to be excluded from these analyses, further reducing the phylogenetic inferences that could be made. It thus appears that all of these gene order algorithms are sensitive to the disparate rates of change present in our dataset.

DISCUSSION

The present study covers all major recognized clades of annelids (Rouse and Fauchald 1997). Annelid mitochondrial gene order appears to be evolutionarily conserved. With the exception of *trnK*'s placement in *C. torquata*, and as far as could be determined for *R. pachyptila*, both genomes examined here have the same gene order as *Lumbricus terrestris* and the fragments of *Galathealinum brachiosum* and *Helobdella robusta*. *Platynereis dumerilii* differs in the placement of the *UNK* region and a few tRNAs (Figure 1.1). In contrast to annelids, mollusks display considerable gene order variation over a similar timescale (e.g. Dreyer and Steiner 2004). For example, even within Gastropoda and Cephalopoda large numbers of rearrangements are common (e.g.

Kurabayashi and Ueshima 2000, Serb and Lydeard 2003). The three brachiopods also display very dissimilar gene orders. The origins of major taxa in these groups date back to the Cambrian (approximately 540 MYA) (Knoll and Carroll 1999). Thus, it appears that there may be a considerable difference in how annelid, mollusk, and brachiopod mitochondrial genomes evolve. This difference is interesting because of the apparent close relationship of these lophotrochozoan taxa. These results raise the possibility that gene order is highly variable in general across lophotrochozoan taxa, and that only select subgroups exhibit conserved gene orders (e.g., Annelida). If true, this situation may have considerable repercussions on how mtDNA gene order data can be used to infer evolutionary history among different animal clades.

Phylogenetic Relationships

The AA parsimony and DNA likelihood phylogenetic analyses are consistent with previous findings that place Clitellata (McHugh 1997; Rota, Martin and Erséus 2001; Bleidorn, Vogt and Bartolomaeus 2003) and siboglinids (McHugh 1997; Rouse and Fauchald 1997; Kojima 1998; Halanych et al. 1998; 2001) as derived “polychaetes”. Thus, the last common ancestor of “Polychaeta” and Annelida are one and the same. However bootstrap values (67% likelihood, <50% for AA parsimony) for this result were weak and Shimodaira-Hasegawa tests (Shimodaira and Hasegawa 1999), fell short of significant values (in both cases, $p=0.14$, 1000 replicates with RELL option). An alternative topology in the nucleotide parsimony analysis was not well supported. Clearly, considerably more taxa need to be sampled to understand the robustness of these results and placement of these groups within annelids. The groupings *R. pachyptila* + *G.*

brachiosum and *H. robusta* + *L. terrestris* were highly supported in all sequence analyses in agreement with morphological expectations. An additional result consistently recovered by sequenced-based analyses was placement of the sipunculan as sister to or inside Annelida (Shimodaira-Hasegawa test $p = 0.003$). Boore and Staton (2002) first reported this result using many of the same mtDNA sequences used herein. Thus, although gene order may be uninformative in this case, there is high support from both DNA and amino acid sequences for an Annelida/Sipuncula clade to the exclusion of mollusks and brachiopods. Interestingly, nuclear large ribosomal subunit data also weakly supports sipunculans as the sister clade to annelids (Passamaneck and Halanych, in prep). The likelihood tree provides the first suggestion that Sipuncula is within Annelida, but this finding requires additional verification.

In contrast to the sequence-based data sets, the gene order analysis offers little resolution. This result is to be expected with the limited observed variation in annelid gene order. Nonetheless, annelids and the sipunculan cluster together because of identical arrangement of the 11 genes between *cox1* and *cob* (inclusive) and the sequence *mSSU—trnV—mLSU*. The latter sequence appears to be somewhat conserved across lophotrochozoan clades (it is found in 10 of the 23 lophotrochozoan taxa for which data are currently available in GenBank), and potentially in other protostomes as well. Further, the subsequence *trnV—mLSU* is found in 16 of the 23 lophotrochozoan genomes, and some protostomes. In any case, based on the available data, gene order appears to be of limited utility for relationships within the annelids because of its highly conserved nature. All rearrangements seen so far are minor and found in single taxa only, although

with greater taxonomic coverage potential synapomorphic gene orders may emerge. Apparently, both within annelids and between annelids and other lophotrochozoans, there is no consistent mechanism controlling the rate or types of gene order modifications. In contrast to the lack of phylogenetic signal in gene order among annelids, the resolution offered by sequence-based analyses holds promise.

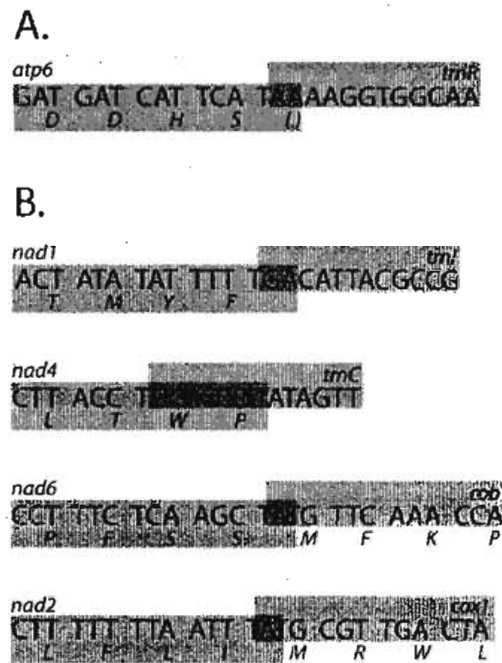


Figure 1.6 Overlapping genes and post-transcriptional splicing. A., An example of a complete in-frame stop codon from *C. torquata*. B., Four overlaps where no in-frame stop codon exists. First two overlaps from *C. torquata*; last two from *R. pachytila*.

Mitochondrial Genome Organization and Structure

The two genomes presented here also exhibit the pattern of post-transcriptional modification and splicing described by Ojala, Montoya, and Attardi *et al.* (1981), in which many stop codons are incomplete in the transcript and are filled in by post-

transcriptional editing machinery. This type of splicing is presumed to occur in several genes in both the *C. torquata* and the *R. pachyptila* genomes. In the majority of these cases, the overlap in question contains an in-frame stop codon (TAA or TAG), but it is not presumed to be functional. Moreover, in several cases there is no in-frame stop codon at or near the end of the protein-encoding gene, making post-translational addition of a stop codon the only plausible mechanism (Figure 1.6). One example is the *nad1/trnI* junction in *C. torquata*, where *nad1* presumably ends with T__, and *trnI* begins with GA, such that assigning more of the codon (TG_ or TGA) to *nad1* still does not produce a stop codon. Additionally in *C. torquata*, the last six bases of *nad4* (GGCCCT) appear to be used as the first six of *trnC*; a seven-base overlap could give *nad4* an incomplete TA_ stop codon, but the next base is a T, and therefore it is not possible to generate a full stop codon from the primary sequence.

The AT-bias seen in both genomes seems to be contributing to a strong codon bias in protein-coding genes. Although the *R. pachyptila* genome is incomplete, the absence of two GC-rich codons (CCG, encoding proline, and CGG, encoding arginine) may be linked to the low percentage of G and C. However, even given these low frequencies in the protein-coding genes as a whole, the probability of never observing CCG (Pro) in 160 proline codons given an average G content of 12% is $(0.12)^0(1-0.12)^{160} = 1.31 \times 10^{-9}$, and the probability of never seeing CGG (Arg) in 53 arginine codons is $(0.12)^0(1-0.12)^{53} = .0011$ (both assuming independence of codons). Thus, the amount of AT-bias alone does not adequately explain the lack of these two codons and suggests that some other mechanism(s) is responsible for the observed codon

bias. Cardon *et al.* (1994) discuss the paucity of CG dinucleotides in metazoan mitochondrial genomes regardless of their position in codons (i.e. positions (I,II), (II,III), and (III,I)) and overall low usage of arginine (CGN) in mitochondrial proteomes. Indeed, arginine is the least frequent of all amino acids possessing four-fold degenerate codons in both *C. torquata* and *R. pachyptila*, and is even less frequent than some two-fold degenerate amino acids. Based on the symmetrized odds-ratios (ρ_{NN} , where NN is the dinucleotide in question) of Cardon *et al.* (1994), *R. pachyptila* does show CG suppression ($\rho_{CG}=0.5299$; $0.78 \leq \rho_{NN} \leq 1.23$ is considered the normal range). Suppression of CG dinucleotides in vertebrate nuclear genomes has been linked to mutation to TG by methylation of the C followed by deaminization to T. This cannot underlie CG suppression in mtDNAs because mitochondria lack the methylation pathway, and because mtDNAs do not usually contain an excess of TG dinucleotides (*R. pachyptila* $\rho_{TG}=0.83$). Although no simple explanation has been found, the authors suggest that CG suppression is correlated with small genome size and "streamlined" mtDNA organization.

R. pachyptila is a large tubeworm found at Eastern Pacific hydrothermal vent fields. Early genetic analyses on this species led to speculation that hydrothermal vent animals would harbor a high GC nucleotide composition because the extra hydrogen bond, when compared to AT base pairing, would confer additional stability in the potentially high-temperature and reducing environment (Dixon, Simpson-White, and Dixon 1992). Although high GC content has been documented in thermophilic microbes (Woese *et al.* 1991), *R. pachyptila*'s low GC content (a pervasive feature of metazoan

mtDNAs in general) argues against such temperature-driven evolution in *R. pachyptila*. Possibly, the higher GC content in *R. pachyptila* postulated by Dixon and colleagues is restricted to the nuclear genome; however, it is unclear why mitochondrial and nuclear genomes would respond in different ways to the same environmental pressure if this were true.

Genomic Amplification and Sequencing

Our difficulties in amplifying and cloning the *UNK* region of *C. torquata* and *R. pachyptila* likely stem from regulatory aspects of this region of the molecule. In *R. pachyptila*, our long PCR reactions for the region *cob-nad4* repeatedly generated 3-5 bands, even though the reactions employed two ~30mer species-specific primers. Attempts to clone the band of the expected size resulted in very low transformation efficiencies. Of three clones sequenced, each contained an apparent splice in a similar, but not exact, position just downstream of *atp6*, indicating host removal of the genes between *atp6* and *nad4* (presumably containing *trnW*, *UNK*, *trnH*, *nad5*, *trnF*, *-E*, *-P*, and *-T*). Sequencing of the 3' end of these clones provided the complete gene sequences for *trnW* and *atp6* but the splice prevented accurate determination of what lay farther downstream. A similar region was apparently unclonable in the sheared fragments of *C. torquata*'s mt-genome and had to be obtained by direct sequencing. Boore and Brown (2000) had similar problems when obtaining the similar region in *Platynereis dumerilii*, and suggested that the presence of signaling elements in *UNK* disrupted PCR. Our observations suggest the *UNK* region is identifiable as a foreign origin of replication and

is spliced out by at least some *E. coli* cell types (in this case DH5a and JM109—both of which are *recA*⁻) in addition to possibly interfering with PCR. Alternative strategies may need to be developed to completely sequence large numbers of complete mitochondrial genomes in order to avoid the need to direct-sequence and primer-walk the region containing *UNK*.

CONCLUSIONS

I have expanded the phylogenetic spread of annelid taxa whose mitochondrial genomes have been sequenced. The high similarity of gene order across annelids provides sharp contrast to the variation observed in mollusks and brachiopods. In both cases, the phylogenetic utility of gene-order data may be limited. The nucleotide and amino acid data, however, produced informative trees with some measure of support. Our results are concordant with the findings of Boore and Brown (2000) and Boore and Staton (2002) on annelid relationships and the relation of Sipuncula to Annelida.

References

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.
- Bader, D. A., Moret, B. M. E. and Yan, M. 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.* **8**: 483-491.
- Blanchette, M., T. Kunisawa, and D. Sankoff. 1999. Gene order breakpoint evidence and animal mitochondrial phylogeny. *J. Mol. Evol.* **49**:193-203.
- Bleidorn, C., L. Vogt, and T. Bartolomaeus. 2003. A contribution to sedentary polychaete phylogeny using 18S rRNA sequence data. *J Zool. Syst. Evol. Res.* **41**:186-195.
- Boore, J. L. 1999. Animal mitochondrial genomes. *Nucl. Acid. Res.* **27**:1767-1780.
- Boore, J. L., and W. M. Brown. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: Sequence and gene rearrangement comparisons indicate the Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Mol. Biol. Evol.* **17**:87-106.
- Boore, J. L., and W. M. Brown. 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.* **8**:668-674.
- Boore, J. L., and W. M. Brown. 1994. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. *Genetics* **138**:423-443.
- Boore, J. L., and J. L. Staton. 2002. The mitochondrial genome of the Sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. *Mol. Biol. Evol.* **19**:127-137.

- Boore, J. L., M. Medina, and L.A. Rosenberg. 2004. Complete sequences of the highly rearranged molluscan mitochondrial genomes of the scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis*. *Mol Biol Evol* **21**: 1492-1503.
- Bremer, K. 1994. Branch support and tree stability. *Cladistics* **10**:295-304.
- Cardon, L. R., C. Burge, D. A. Clayton and S. Karlin. 1994. Pervasive CpG suppression in animal mitochondrial genomes. *Proc. Nat. Acad. Sci. USA* **91**: 3799-3803.
- DeJong, R. J., A. M. Emery, and C. M. Adema. 2004. The mitochondrial genome of *Biomphalaria glabrata* (Gastropoda, Basommatophora), intermediate host of *Schistosoma mansoni*. *J. Parasitol.* **in press**.
- Dirheimer, G., G. Keith, P. Dumas and E. Westhof. 1995. Primary, secondary, and tertiary structures of tRNAs. Pp. 93-126 *in* D. Söll and U. RajBhandary, eds. *tRNA: Structure, Biosynthesis, and Function*. ASM Press, Washington, DC.
- Dixon, D. R., R. Simpson-White, and L. R. J. Dixon. 1992. Evidence for thermal stability of ribosomal DNA sequences in hydrothermal-vent organisms. *J. Mar. Biol. Assoc. U.K.* **72**:519-527.
- Dreyer, H., and G. Steiner. 2004. The complete sequence and gene organization of the mitochondrial genome of the gadilid scaphopod *Siphonodontalium lobatum* (Mollusca). *Mol. Phylogenet. Evol.* **31**:605-617.
- Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech.* **3**:294-299.

- Grande, C., J. Templado, J. L. Cervera, and R. Zardoya. 2002. The complete mitochondrial genome of the nudibranch *Roboastra europaea* (Mollusca: Gastropoda) supports the monophyly of opisthobranchs. *Mol. Biol. Evol.* **19**:1672-1685.
- Halanych, K. M. 2004. The new view of animal phylogeny. *Ann. Rev. Ecol. Evol. Syst.* **35**:229-256.
- Halanych, K. M., T. G. Dahlgren, and D. McHugh. 2002. Unsegmented annelids? Possible origins of four lophotrochozoan worm taxa. *Integrat. Compar. Biol.* **42**:678-684.
- Halanych, K. M., R. A. Feldman, and R. C. Vrijenhoek. 2001. Molecular evidence that *Sclerolinum brattstromi* is closely related to vestimentiferans, not to frenulate pogonophorans (Siboglinidae, Annelida). *Biol. Bull.* **201**:65-75.
- Halanych, K. M., R. A. Lutz, and R. C. Vrijenhoek. 1998. Evolutionary origins and age of vestimentiferan tube-worms. *Cah. Biol. Mar.* **39**:355-358.
- Hatzoglou, E., G. C. Rodakis, and R. Lecanidou. 1995. Complete sequence and gene organization of the mitochondrial genome of the land snail *Albinaria caerulea*. *Genetics* **140**:1353-1366.
- Helfenbein, K. G., and J. L. Boore. 2004. The mitochondrial genome of *Phoronis architecta* - Comparisons demonstrate that phoronids are lophotrochozoan protostomes. *Mol. Biol. Evol.* **21**:153-157.
- Helfenbein, K. G., W. M. Brown, and J. L. Boore. 2001. The complete mitochondrial genome of the articulate brachiopod *Terebratalia transversa*. *Mol. Biol. Evol.*

18:1734-1744.

- Hoffmann, R. J., J. L. Boore, and W. M. Brown. 1992. A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. *Genetics* **131**:397-412.
- Knoll, A., and S. B. Carroll. 1999. Early animal evolution: Emerging views from comparative biology and geology. *Science* **284**:2129-2137.
- Kojima, S. 1998. Paraphyletic status of Polychaeta suggested by phylogenetic analysis based on the amino acid sequences of elongation factor-1-alpha. *Mol. Phylogenet. Evol.* **9**:255-261.
- Kurabayashi, A., and R. Ueshima. 2000. Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. *Mol. Biol. Evol.* **17**:266-277.
- Lavrov, D. V., W. M. Brown, and J. L. Boore. 2004. Phylogenetic position of the Pentastomida and (pan)crustacean relationships. *Proc. R. Soc. Lond. B.* **271**:537-544.
- Maddison, D. R., and W. P. Maddison. 2000. *MacClade*. Sinauer Associates, Inc., Sunderland, MA.
- Mangum, C. P. 1964. Studies on speciation in Maldanid polychaetes of the North American Atlantic Coast. II. Distribution and competitive interaction of five sympatric species. *Limnol. Oceanogr.* **9**: 12-26.
- McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proc. Natl. Acad. Sci. USA* **94**:8006-8009.
- McHugh, D. 2000. Molecular Phylogeny of the Annelida. *Can. J. Zool.* **78**:1873-1884.
- Noguchi, Y., Endo, K., Tajima, F. and Ueshima, R. 2000. The mitochondrial genome of

- the brachiopod *Laqueus rubellus*. *Genetics* **155**: 245-259.
- Ojala, D., J. Montoya, and G. Attardi. 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**:470-474.
- Palumbi, S. R. 1996. Nucleic acids II: The polymerase chain reaction. Pp. 205-248 in D. M. Hillis, C. Mortiz, and B. K. Mable, eds. *Molecular Systematics*. Sinauer Associates, Inc., Sunderland, Massachusetts.
- Perna, N. T. and Kocher, T. D. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J Mol. Evol.* **41**: 353-358.
- Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**:817-818.
- Rota, E., P. Martin, and C. Erséus. 2001. Soil-dwelling polychaetes: enigmatic as ever? Some hints on their phylogenetic relationship as suggested by a maximum parsimony analysis of 18S rRNA gene sequences. *Contri. Zool.* **70**:127-138.
- Rouse, G. W., and K. Fauchald. 1997. Cladistics and polychaetes. *Zool. Scripta* **26**:139-204.
- Rouse, G. W., and K. Fauchald. 1995. The articulation of annelids. *Zool. Scripta* **24**:269-301.
- Rouse, G. W., and F. Pleijel. 2001. *Polychaetes*. Oxford University Press, New York.
- Serb, J. M., and C. Lydeard. 2003. Complete mtDNA sequence of the North American freshwater mussel, *Lampsilis ornata* (Unionidae): An examination of the evolution and phylogenetic utility of mitochondrial genome organization in Bivalvia (Mollusca). *Mol. Biol. Evol.* **20**:1854-1866.

- Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of Log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**:1114-1116.
- Southward, A. J. and E. C. Southward. 1988. Pogonophora: Tube-worms dependent on endosymbiotic bacteria. *Anim. Plant Sci.* **1**: 203-207.
- Stechmann, A., and M. Schlegel. 1999. Analysis of the complete mitochondrial DNA sequence of the brachiopod *Terebratulina retusa* places Brachiopoda within the protostomes. *Proc. Roy. Soc. Lond. Ser. B.* **266**:2043.
- Swofford, D. L. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Inc., Sunderland, MA.
- Terrett, J. A., S. Miles, and R. H. Thomas. 1996. Complete DNA sequence of the mitochondrial genome of *Cepaea nemoralis* (Gastropoda: Pulmonata). *J. Mol. Evol.* **42**:160-168.
- Thompson, J., T. Gibson, F. Plewniak, F. Jeanmougin, and D. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acid. Res.* **25**:4876-4882.
- Tomita, K., Yokobori, S. I., Oshima, T., Ueda, T. and Watanabe, K. 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: multiplied noncoding regions and transposition of tRNA genes. *J Mol. Evol.* **54**: 486-500.
- Wilding, C. S., P. J. Mill, and J. Grahame. 1999. Partial sequence of the mitochondrial genome of *Littorina saxatilis*: Relevance to gastropod phylogenetics. *J. Mol. Evol.* **48**:0348-0359.

Woese, C.R., L. Achenbach, P. Rouviere and L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14:364-371.

Chapter 3: Assessment of the Cape Cod phylogeographic break using the bamboo worm *Clymenella torquata* (Annelida: Maldanidae)

ABSTRACT

Phylogeographic breaks are important in creating and maintaining genetic structure in populations of coastal marine benthic invertebrates. Previous genetic studies have suggested Cape Cod, Massachusetts, USA as a phylogenetic break; however, diffuse sampling in this area has hindered fine-scale determination of the break's location, with different species exhibiting breaks in different places, and others exhibiting no breaks in this region. I present phylogeographic patterns based on two mitochondrial genes from ten populations of the bamboo worm *Clymenella torquata* (Annelida: Maldanidae) focused around Cape Cod but extending from the Bay of Fundy, Canada to central New Jersey, USA. A common invertebrate along the US coast, *C. torquata* possesses a short planktonic larval period of about 3 days, a short lifespan of only a few years, and synchronous reproduction, making it sensitive to factors such as dispersal barriers, bottlenecks, and founder events. Both gene regions show a cline of haplotype frequencies from north to south and a phylogenetic break south of Cape Cod. The closer spacing of sampled populations on Cape Cod, combined with other sampled populations, place this break to the south of Cape Cod instead of at the tip of the peninsula. In addition, an imprint of founder events, presumably caused by glacial eradication of *C. torquata* populations in the north, can be seen in the reduced genetic diversity of northern sites.

