

UNIVERSITY OF CALIFORNIA  
Santa Barbara

Molecular Approaches to Assessing  
Red Sea Urchin (*Strongylocentrotus franciscanus*) Populations: Implications of  
Sequence Variability for Evolution and Population Genetics of the Species

A dissertation submitted in partial satisfaction  
of the requirements for the degree of

Doctor of Philosophy

in

Biological Sciences

by

Patty Debenham

Committee in charge:

Dr. Mark Brzezinski, co-chairperson  
Dr. Kathleen R. Foltz, co-chairperson  
Dr. Steven D. Gaines  
Dr. Robert R. Warner

Dr. Stephen R. Palumbi (Harvard University)

August 1997

Signature page

August 1997

Copyright by  
Patty Debenham  
**1997**

## VITA

October 11, 1963, Born, Stanford California

Education: **Stanford University** **Stanford, CA**  
 1982-1985 A.B. in Human Biology with a concentration in marine biology.

summer 1983 **Hopkins Marine Station** **Pacific Grove, CA**  
 Studied sub-tidal ecology through daily scuba diving and lab work.

### Scholarships to support graduate studies

- \$ 1,000 UC Santa Barbara General Affiliates, 1995/1996.
- \$ 2,100 UCSB Biology Departmentt Regents Fellowship, 1995/1996.
- \$ 2,500 PADI Foundation, 1994/1995.
- \$ 4,000 Seaspace Houston Underwater Club Scholarship, 1993/1994.
- \$ 500 International Women's Fishing Association, 1993/1994.
- \$ 3,000 Soroptomist Society International, 1992/1993.
- \$ 1,000 Seaspace Houston Underwater Club Scholarship, 1993/1994.
- \$ 800 International Women's Fishing Association, 1992/1993.

### Experience:

1992-1996 **UC, Santa Barbara** **Santa Barbara, CA**  
 Teaching Assistant, Department of Ecology, Evolution, Marine  
 Biology University of California, Santa Barbara

1991-1997 **UC, Santa Barbara** **Santa Barbara, CA**  
 Research Assistant, Marine Sciences Institute, UC, Santa Barbara

April 1992 **Research Cruise** **Coastal California**  
*Diver and scientific crew.* Part of 8-person scientific team to  
 examine biological oceanographic processes off the coast of  
 California.

4/88-6/91 **Center For Marine Conservation** **Washington, DC**

*Manager Marine Debris Program.* Wrote proposals and received funding (over \$400,000) from National Oceanic and Atmospheric Administration and EPA to conduct educational programs. Managed contracts and reported to government and industry grants officers. Produced newsletter to relay data collection information to general public and policy makers. Coordinated National Beach Cleanup (100,000 volunteers in 1990). Managed staff of five.

- 8/87-4/88     **Science Applications International Corporation (SAIC)**     **McLean, VA**  
*Research Assistant.* Contracted by the Environmental Protection Agency to complete assignments for the Office of Marine and Estuarine Protection. Contributed to "Permit Writer's Guide to Ocean Dumping Manual." Created database for species information for West Coast Ocean Incineration site designation project. Reviewed floatables study for New Jersey Department of Environmental Protection.
- summer 1986     **WQED/National Geographic**     **Los Angeles, CA**  
*Production Intern.* Worked on production of National Geographic Specials.
- winter 1986     **Operation Raleigh U.S.A.**     **Patagonia, Chile**  
*Venturer.* Awarded scholarship (\$5,500) for 3-month science and service expedition. Built bridge over 30-foot glacial river in Queulat, IV Region.
- 1985     **1985 Our World Underwater Scholarship**  
*Scholar.* Chosen for year-long internship program with the leaders of the underwater industry. Projects included:
- Deep Ocean Technology**     **San Leandro, CA**  
 Assisted the Director, Dr. Sylvia Earle. Initiated study of deepwater submersible manufacturer. Catalogued audio-visual materials.

**Bob Soto's Diving Ltd & Fisheye  
Photographic Services** **Grand Cayman**  
Studied underwater video, photography, taught SCUBA  
diving.

winter 1984 **Marine World Africa U.S.A** **Redwood City, CA**  
*Intern.* Instructor for after-school programs at marine and land  
animal theme park.

fall 1983 **Stanford University** **Stanford, CA**  
*Teaching Assistant.* Assistant instructor of Open Water SCUBA  
certification course.

**ABSTRACT**

**Molecular Approaches to Assessing  
Red Sea Urchin (*Strongylocentrotus franciscanus*) Populations: Implications of  
Sequence Variability for Evolution and  
Population Genetics of the Species**

by

Patty Debenham

I evaluated sequence variability in the red sea urchin, *Strongylocentrotus franciscanus*, using DNA sequence data from 134 adult individuals collected between Alaska and Baja California in 1995 and 1996. DNA sequence data was obtained from direct sequencing of a 273 base pair region of bindin—a sperm protein required for fertilization. These data were used to evaluate genetic subdivision as well as natural selection at the bindin locus operating within and between species.

Results indicate that bindin is sufficiently polymorphic to serve as a genetic marker. I identified 14 unique alleles present in the entire range sampled with a maximum of eight alleles at a specific site. Analysis of allele frequencies indicate that the bindin locus in the red sea urchin is panmictic throughout the entire range examined, suggesting there is a high exchange of genetic material.

To evaluate selection operating on a marine invertebrate fertilization protein, I conducted both an intra- and interspecific analysis of bindin sequence variation. Based on our analyses, it is not possible to reject the null hypothesis that sequence variation observed in *S. franciscanus* bindin is a result of neutral evolution. Several statistical tests suggest that random mutation followed by genetic drift created both the polymorphism observed within the 273 base pair region of *S. franciscanus* bindin and the divergence observed between several species.

Finally, I explored possible explanations for the different patterns of sequence divergence in *Strongylocentrotus* compared to *Echinometra* bindin. One analysis included a simulation to evaluate the probability that the degree of sequence variation observed in *Echinometra* could be the result of chance alone and not natural selection. Our simulation suggests that a sliding window analysis can incorrectly identify a significant spatial clustering of replacement substitutions.

## TABLE OF CONTENTS

Dedication		<i>viii</i>
Acknowledgments		<i>ix</i>
List of Figures		<i>x</i>
List of Tables		<i>xii</i>
Chapter 1:	Introduction	1
Chapter 2:	Materials and Methods Applicable to Entire Dissertation	6
Chapter 3:	Results Applicable to Entire Dissertation	21
Chapter 4:	Population Genetics of ohe Red Sea Urchin, <i>S. franciscanus</i> , Along the West Coast of North America	99
Chapter 5:	Intraspecific Sequence Variation in the Bindin Locus of <i>S. franciscanus</i> : Unusual Nucleotide Variation in a Marine Invertebrate Gamete Interaction Molecule	135

Chapter 6:	Interspecific Sequence Variation In Four Species Of Sea Urchins At The Bindin Locus As An Indicator Of Adaptive Evolution In A Marine Invertebrate Gamete Interaction Molecule.	177
------------	--	-----

## **Dedication**

*With love I dedicate this work to my parents, Sally and Warren Debenham. They have given me the foundation to grow, learn, and love. I could not have had the courage to take on this challenge without the stability and guidance they offer. Without their compassion and concern for the world, I would not have had the example to use what I have learned in an effort to contribute to the world around me.*

## Acknowledgments

Nearly six years is a long time to work on one project and there have been many people that have helped me along the way. I feel the most gratitude to my friends that have kept me sane, laughing, and helped me to put this process in perspective. Thank you **Beth Mitchner, Jennifer Vettel, Aniko and Chris Somogyi, John Tilton, Steve Haddock, George Polchin, Danny and Myla Kato, Brad Burk, Amanda Thomas, Dave Kapolnek, Amanda Rass, Molly Cummings, Lisa Wooninck**, and many more.

**Steve Haddock** really deserves second author on this dissertation for all of his technical and theoretical help. Thank you.

I could not have had nearly as much joy in my life in the last several years without the companionship of my dog, **Ka Nai'a**. I appreciate all the people who genuinely care about Ka Nai'a and help her to have a comfortable home around the marine lab.

I have a terrific committee who was challenging, punctual, respectful, and often right. Thank you **Kathy Foltz, Mark Brzezinski, Bob Warner, Steve Gaines**, and **Steve Palumbi**. **Kathy and Mark** deserve extra thanks for fulfilling a commitment to me as their student despite the fact that the topic of my thesis lay outside of their research areas.

I would also like to thank the current and previous members of the **Foltz lab**. Good luck with your work and future plans. **Scott Aiken, Mel Kelley**, and **Richelle Feldan** were especially helpful with different stages of my research. **Ken Hoang** and **Yama Abassi** have helped the work day be enjoyable. Thank you **Andy Giusti** for taking the trip with me down to Baja California.

More than once I thought my technical problems had become insurmountable. Thanks to extra guidance from the following individuals, I was able to get the job

done: **Rafael Jovine, Kelley Thomas, Ed DeLong, Lisa Wooninck, Brad Phillips, Steve Poole, Dorothy Mederios-Bergen, and Steve Karl.**

I received a great deal of help from many individuals when the idea for this project first germinated. Thank you for your assistance in the proposal writing stage and during subsequent discussions: **Steve Palumbi, Sandie Degnan, Steve Schroeter, John Dixon, John Duffy, Bruce Steele, Gary Davis, and Bruce Steele.**

### **List of Figures**

- Figure 2-1. Map of six geographic sampling locations as source of adult urchin populations.
- Figure 2-2. Schematic diagram of bindin gene and region sequenced.
- Figure 2-3. Agarose gel as example of multiple stages of sample preparation.
- Figure 2-4. Sequencing gel illustrating results of direct sequencing and example of heterozygous individual.
- Figure 3-1. Six graphs of bootstrapped simulation to determine efficacy of sample size.
- Figure 3-2. List of DNA sequence of 14 unique alleles observed throughout the entire population sampled.
- Figure 3-3a. Sequences of all alleles observed in the Alaska population.
- Figure 3-3b. Sequences of all alleles observed in the Washington population.
- Figure 3-3c. Sequences of all alleles observed in the Oregon population.

- Figure 3-3d. Sequences of all alleles observed in the Northern California population.
- Figure 3-3e. Sequences of all alleles observed in the Santa Barbara population.
- Figure 3-3f. Sequences of all alleles observed in the Baja California population.
- Figure 4-1. Map of six geographic sampling locations as source of adult urchin populations.
- Figure 4-2. Parsimony tree showing the phylogenetic relationship of 14 unique alleles observed in entire population.
- Figure 4-3. Allele frequency data and geographic correlation for 14 unique alleles.

**List of Figures (cont.)**

- Figure 5-1. Schematic diagram of bindin gene and region sequenced.
- Figure 5-2. Phylogenetic relationship of 14 unique bindin alleles observed.
- Figure 5-3. List of DNA sequence of 14 unique alleles observed throughout the entire population sampled.
- Figure 6-1. Alignment of mature bindin from four species of sea urchin
- Figure 6-2a. Common histogram resulting from simulation to test clustering of replacement substitutions in *Echinometra* data set.
- Figure 6-2b. Histogram resulting from simulation to test clustering of replacement substitutions in *Echinometra* data set where  $d_n$  is significantly greater than  $d_s$ .



**List of Tables**

- Table 2-1. Names and address of individuals that assisted in sample collection and sample dissection.
- Table 2-2. Identification of representative genotypes cloned to confirm sequence of component alleles.
- Table 3-1. Summary information for all adult urchins collected including habitat information, sex, test size, and color morphology.
- Table 3-2. Detailed information for sampling effort # 1.
- Table 3-3. Detailed information for sampling effort # 2.
- Table 3-4. Total quantity of adults sequenced, number of alleles and genotypes observed for each geographic sampling location.
- Table 3-5. Genotypes observed for each individual sequenced.
- Table 4-1. Identification of polymorphic positions for each of the 14 alleles observed in all populations.
- Table 4-2. Measurements of heterozygosity and deviations from Hardy-Weinberg equilibrium for all populations.
- Table 5-1. Intraspecific comparison of nonsynonymous substitutions per nonsynonymous site (dn) and synonymous substitutions per synonymous site (ds) for all populations and for theoretical data set created by DNA Evolve.
- Table 5-2. Summary of results of McDonald-Kreitman test for neutral evolution.

Table 6-1. Interspecific comparison of nonsynonymous substitutions per nonsynonymous site (dn) and synonymous substitutions per synonymous site (ds) for all populations and for theoretical data set created by DNA Evolve.

**List of Tables (cont.)**

- |           |   |
|-----------|---|
| Table 6-2 | Detailed information utilized to use the McDonald-Kreitman test for neutral evolution.                          |
| Table 6-3 | Summary results from simulation to test clustering of replacement substitutions in <i>Echinometra</i> data set. |

## **Chapter 1**

### **Introduction**

Many marine invertebrates are external fertilizers (Strathmann, 1978). For these organisms, there is a strong contrast between the limited mobility of the adult phase and the potential for high vagility in the larval stage. Males and females of these often synchronous broadcast spawners, simultaneously release their gametes into the water column (Strathmann, 1978). Commonly, after fusion of the egg and sperm, a free swimming larvae develops and remains in the water column until it either dies or settles onto an acceptable substrate (Thorson, 1961). As opposed to organisms with mating interactions between conspecific adults, all interactions between potential mates are isolated to the egg and sperm for broadcast spawners. Small changes in the interaction between egg and sperm theoretically can have large impacts on the outcome of fertilization. Subsequently, the gamete interaction molecules in these organisms possess a unique potential to be determinants of reproductive isolation (Palumbi, 1992).

There is no clear understanding of what determines reproductive isolation in marine environments. As an open system (Roughgarden, 1985), there are few distinct geographic barriers that would limit either adult migration or larval transport. Yet, reproductive isolation must occur in broadcast spawners as evidenced by the record of speciation events (Jablonski, 1986). However, the connection between gamete interaction molecules and reproductive isolation is unclear. In order to probe this connection, we examined DNA sequence variability of bindin, a sea urchin sperm protein necessary for fertilization. As a result of being a model system of reproduction, there is a large amount of information available on both sea urchin fertilization including sequence data of many species for bindin, a sperm protein necessary for successful fertilization (Vacquier *et al.*, 1995; Minor *et al.*, 1989). Many investigators have demonstrated species-specific differences in bindin (see Vacquier *et al.*, 1995). Therefore, it is possible to assume that reproductive isolation must have existed in bindin. It is also a plausible assumption that forces that created reproductive isolation in the past are likely to be operating in the present. However, the mechanism that would mediate the reproductive incompatibilities is unclear. For example, perhaps geographic isolation followed by

genetic drift creates incompatible alleles. On the other hand, directional selection in these isolated environments can create the same pattern of divergence.

Our approach to evaluate possible correlations between marine invertebrate gamete interaction molecules and reproductive isolation was to examine sequence variability in a portion of the *bindin* locus in the red sea urchin species, *Strongylocentrotus franciscanus*. If no intraspecific variability exists in the *bindin* locus, there is no potential for reproductive isolation. Therefore, our first line of inquiry explored the level of sequence variation in the *bindin* locus. However, variability within the red urchin provides the potential for, yet not the explanation of, species-specific differences. If the variability is a result of random mutations created by neutral evolution, reproductive isolation will not ensue. On the other hand, variability could be the result of a selective force such as frequency dependent selection that creates incompatibilities in alternate forms of *bindin*. Subsequently, we used several statistical methods to test the null hypothesis of neutral evolution. We looked for evidence of selection both throughout the species range examined and at each of six subpopulations.