INTRODUCTION

Physical features of the environment are known to be important in creating and maintaining genetic structure in coastal marine benthic invertebrates. Well-documented genetic breaks include the Floridean peninsula on the east coast of the U.S.A. (e.g. Avise 1992) and Point Conception on the west coast (reviewed in Burton 1998).

Phylogeographic investigations of coastal invertebrate species in the northwestern Atlantic Ocean have also pointed to barriers to gene flow near Cape Cod, Massachusetts (reviewed in Wares 2002). Species with a broad range of life history characteristics, from strongly dispersive (e.g. Bastrop et al. 1998) to weakly dispersive

(e.g. Wares and Cunningham 2001), show genetic discontinuities in this region. However, sampling around Cape Cod has typically been diffuse, hindering fine-scale determination of the location of breaks. Samples often include only a single population representing Cape Cod, with the next sample to the south as far away as Chesapeake Bay or North Carolina. A consensus has yet to emerge among these studies; however, several species have exhibited breaks between Cape Cod and the next sampled site to the south, rather than on the peninsula itself (e.g. Franz et al. 1981, Dillon and Manzi 1992, Vogler and Desalle 1993, Bastrop et al. 1998, Lee 1999). Not all species show evidence of a phylogeographic break in this region at all; no break was observed from Nova Scotia to Virginia in the ocean quahog *Arctica islandica* (Dahlgren et al. 2000), from north of Nova Scotia to Florida in the slipper snail *Crepidula fornicata* (Collin 2001), or from Iceland to Virginia in the surfclam *Spisula solidissima solidissima* (Hare and Weinberg 2005). Given the wide spacing of many samples, it is unclear whether a phylogeographic break near Cape Cod is present in the majority of species or not. Alternatively, the peninsula may mark a transition zone where northern genetic types grade into southern types. Understanding why some species exhibit phylogenetic breaks in a region while others do not can help reveal the historical processes that create them and the present day forces that maintain them.

Several studies have found, especially for hard-substrate intertidal invertebrates, that populations in the northwestern Atlantic are a genetic subset of European populations (reviewed in Wares and Cunningham 2001). Clines in genetic diversity have also been found within the northwestern Atlantic, with northern populations (e.g. Iceland, Nova

Scotia) having higher genetic diversity than southern populations (e.g. in the Gulf of Maine and on Cape Cod; Dahlgren et al. 2000, Govindarajan et al. 2004); however, some clines show the opposite orientation (Bernatchez and Wilson 1998, Cunningham and Collins 1998, Hare and Weinberg 2005). Given these variations, fundamental questions remain as to how physical features of the environment, historical processes, and organisms' life histories interact to create genetic structure.

Although in a few cases these phylogeographic patterns have been attributed to selective gradients in particular genes, (e.g. Koehn et al. 1976, Smith et al. 1998), in general three broad environmental hypotheses have been suggested to explain observed phylogeographic patterns in the Northwestern Atlantic (cf. Wares 2002): 1) circulation patterns of coastal currents, 2) differences in water mass characteristics north and south of Cape Cod, and 3) the historical influence of glaciation during the Pleistocene era. Because organisms' interactions with their environment are seldom simple, some combination of these (and other) hypotheses may ultimately provide the best explanation of observed genetic structure. Nevertheless, because these three hypotheses predict different phylogeographic patterns, they are useful end-members for interpreting more complicated scenarios.

Coastal currents hypothesis: Currents are frequently thought to underlie genetic structure in benthic marine invertebrates, because many benthic species possess planktonic larvae. A current that flows perpendicular to the path connecting two populations may also impede the flow of organisms between them (but see Palumbi et al. 1997). In the northwestern Atlantic, the Gulf of Maine circulation brings cold water

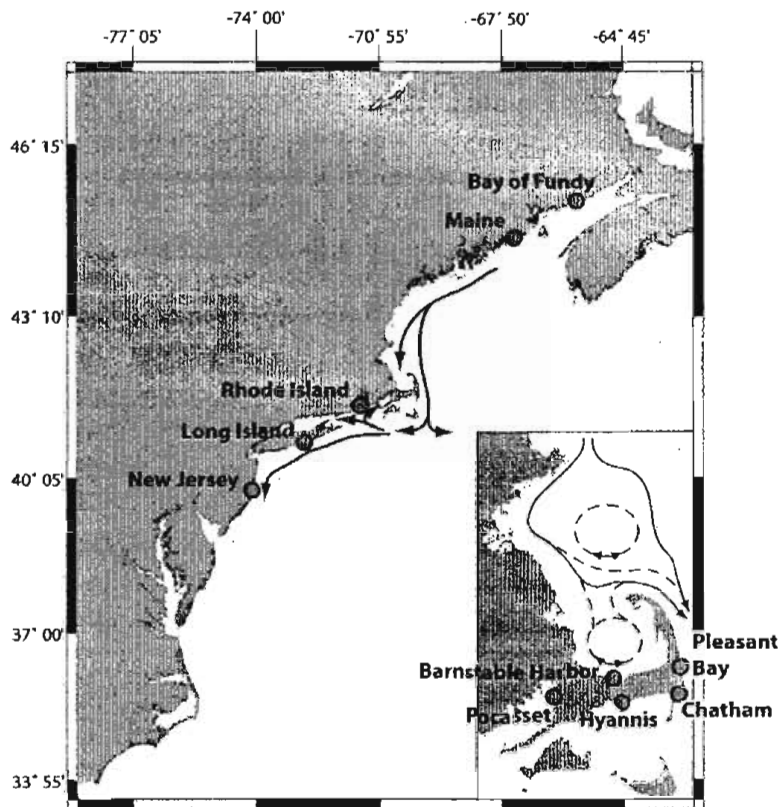


Figure 2.1 Sampling locations and coastal currents in the Northwest Atlantic. Gulf of Maine circulation adapted from Lynch et al. (1997); Massachusetts Bay and Cape Cod Bay circulation redrawn from Lermusiaux et al. (2001); Middle Atlantic Bight circulation based on Churchill (1985) and Spaulding and Gordon (1982).

south along the New England coast, into Cape Cod Bay, and south along the Atlantic coast of Cape Cod (Lynch et al. 1996, Lynch et al. 1997, Lermusiaux 2001). Although coastal currents south of Cape Cod are tidally driven and strongly affected by weather conditions, the mean flow is southwestward along the south coast of Cape Cod, and westward along both coasts of Long Island (Figure 2.1; Spaulding and Gordon 1982, Churchill 1985, Vieira 2000). Since the mid 1930's, the Cape Cod Canal has provided a potential alternative path between sites on either side of Cape Cod. Although no studies have explicitly examined the effect the Canal might have on population genetics, any effects are likely to be small because there is likely little net water flow through the Canal

(Anraku, 1964). If currents are structuring gene flow, a distinct polarity should be present, with southward migrations occurring more frequently than northward migrations. Thus, the coastal currents hypothesis can be distinguished from the water-mass hypothesis by increased occurrence of gene flow from north to south.

Water mass hypothesis: Characteristics such as temperature and salinity have often been hypothesized to affect organismal distributions (Hutchins 1947, Valentine 1966). The correlation of water temperature and salinity differences with species boundaries led researchers (e.g. Hutchins 1947, Hayden and Dolan 1976, Engle and Summers 1999) to define Cape Cod's southern coast as the boundary between the Acadian coastal biogeographical province to the north (extending to the northeast corner of Maine) and the Virginian province to the south (extending to the southern edge of Virginia). Waters in the Acadian province are uniformly colder than those of the more variable Virginian, and differences in average salinity are also known (Hayden and Doylan 1976, Engle and Summers 1999). Although this biogeographic boundary could coincide with a phylogeographic boundary, the two are not identical *a priori*, and since some species' ranges cross the Acadian/Virginian boundary, even the biogeographic boundary in this region may not be absolute. Nevertheless, the Acadian/Virginian line seems to mark the northern limit of species' ranges more frequently than it marks species' southern limits (Engle and Summers 1999), suggesting that species from the Virginian are unable to persist in the colder, saltier Acadian waters. If gene flow patterns within species found on both sides of Cape Cod mirror these patterns in species distribution,

populations on Cape Cod should be more closely related to northern than to southern populations, and gene flow from Cape Cod southward should be limited.

Historical glaciation hypothesis: If population genetics in the northwestern Atlantic have been shaped by Pleistocene ice sheets, differences in genetic diversity between locations should be apparent. Intertidal invertebrates that could not survive glaciation and the associated climatic changes would have disappeared from Long Island and Cape Cod northwards, or they would have been forced into refugia outside of the ice sheets (Pielou 1991, Holder et al. 1999). Soft-substrate invertebrates might also have shifted ranges south (or contracted their ranges if already present in the south), but in general, organisms eradicated from northern sites by glaciation would have been able to recolonize glaciated sites only since the ice sheets melted (20,000-18,000 years ago; Pielou 1991). The founder event resulting from recolonization would result in lower genetic diversity in modern populations located within the northern extent of glacial ice, as melt-back allowed reintroductions to advance northward (e.g., Bernatchez and Wilson 1998, Cunningham and Collins 1998, Hare and Weinberg 2005). Alternatively, if northern sites were more easily recolonized from refugial pockets within the glacial region (e.g. Nova Scotia), present-day locations within the southern end of the glaciers' range would exhibit lower diversity (e.g., Dahlgren et al. 2000, Govindarajan et al. 2004). In this case, populations beyond the southern extent of glaciation would not show reduced diversity. Regardless of the direction of a diversity gradient, intertidal invertebrates that prefer sandy bottoms are unlikely to have survived the extreme

reduction of suitable habitat in glaciated northern regions dominated by rocky habitat, leading to gradients with reduced genetic diversity in northern populations.

I tested predictions of these hypotheses by examining the phylogeographic patterns of the intertidal bamboo worm *Clymenella torquata* (Annelida: Maldanidae) around Cape Cod. Several aspects of the life history of *C. torquata* make it well suited to discerning among the three gene flow hypotheses. *Clymenella torquata* is a common coastal invertebrate from New Brunswick, Canada to Florida, USA (Mangum 1962). Often forming dense aggregations, this worm lives head-down in tubes constructed of sand grains, and consumes ingested sediment and associated interstitial organisms near the anterior end of the tube. *Clymenella torquata's* short larval lifespan (only a few days, Newell 1951) likely allows only limited dispersal, which increases the chances of observing genetic discontinuities around any geographical impediments to dispersal. The short larval lifespan also implies that the rate of initial reintroduction of the species to previously glaciated locations would be slow, as well as the rate of subsequent spread of additional genetic types to those locations. Thus, I expect founder effects to be pronounced and persistent in *C. torquata*. *Clymenella torquata* also reproduces synchronously once per year and probably lives for only a few years (Mangum 1964), conditions which approximate conventional population genetics models. I have combined samples from several closely spaced populations along the coasts of Cape Cod with multiple samples to the north and south, in order to determine the location of phylogeographic breaks near Cape Cod at a finer scale. Analysis of these samples also

offers the opportunity to explore more fully the importance of physical barriers, water mass differences, and historical effects on phylogeography in the northwestern Atlantic.

METHODS AND MATERIALS

Sample collection and sequencing

The majority of *Clymenella torquata* populations were sampled from May to November 2002 (Table 2.1); these include the five Cape Cod populations and a

Location	Collection Date	Latitude	Abbreviation	N		
		Longitude		<i>atp6</i>	<i>nad4</i>	both
Bay of Fundy	July 2003	45° 06' 00.0" N 66° 24' 00.0" W	BF	15	15	15
Maine	January 2003	44° 57' 13.0" N 67° 09' 45.0" W	ME	28	28	28
Barnstable Harbor	May 2002	41° 42' 39.6" N 70° 19' 29.4" W	BH	28	28	27
Pleasant Bay	August 2002	41° 42' 24.7" N 69° 58' 24.1" W	PB	24	23	22
Chatham	August 2002	41° 40' 00.0" N 69° 58' 34.0" W	CH	9	9	9
Hyannis	August 2002	41° 37' 57.9" N 70° 19' 18.3" W	HY	14	13	11
Pocasset	September 2002	41° 40' 27.9" N 70° 38' 27.7" W	P	6	4	4
Rhode Island	September 2003	41° 26' 57.1" N 71° 27' 04.7" W	R	24	24	22
Long Island	September 2003	40° 47' 06.8" N 72° 47' 26.0" W	L	17	18	15
New Jersey	September 2002	40° 11' 11.6" N 74° 01' 50.8" W	NJ	24	14	13
Total				189	176	166

Table 2.1 Sampling and gene amplification information for sites in this study. N, number of individuals sequenced for *atp6*, *nad4*, or both genes.

population from New Jersey. A population from Maine was sampled in January 2003, one from the Bay of Fundy in July 2003, and one population each from Rhode Island and Long Island in September 2003.

Worm tubes were obtained by shovel, separated from sediment, and kept cool on ice until they were sorted in the laboratory. The worms were removed from their tubes and transferred to finger bowls containing isotonic magnesium chloride ($MgCl_2$) in seawater for species diagnosis. *C. torquata* was almost always found in monospecific stands; however, every individual used in this study was confirmed to be *C. torquata* by the presence of a collarette on setiger four and the absence of red bands in the mid-region (Mangum 1962). Incomplete anterior ends were only used if low numbers of complete worms were collected, and only if they were intact to setiger four. Two complete worms from each location were preserved in formalin then stored in ethanol as vouchers, and deposited in the Smithsonian Museum's collections. For DNA extraction, tissue samples (~25 mg) were removed from the midsection of individual worms using sterile techniques. Genomic DNA was extracted with the DNeasy extraction kit (Qiagen) following the manufacturer's protocol.

To determine which markers were best suited for phylogeographic analyses, a number of genes were screened, using primers designed from *C. torquata*'s complete mitochondrial genome (Table 2.2). I used the DNA sequences from small samples of

Gene	<i>N</i> total	<i>N</i> each site	Fragment Length	Included bp	Variable bp*	Pars inf bp*
atp6	66	ME 28 HY 14 NJ 24	530 bp	409	18 (4.4%)	11 (2.7%)
12S	16	HY 3 NJ 7 ME 6	578 bp	574	3 (.52%)	1 (.17%)
16S	24	HY 11 BH 5 NJ 6 PB 2	376 bp	371	12 (3.2%)	12 (3.2%)
ITS1, 5.8S, ITS2	11	HY 4 NJ 2 PB 1 ME 4	608 bp	574	6 (1.0%)	1 (.17%)
nad1	17	BH 6 HY 5 NJ 6	593 bp	549	80 (14.6%)	4 (.73%)
nad4	14	BH 1 HY 8 NJ 5	456 bp	386	8 (2.1%)	6 (1.6%)
nad6	43	ME 19 HY 12 NJ 9 BF 2 PB 1	474 bp	456	53 (11.6%)	6 (1.3%)

Table 2.2 Genes screened to select population genetic markers. *N*, number; bp, basepairs; pars inf, parsimony informative. *Percentages of variable and parsimony informative base pairs were calculated based on included characters.

worms from Maine in the north, a Cape Cod population (usually Hyannis or Barnstable Harbor), and New Jersey in the south, to compare genetic diversity among these markers. Two mitochondrial genes, the ATPase F0 subunit 6 (*atp6*) and the NADH dehydrogenase subunit 4 (*nad4*), were selected because they consistently amplified well and possessed the greatest sequence divergence between these populations. Nuclear ITS1 and ITS2 were also screened but exhibited almost no variation in this range. Primers were designed from the mitochondrial genome data to amplify 650 bp of the *atp6* gene (Ct*atp6*f, 5'-GACCCTGCTACTAACTCTTTT-3' and CtArgR, 5'-TTGCCACCTTTTAATGAATGA-3') and 630 bp of the *nad4* gene (Ctnad4PGf, 5'-TATTTCTTATTCTAGGGYGAGGTT-3' and Ctnad4PGr, 5'-TCTTCGTGATTGGGYGGTTTC-3'). Each fragment was amplified in separate 50µL PCR reactions consisting of 30.4µL of sterile water, 5µL of 10X buffer, 5µL of 25mM MgCl₂, 5µL of 4mM dNTPs, 1.2µL of each primer, 2µL extracted DNA template, and 0.2µL of Taq polymerase (Promega). PCR conditions for both genes consisted of: initial denaturation, 94°C, 1 min.; 35 cycles of (94°C, 45 sec; 49°C, 45 sec, 72°C, 1 min); final extension 72°C, 5 min; final hold 4°C. Successfully amplified DNA was purified directly from the reactions using the SV Gel and PCR Cleanup Kit (Promega) and eluted in 30µL sterile water. Standard one-eighth format sequencing reactions were performed using Big Dye Terminators (version 3, Perkin-Elmer) in 96-well plates. Sequencing reactions were purified by isopropanol precipitation and sequenced on an ABI 377 or an ABI 3730 Capillary Sequencer. PCR products were sequenced bi-directionally and proofread in

Auto Assembler (Applied Biosystems). Sequences from all individuals were aligned in MacClade 4.06 (Maddison and Maddison 2000), with the aid of inferred amino acid translations using the *Drosophila* mitochondrial code. Sequences were trimmed to minimize gaps at the beginning and end of the alignments, and all differences between individuals were verified in the electropherograms.

Population Structure

To investigate the relatedness of haplotypes, parsimony networks with 95% connection limits were constructed using the program TCS 1.17 (Clement et al. 2000). Optimal networks were obtained singly for each gene and for the concatenated sequences of both genes. In all cases, gaps (only present at the beginning and end of the alignments) were treated as missing data. To determine the most likely ancestral haplotypes, outgroup weights were also computed in TCS using the algorithm of Castelleo and Templeton (1994). The networks were then transformed into geographic maps of haplotype distributions (i.e., haplotype maps).

To evaluate the geographic isolation of populations and regional groups, an analysis of molecular variance (AMOVA) was conducted on each gene separately and both genes together using Arlequin (Schneider et al. 2000). The AMOVA was performed on pairwise genetic distances using 3 hierarchical levels: within each sampling location, between locations but within regional groups, and between regional groups. The regional groups were defined based on shortest over-water distances as follows. The five Cape Cod sites are clustered (i.e., closer to each other than to their nearest non-Cape neighbors)

and were placed into a group with Rhode Island because the latter is closer to the nearest Cape Cod site (Pocasset) than it is to the nearest non-Cape site (Long Island). Placing Maine and the Bay of Fundy into a second group, and Long Island and New Jersey into a third, resulted in three groups, all of which consisted of smaller intra-group distances than inter-group distances. Over-water routes through the Cape Cod Canal were used in calculating distances if they were shorter than routes around the peninsula. This operational definition, though perhaps crude, is based on geographic proximity in a straightforward manner, and makes no further assumptions about population genetics or coastal currents. All AMOVA analyses were bootstrapped 10,000 times to assess significance. Traditional F_{ST} values, estimates of the number of migrants per generation (Nm), and significance values for all population pairs were also computed in Arlequin, and significance determined by the sequential Bonferroni correction procedure of Rice (1989). The relationship between genetic and geographic distances was analyzed by performing a Mantel test (Mantel 1967) on the matrix of estimated number of migrants (Nm) and shortest over-water distance between sampled populations, with significance tested after 10,000 random permutations.

The importance of currents as dispersal vectors was analyzed using the program MIGRATE version 2.0.3 (Beerli 2002) to compare models employing asymmetric versus symmetric migration rates. MIGRATE uses a Metropolis-coupled Markov chain Monte Carlo (MCMCMC) strategy to maximize a likelihood function for migration rates (and other parameters if desired). The underlying population genetic model is a coalescent model originally developed by Hudson (1990) and modified by Notohara (1994) to

include migration between discrete populations. Mutation rates are estimated from the data using Felsenstein's (1984) model, in which mutations between all DNA bases are equally probable and time-reversible (i.e., A to T has the same rate as T to A). In the symmetric model, migration from population i to population j (M_{ij}) is constrained to be the same as migration from j to i (M_{ji}), whereas in the asymmetric model they are estimated separately. The likelihood function thus calculates the probability of seeing the observed genealogy given the parameter estimates, times the probability of the sample data given the genealogy, and integrates over all possible genealogies (see MIGRATE manual and Beerli 2002).

Searches for maximum likelihood estimates (MLEs) of migration rates under both symmetric and asymmetric models were performed on concatenated *atp6* and *nad4* sequences with initial parameters (T_i/T_v of 15:1 and gamma distribution shape parameter $\alpha=0.2701$) estimated empirically in PAUP* 4.0b (Swofford 2002). Migration rates and their likelihoods were estimated using the "quick and dirty" option per the MIGRATE user's manual. Fifteen short chains were run with an increment of 20 and a sample of 5,000 trees, followed by three long chains with an increment of 20 and a sample of 50,000 trees after a burn-in of 10,000 trees. Because trial runs with these settings and no heating failed to converge quickly, adaptive heating was employed on short chains with initial temperatures of (1.0, 1.5, 3, 6, 12), and the "LastChains" replication option. Two runs were performed for each model, with the second run starting with the MLEs from the previous run. Since the symmetric model is a nested case of the asymmetric model, a

likelihood ratio test (LRT) was performed on the final maximum likelihood scores to select the model that best fit the data.

Genetic Diversity

To identify patterns in genetic diversity, Nei's measure of gene diversity (Nei, 1987) was computed in Arlequin. In the absence of any information on the rate of invasion of *C. torquata* into exposed habitat post-glaciation or the effects this might have had on the shape of a genetic diversity gradient, the relationship between Nei's gene diversity and degrees north latitude of the sampled populations was analyzed by simple linear regression.