Finally, we evaluated how genetic variability correlated with geography as an indicator of incipient reproductive isolation. To test the null hypothesis of complete

genetic homogeneity, we evaluated inbreeding coefficients as well as conformance to Hardy-Weinberg equilibrium. If genetic substructuring exists, it is possibly the result of geographic isolation followed by random mutation and genetic drift, or directional selection in different environments.

This dissertation is organized into six chapters. Following this introduction is a description of materials and methods that apply to the entire dissertation (Chapter 2). Similarly, Chapter 3 presents the results that apply to the entire dissertation. An examination of population subdivision is presented in Chapter 4. To evaluate the operation of selection, we evaluated evidence for selection operating on the binding locus within *S. franciscanus* (Chapter 5) and between four related urchin species (Chapter 6). In combination, these last three chapters investigate how variation observed in gamete interaction molecules relates to reproductive isolation.

### References

Jablonski, D. (1986). Larval ecology and macroevolution in marine invertebrates. *Bull. Mar. Sci* 39:(2) 565-587.

Minor, J. E., Gao, B., and Davidson, E. H. (1989). The molecular biology of bindin. In *The Molecular Biology of Fertilization*, H. Schatten and G. Schatten, eds.: Academic Press, pp. 73-88.

Palumbi, S. R. (1992). Marine speciation on a small planet. *TREE* 7:(4) 114-117.

Roughgarden, J., Iwasa, Y., and Baxter, C. (1985). Demographic theory for an open marine population with space-limited recruitment. *Ecology* 66: 54-67.

Strathmann, R. R. (1978). Length of pelagic period in echinoderms with feeding larvae from the Northeast Pacific. *J. Exp. Mar. Biol. Ecol.* 34: 23-27.

Thorson, G. (1961). Length of pelagic life in marine invertebrates as related to larval transport by ocean currents. In *Oceanography*, edited by M. Sears, AAAS Washington, DC 455-474.

Vacquier, V. D., Swanson, W. J., and Hellberg, M. E. (1995). What have we learned about sea urchin sperm bindin? *Develop., Growth and Differ.* 37: 1-10.

## Chapter 2

### Materials and Methods Applicable to Entire Dissertation

#### *Sample Collection*

In total, gonad tissue samples from 479 adult red sea urchins representative of six different geographic regions were collected. The six sampling sites span the geographic range of the species from Alaska to Baja California. Approximately 30 samples from each site (a total of 181) were shipped to UCSB in 1993 from the following locations: Ketchikan, Alaska; the Straits of Juan de Fuca, Washington; Depoe Bay, Oregon; Ft. Bragg, California; Anacapa Island, California; and Punta Baja, Baja California, Mexico (Figure 2-1). These individuals are all designated as part of sampling effort #1 and identified by a number from 1 - 30.

A second sampling effort was conducted in 1995 to collect 298 urchins, approximately 50 samples from each site. The locations for sampling effort #2 were Ketchikan, Alaska; Port Townsend, Washington; Port Orford, Oregon; Ft. Bragg, California; Santa Barbara, California; and Ensenada, Baja California, Mexico. These individuals are all designated as part of sampling effort #2 and identified by a number from 101 - 154.

All samples were collected on SCUBA either by myself or by commercial fishermen. Table 2-1 lists the names and address of commercial fishermen and government agency representatives that assisted in the sample collection. In addition, this table lists the numerous individuals who assisted in sample dissection. In several cases, the samples were shipped to UCSB via overnight express airmail service. To keep the animals alive during shipping, each animal was wrapped individually in newspaper and packed in a Styrofoam container with ice. During the tissue preparation, all necessary sterile precautions were taken to avoid cross-contamination of samples. Examples of these precautions included changing gloves, rinsing cutting implements in bleach and water, and using sterile weigh boats for each new sample.

To eliminate possible genetic variation associated with the depth of habitat, the depth at which urchins were collected was chosen to represent a depth of common occurrence for the species and to be standardized between the sampling sites. All individuals came from as close to a 10 - 15 meter depth range as possible. In several cases it was not possible to collect animals from this depth range, however, it is unlikely that these small variations in depth significantly affect the results of the work. Actual depths for each collection are listed in Tables 3-2 and 3-3 (Chapter 3). To insure that all samples were adults, all individuals collected were at least 80 mm in test diameter.

Where possible, we identified the sex of each sample: males were indicated by a white exudate of sperm from the gonads and females were identified by an orange exudate of eggs. Because the exudate produced by females is the same color as the gonad, in some cases it was difficult to distinguish between a female sample and a reproductively immature sample that could be male or female. In this case, we examined the gonad under a light microscope at 40X magnification. In most cases it was possible to see either the existence of eggs or sperm. In some cases in which the gonad contained reproductive cells in the early stages of gametogenesis, it was still not possible to determine the sex of the sample using the light microscope. An aliquot of each tissue sample was stored in sterile containers at -70 °C and additional reserves of tissue were stored at -20°C.

### ***DNA Extraction***

DNA extractions were performed as described in Milligan (1992). Approximately 25 -100 ug of gonadal tissue was homogenized in 700 ul CTAB buffer prewarmed to 60°C (100 mM Tris-HCl, pH 8.0; 1.4 M NaCl; 20 mM EDTA; 2% hexadecyltrimethylammoniumbromide (CTAB, w/v)); 1% polyvinylpyrrolidone (PVP-360, w/v); 0.2% 2-mercaptoethanol (v/v added just before use). At the end of a 30-60 minute incubation at 60°C (with periodic swirling) 700 ul of chloroform:isoamyl alcohol (24:1) was added followed by vortexing.

Centrifugation for 10 minutes at 1500 x g separated the aqueous and the organic phases. One extraction with an equal volume of chloroform:isoamyl alcohol (24:1) was followed by extractions with phenol:chloroform:isoamyl until the interface was clear. The aqueous phase was transferred to a sterile tube followed by addition of 2/3 volume of ice-cold isopropanol and vortexing. The pelleted DNA was washed with 500 ul wash buffer (76% ethanol, 10 mM ammonium acetate), dried, and resuspended in 20-30 ul TE containing 1 ul RNase (10 mg/ml). It was imperative to conduct the extraction protocol on fresh ovary tissue. Ovary samples frozen for as short as one day showed reduced yield of DNA and reduced success in subsequent PCR amplification.

Three methods were used to determine the success of DNA extraction.

1. Resolve 1-5 ul of extracted genomic DNA on a 1% agarose gel. A bright, high molecular weight band indicated successful extraction and acceptable yield. In many cases a large smear occupied the mid and lower portions of the lane which most likely represented degraded DNA. These samples were acceptable for PCR amplification.
2. Perform double-stranded PCR amplification using approximately 100-200 ng of extracted DNA. It was not always necessary to quantify the DNA and was often sufficient to use 0.5 ul of extracted DNA in a 25 ul PCR reaction.

3. Use spectrophotometer to quantify absorbency at 260 and 280 nm and subsequently calculate purity and DNA concentration (see Sambrook *et al.* , 1989). This method was time consuming and was not necessary for most samples.

## ***PCR Amplification***

### *Double-stranded PCR amplification*

The primers FNbindin 5' (5'-AGTCGACGTTTCGACAGACGAC-3') and FNbindin 3' (5'-TTACATGGTCCATTATAGTATGCC-3') amplify a 431 base pair region of the 5' end of the bindin gene (Figure 2-2). Amplification followed standard procedures (Saiki *et al.*, 1988) using a reaction volume of 25  $\mu$ l and final magnesium chloride concentration of 2 mM. The thermocycler (Perkin Elmer Cetus) profile for all double stranded reactions was: 1 cycle of 95 °C, 5 min. followed by 30 cycles of 94 °C, 1 min.; 60 °C, 1 min.; 72 °C, 2 min. Five  $\mu$ l of each PCR product was resolved on a 2% Nu Sieve agarose, low melting temperature TBE gel. Gel isolates were removed with sterile wide-bore, disposable polyethylene transfer pipettes and stored in a microfuge tube with 200  $\mu$ l of water at -20 °C.

### *Single-stranded PCR amplification*

Each gel isolate was heated to 65°C for approximately 5-10 minutes and used as template for the single-stranded PCR amplification. To amplify a single-stranded product of the 5' strand, the PCR amplification conditions were identical to those identified above for amplification of the double-stranded product except the 5' primer (FNbindin 5') was used at a final concentration of 0.5  $\mu$ M.

The PCR amplification conditions to amplify the 3' single-stranded product required a lower annealing temperature (58°C) and lower MgCl<sub>2</sub><sup>+</sup> concentration (1.2 mM final). The concentration of the 3' limiting primer was 2.5 uM (final). Some samples required additional adjustments to annealing temperature (between 57-63°C) and limiting primer concentration (0.5-2.5 uM final concentrations).

#### ***Direct sequencing of single-stranded PCR product***

The single-stranded PCR products were washed in Centricon filter units (30,000 MWCo) and resuspended in 7 ul H<sub>2</sub>O for Sanger dideoxy sequencing (Sequenase version 2.0, U.S. Biochemical) using internal sequencing primers KTseq 5' (5'-GGAGCGCGTAAGAAGCGTTAT-3') and KTseq 3' (5'-ATACACACGATGGTCAAG-3') at 10uM.

Figure 2-3 shows an example of all reactions (DNA extraction, Double-stranded PCR product, single-stranded PCR product, and Centricon-purified DNA template) performed prior to DNA sequencing and resolved on a 1% agarose gel. Figure 2-4 shows an autoradiogram resulting from direct sequencing of homozygous and heterozygous individuals. An individual is heterozygous when two bands are present at the same nucleotide position on the sequencing gel (Figure 2-4).

#### ***Cloning of bindin DNA from heterozygous individuals***

In order to confirm the exact sequence of representatives of all alleles, the PCR products were cloned. The samples cloned and their representative genotype are listed in Table 2-2. Double-stranded PCR products were amplified using primers KBRS 5' (5'CGCGGATCCAGT CGACGTTTCGACAGACGAC-3') and 3' (5'GCCAAGCTTTTACATGG TCCATTATAGTATGCC-3') and the double-stranded PCR protocol. These primers incorporate the restriction sites BamHI and HindIII, respectively on their 5' ends to facilitate directional cloning into the pBMKS bluescript vector. The PCR products were resolved on a 2% agarose gel and the excised gel fragment was purified using Quiaquick spin columns (Quiagen) and then digested with BamHI and HindIII. The gel-purified fragment was ligated into pBMKS and transformed into *E. coli* DH5 $\alpha$ . The DNA from a minimum of four transformants were sequenced for each PCR product. Several individuals with the same genotype, but from different geographic locations, were cloned and sequenced to verify sequence consistency among alleles found at different geographic locations. Plasmid DNA was purified using the alkaline lysis method (Sambrook *et al.*, 1989). The double-stranded plasmid DNA was sequenced using Kgseq 5' (5'GTTTCTGACG ATTCGGAAAGA-3') and Kgseq 3' (5'-GAAACAACCAATTTAAAATA-3') as internal sequencing primers.

Cloning of the PCR products also allowed us to assess nucleotide changes due to PCR incorporation error to be detected and eliminated. Cloned PCR products from four different reactions were sequenced on both strands. A substitution in a sequence was determined to be PCR incorporation error based on one or both of the following observations: (1) the nucleotide present in the sequence was not present in either the 5' or 3' strand produced by direct sequencing; or (2) the existence of this new sequence now created a total of three alleles for one genotype which is biologically impossible for a diploid organism. We found what appeared to be a substitution resulting from PCR incorporation error in approximately one out of 2,000 nucleotides. The sequence of a particular allele was considered to be an accurate representation of an individual genotype when it appeared in at least two out of three of the cloned sequences.

## Chapter 3

### Results Applicable to Entire Dissertation

#### *Results of Sample Collection*

Table 3-1 presents a summary of sampling data from each site for both sampling effort #1 and #2 including: the number of *Strongylocentrotus franciscanus* collected from each site, sex of samples, test size, and color morphology. In general there were two types of color morphologies found in all locations. A red color morphotype indicates individuals that were a red, and in some cases, pink color. Adults that had a purple color morphology were a deep shade of purple/maroon. In addition, some individuals from Port Orford, Oregon (sampling #2) had partly green tests. Table 3-2 contains detailed information of each sample collected from sampling effort #1 and Table 3-3 contains detailed information of each sample collected from sampling effort #2.

Of the total 479 urchins that were received from the two sampling efforts, there were 218 females and 182 males. It was not possible to determine the sex of an additional 79 samples. The average size of all samples was 98.39 mm. There

were 149 samples with a red color morphology, 295 purple morphotypes, 25 green, four that were a combination of red and green, four that were a combination of red and purple, and two samples where the information about the color morphology was not recorded.

### ***Sampling Effort #1***

Specifically in sampling effort #1, out of 181 urchins collected, 59 were female and 72 were male. It was not possible to determine the sex of 50 of the samples. The average size of these samples was 95.62 mm. There were 66 samples with a red color morphology, 109 purple morphotypes, and four that were a combination of red and purple, and two samples where the information about the color morphology was not recorded.

Table 3-2 includes detailed information about sampling effort #1 including information about habitat, sample collectors, and shipping information. There were no unusual circumstances for urchins received from Washington, Northern California, or Santa Barbara. The following is a summary of unusual circumstances from specific collection sites.

Because of the early stage of the project, information about the sex of the animals received from Alaska in sampling effort #1 was not recorded and measurements of test diameters are rough estimates rather than exact measurements.

For samples received from all other sites, the sex of each sample and a precise measurement of test diameter (mm) was recorded.

Urchins received from Depoe Bay, Oregon had very small gonads. The collectors reported that it appeared as if the urchins were eating the tube worms (possibly *Phragmatopoma sp.*) that covered the rocks within the urchin habitat. Further support that the urchins were ingesting tube worms came from the fact that there were pieces of calcareous shell as excrement inside the test rather than the more common pieces of processed kelp. It is possible that this alternate food source is connected to fact that these samples had very small gonads.

The original plan was to collect urchins from their most southerly distribution which is published to be Cedros Island, Baja California, Mexico (Morris *et al.*, 1980). Local representatives of the fisheries departments in Baja were unable to convincingly confirm or deny these reports. One fishery government official gave us false information that it was possible to find *S. franciscanus* in an eelgrass bay slightly north of Guerrero Negro however, we were unable to find any *S. franciscanus* adults in this location and the species present was probably *Arbacia stellata* (Gmelin) (Morris *et al.*, 1980).

Although local fishermen seemed to be the most knowledgeable about species distribution, we were never able to conclusively determine if red sea urchins were present on Cedros Island. Fishermen on the mainland indicated that the urchins on

Cedros Island "are the same species that they fish in El Rosario." El Rosario is the major area of *S. franciscanus* harvest in Baja California. Based on the lack of confirmation that *S. franciscanus* exists in sufficient numbers on Cedros Island, the cost (approximately \$300), and the time to get the samples (approximately 3 days), we decided to collect samples from Punta Baja, a location on the coast approximately 300 miles north of Cedros Island.

### ***Sampling Effort #2***

For sampling effort #2, 50 individuals were collected from each site. This number ensured sufficient sample sizes for a thorough population study and to assess diversity enhancing selection even if we would be forced to conduct the entire study on only male samples. To increase the probability DNA extractions from female samples would be successful, DNA from ovaries were performed within four hours of sacrifice. In addition, all tissue samples were stored at -70°C with additional aliquots of these samples stored at -20°C.

Out of 298 urchins collected in sampling effort #2, 159 were female and 110 were male; it was not possible to determine the sex of 29 of these individuals. The average test size of these adults was 101.16 mm diameter. There were 83 samples with a red color morphology, 186 were purple in color, 25 were green and four were a combination of red and green.

Table 3-3 includes detailed information about sampling effort #2 including information about habitat, sample collectors, and shipping information. There were no unusual circumstances for urchins received from Alaska, Northern California, Santa Barbara, or Baja California. For urchins collected in Washington, it is not possible to describe the exact location, depth, or habitat of this population.

Samples received from Port Orford, Oregon, had a longer than average travel time (approximately 2 1/2 days). Upon receipt the urchins showed very few signs of life and also had a very strong odor. In addition, the gonads of nearly all the urchins were very small and recessed. Finally, it was difficult to determine the sex of many of these urchins and in many cases the sex indicated for a specific sample may be incorrect. Because it was difficult to be certain about the sex, DNA was extracted immediately from all individuals that appeared to be female as well as any individual that could not definitely be determined as male.

### ***Results of DNA Extraction and PCR Amplification***

In many cases, the result of PCR amplification was used to determine the success of DNA extraction (see Materials and Methods) and for this reason, the results of these two procedures are presented together. Gonad tissue from sampling effort #1 were stored in 50% ethanol, and extracted using a phenol chloroform extraction protocol followed by ethanol precipitation of DNA (see Materials and

Methods and Sambrook *et al.*, 1989). These samples had a very low percentage of successful DNA extractions. Out of 181 samples collected at all six sites, 35% (63 urchins) produced DNA acceptable for PCR. In addition, there appears to be a correlation in the success of PCR amplification and the sex of the sample—with extraction of male tissue appearing to produce DNA more suitable for PCR than female tissue. Of all the ovary tissue extracted, only 9% were successful; however, of all the testis extracted, 44% produced DNA acceptable for PCR. In general, extraction from sampling effort #1 was very poor, but extraction from male tissue was substantially more effective than female tissue. It is possible that storing tissue samples in ethanol increased dehydration and degradation of gonadal material and DNA. Perhaps this explains why storing tissue samples at -20°C was a more successful preservation technique.

DNA extraction procedures were optimized prior to sampling effort #2. Gonad tissue collected from sampling effort #2 was stored as frozen aliquots. In addition, DNA was extracted immediately from female tissue samples whereas male samples were stored at -70°C. from one to four days before extraction. Both male and female samples were extracted by using a CTAB/PVP lysis buffer (see Materials and Methods). There does not appear to be a correlation between success of DNA extraction and sex of the sample. Only six percent of all extractions did not work. It is interesting to note that of the thirteen samples that did not yield DNA, eleven

were female, one was male, and one was of unknown sex. In other words, although extraction of sampling method #2 was very successful, there still was a trend for ovarian DNA to be more difficult to extract than testis DNA.

Regardless if the tissues were stored in ethanol or frozen, there was a consistent trend that it was more difficult to prepare DNA extraction from ovary DNA compared with testis DNA. Perhaps the polysaccharide content in ovary is higher than in testis and somehow inhibits DNA extraction and/or subsequent PCR reactions. The method that was ultimately successful extracting DNA from both ovary and testis was originally developed to extract DNA from plant tissues with very high polysaccharide content (Milligan, 1992).

### ***Results of DNA Sequencing***

Table 3-4 summarizes the sample size, the number of alleles, and the number of genotypes present at each site. The range in number of alleles at each site was between four and eight with a total number of alleles throughout the region as 14. The number of genotypes found at a specific site was lowest in Oregon (6) and highest in Santa Barbara (12). In all sites combined, there were 21 genotypes.

It was necessary to estimate whether the allele composition in our samples were representative of the number and frequency of all alleles in the population. Figure 3-1 plots the sample size (n) against the total number of alleles at that value of n.

Data points are the mean number of 100 randomizations of data resampling. The point at which the curve asymptotes indicates a sample size at which essentially all the alleles in the population have been sampled. At a sample size that falls within the range of the asymptote, it is unlikely to identify additional alleles in the population. For all sites, the data indicate that most have been identified and in general, it would be necessary to sequence from 10-15 additional samples to find another unique allele. In addition, these rare alleles represent one allele out of the entire population ( $n$  for allele number = 268) or 0.3% of the population. Based on these results, the alleles present in the samples probably represent a random sample of each geographic location as well as all but the rarest alleles in each location.

There are a total of 14 unique alleles throughout the species range (Figure 3-2). Different combinations of these alleles make up 21 unique genotypes (Table 3-5). Figures 3-3 a-f shows the nucleotide sequence data for all 134 individuals sampled for each site. These figures also include information about the sex and genotype of each sample. These data were generated by sequencing both strands of DNA. In some cases it was not possible to read the entire 282 base pair region with both strands. However, for all samples, all variable positions in the sequence have been confirmed by sequencing both strands of DNA.

## Chapter 4

### **Population Genetics of the Red Sea Urchin, *S. franciscanus*, Along the West Coast of North America**

#### **Introduction**

Barriers to gene flow are one of the most basic mechanisms creating genetic population structure within a species (Hartl and Clark, 1989). In the marine environment, these barriers could result from interruptions in species distribution (see Bermingham and Avise, 1986), oceanic current patterns (Tracey *et al.*, 1975; Saunders *et al.*, 1986), paleogeography, such as the rise of the Isthmus of Panama (Bermingham and Lessios, 1993), or dispersal limitations of planktonic larvae (reviewed by Avise, 1994; see also: Janson, 1987; McMillan *et al.*, 1992; Waples, 1987). Of these mechanisms, the role of dispersal in creating population structure is particularly uncertain (Palumbi, 1995).

In studies of genetic differentiation of marine species, all permutations of dispersal potential and corresponding population subdivision exist (see Gooch, 1975; Burton, 1983; Hedgecock, 1986; Palumbi, 1995). For example, organisms

with restricted dispersal can have either a great degree of genetic differentiation (Duffy, 1993; Berger, 1973; Day and Bayne, 1988; Gaines *et al.*, 1974) or show no genetic substructuring (France *et al.*, 1992; Selander *et al.*, 1970). Long distance dispersers can be both homogeneous throughout their entire range (Buroker *et al.*, 1983; Selander *et al.*, 1970; Ovenden *et al.*, 1992) or have distinct subpopulations (Burton and Feldman, 1981; Burton and Lee, 1994; Planes, 1993). Of these four scenarios, the first and the third patterns represent those expected for populations of marine species primarily as a result of the ability of dispersal to facilitate or limit gene flow. In the absence of dispersal and subsequent gene flow, populations of genes will differentiate as a result of natural selection and genetic drift (Nei, 1987).

There are numerous examples that support the expectation that species with low dispersal potential should have a high amount of genetic differentiation. Organisms with dispersal restricted to tens of kilometers such as limited dispersing shrimp (Duffy, 1993) and gastropods (Berger, 1973; Day and Bayne, 1988; Gaines *et al.*, 1974) show a high degree of genetic differentiation. The opposite pattern of species with high dispersal potential exhibiting a low amount of genetic differentiation is seen in oysters (Buroker *et al.*, 1983) crabs (Selander *et al.*, 1970) and rock lobsters (Ovenden *et al.*, 1992). These species all have a pelagic life history phase that is

capable of long distance dispersal with ocean currents over hundreds and possibly thousands of kilometers.

There are also several examples of marine organisms that do not fit the expected patterns. For example, the splashpool copepod *Tigriopus californicus* (Burton and Feldman, 1981, Burton and Lee, 1994) and a Polynesian surgeonfish with a 60-day planktonic phase (Planes, 1993), have a potential for long distance dispersal, yet exhibit high genetic differentiation over tens to hundreds of kilometers. The converse scenario of species with limited dispersal and high genetic homogeneity over a large regional scale are seen in hydrothermal vent amphipods (France *et al.*, 1992) and horseshoe crabs (Selander *et al.*, 1970). However, these species are not wholly panmictic and do show some genetic differences within the species' range.

This report examines the population structure of a sea urchin with an extremely long dispersal stage. *Strongylocentrotus franciscanus*, the red sea urchin, has a continuous distribution along the Pacific coast of North America and a planktotrophic larval phase of 61 to 131 days (Strathmann, 1978) that would allow for long distance dispersal within the predominantly southward moving California Current (Hickey, 1979). Published reports identify the range of *S. franciscanus* to be from northern Japan and Alaska to Cedros Island, Baja California (Morris *et al.*,

1980). The red sea urchin is a tremendously fecund broadcast spawner (Morris *et al.*, 1980) with one female releasing over eight million eggs (Strathmann, 1987). Spawning occurs year-round, with a strong seasonal maximum in late winter and early spring (Morris *et al.*, 1980). Studies of red sea urchin larvae collected from settlement brushes indicate that settlement occurs both at low levels year-round superimposed on large episodic pulses (Ebert *et al.*, 1994).

These life history characteristics, especially the ability of larvae potentially to travel thousands of kilometers, predict high gene flow and low or absent population subdivision in *S. franciscanus*. This prediction is supported in theory by Wright (1931) who modeled that for selectively neutral genes, one reproductively successful immigrant in each subpopulation every other generation is sufficient to homogenize allele frequencies across all sub populations ( see also Hedgecock *et al.*, 1994). Yet without knowledge of isolating mechanisms that may affect this species such as selection, dispersal, biogeography, etc., it is not possible to accept the de facto prediction of panmixia in *S. franciscanus* without experimental support. In addition, as an extremely valuable fishery (Leet *et al.*, 1992), it is important to understand the population structure of the red sea urchin.

*S. purpuratus*, which has distribution and life history characteristics similar to *S. franciscanus*, may have similar population structure. Examination of the population structure of the congener *S. purpuratus* have produced conflicting results. Britten *et al.* (1978) reported no difference in the thermal renaturation of DNA from two urchins separated by 2,000 km compared to the reassociation of DNA of each urchin compared to itself, suggesting a lack of genetic subdivision. Expanding the sample size and using a more sensitive technique, Palumbi and Wilson (1990) used restriction fragment length polymorphism (RFLP) of mitochondrial DNA (mtDNA) and saw no genetic differentiation in 28 *S. purpuratus* individuals from populations separated by 1,500 km along the Pacific Coast of the USA. Additionally, sequence data of cytochrome oxidase I (COI; a mitochondrial gene) in 30 individuals collected from three sites between Washington and Santa Barbara, CA (approximately 2,500 km) showed approximately 1% sequence variation yet no genetic heterogeneity among populations and no population subdivision (Palumbi and Kessing, 1991).

In contrast to these three studies that conclude large scale genetic homogeneity in the purple sea urchin, Edmands *et al.* (1996) found distinct genetic differentiation among subpopulations. Allozymes examined in *S. purpuratus* collected over 1,000

km in the southern range of the species distribution showed significant genetic differentiation among populations. Additionally, sequence data of the mtDNA gene investigated by Palumbi and Kessing (1991) yet using a large sample size and focusing on the southern range of the distribution, revealed a significant heterogeneity among locations with a contingency chi-square analysis (Edmands *et al.*, 1996 ;  $n=147$ ,  $X^2 = 115.05$   $df = 90$ ,  $p < 0.05$ ). Analyzing the same data with an Analysis of Molecular Variance, AMOVA (Excoffier *et al.* 1992), Edmands *et al.* (1996) did not find significant differentiation among locations ( $F_{ST} = 0.017$ ,  $p > 0.05$ ). However, based on a regional analysis, there was a statistically significant genetic break approximately 300 km south of Pt. Conception in central California ( $F_{RT} = 0.064$ ,  $p < 0.05$ ). All other regions and populations examined by Edmands *et al.* (1996) were not genetically differentiated.

The result of Edmands *et al.* (1996) is in conflict with those of Palumbi and Kessing (1991) and Palumbi and Wilson (1990) that conclude complete genetic homogeneity throughout the species range. Based on the conflict in these two data sets, it is not certain whether there are barriers to gene flow that result in genetically isolated populations in *S. franciscanus*. Although it is common for different genetic markers to display different portraits of genetic variation, clearly more work is

needed to resolve the discrepancy in information regarding population structure in *S. purpuratus*.

The current study examines genetic differentiation of sequence data in *S. franciscanus* using relatively large sample sizes from nearly the entire geographic range of the species. Population structure was examined using a 273 base-pair region of the nuclear gene coding for the sperm protein called bindin. The bindin protein is localized to the acrosomal tip of sea urchin sperm and mediates binding to the surface of urchin eggs (Vacquier *et al.*, 1995; Minor *et al.*, 1989). Because portions of urchin single copy nuclear DNA (scnDNA) are suggested to have 8-20 times more variability than mtDNA (Palumbi and Wilson, 1990; Palumbi and Metz, 1991; Palumbi, 1995; Britten *et al.*, 1978), it is possible that the population signal would be greater from a nuclear molecule than a mitochondrial marker. Additionally, other work evaluating variation in recognition proteins (Vacquier *et al.*, 1995; Metz and Palumbi, 1996; Hughes and Nei, 1988) suggests gamete interaction molecules such as bindin may exhibit a high degree of genetic variation both between and within species (Metz and Palumbi, 1996). The 273 base pair region of bindin examined here coincides with the region of greatest variation in *Echinometra sp.* bindin DNA (Metz and Palumbi, 1996). Finally, new evidence

suggests that this gene has evolved as a result of neutral evolution (see Chapter 5) which increases its value as a genetic biogeographic marker (Hillis and Moritz, 1990). As a result of potential increased variation in sea urchin nuclear DNA and the fact that *bindin* is a gamete interaction molecule evolving as a result of neutral evolution, *bindin* is potentially an informative population marker.

## **Materials and Methods**

### ***Sample Collection***

Between August 1995 and February 1996, 298 adult *S. franciscanus*, approximately 50 animals from each site, were collected from the following six locations: Ketchikan, Alaska; the Port Townsend, Washington; Port Orford, Oregon; Ft. Bragg, California; Santa Barbara, California; and Ensenada, Baja California, Mexico (Figure 4-1). All animals were collected on SCUBA and then shipped via overnight express to our laboratory. All individuals came from as close to a 10 to 15 meters foot depth range as possible. To insure that all samples were adults, all individuals collected were at least 80 mm in test diameter.

### ***DNA Extraction***

DNA extractions were performed as described in Milligan (1992). Approximately 25 -100 ug of gonadal tissue was homogenized in 700 ul CTAB buffer prewarmed to 60°C (100 mM Tris-HCl, pH 8.0; 1.4 M NaCl; 20 mM EDTA; 2% hexadecyltrimethylammoniumbromide (CTAB, w/v)); 1% polyvinylpyrrolidone (PVP-360, w/v); 0.2% 2-mercaptoethanol (v/v added just before use). At the end of a 30-60 minute incubation at 60°C (with periodic swirling) 700 ul of chloroform:isoamyl alcohol (24:1) was added followed by vortexing. Centrifugation for 10 minutes at 1500 x g separated the aqueous and the organic phases. One extraction with an equal volume of chloroform:isoamyl alcohol (24:1) was followed by extractions with phenol:chloroform:isoamyl until the interface was clear. The aqueous phase was transferred to a sterile tube followed by addition of 2/3 volume of ice-cold isopropanol and vortexing. The pelleted DNA was washed with 500 ul wash buffer (76% ethanol, 10 mM ammonium acetate), dried, and resuspended in 20-30 ul TE containing 1 ul RNase (10 mg/ml). It was imperative to conduct the extraction protocol on fresh ovary tissue. Ovary samples frozen for as short as one day showed reduced yield of DNA and reduced success in subsequent PCR amplification.