Test of Neutrality

To test the assumption that these phylogenetic analyses were conducted on selectively neutral genes, Tajima's D (Tajima 1989, 1996) was calculated for each population separately, and for all populations treated as a single population. Significance values for neutrality tests were calculated from 1000 bootstrap replicates.

RESULTS

From the 10 populations, a total of 189 and 176 individuals were amplified for *atp6* and *nad4* respectively (Table 2.1); 23 individuals in the *atp6* dataset did not amplify for *nad4*, and 10 individuals in the *nad4* dataset did not amplify for *atp6*, for a total of

166 individuals for which both gene sequences are available. The aligned data sets have been deposited to TREEBASE (www.treebase.org).

Population Structure

The estimated parsimony network for *atp6* contained two cycles (groups of haplotypes with multiple connection paths) involving three and four haplotypes, respectively (Figure 2.2A). The two most common haplotypes encountered in the samples occupied central positions, and were shared by the northern and Cape populations (one shared by Bay of Fundy, Maine, Barnstable, Pleasant Bay, Hyannis, and Rhode Island; the other shared by all five Cape Cod sites). One common haplotype was shared by New Jersey and Long Island, and all other haplotypes were rare and found in a restricted number of locations. Three intermediate haplotypes not found in any sample were required to completely connect the network. The most likely ancestral haplotype was found only in two individuals from New Jersey (Figure 2.2A arrow). Three of the five New Jersey haplotypes clustered with all of the Long Island haplotypes. A geographical representation of the *atp6* data (the haplotype map, Figure 2.2B) showed that the most common shared haplotype was dominant in the north, declined in frequency on Cape Cod, and disappeared to the south. The Bay of Fundy sample exhibited markedly lower diversity than others of comparable size (Chatham and Pocasset possessed low diversity, but sample sizes were small).

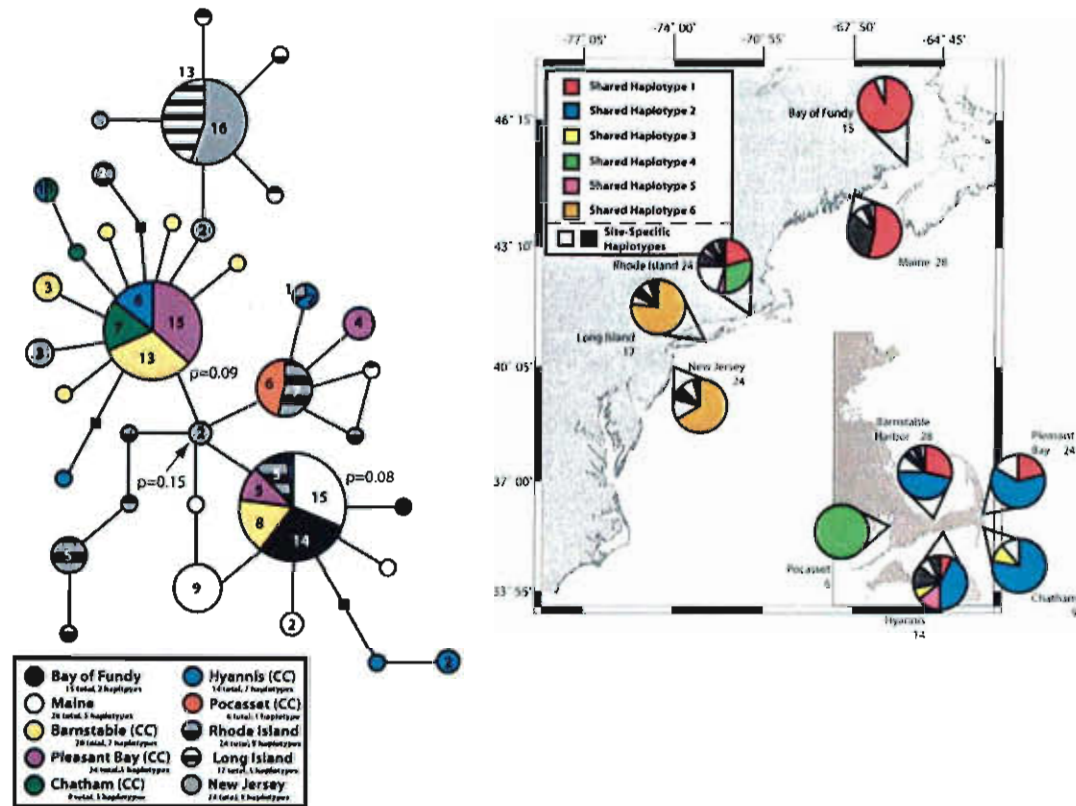


Figure 2.2 A, parsimony network for *atp6*. Circles represent the different haplotypes observed, with circle areas proportional to the number of individuals possessing that haplotype. Haplotypes shared between geographical locations are further broken down into pie graphs. Connecting lines represent single base pair differences, and missing intermediates are represented by small squares. P-values are the probabilities of haplotypes being the ancestral type; the haplotype with the highest outgroup probability is marked with an arrow. B, haplotype map for *atp6*. Pie graphs indicate the haplotypic composition at each location; colored haplotypes are shared between multiple locations, and shades of gray indicate site-specific haplotypes found only in a single location.

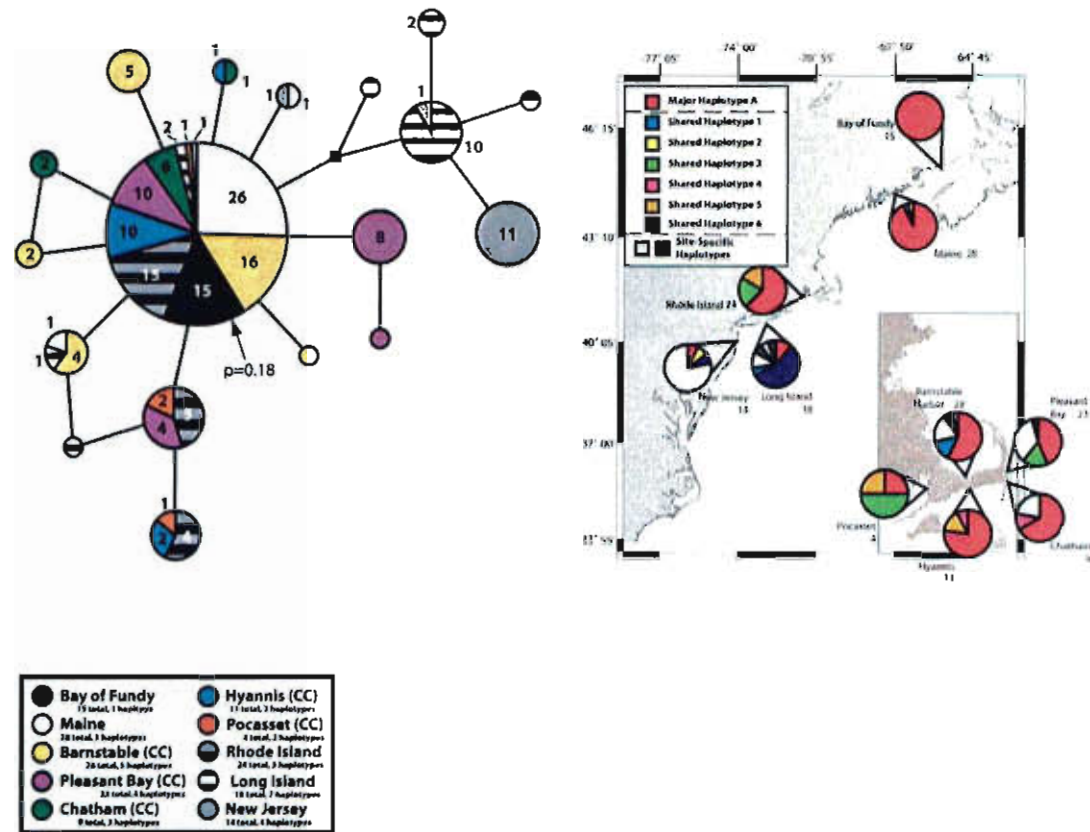


Figure 2.3. A, parsimony network for *nad4*. Circles represent the different haplotypes observed, with circle areas proportional to the number of individuals possessing that haplotype. Haplotypes shared between geographical locations are further broken down into pie graphs. Connecting lines represent single base pair differences, and missing intermediates are represented by small squares. P-values are the probabilities of haplotypes being the ancestral type; the haplotype with the highest outgroup probability is marked with an arrow. B, haplotype map for *nad4*. Pie graphs indicate the haplotypic composition at each location; colored haplotypes are shared between multiple locations, and shades of gray indicate site-specific haplotypes found only in a single location.

The *nad4* parsimony network (Figure 2.3A) consisted of a single central haplotype found in all locations, and a star-like structure of site-specific haplotypes similar to that in the *atp6* network. The network could be converted into a completely resolved tree by resolving two cycles (involving three and four haplotypes, respectively). The common, central haplotype was the most likely root (Figure 2.3A arrow). Similar to *atp6*, the most common haplotype of *nad4* was dominant in the north and declined in frequency to the south (Figure 2.3B). Two of the four New Jersey haplotypes occurred in a separate part of the network together with four of seven Long Island haplotypes. New Jersey was dominated by a single haplotype not found anywhere else, and again there was a haplotype shared only between Hyannis and Chatham. In contrast to *atp6*, however, a *nad4* haplotype was found in Maine, Barnstable Harbor, and Long Island (but not Rhode Island or Pocasset in between them). Finally, another haplotype was shared between Maine (the northernmost site) and New Jersey (the southernmost), but not any intermediate location.

Although the two genes exhibited similar genetic structure, their combined parsimony network (not shown) was highly reticular, indicating incongruence between the genes. Eighteen intermediate haplotypes were required to connect the network, and numerous cycles were present. The majority of these intermediates and unresolved connections occurred in the central part of the network, separating peripheral groups similar to those found in both single-gene networks. The haplotypes specific to Rhode Island were also found in the central part of the network. Two haplotypes similar to the

two largest in the *atp6* network were present in the combined network as well, as was a separate substructure of Long Island and New Jersey lineages.

A. Combined AMOVA

Source of Variation	df	SS	Variance Components	Percentage of Variation	Significance
Among groups	2	124.10	1.10	24.72	p=0.003
Among populations					
within groups	7	58.02	0.34	7.68	p≤0.000005
Within populations	169	468.69	3.00	67.60	p≤0.000005
Total	178	650.81	4.44		

B. Locus-by-locus AMOVA

Source of Variation		df	SS	Variance Components	Percentage of Variation	Significance
Among groups	<i>atp6</i>	2	16.13	0.14	27.55	p≤0.000005
	<i>nad4</i>	2	10.95	0.08	22.27	p≤0.000005
Among populations	<i>atp6</i>	7	9.04	0.06	12.96	p≤0.000005
	within groups <i>nad4</i>	7	8.55	0.06	16.92	p≤0.000005
Within populations	<i>atp6</i>	169	46.13	0.30	59.50	p≤0.000005
	<i>nad4</i>	169	35.92	0.23	60.81	p=0.0130
Total	<i>atp6</i>	178	71.30	0.50		
	<i>nad4</i>	178	55.43	0.38		

Table 2.3 AMOVA results and significance from 10,000 bootstrap replicates. df, degrees of freedom; SS, sum of squares.

The AMOVA analysis revealed highly significant differences at all hierarchical groupings tested (all p-values ≤0.003, Table 2.3). Differences within populations explained the largest amount of total variance for both genes, followed by differences between the regional groups. These patterns were consistent when the AMOVA was conducted on each gene separately. When gene flow was further analyzed between all population pairs for both genes, 36 of the 45 F_{ST} values were significant after sequential

Bonferroni correction, indicating that most population pairs exchanged few migrants (Table 2.4). Of the nine F_{ST} values that were not significantly different from zero, seven involved the small sample from Pocasset. Hyannis was not significantly differentiated from Pleasant Bay or Rhode Island.

	Bay of Fundy	Maine	Barnstable Harbor	Pleasant Bay	Chatham	Hyannis	Pocasset	Rhode Island	Long Island
Maine	0.419								
Barnstable Harbor	0.329	0.180							
Pleasant Bay	0.333	0.179	0.092						
Chatham	0.493	0.281	0.183	0.182					
Hyannis	0.350	0.169	0.074	0.072	0.170				
Pocasset	0.451	0.206	0.096	0.094	0.218	0.072			
Rhode Island	0.315	0.159	0.071	0.069	0.160	0.049	0.069		
Long Island	0.402	0.227	0.136	0.135	0.238	0.119	0.150	0.113	
New Jersey	0.337	0.176	0.087	0.085	0.178	0.066	0.087	0.063	0.130

Table 2.4 Pairwise F_{ST} values for all populations. Boldface numbers indicate F_{ST} values not significantly different from zero. Significance is assessed after sequential Bonferroni correction.

The symmetric migration rates estimated from MIGRATE ($-\ln L_S = 305.991$, data not shown) were congruent to those estimated from Arlequin, with most non-zero migration rates between sites on Cape Cod (Table 2.5). Migration rates produced by MIGRATE were generally either on the order of $>10^{-2}$ or $<10^{-9}$. Because the latter were unrealistically small (i.e., one migrant on average every billion years), they were considered to be effectively zero. Similarly to symmetric rates, non-zero migration rates estimated from the asymmetric model ($-\ln L_A = 213.966$, data not shown) were largely between Cape Cod sites; high migration was also inferred southward from the Bay of Fundy to Maine and northern Cape Cod sites, from Long Island to New Jersey, and from

	ME	BH	PB	CH	HY	P	R	LI	NJ
ME	North from ME 0 South to ME 4.94	North from BH 0 South to BH 14.59	North from PB 0 South to PB 0	North from CH 0 South to CH 0	North from HY 0 South to HY 1.65	North from P 0 South to P 0	North from R 0 South to R 0	North from LI 0 South to LI 0	North from NJ 0 South to NJ 0
MR		North from ME 0 South to ME 0	North from PB 0 South to PB 0	North from CH 0 South to CH 0.19	North from HY 0 South to HY 0	North from P 0 South to P 0	North from R 0 South to R 0	North from LI 0 South to LI 0	North from NJ 0 South to NJ 0
BH			North from PB 0 South to PB 0	North from CH 123.47 South to CH 6.50	North from HY 0 South to HY 30.31	North from P 0 South to P 0	North from R 0 South to R 2.55	North from LI 0 South to LI 0	North from NJ 0 South to NJ 0
PB				North from CH 0 South to CH 0	North from HY 1.91 South to HY 0	North from P 0 South to P 0	North from R 0 South to R 0	North from LI 0 South to LI 0.40	North from NJ 0 South to NJ 0
CH					North from HY 5.98 South to HY 0	North from P 0.19 South to P 0	North from R 0 South to R 0	North from LI 0 South to LI 0	North from NJ 0 South to NJ 0
HY						North from P 0 South to P 0	North from R 4.80 South to R 0	North from LI 0 South to LI 0	North from NJ 0 South to NJ 0
P							North from R 1.21 South to R 0	North from LI 0 South to LI 0	North from NJ 0 South to NJ 0
R								North from LI 0.80 South to LI 0	North from NJ 0 South to NJ 0.25
LI									North from NJ 0 South to NJ 0.50
NJ									

Table 2.5 Estimated number of migrants between populations from the asymmetric MIGRATE model.

Rhode Island to southern Cape Cod sites, Long Island, and New Jersey. The asymmetric migration model was a significantly better fit to the data ($2(\ln L_A - \ln L_S) = 184.049$, $df=45$, $p=9.9 \times 10^{-19}$). About 60% of the estimated nonzero migration rates (10 of 17) were in a southward direction (Table 2.5), following the predominant coastal current patterns. For the northern group, southward migrations dominated (5 of 5 nonzero migration rates) whereas migration rates involving Cape Cod were more evenly split between northward and southward (7 southward of 12 total), particularly migration rates to other sites on the Cape (3 southward of 7 total). Migrations involving the southern group were also evenly split (4 southward out of 7 total). This pattern of mostly southward migration over large distances but mixed migration at smaller scales makes sense given the more variable nature of small scale coastal currents around Cape Cod and the southern sites.

The Mantel test revealed a significant negative correlation between shortest over-water distance between populations and the estimated number of migrants exchanged between them (correlation coefficient = -0.5408 , $r^2=29.25$, $p=0.0280$). Thus, a given population appears to have received significantly fewer migrants from distant populations than more nearby populations (isolation by distance). An alternative Mantel test comparing the same estimated number of migrants, but this time using shortest over-water distances ignoring the Cape Cod Canal, exhibited a slightly smaller, but still significant, correlation (correlation coefficient = -0.5221 , $r^2=27.26$, $p=0.0282$).

Genetic Diversity

Linear regressions of gene diversity for each population on degrees north latitude were statistically significant after removing the small samples of Chatham and Pocasset (Figure 2.4). Both relationships have negative slopes, indicating higher diversity in southern locations than in northern, with latitude explaining 70% of the variation in *atp6* gene diversity and 65% of the variation in *nad4*.

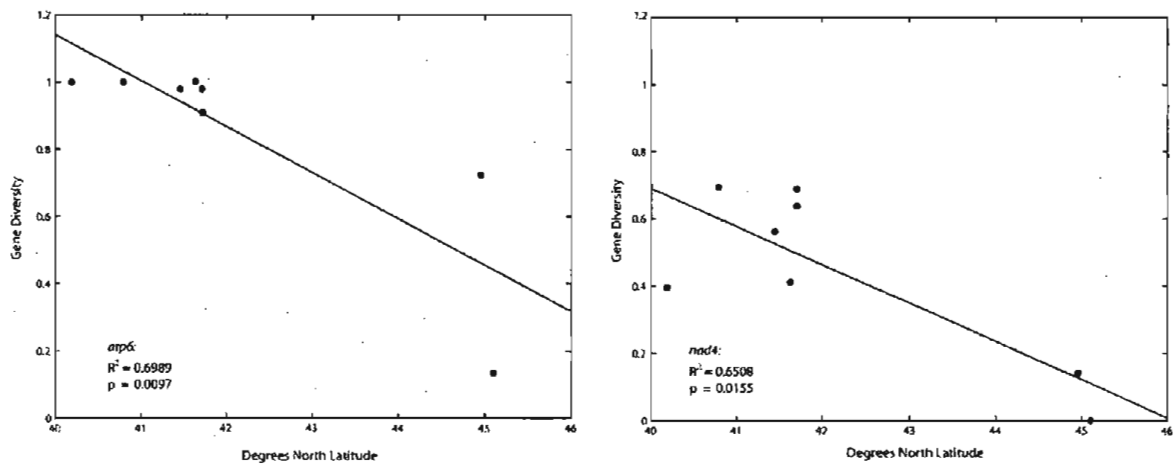


Figure 2.4 Molecular diversity indices at collection sites ordered by latitude for *atp6* (A) and *nad4* (B). r^2 and p-values are shown for linear regression.

Test of Neutrality

No convincing evidence of a departure from selective neutrality was found (Table 2.6). Tajima's D was negative in eight populations (Bay of Fundy, Maine, Barnstable Harbor, Chatham, Hyannis, Pocasset, Long Island, and New Jersey) and positive in two populations (Pleasant Bay and Rhode Island). Only the Long Island value was statistically significant. When all populations were combined into one total population, Tajima's D was negative, but not significant.

Location	Tajima's D	Significance
Bay of Fundy	-1.16	0.146
Maine	-1.21	0.131
Barnstable Harbor	-0.96	0.186
Pleasant Bay	0.81	0.812
Chatham	-0.69	0.292
Hyannis	-1.36	0.088
Pocasset	-0.61	0.378
Rhode Island	0.58	0.751
Long Island	-1.80	0.021
New Jersey	-1.34	0.099
Total population	-1.44	0.051

Table 2.6 Test of neutrality for each population and significance estimated from 1000 replicates.

DISCUSSION

The parsimony networks and haplotype maps of both gene regions show clear evidence of a phylogeographic break between Rhode Island and Long Island, consistent with both the water-mass and coastal currents hypotheses. The most striking evidence of this discontinuity in both genes is the marked decrease in frequency of the most common haplotypes on Cape Cod and their virtual disappearance further south. In addition, both parsimony networks have substructures of haplotypes found only in Long Island and New Jersey, and both contain substructures of haplotypes found only near the southern Cape. Thus, the Cape Cod sites appear to be more similar to northern sites than to southern, as predicted by the water-mass hypothesis. The greater similarity of water temperature in the Gulf of Maine and all around Cape Cod may facilitate gene flow among these populations, while prohibiting the southern haplotypes from persisting in the north even if they are transported there. However, an effect of coastal currents is strongly implied

by the highly significant, southward-biased asymmetric migration rates. Note that, within the closely spaced Cape Cod sites, southward migrations are less dominant, consistent with the variable nature of local currents in this region.

The phylogeographic barrier is not absolute. The observations of a *nad4* haplotype shared between New Jersey and Maine, and another shared between Long Island and Barnstable Harbor (with neither found in intervening locations) imply that dispersal can occur between the regional groups. Since both haplotypes were found in two individuals only, parallel mutation (i.e. homoplasy) is also a possibility, particularly for the New Jersey/Maine haplotype. Regardless, differences between the regional groups were significant in the AMOVA ($p \leq 0.0001$), and support the presence of a fairly strong dispersal barrier between the "Cape Cod" and "South of Cape" groups. In addition, the presence of a common, shared haplotype on Cape Cod that is not found in either the northern or southern sites implies that all three regional groups in the AMOVA are significantly differentiated. Thus, between Cape Cod and the northern sites, similar haplotypes occur at different frequencies, whereas between these two groups and the southern group, different haplotypes occur altogether.

Several other factors indicate that, although migration rates are not zero across the barrier nor within the groups on either side of it, the overall level of gene flow is very low. The centrality of the shared haplotypes in the networks and the rarity of other widely shared haplotypes support this scenario, and are consistent with *C. torquata's* short planktonic period and the barriers to gene flow revealed by the AMOVA. Most of the low-frequency shared haplotypes occur in neighboring populations (e.g. Long Island

and New Jersey, Rhode Island and Pocasset, Chatham and Hyannis), implying that dispersal occurs only over short distances (within regions). The high F_{ST} values correspond to inferred numbers of migrants low enough to maintain differentiation. The presence of isolation-by-distance revealed by the Mantel test (i.e., fewer expected migrants exchanged between populations farther apart) corroborates low rates of gene flow. Further, the slightly better correlation between geographic distances measured through the Cape Cod Canal (if they are shorter than distances around Cape Cod) indicates that the Canal could provide an alternative, if secondary, dispersal route.

Modern day gene flow directed along coastal currents is not the only force that appears to have shaped population genetics in *C. torquata*; the patterns of genetic diversity from south to north suggest an influence of glacial effects at a deeper level in *C. torquata*'s history. The gradient in genetic diversity is consistent with a wave of northward reestablishment of populations (resulting in founder effects) following the retreat of glacial ice, as has been found in other studies of the northwest Atlantic (e.g. Bernatchez and Wilson 1998, Cunningham and Collins 1998, Hare and Weinberg 2005) and in some studies of the North American Pacific coast (e.g. Marko 2004, Wares and Cunningham 2005). Intertidal species might be expected to show greater genetic evidence of glaciation and recolonization than subtidal species because of habitat reduction and exposure to cold stress (e.g., Dahlgren et al. 2000, Marko 2004).

Interestingly, present day *C. torquata* populations are in general subtidal, but extend into the low intertidal on Cape Cod (pers. obs.) and to the north (Mangum 1964). If these depths are where *C. torquata* populations lived as ice sheets spread just before the LGM,

then northern, intertidal populations may similarly have been exterminated by temperature stress, while more sheltered southern subtidal populations persisted.

The presence of the most likely ancestral haplotype in New Jersey (in the *atp6* network) also implies a northward spread of *C. torquata* just after the last glacial maximum (LGM). The possibility that one of the common shared haplotypes is the true ancestor cannot be dismissed, however, and is consistent with the most likely ancestor inferred from the *nad4* network. Regardless, the dominant haplotype in the Bay of Fundy and in Maine is the ancestral, most widespread haplotype across all populations, which is also the one most likely to disperse to new sites. Under this hypothesis, populations further south might originally have harbored the ancestral haplotype during the LGM, and subsequently lost it after its spread northward. Alternatively, the most common haplotypes in both genes might have arisen after the LGM somewhere on or north of Cape Cod, with the phylogeographic barrier preventing their spread further south. Although selective gradients could also cause such differences in diversity, the only evidence for selection was the negative Tajima's D for Long Island, which cannot explain low diversity in the north. Tajima's D is known to be sensitive to factors such as recent selective sweeps or population expansions (see for example Marko 2004), the latter of which may be especially relevant here.

Although the glaciation hypothesis is by far the one most frequently suggested to explain diversity gradients in the Northwest Atlantic, it is interesting to consider how the observed pattern of genetic diversity might be related to the well-known (if often oversimplified) pattern of species diversity from tropics to pole. Although a myriad of

hypotheses have been put forward to explain the generally higher species diversity in the tropics than at the poles, the most common involve the degree of spatial and/or temporal variability associated with habitats (see Sanders 1968, Crame and Clarke 1997, Williamson 1997). On the spatial axis, highly variable habitats are suggested to create micro-niches that harbor high species diversity. On the temporal axis, highly variable environments (for instance, a large amplitude in salinity over a tidal cycle or highly variable temperature throughout the year) are suggested to reduce the number of species that can tolerate the extremes, leading to low diversity. While both of these hypotheses may apply to patterns of species diversity, only temporal variation should be relevant to genetic patterns within a single species, unless single populations of that species are known to span more than one niche. Furthermore, even if this is true, individuals living in the different environments would have to differ genetically, and samples from a single population would have to cross multiple habitats in order for high genetic diversity to be detected. On the temporal side then, over the range of only five degrees of latitude that this study spans, it is likely that the northern sites are less spatially variable and experience a smaller range of temperatures and perhaps salinities (Engle and Summers 1999). Therefore, although the trend I encountered in genetic diversity in *C. torquata* is in keeping with larger latitudinal diversity trends, the more temporally variable sites in this study exhibited higher, not lower, genetic diversity. In contrast, suggestions that polar climates harbor lower species diversity because they promote lower mutation rates (and thus rates of speciation) are in keeping with the genetic diversity gradient I observed; however, a global gradient in mutation rates has been difficult to establish.

The history of glaciation and founder effects thus seems to be the best hypothesis to explain the observed genetic diversity gradient. On the whole, however, all three phylogeographic hypotheses appear to explain different aspects of the data, with a northward spread occurring after the LGM, followed by a genetic break imposed by water-mass differences and a low-level of modern day gene flow via the coastal currents present today.

It is unclear what is causing the highly reticular network obtained when two seemingly congruent single-gene networks are combined. The many unresolved relationships of the combined network would seem to indicate that differences in phylogeographic signal between the genes, while small, are nonetheless significant. The degree of incongruence between the genes was not significant, however, when measured by an Incongruence Length Difference test (ILD, Michevich and Farris 1981). It is possible that the larger number of haplotypes (and presumed faster rate of evolution) in *atp6* is causing the disagreement, although with few variable characters, any estimate of evolutionary rates would likely be inaccurate. The issue of estimating rates of evolution is further exacerbated by the lack of a geologic date to serve as calibration, as well as by the lack of published estimates in these genes for closely related annelids. To resolve such issues I will need to develop additional markers (especially for understudied taxa such as annelids) and will need to examine multiple species. Apparently in the case of *C. torquata*, there is no single explanation for present day genetic diversity. However, we should perhaps start thinking of Long Island or Rhode Island, rather than Cape Cod, as the genetic breakpoint along the north east coast of North America.

REFERENCES

- Anraku M (1964) Influence of the Cape Cod Canal on the hydrography and on the copepods in Buzzards Bay and Cape Cod Bay, Massachusetts. I. Hydrography and distribution of copepods. *Limnology and Oceanography*, **9**, 46-60.
- Avise JC (1992) Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *OIKOS*, **63**, 62-76.
- Bastrop R, Juerss 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.
- Berli P (2002) MIGRATE: documentation and program, part of LAMARC. Version 1.6, Revised August 23, 2002. Distributed over the Internet, <http://evolution.genetics.washington.edu/lamarc.html> [Downloaded: Version 2.0.3, 12/19/2004].
- Bernatchez L, Wilson CC (1998) Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular Ecology*, **7**, 431-452.
- Burton RS (1998) Intraspecific phylogeography across the Point Conception biogeographic boundary. *Evolution*, **52**, 734-745.
- Castelloe J, Templeton AR (1994) Root Probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution*, **3**, 102-113.
- Churchill JH (1985) Properties of flow within the coastal boundary layer off Long Island, New York. *Journal of Physical Oceanography*, **15**, 898-916.
- Clement M, Posada D, Crandall KA (2000) TCS: a program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657-1659.
- Collin R (2001) The effects of mode of development on phylogeography and population structure of North Atlantic *Crepidula* (Gastropoda: Caplytraeidae). *Molecular Ecology*, **10**, 2249-2262.
- Crame, J.A. and A. Clarke. 1997. The historical component of marine taxonomic diversity gradients. Pp. 258-273 in R. F. G. Ormond, J. D. Gage and M. V. Angel, eds. Marine Biodiversity, Patterns and Processes. Cambridge University Press, Cambridge.

- Cunningham CW, Collins TM (1998) Beyond area relationships: extinction and recolonization in molecular marine biogeography. In: *Molecular approaches to ecology and evolution* (eds. DeSalle R, Schierwater B), pp. 297-321. Birkhäuser Verlag, Basel, Switzerland.
- Dahlgren TG, Weinberg JR, Halanych KM (2000) Phylogeography of the ocean quahog (*Arctica islandica*): influences of paleoclimate on genetic diversity and species range. *Marine Biology*, **137**, 487-495.
- Dillon RT, Manzi JJ (1992) Population genetics of the hard clam, *Mercenaria mercenaria*, at the northern limit of its range. *Canadian Journal of Fisheries and Aquacultural Science*, **49**, 2574-2578.
- Engle VD, Summers JK (1999) Latitudinal gradients in benthic community composition in Western Atlantic estuaries. *Journal of Biogeography*, **26**, 1007-1023.
- Felsenstein J (1984) Distance methods for inferring phylogenies: a justification. *Evolution* **38**: 16-24.
- Franz DR, Worley EK, Merrill AS (1981) Distribution patterns of common seastars of the middle Atlantic continental shelf of the northwest Atlantic (Gulf of Maine to Cape Hatteras). *Biological Bulletin*, **160**, 394-418.
- Govindarajan AF, Halanych KM, Cunningham CW (2005) Mitochondrial evolution and phylogeography in the hydrozoan *Obelia geniculata* (Cnidaria). *Marine Biology*, **146**, 213-222.
- Hare MP, Weinberg JR (2005) Phylogeography of surfclams, *Spisula solidissima*, in the western North Atlantic based on mitochondrial and nuclear DNA sequences. *Marine Biology*, **146**: 707-716.
- Hayden BP, Dolan R (1976) Coastal marine fauna and marine climates of the Americas. *Journal of Biogeography*, **3**, 71-81.
- Holder K, Montgomerie R, Friesen VL (1999) A test of the glacial refugium hypothesis using patterns of mitochondrial and nuclear DNA sequence variation in rock ptarmigan (*Lagopus mutus*). *Evolution*, **53**, 1936-1950.
- Hudson, RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**: 1-44.
- Hutchins LW (1947) The bases for temperature zonation in geographical distribution. *Ecological Monographs*, **17**, 325-335.

- Jennings RM, Halanych KM (2005) Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): evidence for conserved gene order in Annelida. *Molecular Biology and Evolution*, **22**:210-222.
- Koehn RK, Milkman R, Mitton JB (1976) Population genetics of marine pelecypods. 4. Selection, migration and genetic differentiation in the blue mussel *Mytilus edulis*. *Evolution*, **30**, 2-32.
- Lee CE (1999) Rapid and repeated invasions of fresh water by the copepod *Eurytemora affinis*. *Evolution*, **53**, 1423-1434.
- Lermusiaux PFJ (2001) Evolving the subspace of the three-dimensional multiscale ocean variability: Massachusetts Bay. *Journal of Marine Systems*, **29**, 385-422.
- Lynch DR, Holboke MJ, Naimie CE (1997) The Maine coastal current: Spring climatological circulation. *Continental Shelf Research*, **17**, 605-634.
- Lynch DR, Ip JTC, Naimie CE, Werner FE (1996) Comprehensive coastal circulation model with application to the Gulf of Maine. *Continental Shelf Research*, **16**, 875-906.
- Maddison DR, Maddison WP (2000) MacClade. Sinauer Associates, Inc., Sunderland, MA.
- Mangum CP (1962) Studies on speciation in the Maldanid polychaetes of the North American Atlantic Coast. I. A taxonomic revision of three species of the subfamily Euclymeninae. *Yale Peabody Museum Postilla*, No. **65**, 12 p.
- Mangum CP (1964) Studies on speciation in Maldanid polychaetes of the North American Atlantic Coast. II. Distribution and competitive interaction of five sympatric species. *Limnology and Oceanography*, **9**, 12-26.
- Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.
- Marko PB (2004) 'What's larvae got to do with it?' Disparate patterns of post-glacial population structure in two benthic marine gastropods with identical dispersal potential. *Molecular Ecology*, **13**, 597-611.
- Michevich MF, Farris JS (1981) The implications of congruence in *Menidia*. *Systematic Zoology*, **30**, 351-370.
- Nei M (1987) *Molecular Evolutionary Genetics* Columbia University Press, New York, NY.

- Newell GE (1951) The life-history of *Clymenella torquata* (Leidy). *Proceedings of the Zoological Society of London*, **121**, 561-590.
- Notohara M. (1994) The coalescent and the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **29**: 59-75.
- Palumbi SR, Grabowsky G, Duda T, Geyer L, Tachino N (1997) Speciation and population genetic structure in tropical Pacific sea urchins. *Evolution*, **51**, 1506-1517.
- Pielou EC (1991) *After the ice age*. University of Chicago Press, Chicago, IL.
- Rice WR (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223-225.
- Sanders, H.L. 1968. Marine benthic diversity: a comparative study. *American Naturalist* **102**: 243-282.
- Schneider S, Roessli D, Excoffier L (2000) Arlequin: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva.
- Smith MW, Chapman RW, Powers DA (1998) Mitochondrial DNA analysis of Atlantic Coast, Chesapeake Bay, and Delaware Bay populations of the teleost *Fundulus heteroclitus* indicates temporally unstable distributions over geologic time. *Molecular Marine Biology and Biotechnology*, **7**, 79-87.
- Spaulding ML, Gordon RB (1982) A Nested Numerical Tidal Model of the Southern New England Bight. *Ocean Engineering*, **9**, 107-126.
- Swofford DL (2002) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Inc., Sunderland, MA.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585-595.
- Tajima F (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457-1465.
- Valentine JW (1966) Numerical analysis of marine molluscan ranges on the extratropical Northeastern Pacific shelf. *Limnology and Oceanography*, **11**, 198-211.
- Vieira MEC (2000) The Long-Term Residual Circulation in Long Island Sound. *Estuaries*, **23**, 199-207.

- Vogler AP, DeSalle R (1993) Phylogeographic patterns in coastal North American tiger beetles (*Cicindela dorsalis*) inferred from mitochondrial DNA sequences. *Evolution*, **47**, 1192-1202.
- Wares JP (2002) Community genetics in the northwestern Atlantic intertidal. *Molecular Ecology*, **11**, 1131-1144.
- Wares JP, Cunningham CW (2005) Diversification before the most recent glaciation in *Balanus glandula*. *Biological Bulletin*, **208**, 60-68.
- Wares JP, Cunningham CW (2001) Phylogeography and historical ecology of the North Atlantic intertidal. *Evolution*, **55**, 2455-2469.
- Williamson, M. 1997. Marine biodiversity in its global context. Pp. 1-17 in R. F. G. Ormond, J. D. Gage and M. V. Angel, eds. *Marine Biodiversity, Patterns and Processes*. Cambridge University Press, Cambridge.

Chapter 4: Stage-specific selection structures geographic genetic patterns in marine benthic invertebrate populations

ABSTRACT

Population geneticists working on marine invertebrates are increasingly interested in short time scale questions involving ecological processes, for example the dynamics of single dispersal events or periods of high mortality. Although some data have been collected on the genetic composition of the sub-adult life stages involved in these processes (i.e., larvae and recruits), there exists no theoretical context or model in which to understand what causes the encountered patterns. Further, current population genetic models may be lacking for such datasets because they do not explicitly model life stages differently. Finally, there is a growing awareness that neutral processes may not be as pervasive in causing the observed patterns of genetic variation, particularly in sub-adult samples. I have constructed a stage-structured population genetic model involving reproduction, dispersal, settlement, and post-settlement survival, and analyze the patterns of genetic differentiation seen in newly settled individuals versus adults. Further, I have investigated the potential of neutral versus selective processes operating at several life stages to create and maintain such differences.

INTRODUCTION

The relationship between dispersal of marine benthic invertebrates and genetic differentiation of their populations has been the focus of decades of research. To date, the majority of this work has been accomplished by analyzing genetic markers of adults sampled from an array of geographic sites, and using models to infer levels of gene flow and migration among the sites. Often, the goal is measuring gene flow (i.e., the migration of individuals of some genetic type to new populations, and their incorporation into the pool of reproductive adults). In this case, standard models linking adult genetics and gene flow parameters such as Wright's island model (1943), Kimura and Weiss' stepping stone model (1964), Avise's phylogeographic method (see Avise 2000) etc. are the mainstay of population genetic analysis.

In contrast, many marine ecologists and population geneticists are increasingly interested in genetic estimates of dispersal itself (that is, the movement of individuals between populations regardless of their fate afterwards). Information is often sought over short timescales (sometimes single dispersal events); in this circumstance the life history and population biology of a marine invertebrate is extremely important to consider. In the marine environment, many benthic invertebrates, particularly those dwelling in soft-bottoms, produce large numbers of young, very few of which survive to reproduce (reviewed in Hunt and Scheibling 1997). Because this steep type III survival curve could make adult survivors of migrants much more rare than the original pool of migrants, population geneticists are turning to genetic samples of non-adult stages, particularly larvae and newly settled individuals (the latter called "settlers" herein for simplicity).

Samples of sub-adult stages present challenges that adult samples do not. Collections of larvae from the water column over a benthic site of interest can be especially problematic because it is unclear whether these larvae have migrated or were produced locally. It is equally unclear if the larvae would have settled into the population over which they were sampled, or might have migrated further. For this reason, most recent investigations of non-adult population genetics have sampled settlers, where the final destination of sampled individuals is more certain. Taxa for which such investigations have been undertaken include limpets (*Siphonaria jeanae*, Johnson and Black 1984), oysters (*Crassostrea gigas*, Li and Hedgecock 1998), lobsters (*Panulirus cygnus*, Johnson and Wernham 1999), sea urchins (*Strongylocentrotus franciscanus*, Moberg and Burton 2000), barnacles (*Semibalanus balanoides*, Drouin et al. 2002), and

fish (*Pseudopleuronectes americanus*, Crivello et al. 2004). As Moberg and Burton (2000) point out, it seems likely that non-adult samples are only recently being used because of an assumption among population geneticists that settlers are unlikely to reveal any information about gene flow (and/or dispersal) that could not be gleaned from adult samples. However, when these researchers sampled adults and settlers of the red sea urchin *Strongylocentrotus franciscanus*, they found significant differences in genotypes and genetic diversity between settlers and adults in the same population. In addition, they detected geographic structure to the pattern of genetic differentiation in settlers but not in adults.

Interpretation of these findings has sometimes been difficult, perhaps because of the novelty of such datasets. When patterns of genetic differentiation in settlers differ from those of adults, selection acting after settlement is often thought to be the cause (e.g., Hellberg 1996, but see Johnson and Wernham 1999, Moberg and Burton 2000, Drouin et al. 2002, Planes and Romans 2004). In some cases, the existence of an unsampled population outside the study area has also been suggested to underlie settler/adult differentiation (Drouin et al. 2002). This could be caused if the "ghost" population harbored a high frequency of a genetic type that was rare in the sampled adults. The presence of settlers from the ghost population among the settlers from sampled populations would differentiate them from the adult pool, from which the ghost population's genotype had been removed by selection. Although post-settlement mortality varies widely between taxa and is extremely difficult to measure (Hunt and Scheibling 1997), the process is usually assumed to be selectively neutral. If settlers and

adults frequently show different patterns of genetic differentiation due to selection after settlement, it would violate the assumptions of the neutral or "nearly-neutral" (Ohta and Kimura 1971) theories of population genetics, in which neutral processes are the dominant (or sole) forces affecting population dynamics. Given an observed amount of differentiation between settlers and adults (e.g., Moberg and Burton 2000), there is currently no context to identify or quantify the presumed selection. Likewise, there is no model that allows the quantification the number and source of migrants from sub-adult samples. Other evidence (Johnson and Black 1984) suggests that even neutral post-settlement mortality (which can be thought of as severe genetic drift) can cause different patterns of differentiation in settlers and adults, although the differences described in this work were ephemeral. These studies raise the question of whether some selection is required in post-settlement mortality to cause settler/adult differentiation.

Other processes have been suggested that do not act at the post-settlement stage, and/or do not require selection to cause differentiation between stages in a single population, or different spatial patterns of differentiation in different stages. Johnson and Wernham (1999) suggest that seasonal differences in the genetic composition of settling larvae could cause settlers to be genetically differentiated from adults (see also Crivello et al. 2004). The genetic composition of larvae reaching any particular site would also change in time, which would lead to differentiation in adult populations as well. Conversely, timing of reproduction could vary geographically if it follows a variable environmental cue like temperature. If these populations contained distinguishable genotypes (or different frequencies of the same genotypes), a time-varying larval supply

would also result. Finally, even if reproduction occurs on the same date for geographically spaced populations, larvae from some populations (e.g. those farther apart or those in bays with strong local retention) would take longer to be transported and would arrive later than others. As long as the later arrival dates were within the period of competency for the larvae, a genetically varying larval supply would result.

As more data on sub-adult genetic composition are gathered, the assumptions and limitations of current theoretical models become more apparent. While these models have proven extremely useful in spite of, perhaps even because of, their simple assumptions and the generality of their biology, a more ecologically relevant population genetic model is needed to determine how forces acting at different life stages combine to create patterns of genetic differentiation in time and space. Such a model should reflect recent changes in the direction of population genetics: 1) explicit attention to the nature of marine invertebrate life histories, 2) consideration of selection as an important force in differentiating populations, and 3) the use of samples of non-adult stages. Herein I present a simple model towards these ends. Our main goal in constructing and analyzing this model is to determine what additional insight is gained on short time scales by looking at the genetics of sub-adult stages in addition to adults. Specifically, I will address the following questions:

What can be learned about short-term genetic variation by including samples of sub-adults? For instance, is genetic differentiation in sub-adults more temporally variable than it is in adults? If samples of settlers preserve more of the genetic diversity of the initial migrant pool than do adults, they may reveal the importance of migration

events that are not detectable in adults. Do sub-adult samples reveal different spatial patterns of differentiation than adults? Thus, sub-adult samples may exhibit a higher correlation between genetic differentiation and geographic distance (at some or all spatial scales).