### ***PCR Amplification***

#### *Double-stranded PCR amplification*

The primers FNbindin 5' (5'-AGTCGACGTTTCGACAGACGAC-3') and FNbindin 3' (5'-TTACATGGTCCATTATAGTATGCC-3') amplify a 431 base pair region of the 5' end of the bindin gene. Amplification followed standard procedures (Saiki *et al.*, 1988) using a reaction volume of 25 ul and final magnesium chloride concentration of 2 mM. The thermocycler (Perkin Elmer Cetus) profile for all double stranded reactions was: 1 cycle of 95 °C, 5 min. followed by 30 cycles of 94 °C, 1 min.; 60 °C, 1 min.; 72 °C, 2 min. Five ul of each PCR product was resolved on a 2% Nu Sieve agarose, low melting temperature TBE gel. Gel isolates were removed with sterile wide-bore, disposable polyethylene transfer pipettes and stored in a microfuge tube with 200 ul of water at -20 °C.

#### *Single-stranded PCR amplification*

Each gel isolate was heated to 65°C for approximately 5-10 minutes and used as template for the single-stranded PCR amplification. To amplify a single-stranded product of the 5' strand, the PCR amplification conditions were identical to those identified above for amplification of the double-stranded product except the 5' primer (FNbindin 5') was used at a final concentration of 0.5 uM.

The PCR amplification conditions to amplify the 3' single-stranded product required a lower annealing temperature (58°C) and lower MgCl<sub>2</sub><sup>+</sup> concentration (1.2 mM final). The concentration of the 3' limiting primer was 2.5 uM (final). Some samples required additional adjustments to annealing temperature (between 57-63°C) and limiting primer concentration (0.5-2.5 uM final concentrations).

***Direct sequencing of single-stranded PCR product***

The single-stranded PCR products were washed in Centricon filter units (30,000 MWCo) and resuspended in 7 ul H<sub>2</sub>O for Sanger dideoxy sequencing (Sequenase ver. 2.0, U.S. Biochemical) using internal sequencing primers KTseq 5' (5'- GGAGCGCGTAAGAAGCGTTAT-3') and KTseq 3' (5'- ATACACACGATGGTCAAG-3') at 10uM.

***Cloning of bindin DNA from heterozygous individuals***

In order to confirm the exact sequence of representatives of all alleles, the PCR products were cloned. Double-stranded PCR products were amplified using primers KBRS 5' (5'CGCGGATCCAGTCGACGTTTCGACAGACGAC-3') and 3' (5'-GCCAAGCTTTTACATGGTCCATTATAGTATGCC-3') and the double-stranded PCR protocol. These primers incorporate the restriction sites BamHI and HindIII respectively on their 5' ends to facilitate directional cloning into the pBMKS

bluescript vector. The PCR products were resolved on a 2% agarose gel and the excised gel fragment was purified using Quiaquick spin columns (Quiagen) and then digested with BamHI and HindIII. The gel-purified fragment was ligated into pBMKS and transformed into *E. coli* DH5 $\alpha$ . The DNA from a minimum of four transformants were sequenced for each PCR product. Several individuals with the same genotype, but from different geographic locations, were cloned and sequenced to verify sequence consistency among alleles found at different geographic locations. Plasmid DNA was purified using the alkaline lysis method (Sambrook *et al.*, 1989). The double-stranded plasmid DNA was sequenced using Kgsseq 5' (5'GTTTCTGACG ATTCGGAAAGA-3') and Kgsseq 3' (5'-GAAACAACCAATTTAAAAATA-3') as internal sequencing primers.

Cloning of the PCR products also allowed us to assess nucleotide changes due to PCR incorporation error and thus disregard those sequences. Cloned PCR products from four different reactions were sequenced on both strands. A substitution in a sequence was determined to be PCR incorporation error based on one or both of the following observations: (1) the nucleotide present in the sequence was not present in either the 5' or 3' strand produced by direct sequencing; or (2) the existence of this new sequence now created a total of three alleles for one genotype which is

biologically impossible for a diploid organism. We found what appeared to be a substitution resulting from PCR incorporation error in approximately one out of 2,000 nucleotides. The sequence of a particular allele was considered to be an accurate representation of an individual genotype when it appeared in at least two out of three of the cloned sequences.

### ***Sequence and Statistical Analyses***

From the 431 base pair PCR product, we obtained unambiguous sequence information confirmed with both strands of DNA for a 273 base pair region. Sequences were aligned using Seq-App ver 1.9a multiple sequence alignment program for the Macintosh (Gilbert, 1994). Molecular Evolutionary Genetic Analysis ver. 1.01 (MEGA, Kumar, *et al.*, 1993) was used to calculate nucleotide sequence divergence. Individual genotypes were coded as paired alphabetical characters and analyzed with BIOSYS (Swofford, 1989) to obtain estimates of the following: allele frequencies, conformance to Hardy-Weinberg equilibrium, Wright's (1978) F-statistics, and Nei's (1972) minimum genetic distance in pairwise comparisons. Conformance to Hardy-Weinberg proportions were estimated in three ways: (1) contingency chi-square analysis with Leven's (1949) correction for small sample size; (2) chi-square with pooling of rare and common categories; and (3) significance test with exact probabilities. Finally, a Monte Carlo simulation of a

chi-square contingency test as per Roff and Bentzen (1989) was performed using 1000 runs. The CHIRXC program, available from Zaykin and Pudovkin (North Carolina State University), tests conformance to the null hypothesis of homogeneity.

### **Results**

Out of 134 individuals sequenced from six geographic locations along the Pacific coast of North America, 14 alleles and a total of 21 genotypes were identified, suggesting a high degree of polymorphism at this locus (see Table 3-4). The number of alleles per sampling site ranged from four to eight locus (Figure 4-1 and see Table 3-4). The maximum number of genotypes identified at a specific geographic location ranges from six to twelve locus (Figure 4-1 and see Table 3-4). There is no obvious trend in the number of alleles or the number of genotypes with geographic location (locus (Figure 4-1 and see Table 3-4).

There are a total of 13 variable nucleotide positions in all 14 alleles (Table 4-1). All other positions of the 273 base pair region, in all 134 individuals sequenced, are identical. The number of nucleotide differences between any two sequences is low, ranging between one (0.35%) and six (2.1%) nucleotide substitutions in the 273 base pair region. These pairwise comparisons of the 14 unique alleles show a maximum of six nucleotide differences in only one out of 91 (<1%) of the

comparisons. On average, there are 2.9 nucleotide positions that vary between any two unique alleles. Average nucleotide-sequence diversity is 1.06% (Tajima and Nei, 1984, excluding correction for multiple hits). Figure4-2 represents the phylogenetic relationship of the 14 unique alleles as determined by a parsimony analysis (PAUP, Swofford, 1993).

The allele frequencies data were calculated to assess population structure and conformance to Hardy-Weinberg equilibrium. Figure4-3 presents the frequency data for each allele and each geographic location. Four alleles are dominant throughout the range examined; allele A is the most common allele (51%). Allele B, C, and D are the next most common alleles with frequencies of 16%, 20%, and 8% respectively. All other alleles are rare, representing  $\leq 1\%$  of the total population. Allele G is present in three and allele F is present in two individuals out of the total 134 individuals sequenced. All remaining alleles (E, H, I, J, K, L, M, and N) are present in only a single individual out of all urchins sampled.

Using BIOSYS (Swofford, 1989), average heterozygosity was calculated in three ways (Table 4-2 (1) direct count of the proportion of individuals that are heterozygous; (2) an estimate of heterozygosity based on Hardy-Weinberg expectations; and (3) an unbiased estimate based on conditional expectations

(Levene, 1949; Nei, 1978). The results from three different tests for deviation from Hardy-Weinberg equilibrium were all nonsignificant at all geographic locations (Table 4-2). Analysis of allele frequencies in Alaska with pooled data indicated a mildly significant departure from Hardy-Weinberg equilibrium ( $p=0.037$ ). All other measures of allele frequencies in Alaska suggest this population is in Hardy-Weinberg equilibrium (significance test with exact probabilities,  $p=0.08$ , and chi-square test on complete data set,  $p=1.0$ ).

To evaluate population subdivision over all geographic scales, F-statistics were calculated for all possible combinations of geographic locations. All combinations of regional grouping were tested and indicate insignificant values and random mating ( $F_{ST} \leq 0.008$ ). For example, analysis of Alaska as one group versus a second grouping of Washington, Oregon, Northern California, Santa Barbara, and Baja California, indicated insignificant deviations from zero. Population structure was also evaluated using a Monte Carlo simulation of a randomized chi-square test (Roff and Bentzen, 1989). After 1,000 simulations, it is not possible to reject the null hypothesis of homogeneity ( $X^2 = 52.46$ ,  $df = 65$ ,  $p > 0.05$ ).

## **Discussion**

All measures of genetic variation in *bindin* indicated that the red sea urchin, *S. franciscanus*, is panmictic throughout the entire range examined from Baja to Alaska. Regardless of whether the data are pooled, resampled, or evaluated for inbreeding coefficients, no statistically significant differences in allele frequencies among sites were observed (see below for single exception). Conformance to Hardy-Weinberg equilibrium expectations at all sampling sites suggests that urchin populations from Alaska to Baja are mating randomly. Finally, no combination of regional groupings indicated the existence of a genetic break between any sampling sites in the range examined. These data are consistent with a conclusion of sufficient gene flow to prevent genetic divergence within *S. franciscanus* along the Pacific coast.

As was expected based on work with other recognition proteins (Metz and Palumbi, 1996; Vacquier *et al.*, 1995; Hughes and Nei, 1988), *bindin* appears to be an acceptable genetic marker with sufficient polymorphism to detect genetic structure. All six populations of *S. franciscanus* are highly polymorphic with at least four alleles, but as many as twelve alleles at a single geographic site. With the amount of polymorphism present in *bindin*, genetic isolation could readily result in changes in allele frequencies. Given the constancy of *bindin* allele frequencies

throughout the range examined, it seems likely that genetic isolation does not exist for *S. franciscanus*.

Individuals sampled from Alaska show allele frequency patterns that are slightly different than the five other sampling locations. When Alaska genotypes are pooled into one of the following three categories, ( (a) homozygotes for most common allele; (b) common/rare heterozygotes; (c) rare homozygotes and other heterozygotes), there is a statistically significant deviation from chi-square expectations ( $p=0.037$ ). In addition to the fact that the  $p$ -value is only marginally significant, other tests of the same hypothesis using alternate pooling strategies do not detect statistical significance (i.e. Fisher exact test,  $p=0.08$ ). The goal of pooling allele frequency data is to ensure that rare alleles are not overrepresented in the population sample (Cochran, 1954). As a rule of thumb, Cochran (1954) suggested that no expected frequency should be  $<1.0$  and no more than 20% of the expected frequencies should be  $< 5.0$  (Roff and Bentzen, 1989; Cochran, 1954). Although the pooling method described above satisfies these guidelines, there are many alternative ways to pool the data that do not result in deviations from Hardy-Weinberg expectations.

One explanation for the anomaly in the Alaska data set is that mating is non-random. However, a more likely explanation is that signal of selection evident in the pooled Alaska data is incorrect. Typically the most accurate assessment of allele frequencies for small sample sizes is Fisher's exact test (Lessios, 1992) or a resampling technique to compensate for statistically small sample sizes (Roff and Bentzen, 1989). According to these two tests, selection is not detected in Alaska.

Slatkin and Maddison (1990) describe a method to infer isolation by distance using phylogenies of genes. For *S. franciscanus*, the geographic location of alleles does not correlate with the phylogenetic relationships described by the phylogenetic tree (see Figure 4-2 and Figure 4-3). More importantly, the inability to resolve polytomous branches suggests that there are many alternative ways to rearrange the branches on the tree. The uncertainty in the phylogenetic tree is most likely a result of the small number of nucleotide substitutions in the region of the *bindin* gene sequenced (range from one to six when comparing any two alleles). As a result, the variance of sequence diversity is high relative to the mean, hindering the ability to identify relationships among alleles.

Slatkin (1985) also described a method to use rare, "private alleles" existing in only one geographic location to estimate gene flow. However, it was not possible to

use this method here because the rare singleton alleles often occur in several geographic locations and are thus not considered "private alleles." The existence of the same rare allele in populations separated by 1,500 km is additional evidence for extremely high gene flow in *S. franciscanus*.

*S. franciscanus* and *S. purpuratus* are distantly related congeners and most of the life history traits that would determine gene flow appear identical in the two species: long planktotrophic larval stage; habitat range along the Pacific coast of North America (Morris *et al.*, 1980); tremendously high fecundity (Morris *et al.* 1980; Strathmann, 1978); broadcast spawning primarily in winter/spring; and periodic settlement and recruitment (Ebert *et al.*, 1994). The high gene flow identified in *S. franciscanus* is similar to that in *S. purpuratus*, the current study supports the conclusions of limited genetic heterogeneity found in *S. purpuratus* (Britten *et al.*, 1978; Palumbi and Wilson, 1990; Palumbi and Kessing, 1991). However if the two species have similar population structure, the current study contradicts the finding of Edmands *et al.* (1996) that identifies a genetic break in *S. purpuratus* populations. It is possible that there are differences in the two species that can create the mild genetic break seen in purple but not in red urchin populations. However, the genetic break evidenced by mtDNA (Edmands *et al.*, 1996) was not substantiated by the

allozyme data (Edmands *et al.*, 1996). In addition, the allozyme data showed heterogeneity in allele frequencies that was not supported by the mtDNA sequence data (Edmands *et al.*, 1996). Although the sample size for the work presented here using the *bindin* locus is quite large for sequence data, it focuses on a single locus and thus is potentially limited. The value of examining multiple loci in order to formulate assessments of population structure can not be overemphasized.

Selection on the *bindin* locus could also produce the observed polymorphism. In addition, selection acting on *bindin* could influence population structure. For example, a scenario of restricted gene flow with a global selective pressure would result in genetic homogeneity as would a scenario of high gene flow and no selection. However, evaluation of evolutionary forces on the 5' region of *bindin* suggests only purifying selection and neutral evolution act to shape *bindin* variability (see chapter 4). The polymorphism observed in *bindin* is most likely the result of nearly neutral random mutations (see chapter 4) and thus selection is not confounding the pattern of genetic heterogeneity.

Because estimated average scnDNA diversity is 8-20 times greater than mtDNA, the amount of variation in *bindin* was expected to be at least higher than sea urchin mtDNA (Palumbi and Wilson, 1990, Palumbi and Metz, 1991; Palumbi,

1995; Britten *et al.*, 1978). This was not the case. Although bindin sequence variation (1.06%) is low, it is approximately equal to variation in mtDNA (COI) examined in *S. purpuratus*, *S. droebachiensis*, and *Echinometra sp.* (Palumbi and Metz, 1991; Palumbi and Wilson, 1990; Edmands *et al.*, 1996). Population structure has been observed in *S. droebachiensis* (Palumbi and Wilson, 1990) and *Echinometra mathaei*. (Palumbi and Metz, 1991). By analogy, these studies suggest that if significant population structure were present in *S. franciscanus* as well, it would have been detected.

Comparison with these studies also illustrates the role that biogeography can play in facilitating gene flow. All of these species have similar life history characteristics (long pelagic phase, broadcast spawning, etc.). As demonstrated by this and other work (Palumbi and Wilson, 1990; Palumbi and Kessing, 1991), urchin species that have been observed to be homogeneous throughout their range tend to inhabit a continuous coastline. Species that have genetically isolated populations, such as *S. droebachiensis* and *Echinometra sp.*, have interruptions in their species distribution (Palumbi and Wilson, 1990; Palumbi and Metz, 1991). These gaps in species occurrence may be sufficient to limit gene flow and eventually result in population subdivision. Thus, categorization of these species by

one of their most obvious characteristics, biogeography, potentially elucidates the mechanism creating population structure.