Can stage-specific selection create genetic differences between stages in the same population, and different patterns of spatial genetic differentiation in different stages? In exploring this question, I will investigate where and how selection could act to cause these kinds of genetic differentiation, and in particular whether any differences caused are transient or temporally constant. It is known that genetic drift (a neutral process) creates differences between groups by introducing, in effect, sampling error. However, current literature has begun to suggest that the differentiation observed in the field is too great to be caused by drift alone. Because many of these hypotheses (presented in the introduction) involve both neutral and selective processes operating during reproduction and after settlement, I compare and contrast two models. In the first, reproduction is a selective process and post-settlement survival neutral (fecundity selection with neutral survival), and in the second the reverse is true (neutral reproduction with selective survival).

MODELING

The family of models employed here is similar to the model Slatkin (1985) used to develop the rare alleles measure of gene flow. Our model consists of a circle of k populations, each filled with N adults. I keep track of the numbers of individuals, and

their diploid genotypes at a single nuclear locus. Each time step of the model is a generation. In each generation, the adults produce gametes, mutation creates new alleles for some of these gametes, and the gametes randomly unite to form larvae. After larvae are formed, some of them migrate to new populations (and are called juveniles), and a small fraction of the juveniles survive to become the adults of the following generation. Because many of the hypotheses presented in the Introduction involve forces at the reproductive and post-settlement stage, I ran two model variants (Figure 3.1). In the first model (the selective fecundity model (the FEC model), selection acts on reproduction as described below. In the second model (the selective Post-Settlement Mortality model or SPSM), selection acts after settlement in determining which individuals survive to

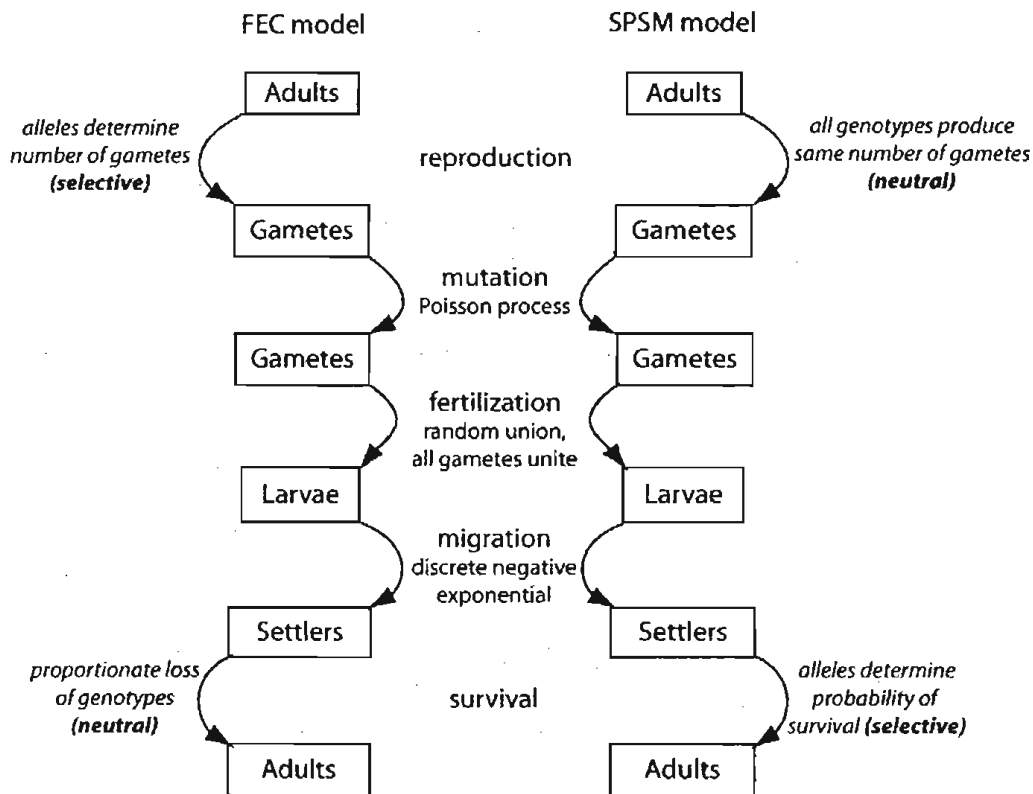


Figure 3.1 Flow chart of one time step in both of the models. Processes common to both models are shown in between them; processes specific to each model are shown to the side.

adulthood.

1. FEC model

Reproduction

The adults produce gametes according to the fitness of their alleles. I define s_i as the effect of allele i on reproduction. The fitness of a diploid individual of genotype A_iA_j is $e^{(s_i+s_j)}$, and its reproductive output is $\phi e^{(s_i+s_j)}$, where ϕ is the number of gametes produced by a (theoretically) neutral genotype. Because $e^x \approx 1+x$ when x is small and the s_i are all small, genotypic fitness is approximately $1+s_i+s_j$ (i.e., the effects of the two alleles are roughly additive). However, the exponential formulation has the advantage of being always nonnegative, as required for reproductive output. Alleles have the same effect in all populations. All of the gametes produced by homozygotes ($i=j$) are of type i . One half of the gametes produced by heterozygotes ($i \neq j$) are of type i and the other half are of type j .

Mutation

The number of mutations occurring in a single time step is a Poisson process with rate $4G\mu$, where G is the vector of gamete pool sizes in each population and μ the mutation rate per generation for the gene. After all mutation has occurred, I set the number of alleles m to the new total ($m \rightarrow m + \sum_k u$). New fitnesses are chosen for the mutant alleles based on a random walk model of evolution (RW, Lande et al. 1975) in which

$$s_{new} = s_{source} + x, \quad x \sim N(0, \sigma) \quad (1)$$

that is, the new allele's fitness is modified from the fitness of the allele it mutated from by adding a normally distributed random variable x with zero mean and variance σ . This allows both advantageous and deleterious mutations to arise.

Larval production

Larval production results from random union of gametes within each population. In each population, this is modeled by picking random pairs of gametes and creating a larva bearing the corresponding genotype until all gametes have been united.

Migration

There are k populations arrayed in a circle to avoid boundary effects. All populations are reachable from any source population, with the probability of migration from population i to population j following a decaying exponential with distance (essentially a discretized Laplace function). Larvae follow the shortest route around the circle between populations; I call this distance n . The dispersal function is thus

$$f(i, j) = \alpha d^{-n} \quad (2)$$

where d sets the level of migration and α is chosen such that $\sum_i f(i, j) = 1$. It can be

shown that choosing

$$\alpha = \left\{ \begin{array}{ll} \frac{(d-1)d^{k/2}}{d^{k/2} - d + d^{1+k/2} - 1} & \text{for } k \text{ even} \\ \frac{(d-1)d^{(k-1)/2}}{d^{(k-1)/2} - 2 + d^{1+(k-1)/2}} & \text{for } k \text{ odd} \end{array} \right\} \quad (3)$$

satisfies this constraint. If there are L larvae per population, L integers are sampled from the set $[1, k]$ with weights given by Eq. 2. Note that the proportion of larvae that do not migrate from population i is $f(i, i) = \alpha$ and thus the proportion of larvae that do migrate from i is $1 - \alpha$. Migration is equally likely for all genotypes (i.e., is selectively neutral). Although larvae do not migrate into a common pool and then get redistributed, all migration happens at once, so a larva does not migrate to more than one population per time step. After migration, the individuals settle on the bottom and are called settlers.

Post-settlement mortality

Settlers are randomly chosen in each population to survive and grow into reproductive adults. Compared to the high number of settlers, the number of survivors (N) is small (generally 0.1%; see Results). Within each population, N individuals are drawn from the settlers to determine the new distribution of the survivors. The surviving settlers become the adults of the next generation, completing one time-step of the model.

2. SPSM Model

In this model, all adults produce the same number of gametes. Gamete union, mutation, and migration operate in the same manner as in the FEC model, but here the selection coefficients determine the survival probabilities of settlers becoming adults. That is, the same RW distribution is used to generate selection coefficients, but in the SPSM model the relative magnitude of an individual's fitness defines its probability of survival.

Results and Discussion

Parameter values for the FEC and SPSM models (Table 3.1) were chosen to be biologically realistic for benthic invertebrates as well as computationally feasible in the simulations. A fecundity of 200 gametes per individual produced on the order of 10^4

Model Component	Parameter	Units	FEC value	SPSM value
Number of populations	k		9	9
Reproduction	ϕ	gametes per individual	200	200
	x mean	gametes	0	—
	x std	gametes	0.1	—
Mutation	μ	mutations/generation/individual	10^{-7}	10^{-7}
Migration	d		10^5	10^5
	α		0.9998	0.9998
Survival	N	individuals	200	200
	x mean		—	0
	x std		—	0.1
Sampling	sample size	individuals	30	30

Table 3.1 Parameter values used in the FEC and SPSM models. Symbols are as defined in the Methods section.

gametes per population. The mutation rate of 10^{-7} mutations per individual per generation produced roughly one mutant in each population in each generation; upwards of 90% of these mutations did not survive more than a single generation. The value of the migration parameter d was set at 10000, corresponding to $\alpha=0.9998$ (that is, 99.98% of larvae stay in the population in which they were spawned). With $\sim 10^4$ larvae, the larval migration rate was $0.9998 \cdot 10^4 \cdot 10^{-5} = 0.1$, or one migrant moving one step

approximately every 10 generations. With an adult population size of 200, this corresponded to an "effective migration rate" (in terms of surviving adults) of 0.02, or one migrant moving one step approximately once every 50 generations (low migration). All model runs consisted of nine populations in a circle; with this configuration there were nine pairs of populations in each category of: one step apart, two steps apart, three steps apart, and four steps apart. The model ran for 1000 time steps; with an adult population size of 200, this translates to greater than $4N$ generations in our simulations (a relevant time scale for many population genetics parameters). I started the model runs with each population containing the same four alleles, but uniformly distributed random numbers of the 10 genotypes. I first describe the general behavior of both models, and then summarize trends analyzed across 30 runs of each type.

1. General behavior

In the FEC model, selection coefficients determined the equilibrium allele frequencies and genetic drift caused small-scale fluctuations around these values (an example is shown in Figure 3.2). When new alleles arose with advantageous selection coefficients (compared to the makeup of the population in which they arose), they quickly ascended in frequency to their own selection-drift balance, causing the other allele frequencies to decrease. The difference in timing between when an allele arose in a single population and when it attained a similar frequency in the total population reflected the low rate of migration. I illustrate this by comparing allele frequency dynamics in population one with the total population (Figure 3.2). Note, for example,

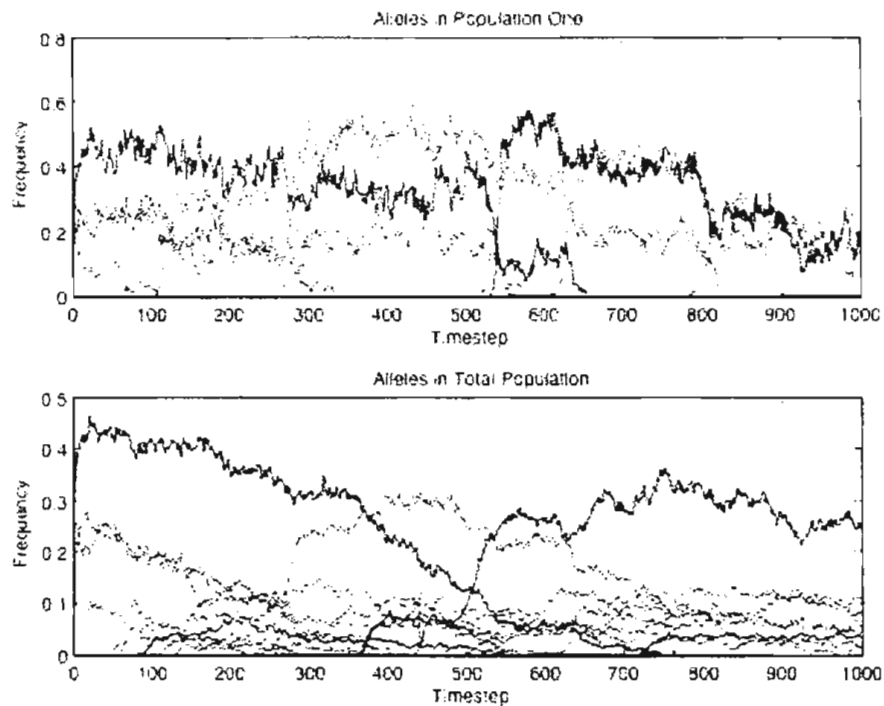


Figure 3.2 Allele frequencies in a typical FEC model run. The different colors indicate different alleles. Allele frequencies in Population One are shown as an example of the dynamics in a single population (top panel); allele frequencies are shown averaged over all 9 populations (Total Population, bottom panel) for comparison.

that the green allele fell in frequency near the beginning of the simulation; however, in population one it almost went extinct around time step 100, while in the total population (i.e., the other populations) it maintained a greater frequency.

In contrast, the PSM model produced different effects on allele frequencies (Figure 3.3). A selection-drift balance was less apparent; the allele frequencies were much closer to neutral expectations (i.e. all frequencies near $1/7$ when seven alleles were present). It is probable that selection produced less divergent allele frequencies in the PSM than in the FEC because in the latter, small differences in fitness were magnified by the neutral fecundity ϕ . PSM runs were also characterized by fewer losses of alleles over the time course. In general, PSM runs tended to be characterized by 1) steeper ascents of

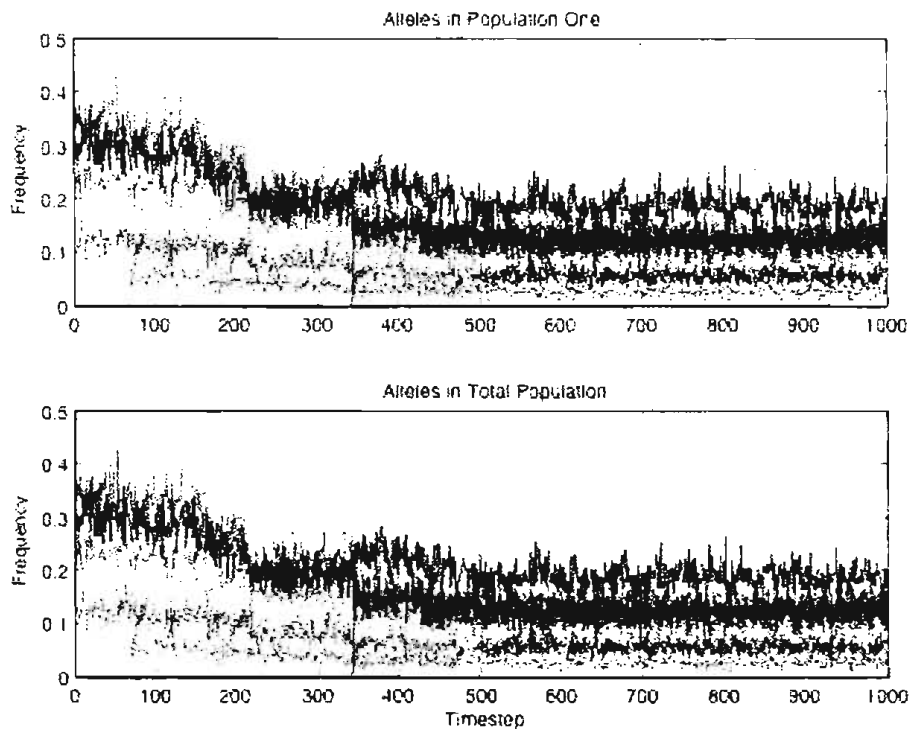


Figure 3.3 Allele frequencies for a typical SPSM run. The different colors indicate different alleles. Allele frequencies in Population One are shown as an example of the dynamics in a single population (top panel); allele frequencies are shown averaged over all 9 populations (Total Population, bottom panel) for comparison.

new alleles to equilibrium frequencies, 2) smaller amplitude fluctuations about that frequency (implying less drift), and 3) longer periods of selection-drift balance before a new allele arose and disrupted the equilibrium.

The value of the selection coefficients (the fitness), averaged over all populations, showed a general increasing trend with sharper increases when an advantageous allele was introduced (Figure 3.4). This behavior was identical for the FEC and PSM models. Occasionally short declines in average selection coefficient were seen in runs of both types (not shown); this occurred when a disadvantageous allele "hitchhiked" in an individual whose other allele was highly advantageous. The advantageous allele ensured

that the individual reproduced, producing enough copies of the disadvantageous allele to lower the average selection coefficient.

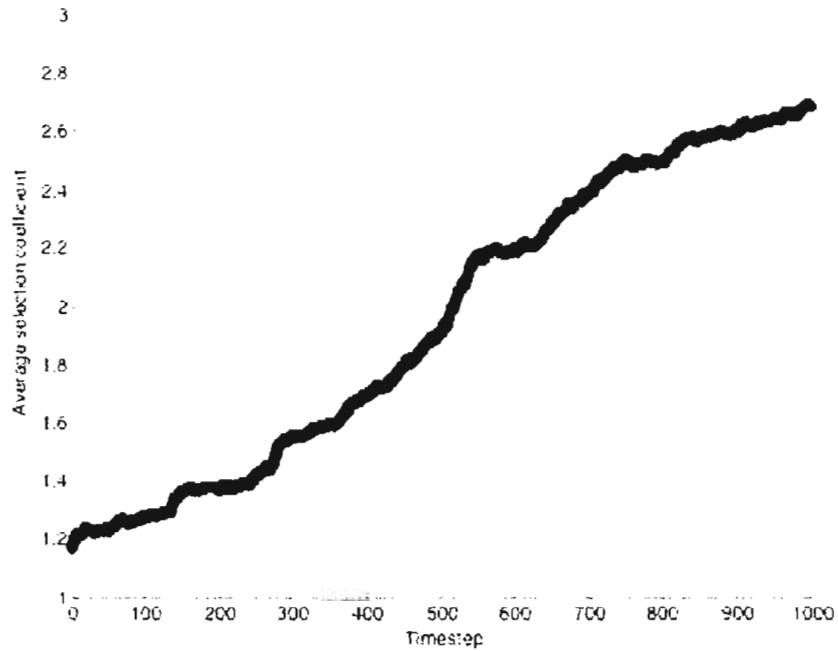


Figure 3.4 The selection coefficient averaged across all populations for the FEC run shown in Figure 3.2.

2. Genetic Differentiation Between Populations

The relationship between "geographic" distance (the number of steps between populations) and genetic distance was measured every 20 generations by sampling settlers and adults and computing F_{ST} values between all pairs of populations. The matrix of pairwise F_{ST} values was compared to the matrix of the model's actual dispersal probabilities using a Mantel test (Mantel 1967) for a negative correlation (higher F_{ST} values imply a lower migration rate). The Mantel statistic is the sum of the terms of entrywise multiplication of the two matrices (i.e., Schur multiplication). That is, if the

F_{ST} matrix is called X and the geographic distance matrix Y, the Mantel statistic is defined as

$$\tilde{Z}_{XY} = \sum_{i \neq j} X_{ij} Y_{ij} \quad (4)$$

Significance is tested by Monte Carlo simulation, where one matrix is held constant, rows and corresponding columns of the other matrix are randomly permuted, and a population of Mantel statistics from these randomized matrices is used as a null distribution against which the observed statistic is compared. Smouse et al. (1986) showed that Mantel analysis is equivalent to regression with the linear model

$$[X_{ij} - \bar{X}] = b_{YX}[Y_{ij} - \bar{Y}] + \epsilon_{ij} \quad (5)$$

where b_{YX} signifies the regression coefficient of Y on X, overbars represent taking means, and epsilon is the residual error. As another measure of similarity between the two matrices, the norm of their difference was also calculated (with smaller norm corresponding to greater similarity). These measures were calculated from a sample of settlers and a sample of adults taken at each sampling point from each population.

The Mantel analysis for the same FEC run as in Figure 3.2 revealed a negative relationship between F_{ST} values and migration, because high F_{ST} values correspond to low migration probabilities (Figure 3.5). Note that the relationship was initially not significant because all populations had similar numbers of individuals bearing the same collection of alleles, which results in low F_{ST} values (i.e., a false conclusion of high ongoing migration rate). As mutation introduced new alleles in single populations, and with migration low enough to slow their spread to other populations, the Mantel slopes

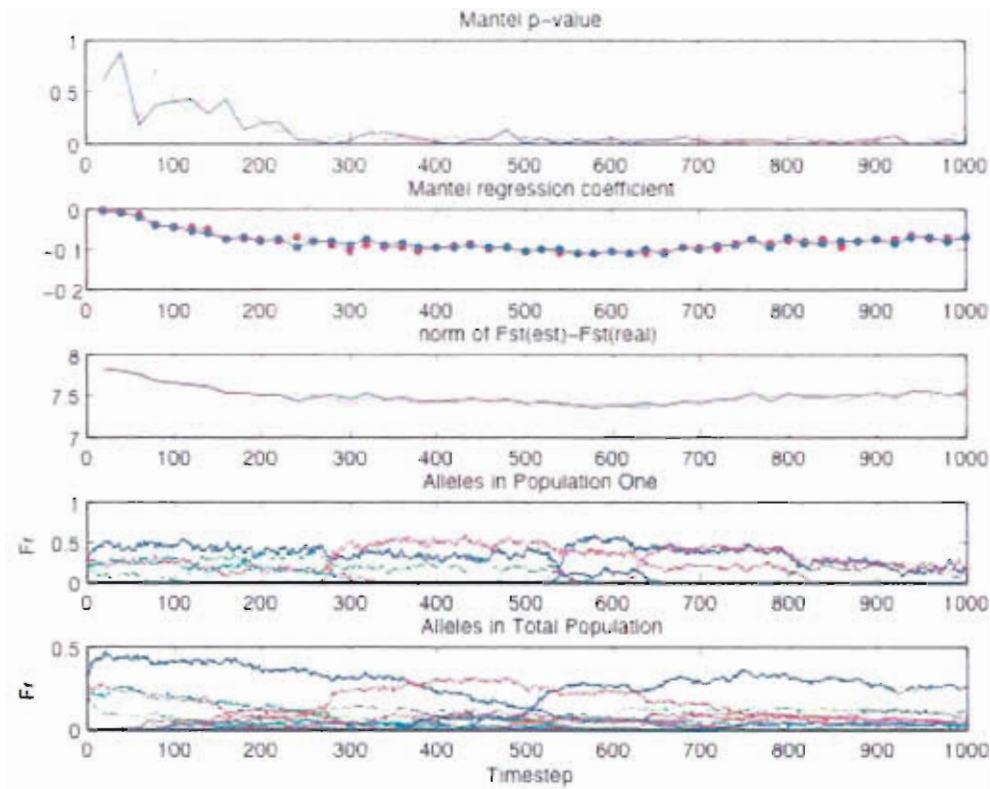


Figure 3.5 Mantel analysis for the FEC run. In the first three panels, blue lines denote quantities measured from settlers and red lines quantities from adults.

decreased and became significantly different than zero. Note also that the slopes calculated from settler and adult samples were extremely similar (mean difference in slope for all runs, $1.623 \times 10^{-4} \pm 1.704 \times 10^{-4}$ SE), as were the norms calculated from these samples. Indeed, the norm and Mantel slopes appeared highly correlated themselves, which was not surprising since both (in our analyses) measured the similarity of one matrix to another. Averaged over all runs, the settler-derived Mantel slope was greater than the adult-derived slope 52.1% of the time (i.e., little more than expected by chance alone).

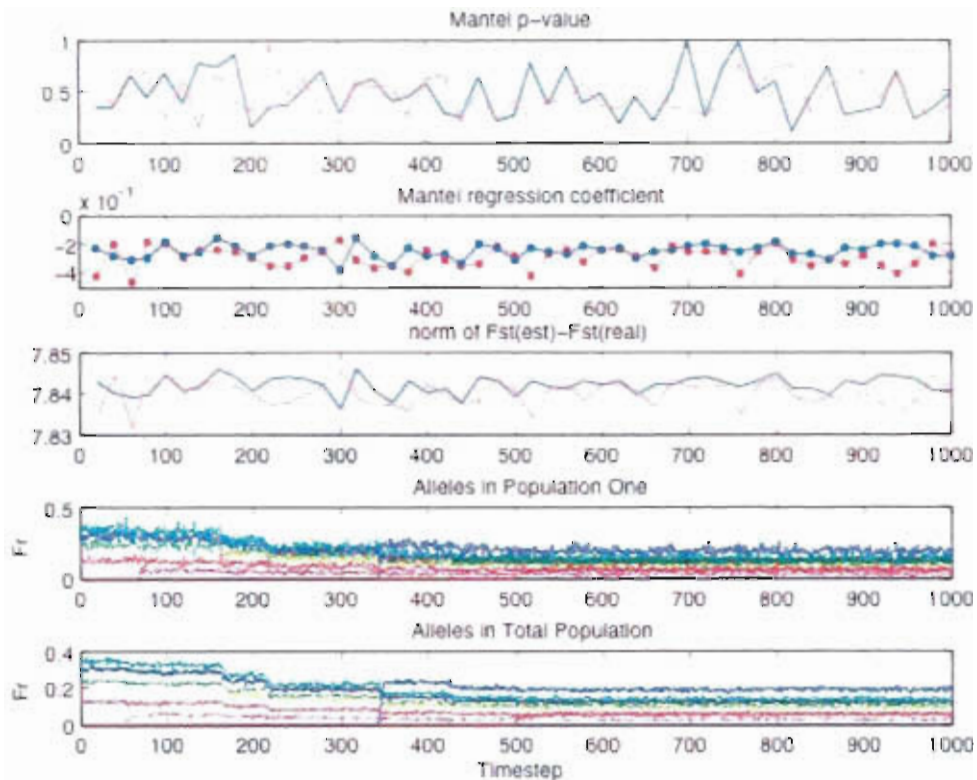


Figure 3.6 Mantel analysis for the SPSM run. In the first three panels, blue lines denote quantities measured from settlers and red lines quantities from adults.

In contrast, the same analyses performed on the PSM run revealed different patterns (Figure 3.6). Here, the settler slope was typically greater (less negative) than the adult slope for the majority of the time course. When all PSM runs were similarly analyzed, the settler slope was less steep on average for 75.6% of the time course, for a mean difference in slope over all runs of $5.816 \times 10^{-4} \pm 1.012 \times 10^{-4}$ SE. This difference, though of the same order of magnitude as in the FEC model, is more consistent. Wilcoxon signed-rank tests confirmed that the difference in slope was significantly different from zero for the PSM runs ($p=4.86 \times 10^{-5}$), but not in the FEC model ($p=0.3519$). This analysis implies a stronger isolation-by-distance effect on adult samples than on settlers in the PSM model. Such an effect was most likely caused by the

selective nature of post-settlement survival differentiating the adult populations from each other in the PSM. Settler populations would remain more similar to each other, resulting in a less pronounced trend of isolation-by-distance. This effect would not have been present in the FEC model, where neutral post-settlement selection resulted in more similar allele frequencies in settlers and adults.

The difference in the genetics of settlers and adults in the FEC and PSM models was also apparent by plotting the mean difference between F_{ST} values calculated from settlers and F_{ST} values calculated from adults (Figure 3.7). In the PSM model there was an indication that the difference between settler and adult F_{ST} values increased with increasing distance. This trend implies that the contrast between settler and adult samples

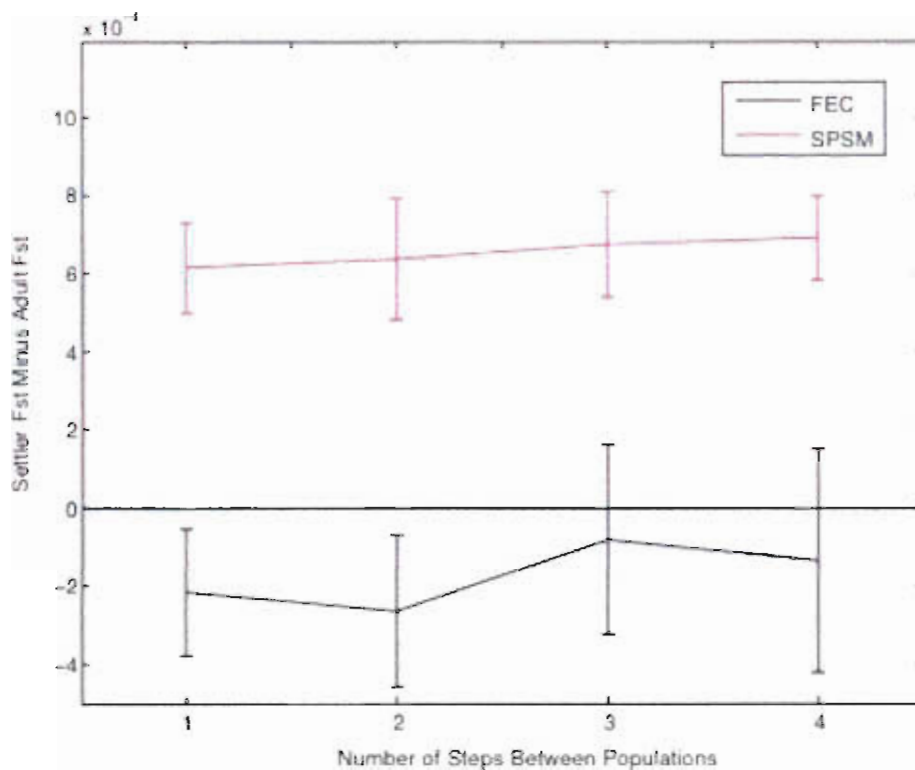


Figure 3.7 Genetic differentiation between populations calculated from settlers versus adults. Error bars indicate standard errors of the mean at each point.

may provide useful dispersal information especially for distant populations. In contrast, such a trend was not apparent in the FEC model, where settler F_{ST} values are lower than adults although only significantly so at short distances. The standard errors of these differences were larger on the whole in the FEC model, especially at larger distances. Thus, in the PSM model, settler F_{ST} values were consistently higher than adult values, whereas in the FEC model, both the sign and the magnitude of the difference were more variable.

These results imply that different types of selection produce different effects on the genetics of stages and populations, and are consistent with some observations from the field. The FEC model produced greater temporal variability in allele frequencies than the SPSM model, but only transient genetic differences between stages. On the other hand, the SPSM model produced more tightly controlled allele frequency dynamics. It was characterized by genetic differences between stages that were more consistent in time, and also different spatial genetic patterns in settlers and adults. These contrasts indicate that selection was more effective in differentiating stages and populations from each other when acting on post-settlement survival than when acting on fecundity (reproduction). Selection was required at the post-settlement stage to create significant, lasting differences between settler and adult F_{ST} values. The stronger effect of selection acting after settlement was also corroborated by the faster approach of new alleles to selection-migration equilibrium in the PSM model than in the FEC. Second, the results presented here fit well with investigations of population genetics at different life history stages. For instance, Johnson and Black (1984) described transient differences in the

genetics of recruit and adult limpets (*Siphonaria jeanae*) and presented evidence that such differences could not have been caused by selective post-settlement mortality. Our findings indicate that selective post-settlement mortality would indeed produce non-transient differences. Based on our models, strong consistent differences in the settler and adult F_{ST} values are generally not caused by variations in fecundity..

There is no evidence in our model, however, that settlers track trends in gene flow more accurately or earlier than they appear in adult samples. I found no consistent lag between settler and adult Mantel statistics after existing alleles went extinct or new alleles arose (Figures 3.5 and 3.6). On the contrary, the smaller norms and Mantel statistics for the adults in the PSM model indicate that, if anything, the adult samples are more congruent to the actual migration matrix.

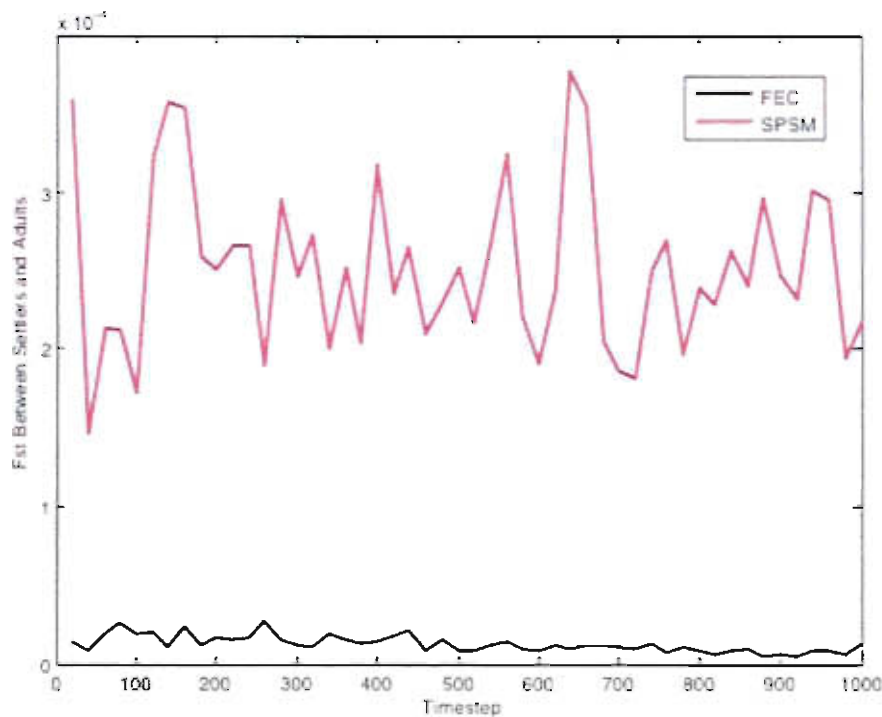


Figure 3.8 Differentiation between settlers and adults at the same site.

3. Genetic Differentiation between Juveniles and Adults in the Same Population

Neither model produced large differentiation between settlers and adults at the same site (Figure 3.8). In the FEC model, there was a slight trend towards non-zero F_{ST} values, but these values were on the order of 10^{-5} . Settler-adult differentiation was an order of magnitude larger ($\sim 10^{-4}$) in the SPSM model than in the FEC. Although the fluctuations in the FEC model appear dampened because they are plotted with the SPSM values, both models produced roughly the same degree of relative variation about the time-average. If these F_{ST} values were converted into Nm (the estimated number of migrants) using the standard equation from Wright's island model, $F_{ST} = \frac{1}{1+4Nm}$ (bearing in mind that there is no migration *sensu stricto* between settlers and adults), one would estimate approximately 10^4 "migrants" per generation in the FEC model and approximately 10^3 "migrants" in the SPSM. These values are extremely high, given the rule-of-thumb that migration on the order of 10^0 individuals per generation is sufficient to maintain genetic homogeneity. It is important to note also that low F_{ST} values lead to notoriously inaccurate Nm estimates because of the hyperbolic relationship between the two; however, the main conclusion is that the FEC model produced settler populations that are genetically more similar to the surviving adults than did the SPSM, which again points to a stronger effect of selection on genetic differentiation in the post-settlement period than during reproduction.

Conclusions

In summary, selection appears to be more effective at maintaining different allele frequencies when operating on fecundity than on the post-settlement period. Selection only created temporally lasting differences between stages and sites when operating on settlement. In addition, a selective post-settlement period seems to create consistently higher F_{ST} values in settlers than in adults, and this effect is most pronounced over larger distances. Selective post-settlement mortality also creates greater genetic differentiation between juveniles and adults at a single location than does fecundity selection. The models therefore indicate that consistent differences in the spatial genetic differentiation of settlers versus adults implicate selection acting after settlement. Further, genetic samples of settlers do not appear to readily offer more accurate estimates of dispersal than do traditional samples of adults. It is possible that more complex analyses than were possible here will confirm the perceived advantages of sub-adult genetic samples. Alternatively, these advantages may not have been obvious in our models because of differences between the way I modeled the biology and the biology of real populations.

The results presented here fit well with some field observations, but also provide alternative explanations for the causes of genetic differentiation. In both the model and in the field (Johnson and Black 1984), ephemeral differences between settlers and adults were not caused by selective post-settlement mortality (in the model they were a result of fecundity selection). In keeping with the suggestion of Moberg and Burton (2000), intransient differences were caused in the model only when post-settlement survival involved selection. In neither case could neutral genetic drift cause differences like those

seen in nature. In contrast to observational work, however, the nature of selection did not have to be spatially variable to create spatial genetic variation. That is, all populations in my model experienced the same selection regime; particularly in the selective SPSM model selection apparently altered the probabilities of survival enough that different populations wound up with measurably different genetic compositions. My models likewise did not incorporate variation in the timing of reproduction between populations, yet still differentiated those populations genetically.

The results of these model analyses represent the first steps towards understanding which aspects of marine benthic invertebrate life histories need to be included in population genetics models. As researchers increasingly focus on the genetics of sub-adult samples, models such as mine will show how forces acting at multiple life stages interact to create the genetic patterns we are used to seeing in adults. Clearly, there is much more work to be done with this type of modeling. A more thorough exploration of the parameter space, coupled perhaps with new methods of genetic analysis, would go far in deepening our understanding of stage-structured population genetics. However, the relevance of these initial findings to hypotheses that are already in the literature is clear, and point towards the utility of such modeling efforts.

References

- Avise JC (2000) *Phylogeography. The history and formation of species* Harvard University Press, Cambridge, MA, 447pp.
- Crivello, J.F., D.J. Danila, E. Lorda, M. Keser and E.F. Roseman. 2004. The genetic stock structure of larval and juvenile winter flounder larvae in Connecticut waters of eastern Long Island Sound and estimations of larval entrainment. *Journal of Fish Biology* **65**: 62.
- Drouin, C.A., E. Bourget and R. Tremblay. 2002. Larval transport processes of barnacle larvae in the vicinity of the interface between two genetically different populations of *Semibalanus balanoides*. *Marine Ecology-Progress Series* **229**: 165.
- Hellberg, M.E. 1996. Dependence of gene flow on geographic distance in two solitary corals with different larval dispersal capabilities. *Evolution* **50**: 1167.
- Johnson, M.S. and R. Black. 1984. Pattern beneath the chaos: the effect of recruitment on genetic patchiness in an intertidal limpet. *Evolution* **38**: 1371-1383.
- Johnson, M.S. and J. Wernham. 1999. Temporal variation of recruits as a basis of ephemeral genetic heterogeneity in the western rock lobster *Panulirus cygnus*. *Marine Biology* **135**: 133.
- Kimura, M. and G.H. Weiss. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**: 561-576.
- Lande R (1975) The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetical Research* **26**: 221-235.
- Li G, Hedgecock D (1998) Genetic heterogeneity, detected by PCR SSCP, among samples of larval Pacific oysters (*Crassostrea gigas*) supports the hypothesis of large variance in reproductive success. *Canadian Journal of Fisheries and Aquatic Science* **55**: 1025-1033.
- Hunt HL, Scheibling RE (1997) Role of early post-settlement mortality in recruitment of benthic marine invertebrates. *Marine Ecology Progress Series*, **155**, 269-301.
- Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.
- Moberg PE, Burton RS (2000) Genetic heterogeneity among adult and recruit red sea urchins, *Strongylocentrotus franciscanus*. *Marine Biology* **136**: 773-784.

- Ohta, T, Kimura M (1971) On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution*, **1**: 18-25.
- Planes, S, Romans P (2004) Evidence of genetic selection for growth in new recruits of marine fish. *Molecular Ecology* **13**: 2049-2060.
- Slatkin, M (1985) Rare alleles as indicators of gene flow. *Evolution* **39**: 53-65.
- Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627-632.
- Wright S (1943) Isolation by distance. *Genetics*, **28**, 114-138.

Chapter 5: Conclusions

General Summary

The general goal of this dissertation was to explore new avenues in marine invertebrate (specifically annelid) genetics, at levels ranging from the phylogenetic to the phylogeographic.

Chapter 1: Annelid phylogenetics based on complete mitochondrial genomes

I have used the complete mitochondrial genome of the bamboo worm *Clymenella torquata* and an estimated 80% of the mt-genome of the hydrothermal vent tubeworm *Riftia pachyptila* to describe patterns of molecular evolution in annelids and to resolve systematic uncertainties among major annelid groups. Several previously described trends for annelid mt-genomes (cf. Boore and Brown 2000, Boore and Staton 2002) were supported by this work, namely 1) annelid mt-genomes show, as do most invertebrate mt-genomes, large AT bias and significant negative GC-skew in the coding strand, especially at third codon positions; 2) all known annelid mt-genes are encoded on a single strand; and 3) mt-gene order is conserved across Annelida to a greater extent than is seen in related taxa (e.g. mollusks and brachiopods). In spite of the fact that *Riftia pachyptila* inhabits strongly reducing, potentially high-temperature habitats and is presumably a derived species, its mt-genome is extremely annelid-like. Specifically, the GC-content of its mt-genome is as low as most other annelids and invertebrates, arguing against environmentally driven molecular evolution at this level as has been previously suggested from nuclear genetic analyses (Dixon et al. 1992).

Although several methodologies were explored to construct sequence-based phylogenetic trees from annelid mt-genomes, I recovered extremely similar topologies with similar support. The phylogenetic trees, which include all major annelid clades for the first time, clearly indicate that siboglinids are derived polychaetes, and suggest that maldanids are among the most basal annelid taxa. Further, they confirm previous work indicating that clitellates (leeches and oligochaetes) are simply highly derived polychaetes, making the last common ancestor of “Polychaeta” and “Annelida” one and the same. Although I present preliminary evidence that Sipuncula may also fall within the polychaete radiation, this arrangement is speculative. More conservatively, I confirm Boore and Staton’s (2002) finding that sipunculans are very closely related to polychaetes.

Chapter 2: Phylogeographic patterns of C. torquata in the Northwest Atlantic using genes obtained from the complete mt-genome

In this work I used the complete mt-genome of *C. torquata* to select quickly evolving genes for high-resolution gene flow analyses in the Northwest Atlantic. This phylogeographic dataset was used to test unresolved hypotheses as to how Cape Cod, MA acts as a barrier to gene flow. Whereas previous work (reviewed in Wares 2002) described evidence from several marine species that there is a phylogeographic barrier in the vicinity of Cape Cod, the closer spacing of my sampled populations combined with *C. torquata*’s low dispersal potential allowed a finer scale determination of this barrier’s location. A sharp, but not complete, shift in mitochondrial haplotypes places the break in

the region between Cape Cod/Buzzard's Bay and Long Island, and not on the Cape peninsula itself. Smaller differences between all Cape Cod sites and sites in the Gulf of Maine and Bay of Fundy provide new evidence that short-distance dispersal on Cape Cod may be more prevalent than previously thought, and appears to be constrained less by the direction of local coastal currents. No significant effect of the Cape Cod Canal increasing dispersal around Cape Cod was found. I demonstrate that, although the patterns of gene flow in *C. torquata* are complex, the differences in salinity and temperature of major oceanographic water bodies near Cape Cod appear to be the dominant force structuring its genetics.