Point Conception is a prominent headland and biogeographic feature along the Pacific coast (Figure 4-1). As a major boundary of the Oregonian and Californian biogeographic provinces (Valentine, 1973), Point Conception divides two regions with disparate sets of co-occurring species and marks the limit of species distribution for many species (see Avise, 1994; Palumbi, 1995; Briggs, 1958, 1974). It is hypothesized that biogeographic boundaries can also be boundaries for gene flow and result in population subdivision (see Avise, 1994; Palumbi, 1995; Briggs, 1958, 1974). The California Current, that flows predominantly southward from Alaska to Baja California, takes a sharp turn seaward at Point Conception potentially demarcating northern and southern water masses (Hickey, 1979). For a species existing on a linear coastline with a unidirectional current, genetic isolation created by Point Conception would predict a greater amount of genetic variation in the south compared to the north. This is not the pattern seen here or for other Strongylocentrotids. Southern populations of *s. franciscanus* and *S. purpuratus* apparently are not genetically isolated from populations north of Point Conception (Edmands *et al.*, 1996; Palumbi and Wilson, 1990; Palumbi and Kessing, 1991).

Although flow of the California Current is predominantly southward, reversals in flow direction are quite common (Pirie *et al.*, 1975). As a result of an El Niño Southern Oscillation (ENSO) event, warm tropical waters can travel past Point Conception as far north as Vancouver Island (Crowe and Schwartzlose, 1972). In addition, every winter, the southern California and Davidson Counter currents send water from southern California northward (Hickey, 1979). Extreme homogeneity throughout the range examined suggests that *S. franciscanus* larvae are able to move south with the California Current as well as north with either a periodic ENSO event or annually with the Davidson Counter current. However, Point Conception is a dramatic genetic break for allozyme loci in the splashpool copepod *Tigriopus californicus* (Burton and Feldman, 1981; Burton and Lee, 1994). Thus, the genetic break at Point Conception suggests that *T. californicus* can not move north across Point Conception with the same currents that move urchin larvae.

Although they share a common biogeography, a major distinction between *T. californicus* and the two Strongylocentrotids is in the duration of the pelagic larval phase. Urchin larvae can spend months in the plankton, yet the copepod eggs hatch and become adults in 14-22 days depending on water temperature and available splashpool habitat (Morris *et al.*, 1980). Even when *T. californicus* eggs hatch

during a northward flow reversal, the larvae may settle before the water mass makes it around Point Conception. On the other hand, *S. franciscanus* and *S. purpuratus* larvae could remain in the current until well northward of the Point. In addition, the shorter larval period of *Tigriopus* allows larval behavior to play a relatively greater role in limiting dispersal (Burton and Feldman, 1982). For urchin larvae that are in the plankton for several months, behavior probably is a relatively smaller factor in determining ultimate dispersal compared to strength and duration of the prevailing current.

Clearly biogeography is not the sole or necessarily the most important determinant of population genetic structure. In the case of the four urchin species that share similar life history characteristics, disparate biogeography resulted in unique patterns of population structure. In the case of two organisms that share a similar biogeography (Strongylocentrotids and *Tigriopus*) dispersal potential apparently influences the limits of gene flow. In fact, another marine organism that has a similar biogeography and spends long periods in the plankton like the Strongylocentrotids, also shows a similar genetic pattern. The northern anchovy (*Engraulis mordax*) showed slight yet significant genetic heterogeneity over small

scales, but the absence of population subdivision at Point Conception (Hedgecock, 1994; Hedgecock *et al.*, 1994).

In conclusion, sequence data of *S. franciscanus* bindin indicate that the red sea urchin represents a marine species where dispersal potential corresponds to the expected portrait of population structure. A dispersal stage of two to four months creates the potential for planktonic larvae to travel 1,000 to 3,000 kilometers with the California Current's average speed of 0.25 m/sec (Pirie, 1975). It is not possible to know whether transfer of the red urchin's genetic material occurs as a result of larval transport over these long distances or as the accumulation of genetic transfer made over small distances. Although the long pelagic phase may allow transmission of genetic information over thousands of kilometers, retention of local water masses could keep larvae close to parental stocks. Even with local retention, high gene flow is possible if genetic material travels up and down the coast in a stepping stone process (Wright, 1978).

The data of *S. franciscanus* bindin allele frequencies does not elucidate whether dispersal potential is rarely, sometimes, or often realized. Examples of shared alleles in urchin populations separated by thousands of kilometers of open ocean (*Echinometra sp.* and *S. droebachiensis*) suggests urchins have the ability to realize

the furthest extent of their dispersal potential. The fact that *Echinometra sp.* and *S. droebachiensis* populations are not panmictic indicates that this very long distance dispersal is not a regular event. The most likely scenario for *S. franciscanus* dispersal is for moderate dispersal (100s of km) to be common, with an occasional delivery of larvae over thousands of kilometers. To be certain of actual dispersal, it will be necessary to identify or introduce population specific markers or to develop techniques to track larval movement.

### References

- Avise, J. C. (1994). *Molecular markers, natural history and evolution* New York, New York: Chapman and Hall.
- Berger, E. M. (1973). Gene-enzyme variation in three sympatric species of *Littorina*. *Biol. Bull.* 145: 83-90.
- Bermingham, E., and Avise, J. C. (1986). Molecular zoogeography of freshwater fishes of the southeastern United States. *Genetics* 113: 939-965.
- Bermingham, E., and Lessios, H. A. (1993). Rate variation of protein and mitochondrial DNA evolution as revealed by sea urchins separated by the Isthmus of Panama. *Proc. Natl. Acad. Sci. USA* 90: 2734-2738.
- Briggs, J. C. (1958). A list of Florida fishes and their distribution. *Bull. Fla. State Mus. Biol. Sci* 2:(223-318) .
- Briggs, J. C. (1974). *Marine Zoogeography* New York: McGraw-Hill.
- Britten, R. J., Cetta, A., and Davidson, E. H. (1978). The single-copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. *Cell* 15: 1175-1186.
- Buroker, N. E. (1983). Population genetics of the American oyster *Crassostrea virginica* along the Atlantic coast and the Gulf of Mexico. *Marine Biology* 75: 99-112.
- Burton, R. S. (1983). Protein polymorphisms and genetic differentiation of marine invertebrate populations. *Mar. Biol. Letts.* 4: 193-206.

- Burton, R. S., and Feldman, M. W. (1982). Population genetics of coastal and estuarine invertebrates: does larval behavior influence population structure? In *Estuarine Comparisons*, V. S. Kennedy, ed. New York: Academic Press, pp. 537-551.
- Burton, R. S., and Feldman, M. W. (1981). Population genetics of *Tigriopus californicus*. II. Differentiation among neighboring populations. *Evolution* 35:(6) 1192-1205.
- Burton, R. S., and Lee, B. G. (1994). Nuclear and mitochondrial gene genealogies and allozyme polymorphisms across a major phylogeographic break in the copepod *Tigriopus californicus*. *Proc. Nat. Acad. Sci. USA* 91: 5197-5201.
- Cochran, W. G. (1954). Some methods for strengthening the common  $X^2$  tests. *Biometrics* 10: 417-451.
- Crowe, F. J., and Schwartzlose, R. A. (1972). Release and recovery records of drift bottles in the California region, 1955 through 1971. In California Cooperative Oceanic Fisheries Atlas #16, Marine Research Committee: State of California.
- Day, A. J., and Bayne, B. L. (1988). Allozyme variation in populations of the dogwhelk *Nucella lapillus* (Prosobranchia, Muricacea) from the south-west peninsula of England. *Mar. Biol.* 99: 93-100.
- Duffy, J. E. (1993). Genetic population structure in two tropical sponge-dwelling shrimps that differ in dispersal potential. *Mar. Biol.* 116: 459-470.
- Ebert, T. A., and Russell, M. P. (1988). Latitudinal variation in size structure of the west coast purple sea urchin: a correlation with headlands. *Limnology and Oceanography* 33: 286-294.
- Ebert, T. A., Schroeter, S. C., Dixon, J. D., and Kalvass, P. (1994). Settlement patterns of red and purple sea urchins (*Strongylocentrotus franciscanus* and *S. purpuratus*) in California, USA. *Mar. Ecol. Prog. Ser.* 111: 41-52.

Edmands, S., Moberg, P. E., and Burton, R. S. (1996). Allozyme and mitochondrial DNA evidence of population subdivision in the purple sea urchin *Strongylocentrotus purpuratus*. *Marine Biology* 126:(3) 443-450.

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479-491.

Felsenstein, J. (1993). *Phylogeny Inference Package (PHYLIP) (Version 3.5)*. Seattle: University of Washington.

France, S. C., Hessler, R. R., and Vrijenhoek, R. C. (1992). Genetic differentiation between spatially disjunct populations of the deep-sea, hydrothermal vent-endemic amphipod *Ventiella sulfuris*. *Mar. Biol.* 114: 551-559.

Gaines, M. S., Caldwell, J., and Vivas, A. M. (1974). Genetic variation in the mangrove periwinkle *Littorina angulifera*. *Mar. Biol.* 27: 327-332.

Gilbert, D. (1994). SeqApp. Multiple sequence alignment program (1.9a157+). Indiana State University.

Gooch, J. L. (1975). Mechanisms of evolution and population genetics. In *Marine Ecology*, O'Kinne, ed. London: Wiley, pp. 349-409.

Hartl, D. L., and Clark, A. G. (1989). *Principles of population genetics*, Second Edition Sunderland, Mass.: Sinauer Assoc.

Hedgecock, D. (1994). Temporal and spatial genetic structure of marine animal populations in the California current. California Cooperative Oceanic Fisheries Investigations (CalCOFI) Report 35: 73-81.

Hedgecock, D., Hutchinson, E., Li, G., Sly, F. L., and Nelson, K. (1994). The central stock of northern anchovy *Engraulis mordax* is not a randomly mating population. California Cooperative Oceanic Fisheries Investigations (CalCOFI) Report 35: 121-136.

- Hickey, B. M. (1979). The California Current system-hypotheses and facts. *Prog. Oceanog.* 8: 191-279.
- Hillis, D. M., and Moritz, C. (1990). *Molecular Systematics* Sunderland, MA: Sinauer Associates.
- Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:(8 September) 167-168.
- Hughes, A. L., Ota, T., and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of Class I Major-Histocompatibility-Complex molecules. *Mol. Biol. Evol.* 7:(6) 515-524.
- Hunt, A. (1993). Effects of contrasting patterns of larval dispersal on the genetic connectedness of local populations of two intertidal starfish, *Patiriella calcar* and *P. exigua*. *Mar. Ecol. Prog. Ser.* 92: 179-186.
- Janson, K. (1987). Allozyme and shell variation in two marine snails (*Littorina*, *Probranchia*) with different dispersal abilities. *Biol. J. Linn. Soc.* 30: 245-256.
- Kumar, S., Tamura, K., and Nei, M. (1993). MEGA: Molecular Evolutionary Genetics Analysis (version 1.01). University Park, PA: The Pennsylvania State University.
- Leet, W. S., Dewees, C. M., and Haugen, C. W. (1992). California's living marine resources and their utilization Davis, California: California Sea Grant.
- Lessios, H. A. (1992). Testing electrophoretic data for agreement with Hardy-Weinberg expectations. *Marine Biology* 112,: 517-523.
- Levene, H. (1949). On a matching problem arising in genetics. *Ann. Math. Stat.* 20: 91-94.

- McMillan, W. O., Raff, R. A., and Palumbi, S. R. (1992). Population genetic consequences of developmental evolution in sea urchins (genus *Heliocidaris*). *Evolution* 46:(5) 1299-1312.
- Metz, E. C., and Palumbi, S. R. (1996). Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* 13:(2) 397-406.
- Milligan, B. G. (1992). Plant DNA Isolation. In *Molecular genetic analysis of populations*, A. R. Hoelzel, ed. Oxford: IRL Press, pp. 71-74.
- Minor, J. E., Gao, B., and Davidson, E. H. (1989). The molecular biology of bindin. In *The Molecular Biology of Fertilization*, H. Schatten and G. Schatten, eds.: Academic Press, pp. 73-88.
- Morris, R. H., Abbott, D. P., and Haderlie, E. C. (1980). *Intertidal invertebrates of California* Stanford, CA: Stanford University Press,.
- Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.
- Nei, M. (1972). Genetic distance between populations. *American Naturalist* 106: 283-292.
- Nei, M. (1987). *Molecular evolutionary genetics* New York: Columbia University Press.
- Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:(5) 418-26.
- Ovenden, J. R., Brasher, D. J., and White, R. W. G. (1992). Mitochondrial DNA analyses of the red rock lobster *Jasus edwardsii* supports an apparent absence of population subdivision throughout Australasia. *Mar. Biol.* 112: 319-326.

Palumbi, S. R. (1995). Using genetics as an indirect estimator of larval dispersal. In *The Ecology of Marine Invertebrate Larvae*, L. R. McEdward, ed.: CRC Press, pp. 369-387.

Palumbi, S. R. (1996). What can molecular genetics contribute to marine biogeography? An urchin's tale. *J. Exp. Mar. Biol. and Ecol.* 203: 75-92.

Palumbi, S. R., and Kessing, B. D. (1991). Population biology of the trans-Arctic exchange: mtDNA sequence similarity between Pacific and Atlantic sea urchins. *Evolution* 45:(8) 1790-1805.

Palumbi, S. R., and Metz, E. C. (1991). Strong reproductive isolation between closely related tropical sea urchins (genus *Echinometra*). *Mol. Biol. Evol.* 8:(2) 227-239.

Palumbi, S. R., and Wilson, A. C. (1990). Mitochondrial DNA diversity in the sea urchins *Strongylocentrotus Purpuratus* and *S. Droebachiensis*. *Evolution* 44:(2) 403-15.

Pirie, D. M., Murphy, M. J., and Edmisten, J. R. (1975). California nearshore surface currents. *Shore and Beach October:* 23-34.

Planes, S. (1993). Genetic differentiation in relation to restricted larval dispersal of the convict surgeonfish *Acanthurus triostegus* in French Polynesia. *Mar. Ecol. Prog. Ser.* 98: 237-246.

Roff, D. A., and Bentzen, P. (1989). The statistical analysis of mitochondrial DNA polymorphisms:  $\chi^2$  and the problem of small samples. *Mol. Biol. Evol.* 6:(5) 539-545.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of *B*-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230: 1350-1354.

- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular cloning: a laboratory manual*: Cold Springs Harbor Laboratory Press.
- Saunders, N. C., Kessler, L. G., and Avise, J. C. (1986). Genetic variation and geographic differentiation in mitochondrial DNA of the horseshoe crab, *Limulus polyphemus*. *Genetics* 112: 613-627.
- Selander, R. K., Yang, S. Y., Lewonton, R. C., and Johnson, W. E. (1970). Genetic variation in the horseshoe crab (*Limulus polyphemus*), a phylogenetic "relic". *Evolution* 24: 402-414.
- Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution* 39:(1) 53-65.
- Slatkin, M., and Maddison, W. P. (1990). Detecting isolation by distance using phylogenies of genes. *Genetics* 126: 249-260.
- Strathmann, M. F. (1987). Reproduction and development of marine invertebrates of North Pacific Coast: Data and methods for the study of eggs, embryos, and larvae. Seattle: University of WA Press 657.
- Strathmann, R. R. (1978). Length of pelagic period in echinoderms with feeding larvae from the Northeast Pacific. *J. Exp. Mar. Biol. Ecol.* 34: 23-27.
- Swofford, D. L. (1989). BIOSYS-1: A computer program for the analysis of allelic variation in population genetics and biochemical systematics (Version 1.7). Champaign, IL: Swofford, D.L. at University of Illinois at Urbana.
- Tajima, F., and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1: 269-285.
- Tracey, M. L., Nelson, K., Hedgecock, D., Shelser, R. A., and Pressick, M. L. (1975). Biochemical genetics of lobsters: genetic variation and structure of American lobster (*Homerus americanus*) populations. *J. Fish. Res. Bd. Can.* 32: 2091-2101.