Imposed upon these gene flow patterns is another phylogeographic pattern: I demonstrate that the impact of glaciation and ice sheet melt-back have shaped levels of genetic diversity in Northwest Atlantic populations of *C. torquata*, as has been described for other species. The lower genetic diversity in northern Gulf of Maine sites than in Cape Cod and southern sites is consistent with the removal of *C. torquata* from glaciated sites, followed by gradual northward reintroduction of *C. torquata* after glacial retreat. This pattern also suggests that glacial dynamics in the Northwest Atlantic may tend to reduce genetic diversity in northern populations of intertidal species, whereas they have a lesser or even reversed effect on subtidal species (many of which are unaffected by intertidal ice scour and may already be adapted to cold environments).

Chapter 3: Modeling the life histories of marine invertebrates in population genetics

Often, marine invertebrates are selected for phylogeographic investigations because their life cycles fit the assumptions of classic population genetic models; however, rarely are the specifics of these life cycles explicitly incorporated into existing models. In this chapter I explored how the emerging field of short time scale population genetics is increasingly turning to sub-adult samples to focus on ecological questions (e.g. single dispersal events or large mortality events). Specifically, I investigated whether a stage-structured model including selection acting at various life stages could produce different genetic patterns in different stages, as has been observed empirically (e.g. Moberg and Burton 2000).

My model analyses indicate that selection may be a more important force in structuring spatial genetic differentiation than has been suggested by the long-popular neutral and nearly-neutral schools of population genetics. Selection acting on either reproduction (i.e., differing fecundities) or post-settlement mortality (i.e., differing probabilities of survival to adulthood) created much larger genetic differentiation between populations than even severe genetic drift (a neutral process).

More specifically, while fecundity selection appeared to exert a stronger control on allele frequency dynamics, selective post-settlement mortality produced larger effects on genetic differentiation between populations and between life stages. Indeed, only with selective post-settlement mortality were consistent differences recovered. As expected, the differences between gene flow estimates of settlers and adults were larger for more distant populations, implying that settler/adult contrasts at multiple spatial scales may

reveal more fine-scale patterns of phylogeography than either could alone. Finally, while only selective post-settlement mortality produced notable differences between settlers and adults in a single site, these differences were small, in contrast to what has been seen in the field. It is likely that more complex models and genetic analyses than were possible here are needed to make best use of genetic analyses of settlers and adults.

Broader impacts and future directions

This thesis spans several areas, from the phylogenetic to the population genetic, and from the empirical to the theoretical, and as such is an example of the ever-broadening reach of molecular biology. The phylogenetic work in Chapter One will contribute to future efforts to resolve fine scale relationships in the Polychaeta, and demonstrate the potential of mt-genomes to do so. Future work incorporating more polychaete mt-genomes will help elucidate how conserved gene order is in worms, as well as which polychaete groups are most closely related to clitellates. This work also offers insight into the methodology of gene order analysis, and will be useful for future efforts to create more sophisticated evolutionary models. Specifically, the development of tree-building procedures and ancestral state reconstructions tailored to gene orders will go far in producing better trees.

This thesis also furthers previous work in the Northwest Atlantic. The increased resolution of the population genetics samples implies more strongly than previous work that differences in oceanic water bodies near Cape Cod are a significant driving force in separating populations in this region. Although I did not conduct a thorough test of an

alternative dispersal route through the Cape Cod Canal, my results would make a more detailed sampling regime an interesting further contribution. Also, similar studies on organisms that disperse over a broader range of scales than *Clymenella torquata* would provide additional insight as to how universal restrictions to gene flow are. Collecting closely-spaced samples of *C. torquata* (or a similarly weakly dispersive species) in Buzzards Bay, Block Island Sound, and along both coasts of Long Island will allow further determination of which areas show genetic isolation.

Finally, the stage-structured population genetics model raises many questions that will require further modeling and analysis to answer. More detailed analysis of stage-structured models will hopefully provide insight as to whether trends and changes in gene flow would appear in sub-adult samples earlier than in adult samples. This work is the first to present preliminary theoretical evidence, in keeping with some empirical observations, that selection during the post-settlement period may have a large effect on genetic structure at all levels. Future work could involve exploring precisely how much selection is required to create a given amount of genetic difference between adults and sub-adults. Perhaps to a greater extent, efforts to compare the change in this differential across spatial scales should go far in providing a more detailed snapshot of dispersal in marine benthic invertebrates.

References

- Boore, J.L. and W.M. Brown. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: Sequence and gene rearrangement comparisons indicate the Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Molecular Biology and Evolution* **17**: 87-106.
- Boore, J.L. and J.L. Ståton. 2002. The mitochondrial genome of the Sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. *Mol. Biol. Evol.* **19**: 127-37.
- Dixon, D.R., R. Simpson-White and L.R.J. Dixon. 1992. Evidence for thermal stability of ribosomal DNA sequences in hydrothermal-vent organisms. *J. Mar. Biol. Assoc. U.K.* **72**: 519-527.
- Moberg, P.E. and R.S. Burton. 2000. Genetic heterogeneity among adult and recruit red sea urchins, *Strongylocentrotus franciscanus*. *Marine Biology* **136**: 773.
- Wares, J.P. 2002. Community genetics in the northwestern Atlantic intertidal. *Molecular Ecology [Mol. Ecol.]*. **11**: 1131-1144.

Appendix 1: PCR amplification and primer information for Chapter One

Table A1.1 PCR cycling profile for *C. torquata* long PCRs (EXL polymerase)

initial denaturation	92°C	2 min.
10 cycles of:	92°C	30 sec.
	xx	1 min.
	68°C	1 min. per kb target length
20 cycles of:	92°C	30 sec.
	xx	1 min.
	68°C	1 min. per kb target length
		+30 sec. per cycle
hold at 4°C.		
annealing temperatures:		
<i>mLSU-cox1</i> :	48°C	
<i>cox1-cox3</i> :	45°C	
<i>cox3-cob</i> :	45°C	
<i>cob-mLSU</i> :	42°C	

Table A1.2 Primers used in mt-genome PCR and sequencing.

Species	Primer	Sequence	PCR (P) or walking (W)	5' pos.*	Gene
"Universal" primers					
	LCO1490†	GGTCAACAAATCATAAAGATATTGG	P,W	14	<i>cox1</i>
	HCO2198†	TAAACTTCAGGGTGACCAAAAAATCA	P,W	722	<i>cox1</i>
	COIII ^o	TGGTGGCGAGATGTTKKTNCNGA	P,W	2803	<i>cox3</i>
	COIII ^r	ACWACGTCKACGAAGTGTCARTATCA	P,W	3377	<i>cox3</i>
	nad4f	TGRGGNTATCARCCNGARCG	P,W	8727	<i>nad4</i>

nad4r	GCYTCNACRTGNGCYTTNGG	P,W	8980	nad4
16Sar-L0	CGCCTGTTTATCAAAAACAT	P,W	10848	mLSU
16SbrH0	CCGGTCTGAACTCAGCTCATGT	P,W	11889	mLSU
<i>C. torquata</i>				
Ctcox1f1	CACAGCATTCTTTGACCCAGCAGG	W	639	cox1
Ctcox1r1	GATGAGCCCATAACAATAAACCC	W	841	cox1
Ctnad6r	GTGGATGGGATTTTCGTATAGGCCTAAAC	W	3959	nad6
Ctcobr1	GTAATAACTGTAGCGCCCCA	W	4380	cob
Ctcobf5	TCACGACGATCTACCTCATTCTACCCG	W	4890	cob
CtWr	GCATCAGGAATTAAGAATCTATC	W	5138	trnW
Ctatp6f1	GGACTATCTCTATGATTTGCCATCCTATTATC	W	5487	atp6
Ctatp6f2	TCTGCCTGATTCAAGCTTATATTTTACT	P,W	5791	atp6
CtRf	CAATTTATGCATTTTGGTTTCGG	W	5863	trnR
CtArgR	TTGCCACCTTTTAATGAATGA	W	5886	trnR
Ctnad5r7	GATTGTTGTTTTATTATAT	W	6144	nad5
Ctnad5rrc	GCTAGGTTTAACTTCTTTCTTAC	W	6411	nad5
Ctnad5r6	GGATTTAGCATTTTGGTAGTAA	W	6441	nad5
Ctnad5r4	GGAGGTCTTGATTATGGAGGTGAACATG	W	7000	nad5
Ctnad5r4rc	TGTTACCTCCATAATCAAGACCTCCGG	W	7028	nad5
Ctnad5r2	GAAATTAAGAGATAGACAAAGAACC	W	7232	nad5
CtFr1	GGGAATATCTTCATCTAAACAGCTTCAGTG	W	7801	trnF
Ctnad4r2	CACCCTAGAATAAGAAATAATG	W	8707	nad4
CtMf1	ATGCCCCGAAAATGGTT	W	9793	trnM
CtmSSUf2	TTTCCGTCTAATTTATGCTGTGA	W	10373	mSSU
CtValr1	CAGCGTAAGTGCAATGTGTCTCC	W	10682	trnV
CtmLSUr1	AAGTAGGGATTTGCCGAGTTC	W	11324	mLSU
Ctnad2f1	GGTGGAATAGGGGTATAAAATCAAACACAACCT	W	14178	nad2
<i>R. pachytila</i>				
Rpnad4bf	CCCATATTCTCTCTCCACCTTTGACTTCC	P,W	474	nad4
Rpnad4br	GGAAGTCAAAGGTGGAGAGAGAGAATATGGG	P,W	503	nad4
Rpnad43f	GCTTATTCCTCTATTGGACAT	W	712	nad4
Rpnad43r	ATGTCCAATAGAGGAATAAGC	W	732	nad4

Rpnad44f	CCAGAGTTAATTTGTAAGTGA	W	1204	<i>nad4</i>
Rpnad44r	TCAGTTACAAATTAAGTCTGG	W	1224	<i>nad4</i>
Rp12Sf	GGCACTACAAACACAGGTTAAAAC	W	1788	<i>mSSU</i>
Rp12Sr	GTTTTAAACCTGTGTTTGTAGTGCC	W	1812	<i>mSSU</i>
Rprnlf	GCTAACTTTTAAATAACGTTATAG	W	2670	<i>mLSU</i>
Rprnlr	CTATAACGTTATTTAAAAGTTAGC	W	2693	<i>mLSU</i>
RpmLSUr	CCATTGAACTAAGAGTCATTGGGCAG	W	2932	<i>mLSU</i>
Rp-50f	GCACCTCGATGTTGGCTTA	W	3335	<i>mLSU</i>
Rp-50r	TAAGCCAACATCGAGGTGC	W	3353	<i>mLSU</i>
Rp431f	TATTCAAATTCGAAAAGGACC	W	3926	<i>nad1</i>
Rp431r	GGTCCTTTTCGAATTTGAATA	W	3946	<i>nad1</i>
Rp920f	CCGAACTCCTTTTCGATTTATC	W	4415	<i>nad1</i>
Rp920r	GATAAATCGAAAGGAGTTCGG	W	4435	<i>nad1</i>
Rp1296f	GATGATGTTAATCACGAGAGC	W	4791	<i>trn1</i>
Rp1458f	ATTACTAAGAAACCGACC	W	4976	<i>nad3</i>
Rp1458r	GGTCGGTTTCTTAGTAAT	W	4993	<i>nad3</i>
Rp1860f	GAACCTTCCTCTCTATTTTCAGC	W	5380	<i>nad2</i>
Rp1860r	GCTGAAATAGAGAGGAAGGTTC	W	5401	<i>nad2</i>
Rp2518f	CAATTAATCACTGCAAGCC	W	6033	<i>nad2</i>
Rp2518r	GGCTTGCAGTGATTAATTG	W	6050	<i>nad2</i>
COIRp1536f	GTAGCCACTAGAATAAGACTCTTAATT	P,W	6937	<i>cox1</i>
COIRp1536r	CTCGAATTAAGAGTCTTATTCTAGTGGCTAC	P,W	6423	<i>cox1</i>
COIRp2161f	ATCTAAACACTTCTTTCTTCGATCCTGCAGG	P,W	6953	<i>cox1</i>
COIRp2161r	CCTGCAGGATCGAAGAAAGAAGTGTTTAGAT	P,W	6983	<i>cox1</i>
Rp3845f	GTAGGAGGATTAACAGGAATTG	W	7360	<i>cox1</i>
Rp3845r	CAATTCCTGTTAATCCTCCTAC	W	7381	<i>cox1</i>
Rp4489f	CTTCACGACCATGCTCTAACCATC	W	8006	<i>cox2</i>
Rp4489r	GATGGTTAGAGCATGGTCGTGAAG	W	8029	<i>cox2</i>
Rp5222f	CTAACCTACTCCTCCTTCTATCGAC	W	8741	<i>atp8</i>
Rp5222r	GTCGATAGAAGGAGGAGTAGGGTTAG	W	8766	<i>atp8</i>
Rp5674r	GAATTATTGCTCAGCGTAGGCCTC	W	9196	<i>cox3</i>
Rp6010f	GCCGATAGAGCTTATGGCACC	W	9531	<i>cox3</i>

Rp6816f	CCGCATTAGTAGATCTTCCAGC	W	10346	<i>cob</i>
Rp6816r	GCTGGAAGATCTACTAATGCGG	W	10367	<i>cob</i>
Rpcobr354 ^{††}	CCAATATTTTCATGTTTC	W	10646	<i>cob</i>
CytBRpf	CAATGATTGTGAGGTGGATTTAGAGTAAGA	P,W	10783	<i>cob</i>
CytBRp	TCTTACTCTAAATCCAGGTCACAATCATTG	P,W	10812	<i>cob</i>
Rp5029f	CCTTGGAGCTATATTTACCGCC	W	11343	<i>cob</i>
Rpcobr825	AAGTATCATTCTGGTTTAATATG	W	11120	<i>cob</i>

* Locations for universal primers reflect positions in the genome of *C. torquata*.

[†]Folmer *et al.* 1994

[°]Boore and Brown 2000

[°]Palumbi *et al.* 1991

^{††}*R. pachyptila*-specific version of *cob354* from Boore and Brown (2000)

Table A1.3 PCR cycling profile for *R. pachyptila* long PCRs (vent/rTth polymerases)

initial denaturation 94°C 1 min.
 35 cycles of: 94°C 20 sec.
 xx°C 20 sec.
 72°C 1 min. per kb target length
 final extension: 72°C 7 min.
 hold at 4°C.

annealing temperatures:

mLSU-cox1: 50°C

cox1-cox3: 45°C

cox3-cob: 45°C

nad4-mLSU: 42°C

Appendix 2: MATLAB model codes used in Chapter Four

A. Files used in FEC Model

```
function FEC(A,gamma,TOT)
% PRELIMINARIES
%   TOT=number of timesteps to run
%   gamma=0 :: house of cards model (any state
%   acheivable from any other state)
%   gamma=1 :: random walk (new state is a normal rv
%   based on old state)
phi=200;           % "neutral" fecundity
mu=1e-7;          % mutation rate
d=10e5;           % strength parameter for
migration exponential decay  $f=\alpha*d^{(-distance)}$ 

ae=0;             % mean of the normal distribution of
allelic effect (under HOC) OR
                 % mean of the normal distribution of
the change in allelic effect (under RW)
stdae=.1;        % std of the normal distribution of
allelic effect (under HOC) OR
                 % std of the normal distribution of the
change in allelic effect (under RW)
N=200;           % number of surviving juveniles... so
also the # of adults per pop
sampsiz=30;      % number of individuals to draw for Fst
samples
p=size(A,3);    % number of populations
nummut=0;       % recorder for number of mutations
juvnum=[0 0];   % recorder for number of juveniles
space=20;       % spacing to record juvie and adult
genetics
gfrequ1=zeros(size(A,1),TOT); % allele
frequencies from population 1
gfrequT=zeros(size(A,1),TOT); % allele
frequencies averaged over total population
FMonA=zeros(size(A,3),size(A,3)); % recorder for
sampled adult Fst's
FfullA=zeros(size(A,3),size(A,3)); % recorder for
complete adult Fst's
FMonJ=zeros(size(A,3),size(A,3)); % recorder for
sampled juvenile Fst's
FfullJ=zeros(size(A,3),size(A,3)); % recorder for
complete juvenile Fst's
rand('state',sum(100*clock)) % reset random
```

```

number generator
cutoff=0; % proportion of
lowest fecundities that do not mate

% initial exponential setup for migration component
if rem(p,2)==0; % even number of populations... use
alpha for even configuration
    alpha=((d-1)*d^(p/2))/(d^(p/2)-d+d^(p/2+1)-1);
    realJMon=zeros(1,p/2+1);
else % odd number of populations... use
alpha for odd configuration
    alpha=((d-1)*d^((p-1)/2))/(d^((p-1)/2)-2+d^((p-
1)/2+1));
    realJMon=zeros(1,(p-1)/2+1);
end

% make the shift matrix 'short', where short(i,:) contains
the shortest
% number of steps (the shift) between population i and the
other populations.
% Then make the matrix MIG, where MIG(i,:) contains the
fraction of larvae
% from population i migrating to all populations (including
those that
% "migrate" to itself).
for k=1:p;
    short(1,k)=min(abs(1-k),abs(p+1-k));
end
short1=alpha*d.^(-short);
for k=1:p;
    MIG(k,:)=circshift(short1,[0,k-1]);
end
cMIG=[zeros(p,1) cumsum(MIG,2)];

% choose initial allelic affects
%s=ones(size(A,1),1); %neutral alleles
s=randn(size(A,1),1).*stdae+ae; %non neutral alleles
S=zeros(length(s),length(s));
for i=1:size(A,1);
    for j=1:size(A,2);
        S(i,j)=exp(s(i)+s(j));
    end
end
Stemp=reshape(S,1,numel(S));

```

```

Stemp(Stemp<=quantile(Stemp,cutoff))=0;
S=reshape(S,length(s),length(s));

C=mean(A,3);
s1=sum(sum(S.*C))./sum(sum(C));
Strend=mean(s1);

file=input('Enter name of file to save data to (.mat
file)\n','s');
mypath='newoutput/';
file=strcat(mypath,file);

% BEGIN COMPONENTS FOR EACH STEP OF MODEL
for T=1:TOT;          % master loop for timesteps;
T

% 1. GAMETES
% produces gamete matrix G (m x K) from adult matrix A
disp('gametes')
[G,s,S]=gametes(A,s,phi,cutoff);

% 2. MUTATION
disp('mutation')
[G,s,S,muvec]=mutation(G,s,gamma,mu,stdae,ae,cutoff);
nummutts(T+1)=nummutts(T)+sum(muvec);

%% ALLELE TRACKING
gfreq1(1:size(G,1),T)=G(:,1)./sum(G(:,1));
gfreqT(1:size(G,1),T)=mean(G,2)./sum(mean(G,2));

% 3. LARVAE
disp('larvae')
L=larvae(G);
larvnum(T,:)=shiftdim(sum(sum(L)))';

% 4. MIGRATION
disp('migration')
J=migration2(L,MIG);
if rem(T,space)==0;
    juvnum(end+1,1:2)=[min(sum(sum(J,2)))]';
    max(sum(sum(J,2)))]';

```

```

    sJ=survive(J,sampsize);
    FfullJ(:,:,end+1)=Fst(J);
    FMonJ(:,:,end+1)=Fst(sJ);
    Jcell(T/space)={[J]};
    %realJMon(end+1,:)=realJ;
end

if rem(T,space)==0;
    del=size(J,1)-size(A,1);
    PreA=padarray(A,[del,del],'post');
    PrePreComp(:,:,1,1:9)=J;
    PrePreComp(:,:,2,1:9)=PreA;
    for i=1:9;

[d1,d2,PreComp(T/space,i)]=find(Fst(PrePreComp(:,:,i)));
        end
    end
clear PrePreComp

% 5. SURVIVE
disp('survival')
%keepJ=survive(J,9*N/10); %these
lines for iteroparity
%iteroA=survive(A,N/10);
%newal=size(J,2)-size(A,2);
%newA=padarray(iteroA,[newal,newal],'post');
%A=keepJ+newA;
A=survive(J,N); %this line
for semelparity

if rem(T,space)==0;
    AJComp(:,:,1,1:9)=J;
    AJComp(:,:,2,1:9)=A;
    for i=1:9;

[d1,d2,FComp(T/space,i)]=find(Fst(AJComp(:,:,i)));
        end
    end
clear AJComp

% matrix cleanup... find genotypes that didn't survive

clear a
a=find(sum(J(:,:,1))==0);
for k=2:size(J,3);

```

```

        anew=find(sum(J(:,:,k))==0);
        a=intersect(a,anew);
    end
    diff(T)=length(a);
    A(a,:,:)=[];
    A(:,a,:)=[];
    s(a)=[]; % this should remove extinct alleles' selection
    coeff's from s
    S(a,:)=[];
    S(:,a)=[];
    J(a,:,:)=[];
    J(:,a,:)=[];

    numal(T+1)=size(A,2);
    if rem(T,space)==0;
        sA=survive(A,sampsize);
        FfullA(:,:,end+1)=Fst(A);
        FMonA(:,:,end+1)=Fst(sA);
        Acell(T/space)={[A]};
    end

    popsize(:,T)=shiftdim(sum(sum(A)),2);
    C=mean(A,3);
    s1=sum(sum(S.*C))./sum(sum(C));
    Strend(T+1)=mean(s1);

end % end for master T loop starting on line 66

%% Fst and Nm block
%realJMon(1,:)=[];
%juvnum(end+1,1:2)=[min(sum(sum(J,2)))/
max(sum(sum(J,2)))]];
%juvnum(1,:)=[];
fj=survive(J,sampsize); % sample of Juveniles
fa=survive(A,sampsize); % sample of Adults
FstJ=Fst(fj); % compute Fst's for both
samples
FstA=Fst(fa);
FJ=0;
FA=0;
for i=1:p-1;
    for j=i+1:p;
        FJ(end+1)=FstJ(i,j);
        FA(end+1)=FstA(i,j);
    end
end