Vacquier, V. D., Swanson, W. J., and Hellberg, M. E. (1995). What have we learned about sea urchin sperm binding? *Develop., Growth and Differ.* 37: 1-10.

Valentine, J. W. (1973). *Evolutionary Paleocology of the Marine Biosphere* Englewood Cliffs, New Jersey: Prentice-Hall.

Waples, R. S. (1987). A multispecies approach to the analysis of gene flow in marine shore fishes. *Evolution* 41:(2) 385-400.

## Chapter 5

### **Intraspecific Sequence Variation in the Bindin Locus of *S. franciscanus*: Unusual Nucleotide Variation in a Marine Invertebrate Gamete Interaction Molecule**

#### **Introduction**

Among sexually reproducing organisms, many marine invertebrates have unique life history and reproductive strategies. For organisms such as urchins, abalone, polychaetes, bivalves, etc., an adult phase releases gametes into the water column (Strathmann, 1987). The gametes fuse species-specifically to create a zygote that develops into a motile larvae (Strathmann, 1987). In most cases, there is limited, if any, interaction between the sexual adults. If courtship or any type of signals for assortative mating exist in these organisms, it must be limited to chemical or physical interactions between the gametes.

These unique life history strategies and recent data concerning the gamete interaction molecules of two marine invertebrates prompted speculation that selection may be creating unexpected patterns of sequence divergence in those molecules. For example, lysin is an abalone sperm molecule that dissolves the egg

vitelline layer, allowing the sperm to penetrate the egg (Lee *et al.*, 1995; Vacquier *et al.*, 1993; Lee and Vacquier, 1993). Closely related abalone species have lysin sequence patterns that suggest selection is operating to create rapid sequence divergence at this locus. Specifically, lysin sequences exhibit a much higher number of replacement versus silent substitutions. According to prediction of the neutral theory of evolution (Kimura, 1983), the proportion of nonsynonymous substitutions per nonsynonymous site ( $d_n$ ) should equal the proportion of synonymous substitutions per expected synonymous site ( $d_s$ ) (Nei and Gojobori, 1986). When  $d_n$  is significantly greater than  $d_s$ , the action of divergent evolution is suggested in comparisons between species and diversity enhancing selection is suggested from comparisons within species. The term positive selection is often applied to both processes (see Endo *et al.*, 1996; Hughes and Nei, 1988; Vacquier *et al.*, 1997; Lee *et al.*, 1995 ). Values of  $d_n:d_s$  significantly less than 1.0 indicate the action of purifying selection eliminating replacement substitutions within a species. In six pairwise comparisons of abalone lysin sequences,  $d_n:d_s$  is significantly greater than 1.0.

In a second marine invertebrate, the widespread tropical sea urchin *Echinometra*, significantly high sequence divergence in the sperm protein bindin was reported among species (Metz and Palumbi, 1996). Bindin binds species specifically to the receptor ligand on the egg (reviewed by Vacquier *et al.*, 1995;

Vacquier and Moy, 1977). A successful interaction between bindin and the receptor is necessary for sea urchin fertilization (see Vacquier *et al.*, 1995; Foltz and Lennarz, 1992). Palumbi and Metz (1996) reported a significantly high ratio of dn:ds (based on a one-tailed t test) in a 39 codon subset of the 276 codon bindin molecule and concluded rapid interspecific divergence occurred in three recently speciated congeners.

High dn:ds values in marine invertebrate gamete interaction molecules has prompted speculation that selection may result in rapid sequence divergence among related species (see Chapter 6; Lee and Vacquier, 1993; Vacquier *et al.*, 1993; Palumbi, 1994; Metz and Palumbi, 1996) and within species (Metz and Palumbi, 1996). An intraspecific pressure creating sequence diversity within species at marine invertebrate gamete interaction molecules is particularly intriguing because it opposes the standard evolutionary thinking that traits closely related to fitness are strictly conserved (Nei, 1987 ). A 60 base pair region in the center of the bindin coding region, designated the "conserved core," is highly conserved. An interspecific analysis of three *Strongylocentrotus* species and *Lytechinus variegatus* (see Chapter 6; Minor *et al.*, 1991) shows extremely high amino acid similarity between the four species in this region (Minor *et al.*, 1991; Vacquier *et al.*, 1995; Chapter 6). But the interspecific variability in bindin observed by Metz and

Palumbi (1996) exists in a 39 codon region just 5' of the conserved core. It has yet to be demonstrated that this variability also exists at the species level.

There are three systems where selection has been shown to create allelic diversity within species. The major histocompatibility complex (MHC) immune recognition molecules in humans and mice, the merozoite surface antigen in the malarial parasite *Plasmodium falciparum*, and the self-incompatibility locus (S-alleles) in plants. All have high allelic diversity as a result of diversity enhancing selection (Hughes and Nei, 1988; Hughes and Nei, 1989, Hughes *et al.*, 1990; Hughes, 1992). For MHCs and *Plasmodium falciparum* there is an adaptive value to increase the ability to recognize a wide range of proteins (Hughes and Nei, 1989, Hughes, 1992). For S-alleles, allelic diversity coupled with self-incompatibility promotes outcrossing and reduces inbreeding depression (Charlesworth, 1987; Haring, 1990). In all three systems there is an adaptive advantage to heterozygosity and high intraspecific polymorphism. Although diversity enhancing selection has not been observed in a marine invertebrate gamete interaction molecule, sequence variation in sea urchin bindin showed higher than expected levels of polymorphism (Palumbi and Metz, 1996).

Despite several "hints" that selection may favor allelic diversity in marine invertebrate fertilization proteins, the mechanisms for such diversifying selection are not evident. Without courtship behavior perhaps mate matching exists in these

broadcast spawners at the molecular level as a means to convey fitness information to a "choosy" gamete. This mechanism would require that the binding locus is linked to other traits, a possibility that has not been explored. It has also been suggested, but not demonstrated, that allelic diversity in the receptor and by necessity binding, may be a means to avoid microbial pathogens that bind to the egg surface receptor as a way to invade the egg (Vacquier *et al.*, 1993).

The similarities between plant and marine invertebrate mating systems suggests a second possible mechanism favoring diversity enhancing selection. The action of free floating marine gametes that after fertilization become larvae, is similar to plant gametes that after fertilization disperse as seeds. Both systems have mating interactions with the opposite sex that do not involve courtship yet rely on the interaction of specific gamete recognition molecules. In addition, both plant seeds and marine invertebrate larvae have only limited control over their ultimate dispersal. Possibly the analogy between plants and marine invertebrates continues such that marine invertebrates possess a prezygotic isolating mechanism similar to that of the S-allele system. However, it is important to keep in mind that although inbreeding depression could exist in marine invertebrate broadcast spawners, the generally longer dispersal distances of marine invertebrate larvae over plant seeds may reduce the probability of mating with a close relative and eliminate the need for a mechanism to favor outcrossing.

Despite considerable speculation about the existence of positive selection within a marine invertebrate species, to date there have been no published examples of such a selective force creating intraspecific sequence variation in marine organisms. Before it is possible to identify the mechanisms of diversity enhancing selection, it is necessary to demonstrate that sequence variation exists in these molecules and that the variation is the result of natural selection.

To test the hypothesis that positive selection in marine invertebrates creates an unusual pattern of sequence variation in gamete recognition molecules, we conducted a large-scale intraspecific examination of DNA sequence variation in the a 273 nucleotide region of bindin locus of the red sea urchin (*Strongylocentrotus franciscanus*). The region examined (Figure 5-1) is the variable region just 5' of the conserved core, corresponding to the site where Metz and Palumbi (1996) observed a "hotspot" of interspecific divergence in *Echinometra*. If prezygotic isolation occurs at the molecular level in the red urchin, it is likely to influence the nucleotide sequence of bindin. This "hot spot" is a logical place to look for variation in the bindin sequence.

The current study aims to determine the nature of sequence variation in the red urchin bindin locus. Intraspecific variation could result from either genetic drift (if the allele is neutral relative to fitness) or selection (if diversification is somehow beneficial). replacement substitutions are absent within a species, this would

suggest the action of purifying selection. Regardless of the result, each possibility has interesting implications relative to newly emerging ideas about the unique operation of marine invertebrate gamete interaction molecules.

## **Materials and Methods**

### ***Sample Collection***

Between August 1995 and February 1996, 298 adult *S. franciscanus*, approximately 50 animals from each site, were collected from the following six locations: Ketchikan, Alaska; the Port Townsend, Washington; Port Orford, Oregon; Ft. Bragg, California; Santa Barbara, California; and Ensenada, Baja California, Mexico. All animals were collected on SCUBA and then shipped via overnight express to our laboratory. All individuals came from as close to a 10 to 15 meters depth range as possible. To insure that all samples were adults, all individuals collected were at least 80 mm in test diameter.

### ***DNA Extraction***

DNA extractions were performed as described in Milligan (1992). Approximately 25 -100 ug of gonadal tissue was homogenized in 700 ul CTAB buffer prewarmed to 60°C (100 mM Tris-HCl, pH 8.0; 1.4 M NaCl; 20 mM EDTA; 2% hexadecyltrimethylammoniumbromide (CTAB, w/v)); 1% polyvinylpyrrolidone

(PVP-360, w/v); 0.2% 2-mercaptoethanol (v/v added just before use). At the end of a 30-60 minute incubation at 60°C (with periodic swirling) 700 ul of chloroform:isoamyl alcohol (24:1) was added followed by vortexing. Centrifugation for 10 minutes at 1500 x g separated the aqueous and the organic phases. One extraction with an equal volume of chloroform:isoamyl alcohol (24:1) was followed by extractions with phenol:chloroform:isoamyl until the interface was clear. The aqueous phase was transferred to a sterile tube followed by addition of 2/3 volume of ice-cold isopropanol and vortexing. The pelleted DNA was washed with 500 ul wash buffer (76% ethanol, 10 mM ammonium acetate), dried, and resuspended in 20-30 ul TE containing 1 ul RNase (10 mg/ml). It was imperative to conduct the extraction protocol on fresh ovary tissue. Ovary samples frozen for as short as one day showed reduced yield of DNA and reduced success in subsequent PCR amplification.

### ***PCR Amplification***

#### *Double-stranded PCR amplification*

The primers FNbindin 5' (5'-AGTCGACGTTTCGACAGACGAC-3') and FNbindin 3' (5'-TTACATGGTCCATTATAGTATGCC-3') amplify a 431 base pair region of the 5' end of the bindin gene. Amplification followed standard procedures (Saiki *et al.*, 1988) using a reaction volume of 25 ul and final magnesium chloride

concentration of 2 mM. The thermocycler (Perkin Elmer Cetus) profile for all double stranded reactions was: 1 cycle of 95°C, 5 min. followed by 30 cycles of 94°C, 1 min; 60°C, 1 min; 72°C, 2 min. Five ul of each PCR product was resolved on a 2% Nu Sieve agarose, low melting temperature TBE gel. Gel isolates were removed with sterile wide-bore, disposable polyethylene transfer pipettes and stored in a microfuge tube with 200 ul of water at -20 °C.

#### *Single-stranded PCR amplification*

Each gel isolate was heated to 65°C for approximately 5-10 minutes and used as template for the single-stranded PCR amplification. To amplify a single-stranded product of the 5' strand, the PCR amplification conditions were identical to those identified above for amplification of the double-stranded product except the 5' primer (FNbindin 5') was used at a final concentration of 0.5 uM.

The PCR amplification conditions to amplify the 3' single-stranded product required a lower annealing temperature (58°C) and lower MgCl<sub>2</sub><sup>+</sup> concentration (1.2 mM final). The concentration of the 3' limiting primer was 2.5 uM (final). Some samples required additional adjustments to annealing temperature (between 57-63°C) and limiting primer concentration (0.5-2.5 uM final concentrations).

#### *Direct sequencing of single-stranded PCR product*

The single-stranded PCR products were washed in Centricon filter units (30,000 MWCo) and resuspended in 7 ul H<sub>2</sub>O for Sanger dideoxy sequencing (Sequenase ver. 2.0, U.S. Biochemical) using internal sequencing primers KTseq 5' (5'-GGAGCGCGTAAGAAGCGTTAT-3') and KTseq 3' (5'-ATACACACGATGGTCAAG-3') at 10uM.

***Cloning of bindin DNA from heterozygous individuals***

In order to confirm the exact sequence of representatives of all alleles, the PCR products were cloned. Double-stranded PCR products were amplified using primers KBRS 5' (5'-CGCGGATCCAGTCGACGTTTCGACAGACGAC-3') and 3' (5'-GCCAAGCTTTTACATGGTCCATTATAGTATGCC-3') and the double-stranded PCR protocol. These primers incorporate the restriction sites BamHI and HindIII respectively on their 5' ends to facilitate directional cloning into the pBMKS bluescript vector. The PCR products were resolved on a 2% agarose gel and the excised gel fragment was purified using Quiaquick spin columns (Quiagen) and then digested with BamHI and HindIII. The gel-purified fragment was ligated into pBMKS and transformed into *E. coli* DH5 $\alpha$ . The DNA from a minimum of four transformants were sequenced for each PCR product. Several individuals with the same genotype, but from different geographic locations, were cloned and sequenced to verify sequence consistency among alleles found at different geographic

locations. Plasmid DNA was purified using the alkaline lysis method (Sambrook *et al.*, 1989). The double-stranded plasmid DNA was sequenced using Kgsseq 5' (5'GTTTCTGACG ATTCGGAAAGA-3') and Kgsseq 3' (5'-GAAACAACCAATTTAAAAATA-3') as internal sequencing primers.

Cloning of the PCR products also allowed nucleotide changes due to PCR incorporation error to be detected and eliminated. Cloned PCR products from four different reactions were sequenced on both strands. A substitution in a sequence was determined to be PCR incorporation error based on one or both of the following observations: (1) the nucleotide present in the sequence was not present in either the 5' or 3' strand produced by direct sequencing; or (2) the existence of this new sequence now created a total of three alleles for one genotype which is biologically impossible for a diploid organism. We found what appeared to be a substitution resulting from PCR incorporation error in approximately one out of 2,000 nucleotides. The sequence of a particular allele was considered to be an accurate representation of an individual genotype when it appeared in at least two out of three of the cloned sequences.

### ***Sequence and Statistical Analyses***

From the 431 base pair PCR product, we obtained unambiguous sequence information confirmed with both strands of DNA for a 273 base pair region.

Sequences were aligned using Seq-App ver 1.9a multiple sequence alignment program for the Macintosh (Gilbert, 1994). Molecular Evolutionary Genetic Analysis ver. 1.01 (MEGA, Kumar, *et al.*, 1993) was used to calculate nucleotide sequence divergence. Average proportions of replacement substitutions per replacement site (dn) and silent substitutions per silent site (ds) (Nei and Gojobori, 1986), with standard errors (Nei and Jin, 1989), were calculated using the MEGA program (Kumar, *et al.*, 1993). Corrections for multiple hits were not used because of the close relationship of the taxa (i.e. all data are alleles from the same species). A program available from T. Whittam (Penn. State University) was used to conduct the sliding window dn:ds analysis on the sequence data. All amino acids were classified into categories of charge and hydrophobicity according to Lewin (1994; and see Weir, 1996). The cladistic analysis was performed by PAUP 3.1.1 (Swofford, 1993). Search for the shortest tree was made by the exact branch-and-bound algorithm which is guaranteed to find all optimal trees. MacClade version 3.05 was used to graphically present the PAUP tree.