```

```

        end
    end
    FJ(1)=[ ];
    FA(1)=[ ];
    NmJ=(1/4)*(1./FJ-1);
    NmA=(1/4)*(1./FA-1);

    s=sprintf('x%d', p);
    load(s);

%%DATA SAVE
save(file,'file','PreComp','FComp','cutoff','FfullA','Ffull
J','Acell','Jcell','larvnum','space','T','diff','juvnum','s
ampsize','x','d','phi','alpha','MIG','gamma','N','A','FstJ'
,'FstA','NmJ','NmA','popsize','numal','Strend','nummut','g
freq1','gfreqT','FMonA','FMonJ');

%% Fst GRAPHICS
figure
hold on
p1=scatter(x,NmJ,'filled','b');
Jls=robustfit(x,NmJ);
p3=plot(x,Jls(1)+Jls(2)*x,'b');
p4=scatter(x,NmA,'filled','r');
Als=robustfit(x,NmA);
p6=plot(x,Als(1)+Als(2)*x,'r');

if rem(p,2)==0;
    p7=plot(0:p/2,MIG(1,1:p/2+1).*sum(sum(mean(A,3))),'k');
else
    p7=plot(0:(p-1)/2,MIG(1,1:(p-
1)/2+1).*sum(sum(mean(A,3))),'k');
end
legend([p1; p4; p7],'Juvenile Nm data and trend','Adult Nm
data and trend','Real dispersal curve')

%%ALLELE TRACKING GRAPHICS
figure
subplot(2,1,1), plot(1:T,gfreq1)
hold on
subplot(2,1,2),plot(1:T,gfreqT)

%% DEMOGRAPHIC GRAPHICS
figure
scatter(1:T+1,numal,'filled','r')

```



```
ylabel('Total number of alleles'), xlabel('Simulation Time-  
step')
```

```
%SELECTION GRAPHICS
```

```
figure
```

```
scatter(1:T+1,Strend,'filled','b')
```

```
ylabel('Average selection coefficient'), xlabel('Simulation  
Time-step')
```

```
figure
```

```
scatter(1:T+1,nummut,'filled','g'), xlabel('Simulation  
Time-step')
```

```
ylabel('Total number of mutations that have occurred')
```

```
function [G,s,S]=gametes(A,s,phi,cutoff);
```

```
for i=1:length(s);
```

```
    for j=1:length(s);
```

```
        S(i,j)=exp(s(i)+s(j));
```

```
    end
```

```
end
```

```
Stemp=reshape(S,1,numel(S));
```

```
Stemp(Stemp<=quantile(Stemp,cutoff))=0;
```

```
S=reshape(S,length(s),length(s));
```

```
G=[];
```

```
[ar,ac,ap]=size(A);
```

```
for k=1:ap;
```

```
    for i=1:ar;
```

```
G(i,k)=floor(phi/2*(sum(S(i,:).*A(i,:,k))+sum(S(:,i).*A(:,i  
,k))));
```

```
    end
```

```
end
```

```
function A=survive(J,N);
```

```
% picks survivors from the juvenile matrix J and
```

```
% puts them in adult matrix A.
```

```
[jr,jc,jp]=size(J);
```

```
A=zeros(size(J));
```

```
for k=1:jp;
```

```
    imat=[0];
```

```
    jmat=[0];
```

```
    for i=1:jr;
```

```

        for j=1:jc;
            imat(end+1:end+J(i,j,k))=i*ones(1,J(i,j,k));
            jmat(end+1:end+J(i,j,k))=j*ones(1,J(i,j,k));
        end
    end
end
[b,c,p]=find(imat);
[b,c,q]=find(jmat);
x=[p' q'];
[g,d]=size(x);
y=randsample(1:g,min(N,length(x)));
for v=1:length(y);

    A(x(y(v),1),x(y(v),2),k)=A(x(y(v),1),x(y(v),2),k)+1;
end

end
clear imat jmat;

```

B. Files used in the SPSM model

```

function SPSM(A,gamma,TOT)
% PRELIMINARIES
%   TOT=number of timesteps to run
%   gamma=0 :: house of cards model (any state
%   acheivable from any other state)
%   gamma=1 :: random walk (new state is a normal rv
%   based on old state)
phi=200;           % "neutral" fecundity
mu=10e-7;         % mutation rate
d=10e5;           % strength parameter for migration
exponential decay f=alpha*d^(-distance)

ae=0;             % mean of the normal distribution of
allelic effect (under HOC) OR
                 % mean of the normal distribution of
the change in allelic effect (under RW)
stdae=0.1;       % std of the normal distribution of
allelic effect (under HOC) OR
                 % std of the normal distribution of the
change in allelic effect (under RW)
N=200;           % number of surviving juveniles... so
also the # of adults per pop
sampsize=30;     % number of individuals to draw for Fst
samples
p=size(A,3);     % number of populations

```

```

nummut=0;           % recorder for number of mutations
juvnum=[0 0];      % recorder for number of juveniles
space=20;          % spacing to record juvie and adult
genetics
gfreq1=zeros(size(A,1),TOT);           % allele
frequencies from population 1
gfreqT=zeros(size(A,1),TOT);           % allele
frequencies averaged over total population
FMonA=zeros(size(A,3),size(A,3));      % recorder for
sampled adult Fst's
FfullA=zeros(size(A,3),size(A,3));     % recorder for
complete adult Fst's
FMonJ=zeros(size(A,3),size(A,3));      % recorder for
sampled juvenile Fst's
FfullJ=zeros(size(A,3),size(A,3));     % recorder for
complete juvenile Fst's
rand('state',sum(100*clock))           % reset random
number generator

% initial exponential setup for migration component
if rem(p,2)==0;      % even number of populations... use
alpha for even configuration
    alpha=((d-1)*d^(p/2))/(d^(p/2)-d+d^(p/2+1)-1);
    realJMon=zeros(1,p/2+1);
else                % odd number of populations... use
alpha for odd configuration
    alpha=((d-1)*d^((p-1)/2))/(d^((p-1)/2)-2+d^((p-
1)/2+1));
    realJMon=zeros(1,(p-1)/2+1);
end

% make the shift matrix 'short', where short(i,:) contains
the shortest
% number of steps (the shift) between population i and the
other populations.
% Then make the matrix MIG, where MIG(i,:) contains the
fraction of larvae
% from population i migrating to all populations (including
those that
% "migrate" to itself).
for k=1:p;
    short(1,k)=min(abs(1-k),abs(p+1-k));
end
short1=alpha*d.^(-short);
for k=1:p;

```

```

    MIG(k,:)=circshift(short1,[0,k-1]);
end
cMIG=[zeros(p,1) cumsum(MIG,2)];

% choose initial allelic affects
%s=ones(size(A,1),1); %neutral alleles
s=randn(size(A,1),1).*stdae+ae; %non neutral alleles
S=zeros(length(s),length(s));
for i=1:size(A,1);
    for j=1:size(A,2);
        S(i,j)=exp(s(i)+s(j)); %exponential
    end
end
version, ~additive % strictly additive
% S(i,j)=s(i)+s(j);

end

C=mean(A,3);
s1=sum(sum(S.*C))./sum(sum(C));
Strend=mean(s1);

file=input('Enter name of file to save data to (.mat
file)\n','s');
mypath='newoutput/';
file=strcat(mypath,file);

% BEGIN COMPONENTS FOR EACH STEP OF MODEL
for T=1:TOT; % master loop for timesteps;
T

% 1. GAMETES
% produces gamete matrix G (m x K) from adult matrix A
disp('gametes')
%[G]=gametesneutral(A,s,phi);
G=gametesneutral(A,s,phi);

% 2. MUTATION
disp('mutation')
[G,s,S,muvec]=mutation2(G,s,gamma,mu,stdae,ae);
nummut(T+1)=nummut(T)+sum(muvec);

%% ALLELE TRACKING

```

```

gfreq1(1:size(G,1),T)=G(:,1)./sum(G(:,1));
gfreqT(1:size(G,1),T)=mean(G,2)./sum(mean(G,2));

% 3. LARVAE
disp('larvae')
L=larvae(G);
larvnum(T,:)=shiftdim(sum(sum(L))');

% 4. MIGRATION
disp('migration')
[J,realJ]=migration3(L,MIG);
if rem(T,space)==0;
    juvnum(end+1,1:2)=[min(sum(sum(J,2))));
    max(sum(sum(J,2)))]];
    sJ=survive(J,sampsize);
    FfullJ(:, :,end+1)=Fst(J);
    FMonJ(:, :,end+1)=Fst(sJ);
    Jcell(T/space)={[J]};
    realJMon(end+1,:)=realJ;
end

% 5. SURVIVE
disp('survival')
%keepJ=survive(J,9*N/10); %these
lines for iteroparity
%iteroA=survive(A,N/10);
%newal=size(J,2)-size(A,2);
%newA=padarray(iteroA,[newal,newal],'post');
%A=keepJ+newA;
A=surviveselect(J,N,S); %this line for
semelparity

if rem(T,space)==0;
    AJComp(:, :,1,1:9)=J;
    AJComp(:, :,2,1:9)=A;
    for i=1:9;
        [d1,d2,FComp(T/space,i)]=find(Fst(AJComp(:, :, :, i)));
    end
end
clear AJComp

```

```

% matrix cleanup... find genotypes that didn't survive

clear a
a=find(sum(J(:, :, 1))==0);
for k=2:size(J,3);
    anew=find(sum(J(:, :, k))==0);
    a=intersect(a,anew);
end
diff(T)=length(a);
A(a, :, :)=[];
A(:, a, :)=[];
s(a)=[]; % this should remove extinct alleles' selection
coeff's from s
S(a, :)=[];
S(:, a)=[];
J(a, :, :)=[];
J(:, a, :)=[];

numal(T+1)=size(A,2);
if rem(T,space)==0;
    sA=survive(A,sampsize);
    FfullA(:, :, end+1)=Fst(A);
    FMonA(:, :, end+1)=Fst(sA);
    Acell(T/space)={[A]};
end

popsize(:,T)=shiftdim(sum(sum(A)),2);
C=mean(A,3);
s1=sum(sum(S.*C))./sum(sum(C));
Strend(T+1)=mean(s1);

end % end for master T loop starting on line 66

%% Fst and Nm block
realJMon(1, :)=[];
%juvnum(end+1,1:2)=[min(sum(sum(J,2)))]
max(sum(sum(J,2)))]];
%juvnum(1, :)=[];
fj=survive(J,sampsize); % sample of Juveniles
fa=survive(A,sampsize); % sample of Adults
FstJ=Fst(fj); % compute Fst's for both
samples

```

```

FstA=Fst(fa);
FJ=0;
FA=0;
for i=1:p-1;
    for j=i+1:p;
        FJ(end+1)=FstJ(i,j);
        FA(end+1)=FstA(i,j);
    end
end
FJ(1)=[];
FA(1)=[];
NmJ=(1/4)*(1./FJ-1);
NmA=(1/4)*(1./FA-1);

s=sprintf('x%d', p);
load(s);

%%DATA SAVE
save(file,'file','realJ','realJMon','ae','stdae','FComp','F
fullA','FfullJ','Acell','Jcell','larvnum','space','T','diff
','juvnum','sampsiz','x','d','phi','alpha','MIG','gamma','
N','A','FstJ','FstA','NmJ','NmA','popsize','numal','Strend'
,'nummut','gfreq1','gfreqT','FMonA','FMonJ');

%% Fst GRAPHICS
figure
hold on
p1=scatter(x,NmJ,'filled','b');
Jls=robustfit(x,NmJ);
p3=plot(x,Jls(1)+Jls(2)*x,'b');
p4=scatter(x,NmA,'filled','r');
Als=robustfit(x,NmA);
p6=plot(x,Als(1)+Als(2)*x,'r');

if rem(p,2)==0;
    p7=plot(0:p/2,MIG(1,1:p/2+1).*sum(sum(mean(A,3))),'k');
else
    p7=plot(0:(p-1)/2,MIG(1,1:(p-
1)/2+1).*sum(sum(mean(A,3))),'k');
end
legend([p1; p4; p7],'Juvenile Nm data and trend','Adult Nm
data and trend','Real dispersal curve')

```

```

%%ALLELE TRACKING GRAPHICS
figure
subplot(2,1,1), plot(1:T,gfreq1)
hold on
subplot(2,1,2),plot(1:T,gfreqT)

%% DEMOGRAPHIC GRAPHICS
figure
scatter(1:T+1,numal,'filled','r')
ylabel('Total number of alleles'), xlabel('Simulation Time-
step')

%SELECTION GRAPHICS
figure
scatter(1:T+1,Strend,'filled','b')
ylabel('Average selection coefficient'), xlabel('Simulation
Time-step')
figure
scatter(1:T+1,nummut,'filled','g'), xlabel('Simulation
Time-step')
ylabel('Total number of mutations that have occurred')

```

```

function [G]=gametesneutral(A,s,phi);

% 1. GAMETES
% produces gamete matrix G (m x K) from adult matrix A

Sneu=ones(length(s),length(s));
G=[];
[ar,ac,ap]=size(A);
for k=1:ap;
    for i=1:ar;

G(i,k)=floor(phi/2*(sum(Sneu(i,:).*A(i,:,k))+sum(Sneu(:,i).
*A(:,i,k))));
    end
end
% state is now gametes, matrix G size m x k

```

```

function A=surviveselect(J,N,S);

% picks survivors from the juvenile matrix J and
% puts them in adult matrix A.

```



```

S=triu(S);
[jr,jc,jp]=size(J);
A=zeros(size(J));
Swork=reshape(S,1,numel(S));
for k=1:jp;
    Jwork=reshape(J(:,:,k),1,numel(J(:,:,k)));
    Swork=reshape(S,1,numel(S));
    atemp=zeros(size(Jwork));
    limit=min(N,sum(Jwork));
    for i=1:limit;
        if sum(Jwork~=0)>1;
            ja=find(Jwork);
            t=randsample(1:length(Jwork),1,'true',Swork);
            Jwork(t)=Jwork(t)-1;
            atemp(t)=atemp(t)+1;
            Swork(find(~Jwork))=0;
            clear t
        else
            [d1,d2,num]=find(Jwork);
            atemp(1,d2)=atemp(1,d2)+limit-i+1;
            break
        end
    end
    A(:,:,k)=reshape(atemp,jr,jr);
    clear atemp
end
clear Jwork Swork

```

C. Files used in both models

```

function
[G,s,S,muvec]=mutation(G,s,gamma,mu,stdae,ae,cutoff);

% 2. MUTATION
% calculates the number of mutations that occur in one
generation modeled
% as a Poisson process with parameter 4N(mu), N= population
size and
% mu=mutation rate

% calculate the total number of gametes in each population
clear muvec GameteSizes
[gl,K]=size(G);
GameteSizes=sum(G);

```

```

% calculate the number of mutations occurring in each
population
muvec=poissrnd(4*GameteSizes*mu);

% place the mutant gametes in matrix Gnew; create new
alleles in s with
% new fitnesses
Gnew=zeros(1,K);
snew=[0];

c=zeros(K,max(muvec));
for i=1:K;
    if muvec(i)~=0;
        c(i,1:muvec(i))=randsampW(G(:,i),muvec(i));
        for j=1:muvec(i);
            Gnew(end+1,i)=1;
            snew(end+1)=gamma*s(c(i,j))+randn(1)*stdae+ae;
            G(c(i,j),i)=G(c(i,j),i)-1;
        end
    end
end
Gnew(1,:)=[];
[i,j,snew]=find(snew);
snew=snew';

% append the mutant gametes in Gnew to the gamete matrix G
if ~isempty(Gnew);
    G=[G;Gnew];
end
s=[s;snew];
%non neutral alleles
%s=ones(size([s;snew]));
%neutral alleles

for i=1:length(s);
    for j=1:length(s);
        S(i,j)=exp(s(i)+s(j));
    end
end
Stemp=reshape(S,1,numel(S));
Stemp(Stemp<=quantile(Stemp,cutoff))=0;
S=reshape(S,length(s),length(s));

% state is now gametes in matrix G

```

```

% if any population has an odd number of gametes, decrease
the most
% frequent gamete type by 1 to ensure that all gametes get
united into
% larvae.
NumGametes=sum(G);
if sum(rem(sum(G),2)>0);
    odd=find(rem(sum(G),2)>0);
    for i=1:length(odd);
        [x,y]=max(G(:,odd(i)));
        G(y,odd(i))=G(y,odd(i))-1;
    end
end
clear Gnew snw NumGametes GameteSizes

```

```

function L=larvae(G);

% produces larvae in tensor L (m x m x k) from gamete
matrix G (m x k)
[gr,gc]=size(G);
big=max(max(G));
jmat=[0];
L=zeros(gr,gr,gc);
for i=1:gc;
    for j=1:gr;
        jmat(end+1:end+G(j,i))=j*ones(1,G(j,i));
    end
    [x,y,jmat]=find(jmat);
    gammat=[];
    gammat=randsample(jmat,length(jmat),false);
    gammat=reshape(gammat,2,length(gammat)/2)';
    [r,c]=size(gammat);
    gammat=sort(gammat,2);
    for k=1:r;
        L(gammat(k,1),gammat(k,2),i)=L(gammat(k,1),gammat(k,2),i)+1;
    end
    jmat=0;
    gammat=[];
end
clear G

[lr,lc,lp]=size(L);
for i=1:lp;

```

```

    L(:,:,i)=2*triu(L(:,:,i))-diag(diag(L(:,:,i)));
end
% state is now larvae in upper triangular tensor L

```

```

function [J,realJ]=migration2(L,MIG,realJ);

% takes the tensor (m x m x k) of genotyped larvae and
migrates them
% according to an exponential decay function
% first subtract the appropriate number of larvae from
source populations in L
% and store them in matrix M; then add them back to L for
% the number of incoming larvae.

[lr,lc,lp]=size(L);
J=zeros(size(L));
realJ=zeros(lp,lp);
for k=1:lp;
    for i=1:lr;
        for j=1:lc;
            if L(i,j,k)~=0;

v=randsample(size(MIG,1),L(i,j,k),true,[MIG(:,k)']);
                for s=1:size(MIG,1);
                    realJ(k,s)=realJ(k,s)+sum(v==s);
                    J(i,j,s)=J(i,j,s)+sum(v==s);
                end
            end
        end
    end
end

for i=1:lp;
    realJ(i,:)=circshift(realJ(i,:),[0 -(i-1)]);
end

[r,c]=size(realJ);
if rem(lp,2)==0;
    realJ(r+1:2*r,1:(r/2+1))=[NaN*zeros(r,1)
fliplr(realJ(1:r,(r/2+1):k))];
    realJ=realJ(1:2*r,1:(k/2+1));
else
    realJ(r+1:2*r,1:(r+1)/2)=[NaN*zeros(r,1)
fliplr(realJ(1:r,((r+1)/2+1):r))];
    realJ=realJ(1:2*r,1:(r+1)/2);

```

```

end
realJ=nanmean(realJ);

```

```

function Fst=Fst2(A);

% Calculate pairwise Fst's from an N-D array of genotype
counts
% (with array entry (i,j,k) being the number of individuals
% of genotype AiAj in population k).

[r,c,k]=size(A);
popsizes=shiftdim(sum(sum(A)),1);

% check to see if array pages (i.e., (:,:,i) for all i) are
upper
% triangular or full. If upper triangular, revert to full
form.
for i=1:k;
    check(k)=sum(sum(tril(A(:,:,i),-1)));
end
c=[];
d=[];
if sum(check)==0;
    for i=1:k;
        c=triu(A(:,:,i),1)./2;
        d=c';
        A(:,:,i)=A(:,:,i)-c+d;
        clear c,d;
    end
end

%compute allelele frequencies from genotype counts
for i=1:k;
    a(:,:,i)=A(:,:,i)./sum(sum(A(:,:,i)));
    alleleles(:,i)=sum(a(:,:,i),2);
end
alleleles2=alleleles.*alleleles;

% construct a cell array of size r x r (only upper triangle
filled) where
% pbarpops(i,j) contains a r x 1 matrix of the average
allelele frequencies between
% populations i and j. The output of
'cell2mat(pbarpops(i,j))' gives the matrix.
pbarpops=cell(k);

```

```

for i=1:k-1;
    for j=i+1:k;
        pi=[popsizes(i)
popsizes(j)]./(popsizes(i)+popsizes(j));
        pbarpops(i,j)={sum([ repmat([pi(1)
pi(2)],r,1)].*[alleles(:,i) alleles(:,j)],2)};
        p2bar(i,j)= {sum([ repmat([pi(1)
pi(2)],r,1)].*[alleles2(:,i) alleles2(:,j)],2)};
        pbar2(i,j)=
{[cell2mat(pbarpops(i,j)).*cell2mat(pbarpops(i,j))]};
        Fst(i,j)= (sum(cell2mat(p2bar(i,j)))-
sum(cell2mat(pbar2(i,j))))/(1-sum(cell2mat(pbar2(i,j))));
    end
end
Fst(end+1,:)=zeros(1,k);

```