Individual genotypes were coded as paired alphabetical characters and analyzed with BIOSYS (Swofford, 1989) to obtain estimates of the following: allele frequencies, conformance to Hardy-Weinberg equilibrium, Wright's (1978) F-statistics, and Nei's (1972) minimum genetic distance in pairwise comparisons. Conformance to Hardy-Weinberg proportions were estimated in three ways: (1)

contingency chi-square analysis with Levene's (1949) correction for small sample size; (2) chi-square with pooling of rare and common categories; and (3) significance test with exact probabilities.

The *S. franciscanus* polymorphic data were compared to the number of fixed differences in four other species following the McDonald and Kreitman (1991) test for neutral evolution. cDNA sequence was available for the *bindin* locus of *S. purpuratus* (Gao *et al.*, 1986, GenBank Accession Number M14487), *S. droebachiensis* (C. Biermann and W. Eanes, State University of New York-Stony Brook, unpublished data), and *Lytechinus variegatus* (Minor *et al.*, 1991, GenBank Accession Number M59489). We counted replacement and silent sites as described by McDonald and Kreitman (1991) and used a G-test of independence with the Williams correction for continuity (McDonald and Kreitman, 1991; Sokal and Rohlf, 1981).

## Results

### *Nucleotide variation*

Figure 5-2 indicates the phylogenetic relationship of the 14 unique *S. franciscanus* bindin alleles identified by sequencing the 273 base pair region of the 5' bindin gene. The published sequence of the congener *S. purpuratus* bindin cDNA was used as an outgroup (Gao *et al.*, 1986). Superimposed on the tree are indications of the most likely point where specific nucleotide substitutions occurred. Based on the tree, replacement substitutions have occurred both in the distant evolutionary past and more recently and therefore are not limited to a specific time in evolutionary history.

To evaluate the degree to which purifying selection determines nucleotide variation in the region of the bindin locus, we examined interspecific amino acid conservation between three *Strongylocentrotus* congeners and a more distant relative *L. variegatus* (see Chapter 6). Figure 6-1 shows the amino acid alignment for the species (see Chapter 6). Although the nucleotide sequence may vary, there are 34 amino acid positions (12%) that are identical between all four species in the 273 nucleotide bindin region. The amino acid conservation across three urchin families suggests conclude that purifying selection constrains the amino acid

sequence of at least twelve percent of the bindin region examined. The asterisks above the codons in Figure 5-3 indicate how these 34 amino acid positions relate to intraspecific variation. One of the 34 conserved positions coincides with a replacement polymorphism in the bindin sequence and a second conserved position coincides with a silent polymorphism. It is difficult to draw an inference based on these this limited number of polymorphism.

There is a high degree of intraspecific polymorphism at the bindin locus as indicated by a total of 14 alleles and 21 genotypes in 134 individuals sequenced. A gene is considered polymorphic when the most common allele has a frequency less than 0.95 (Hartl, 1988, p. 11). The four most common alleles identified have frequencies of 0.51, 0.20, 0.16, and 0.08 respectively (see Chapter 4). There are a total of 13 variable positions in all 14 alleles (Table 4-1). All other positions of the 273 base pair region, in all 134 individuals sequenced, are identical (Table 4-1). The number of nucleotide differences between any two sequences is low and ranges between one (0.35%) and six (2.1%) nucleotide substitutions in the 273 base pair region. On average, there are 2.9 nucleotide positions that vary between any two unique alleles resulting in average nucleotide-sequence diversity of 1.06% (p-distance calculated with MEGA, Kumar *et al.*, 1993).

### ***Tests of neutral evolution***

#### *Creation of a theoretical data set.*

To assess if bindin's amino acid substitutions were more common than would be expected by chance, we constructed a theoretical data set that mimics the process of neutral evolution creating nucleotide variation at the 273 base pair region of the bindin locus (program available from S.R. Palumbi, Harvard University). We chose one of the 14 unique bindin alleles at random and allowed the program, "DNA Evolve," to create 13 new alleles, each different from the original by four nucleotides (as close an approximation as possible to the average number of nucleotide differences between all 14 observed bindin alleles). All new alleles with a stop codon were eliminated and replaced by an allele with an open reading frame. This new data set provided neutral expectations of random nucleotide substitutions. Using the same random starting sequence, we repeated this process ten times to obtain ten theoretical data sets and then repeated all data analyses described above (see Methods) on the ten randomized data sets in order to compare the results of observed bindin sequence variation to truly random, neutral sequence variation.

#### *Probability of replacement changes.*

It is possible for selection to act on a subset of positions in a gene. For example, the strong signal of positive selection seen in the human MHC was based on the 57 codons in the antigen recognition site (ARS) of the 398 total codons in the MHC molecule (Hughes and Nei, 1988). Subsequently, we explored the possibility that selection is acting on the region of the *bindin* gene examined here. The polymorphic sites in *bindin* appear to have a high proportion of replacement substitutions. Nine of the 13 polymorphic sites (69%) in the *bindin* region analyzed are replacement substitutions. If a nucleotide substitution results in an amino acid change, the new amino acid can have similar physiochemical properties (i.e. a conserved amino acid class change) or change in charge and/or polarity from the original amino acid (i.e. a radical change). Seven of nine (77%) replacement substitutions in our data set result in radical amino acid substitutions that change either charge or polarity. Radical amino acid substitutions are often considered to have a greater impact on protein function than conservative changes (Hughes *et al.*, 1990) If this is so, then the different *bindin* alleles could have slightly different properties.

As a comparison to the observed *bindin* data, we calculated the number and type of amino acid changes in the DNA Evolve data set. The ten DNA Evolve data sets (140 sequences total), had on average 74% replacement substitutions and 63% of these were radical amino acid changes. Those values are quite similar to the

observed values of 69% replacement and 77% radical changes in our observed bindin data set. The mean values for the simulated data set for the proportion of replacement substitutions and the proportion of radical replacement substitutions are respectively within 1.19 and 0.61 standard deviations away from the observed values. This result suggests there is no statistically significant difference between the observed and theoretical data sets.

The previous analysis estimated the proportions of radical and conservative replacement substitutions both observed and expected over the entire 91 codon bindin region examined. We also estimated the probability of obtaining radical replacement substitutions by examining potential mutation at only the nine codons that have replacement substitutions. To evaluate whether the proportion of radical amino acid substitutions was different from the proportion expected by chance, we calculated for each polymorphic site the probability that the resulting new amino acid would be in a different class. In the observed bindin data set, each polymorphic site had only one variable nucleotide position. We calculated the probability that, given that a nucleotide is going to change and result in a new amino acid, this new amino acid will be a radical replacement change. For example, nucleotide position 1,028 is polymorphic (see Figure 5-3 and Table 4-1). Some alleles contain the sequence GGG (Gly) while others are AGG (Arg). This is a radical replacement substitution (Gly is neutral hydrophobic and Arg is basic). Yet,

any other nucleotide substitution (e.g. T or a C resulting in Trp and Arg respectively) would also generate radical replacement substitutions. Therefore if nucleotide position 1,028 varies, the probability of a radical substitution at this codon is 1.0, and the probability of a conservative substitution is 0. We used sequence data of *S. purpuratus* to determine the ancestral state of the sequence and thus the direction of the mutation in the *S. franciscanus* polymorphism (see PAUP tree, Figure 5-2). We then averaged the probabilities over the nine replacement polymorphic positions. In the bindin data, the average probability of a replacement change being radical is 0.85, compared to the observed proportion of radical replacement substitutions of 0.77. The observed value is within 0.68 standard deviations from the calculated probability for an amino acid substitution being radical, suggesting it is more likely than not for nucleotide substitutions at these polymorphic positions to result in a radical amino acid substitution. It is however, important to remember that the power of the statistic is low because of the small number of polymorphic positions evaluated.

#### *Dn:Ds Analysis.*

There is some debate as to which method is the most effective test of diversity enhancing selection (see discussion in Li and Graur, 1991; Whittam and Nei, 1991; McDonald and Kreitman, 1991; Kreitman and Ohta, 1996). The most commonly

used method was first introduced by Hughes and Nei (1988) in their evaluation of selection in the MHC immune recognition molecules. This statistic evaluates the ratio of nonsynonymous nucleotide substitutions per potential nonsynonymous site (dn), to synonymous substitutions per synonymous site (ds) (Nei and Gojobori, 1986). A dn:ds ratio that is significantly different from 1.0 indicates the action of selection. A value significantly greater than 1.0 signifies positive selection for diversity between alleles, and a value significantly less than 1.0 indicates purifying selection, which results in conserved amino acid sequence between alleles despite silent nucleotide substitutions.

In *S. franciscanus*, the values of dn:ds for each geographic location range from 0.64-1.45, when evaluating just the unique binding alleles in a specific population (Table 5-1). Except for Alaska, all geographic locations have dn:ds values greater than 1.0. The overall dn:ds for the 14 unique alleles present in the species is 0.82. All measures of dn:ds are statistically indistinguishable from 1.0 ( $p > 0.05$ ). A sliding window analysis (e.g. Metz and Palumbi, 1996; Ina, 1996) of a variety of window sizes (20,25,30,40 codons) did not reveal any regions containing dn:ds ratios significantly different than 1.0. Analysis of the theoretical DNA Evolve data set displays similar results. For the DNA Evolve data, the range of dn:ds values for the unique alleles in each population is slightly higher than the observed binding data at 0.69-1.58 and likewise for all alleles in each population (dn:ds ranges from 1.12-

1.71). The dn:ds value for the 14 unique DNA Evolve alleles (averaged over ten replicate data sets) is 0.90, a value that is close to the average (0.82) of our empirical bindin data set.

*McDonald-Kreitman test for neutral evolution.*

The second method widely used to evaluate selection is the McDonald and Kreitman (1991) test of neutral evolution. According to McDonald and Kreitman (1991), “if the observed substitutions are neutral, the ratio of replacement to synonymous fixed differences between species should be the same as the ratio of replacement to synonymous polymorphisms within a species.” An unequal ratio indicates the action of selection at the loci examined and supports rejection of the null hypothesis of neutral evolution.

Nachman *et al.* (1994) used the McDonald-Kreitman test to evaluate sequence variation within and between three species of mice. A significant difference in the ratio of the number of replacement to silent polymorphisms (i.e. 11:13 = 0.84) compared to the ratio of the number of replacement to silent fixed differences (i.e. 2:23 = 0.08) suggested an excess of replacement polymorphisms and a rejection of the neutral model of evolution. We a similar approach here to determine whether

the apparent excess of replacement polymorphisms in *S. franciscanus* is greater than neutral expectations.

The McDonald-Kreitman method is a more powerful test than the dn:ds method because purifying selection in portions of a molecule will not dampen the signal of directional selection (see debate in Li and Graur, 1991; Whittam and Nei, 1991; McDonald and Kreitman, 1991, Kreitman and Ohta, 1996). In order to enumerate fixed differences between species, bindin cDNA sequences from *S. purpuratus* (Gao *et al.*, 1986), *S. droebachiensis* (W. Eanes and C. Biermann, State University of New York, Stony Brook, unpublished data ), and *Lytechinus variegatus* (Minor *et al.*, 1991) were used. For all four species, we used the McDonald-Kreitman test to evaluate selection in a 114 base-pair region of bindin, a region where alignment was unambiguous. This region initiates at the beginning of mature bindin and corresponds to base pairs 944 to 1061 in Minor *et al.* (1991) (see Figure 5-1, Figure 5-3 and Table 6-2 and Figure 6-2 for interspecific alignment). These DNA sequence alignments represent at least one-third of the entire region examined within the *S. franciscanus* bindin gene.

In all comparisons, the ratio of replacement to silent differences in fixed versus polymorphic sites are not equal and differ by a factor of three to five. Initially this

suggests a proclivity for replacement polymorphic substitutions; however, none of the comparisons are statistically significant (see Table 5-2). In fact, for the comparison of the ratio of replacement to silent differences *S. franciscanus* to *S. purpuratus* to become significant, there would need to be an additional six (14 total) replacement polymorphisms (nearly twice as many as currently exist).

The McDonald-Kreitman test using the 14 unique alleles created by DNA Evolve as the source of *S. franciscanus* polymorphic data produces ratios that are similar to the observed bindin data (see Table 5-2). In addition to comparing replacement and silent substitutions in *S. franciscanus* to three other species, Table 5-2 also includes data comparing these three species to two types of theoretical data sets produced by DNA Evolve. Previously, all mentions of DNA Evolve referred to unique bindin alleles mutated to be approximately 2% divergent from each other. We also performed the McDonald-Kreitman test using “DNA Evolve 1%” as the source of *S. franciscanus* polymorphic data. “DNA Evolve 1%” represents random mutations to a random *S. franciscanus* bindin allele resulting in 14 new alleles, each approximately 1% divergent from each other. The result is a data set with the same number of polymorphic positions (13) as the *S. franciscanus* data (yet the mean sequence diversity for all pairwise comparisons, p-distance, is lower than observed

in bindin). The results of the McDonald and Kreitman (1991) test using the “DNA Evolve 1%” data set are nearly identical to those for *S. franciscanus* bindin (see Table 5-2). For example, the ratio of replacement to silent polymorphisms is nearly three times as great as replacement to silent fixed differences. Table 5-2 also includes results of the McDonald-Kreitman test using “DNA Evolve 2%,” which has a similar p-distance, yet a greater number of polymorphic positions than the observed bindin data. The results of the McDonald-Kreitman test using “DNA Evolve 2%” are slightly different than the observed bindin data (Table 5-2). Both “DNA Evolve 1%” and “DNA Evolve 2%” have unequal ratios of replacement to silent polymorphisms compared to the ratio of replacement to silent fixed differences. Yet the difference in the ratio with “DNA Evolve 1%” is 1.5-fold as opposed to 3- to 5-fold in the “DNA Evolve 2%” comparisons. It appears that the very low numbers of nucleotide substitutions in “DNA Evolve 1%” may create an artifact in the results due to small sample size. Thus, in all the comparisons, although the *S. franciscanus* data suggest a trend towards replacement polymorphisms, this trend is not statistically significant or different than randomized, neutrally evolved data sets.



## Discussion

Our results indicate that the 273 base pair region of the *S. franciscanus* bindin locus analyzed in this study is not subject to positive selection. Sequence variation in the locus appears to be a combination of purifying selection and neutral evolution. Purifying selection is indicated by interspecific comparisons at the same bindin locus in four urchin species. Neutral evolution is indicated by  $d_n:d_s$  and McDonald-Kreitman statistical tests as well as comparisons of the observed data to a neutrally evolved, simulated data set.

Under purifying selection, mutations resulting in amino acid substitutions occur periodically but are selected out of the population because of functional inferiority (Nei, 1987). Silent substitutions have no impact on fitness are ultimately fixed or lost through drift. Based on interspecific comparison of amino acid conservation, purifying selection seems to be constraining replacement substitutions in approximately twelve percent of the bindin region. Purifying selection has been shown to exist in other Strongylocentrotids such as the mitochondrial gene cytochrome oxidase I (CO1; Edmands *et al.*, 1996; Palumbi and Kessing, 1991). In a stretch of sequence approximately equal in length to that studied in bindin (305 bp), CO1 in the congener *S. purpuratus* has three times as many polymorphic sites

as does the bindin sequence (Edmands *et al.*, 1996). There are also ten times as many silent substitutions in CO1 (42 silent substitutions out of 42 polymorphic sites) as compared to *S. franciscanus* bindin (4 out of 13 polymorphic sites).

An accelerated mutation rate in CO1 is one possible explanation for an increase in the number of polymorphic silent sites in CO1 over bindin. It is possible that mutation rates vary between the two molecules and as a result the rate of silent substitutions is higher in CO1 than in bindin. Mutation is known to be accelerated in mitochondrial DNA (mtDNA) genes such as CO1 compared to single copy nuclear (scn) genes (Hartl and Clark, 1989). The accelerated mutation in mtDNA is the result of less efficient proofreading during DNA replication (Hartl and Clark, 1989). Additionally, maternal inheritance and the fact that mtDNA is haploid reduces the effective population size of mtDNA over scnDNA by one-quarter and subsequently increases the time to fixation for non-deleterious substitutions (Hartl and Clark, 1989). These differences between mtDNA and scnDNA may explain why there is a difference in the proportion of silent substitutions between bindin and CO1. However, there are no amino acid substitutions in the 300 base pair region of CO1 (Edmands *et al.*, 1996), yet nine out of 13 (69%) of the bindin polymorphic sites are

amino acid substitutions. Clearly purifying selection is not the only force operating on the *bindin* locus.

It is possible that weak positive selection does influence the *bindin* locus, but that the statistical tests used are not sensitive enough to detect the selection. In the case of the *dn:ds* analysis, purifying selection that creates conserved regions can dampen an otherwise detectable signal of positive selection. Additionally, the small number of polymorphic sites results in high variances, particularly for calculations of *ds*. In many cases, the standard error of *ds* is equal to its calculated value (Table 5-1). Thus, it is possible that selection operates on the *bindin* gene but the *dn:ds* test is too weak to detect selection in this data set. The McDonald-Kreitman test is considered to be a more sensitive test of selection compared to the *dn:ds* method (Kreitman and Akashi, 1995). However, again it is possible that selection exists but the number of polymorphic positions in the *bindin* region examined is too low such that the McDonald-Kreitman test falsely concludes neutral evolution. In addition, the McDonald-Kreitman test as used here with polymorphic data from one species is more conservative than originally described (McDonald and Kreitman, 1991). If the test had included information about polymorphic substitutions in the second species, *S. purpuratus*, the number of polymorphic replacement substitutions would most

likely be greater than observed. Yet, there is no way to know how much greater that number would be and if it would change the statistical outcome of the McDonald-Kreitman test. However, it is also likely that as the number of replacement polymorphisms increases, the overall count for silent polymorphisms would also increase. Therefore, it remains a formal, though unlikely possibility that inclusion of intraspecific sequence variation in *S. purpuratus* would result in a rejection of the null hypothesis of neutral evolution.

In addition to the statistical tests used, other measures of selection such as the proportion of replacement to silent substitutions as well as the probability values for amino acid class changes also suggest that changes in the *bindin* gene are a result of neutral evolution. Finally, it is a very compelling argument that all analyses performed on the DNA Evolve data created to mimic *bindin* variation resulting from neutral evolution produces nearly identical results as the observed *bindin* data. The DNA Evolve data set also shows that, just as is observed in the empirical data, random mutations will result in a high proportion of replacement changes and a majority of these changes will be radical amino acid substitutions. Thus all tests regardless of statistical power are consistent with the conclusion that purifying

selection constrains approximately twelve percent of the bindin sequence as that neutral evolution results in a small number of polymorphic positions.

Neutral evolution would not necessarily predict the relatively high number of alleles observed in at the bindin locus. Following neutral mutation, new polymorphisms should either go to fixation or be removed from the population by genetic drift (Nei, 1987). Yet, as presented by Ohta (1992), population size can influence the action of drift. In large populations, it will be a long time before the neutral mutations are either fixed or lost through genetic drift. As very fecund broadcast spawners, urchins are known to have large populations (Morris *et al.*, 1980). For *S. franciscanus* in particular, the entire Pacific coast of North America is conceivably one extended inter-breeding population as a result of little genetic subdivision along the linear coastline and very high gene flow (see Chapter 4).

Including this data set, to date there is no example of diversity enhancing selection operating intraspecifically in a marine invertebrate gamete interaction molecule. Sequence data within each *Echinometra* species (Metz and Palumbi, 1996) are similar to that observed for *S. franciscanus* bindin. For example, *E. mathaei* has 17 polymorphic sites in 252 base pairs compared to 13 out of 273 in *S. franciscanus*. Like *S. franciscanus*, a high proportion (76%) of these sites are

replacement substitutions. These numbers do not include insertions and deletions (indels) which are prevalent in *Echinometra sp.* bindin sequences, both within and between species. The bindin sequences in both genera have comparable dn:ds values that are not significantly different from 1.0. Finally, the McDonald-Kreitman analysis on both genera revealed non-significant results that initially appear to have an excess of replacement polymorphisms (Metz and Palumbi, 1996).

There is however, a major discrepancy between the two genera in that bindin sequences in *Echinometra sp.* have insertions and deletions (indels) both within and between species that range from one to ten codons in length. There are no indels in *S. franciscanus* bindin sequences aside from point substitutions. The function, if any, of these indels is unknown, but it does suggest that compared to *S. franciscanus*, *Echinometra sp.* bindin can tolerate relatively large alterations to sequence structure conceivably without impairment to function, a feature that would be consistent with the neutral theory of evolution.

All of the evidence presented here suggests that portions of bindin sequence variation are determined by a combination of purifying selection and neutral mutation. Although diversity enhancing selection is not operating, the picture of neutral evolution seen here is still markedly different from that seen in most

proteins. Most genes are highly conserved such that only silent substitutions are observed within a species. Yet for bindin, the majority of substitutions are replacement substitutions. Neutral evolution present in a gamete interaction molecule is contrary to initial predictions for a trait closely related to fitness. The explanation for this contradiction is not clear.

If bindin is evolving according to a neutral model, it would suggest that this 91 codon region of bindin is non-functional. For several reasons, it is difficult to accept that this bindin region has absolutely no function. The fact that 12% of this region is conserved between four species suggests there is a functional need to maintain these amino acids. In addition, there is clearly a cost to maintaining this region of bindin for example the cost of potential reduction in an individual's fitness resulting from deleterious mutations that impair bindin's ability to successfully bind to the receptor. If there is a cost associated with the variation in this region of bindin, it is reasonable to assume there is a concomitant benefit to maintain this region.

Finally, there has been a great deal of biochemical work demonstrating the species specific nature of sea urchin bindin (reviewed by Vacquier *et al.*, 1995). Species specific recognition is not likely to be conveyed through the conserved core.

It is more probable that the variable 5' (N-terminal) or 3' (C-terminal) region is responsible for species specificity. In fact, Lopez *et al.* (1993) identified that *either* the 5' *or* the 3' region was required for species specific agglutination of urchin eggs. To make more definitive correlations of structure and function it will be necessary to determine the three dimensional conformation of bindin and identify the active site that binds to the egg receptor. Although it is possible that all species specificity is conferred only through the C-terminal region, the data presented in Lopez *et al.* (1993) imply a functional role for the N-terminal region. If so, the signal of neutral evolution is perhaps incorrect. Although it seems unlikely, it is possible that simultaneous action of positive and purifying selection combine to appear as neutral evolution.

Although these ideas are merely conjecture, it is important to continue investigating the possible functions of unusual variation in a marine invertebrate gamete interaction molecule. Bindin is different than many other molecules in that bindin evolution is coupled with a ligand expressed on another individual. Any changes in bindin sequence must be tolerated by any protein on the egg surface interacting with bindin. This interaction is dynamic, with both bindin and the receptor co-evolving. Therefore, these polymorphic positions represent the few

sites where variation is tolerated in both bindin and the receptor. Perhaps these positions could be viewed as highly informative regarding the non-functional regions of bindin or perhaps there is an even more interesting, yet currently not understood mechanism operating to maintain polymorphisms in a gamete interaction molecule.

### References

- Burton, R. S., and Feldman, M. W. (1982). Population genetics of coastal and estuarine invertebrates: does larval behavior influence population structure? In *Estuarine Comparisons*, V. S. Kennedy, ed. New York: Academic Press, pp. 537-551.
- Dobzhansky, T. (1937). *Genetics and the origin of species* New York: Columbia University Press.
- Edmands, S., Moberg, P. E., and Burton, R. S. (1996). Allozyme and mitochondrial DNA evidence of population subdivision in the purple sea urchin *Strongylocentrotus purpuratus*. *Marine Biology* 126:(3) 443-450.
- Foltz, K. R., and Lennarz, W. J. (1992). Identification of the sea urchin egg receptor for sperm using an antiserum raised against a fragment of its extracellular domain. *The Journal of Cell Biology* 116:(3) 647-658.
- Foltz, K. R., and Lennarz, W. J. (1993). The molecular basis of sea urchin gamete interactions at the egg plasma membrane. *Dev. Biol.* 158: 46-61.
- Gao, B., Klein, L. E., Britten, R. J., and Davidson, E. H. (1986). Sequence of mRNA coding for bindin, a species-specific sea urchin sperm protein required for fertilization. *Proc. Natl. Acad. Sci. USA* 83: 8634-8638.
- Gilbert, D. (1994). SeqApp. Multiple sequence alignment program (1.9a157+). Indiana State University.
- Hall, T. J., Grula, J. W., Davidson, E. H., and Britten, R. J. (1980). Evolution of sea urchin nonrepetitive DNA. *J. Mol. Evol.* 16: 95-110.

Haring, V., Gray, J. E., McClure, B. A., Anderson, M. A., and Clarke, A. E. (1990). Self-incompatibility: a self-recognition system in plants. *Science* 250:(16 November) 937-941.

Hartl, D. L., and Clark, A. G. (1989). *Principles of population genetics*, Second Edition Sunderland, Mass.: Sinauer Assoc.

Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:(8 September) 167-168.

Hughes, A. L., Ota, T., and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of Class I Major-Histocompatibility-Complex molecules. *Mol. Biol. Evol.* 7:(6) 515-524.

Ina, Y. (1996). Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. *J. Genet.* 75: 91-115.

Kreitman, M. (1996). The neutral theory is dead. Long live the neutral theory. *BioEssays* 18:(8) 678-683.

Kreitman, M., and Akashi, H. (1995). Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* 26: 403-422.

Kumar, S., Tamura, K., and Nei, M. (1993). MEGA: Molecular Evolutionary Genetics Analysis (version 1.01). University Park, PA: The Pennsylvania State University.

Lee, Y.-H., Ota, T., and Vacquier, V. D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 12:(2) 231-238.

Levene, H. (1949). On a matching problem arising in genetics. *Ann. Math. Stat.* 20: 91-94.

- Levin, D. A. (1984). Inbreeding depression and proximity-dependent crossing success in *Phlox drummondii*. *Evolution* 38:(1) 116-127.
- Lewin, B. (1994). *Genes* V New York: Wiley.
- Li, W.-H., and Graur, D. (1991). *Fundamentals of molecular evolution* Sunderland, MA: Sinauer Assoc.
- Lopez, A., Miraglia, S. J., and Glabe, C. G. (1993). Structure/function analysis of the sea urchin sperm adhesive protein bindin. *Dev. Biol.* 156: 24-33.
- Mayr, E. (1954). Geographic speciation in tropical echinoids. *Evolution* 8:(1) 1-18.
- McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:(20 June) 652-654.
- Metz, E. C., and Palumbi, S. R. (1996). Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* 13:(2) 397-406.
- Milligan, B. G. (1992). Plant DNA Isolation. In *Molecular genetic analysis of populations*, A. R. Hoelzel, ed. Oxford: IRL Press, pp. 71-74.
- Minor, J. E., Britten, R. J., and Davidson, E. H. (1993). Species-specific inhibition of fertilization by a peptide derived from the sperm protein bindin. *Mol. Bio. Cell* 4: 375-387.
- Minor, J. E., Fromson, D. R., Britten, R. J., and Davidson, E. H. (1991). Comparison of the bindin proteins of *Strongylocentrotus franciscanus*, *S. purpuratus*, and *Lytechinus variegatus*: Sequences involved in the species specificity of fertilization. *Mol. Biol. Evol.* 8:(6) 781-795.
- Morris, R. H., Abbott, D. P., and Haderlie, E. C. (1980). *Intertidal invertebrates of California* Stanford, CA: Stanford University Press,.

- Nei, M. (1972). Genetic distance between populations. *American Naturalist* 106: 283-292.
- Nei, M. (1987). *Molecular evolutionary genetics* New York: Columbia University Press.
- Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:(5) 418-26.
- Nei, M., and Jin, L. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* 6: 290-300.
- Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23: 263-286.
- Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annu. Rev. Ecol. Syst.* 25: 547-572.
- Palumbi, S. R. (1992). Marine speciation on a small planet. *TREE* 7:(4) 114-117.
- Palumbi, S. R., and Kessing, B. D. (1991). Population biology of the trans-Arctic exchange: mtDNA sequence similarity between Pacific and Atlantic sea urchins. *Evolution* 45:(8) 1790-1805.
- Palumbi, S. R., and Metz, E. C. (1991). Strong reproductive isolation between closely related tropical sea urchins (genus *Echinometra*). *Mol. Biol. Evol.* 8:(2) 227-239.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular cloning: a laboratory manual*: Cold Springs Harbor Laboratory Press.

- Scheltema, R. S. (1971). Larval dispersal as a means of genetic exchange between geographically separated populations of shallow-water benthic marine gastropods. *Biol. Bull.* 140: 284-322.
- Sokal, R. R., and Rohlf, F. J. (1981). *Biometry: The principles and practice of statistics in biological research*, 2nd Edition New York: W.H. Freeman and Co.
- Strathmann, R. R. (1978). Length of pelagic period in echinoderms with feeding larvae from the Northeast Pacific. *J. Exp. Mar. Biol. Ecol.* 34: 23-27.
- Swofford, D. L. (1989). BIOSYS-1: A computer program for the analysis of allelic variation in population genetics and biochemical systematics (Version 1.7). Champaign, IL: Swofford, D.L. at University of Illinois at Urbana.
- Swofford, D. L., and Begle, D. P. (1993). PAUP Phylogenetic analysis using parsimony (3.1). Laboratory of Molecular Systematics, Smithsonian Institution.
- Thorson, G. (1961). Length of pelagic life in marine invertebrates as related to larval transport by ocean currents. In *Oceanography*, edited by M. Sears, AAAS Washington, DC 455-474.
- Vacquier, V. D., and Lee, Y.-H. (1993). Abalone sperm lysin: unusual mode of evolution of a gamete recognition protein. *Zygote* 1:(August) 181-196.
- Vacquier, V. D., and Moy, G. W. (1977). Isolation of bindin: The protein responsible for adhesion of sperm to sea urchin eggs. *PNAS, USA* 74: 2456-2460.
- Vacquier, V. D., Swanson, W. J., and Hellberg, M. E. (1995). What have we learned about sea urchin sperm bindin? *Develop., Growth and Differ.* 37: 1-10.
- Vacquier, V. D., Swanson, W. J., and Lee, Y.-H. (1997). Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure for their divergence. *J. Mol. Evol.* 44 (*Suppl. 1*): S15-S22.

Weir, B. S. (1996). *Genetic data analysis II: Methods for discrete population genetic data* Sunderland, MA: Sinauer Assoc. Inc.

Wright, S. (1978). *Evolution and the genetics of populations. Vol 4: Variability within and among natural populations* Chicago: University of Chicago Press.

## Chapter 6

### **Interspecific Sequence Variation in Four Species of Sea Urchins at the Bindin Locus as an Indicator of Directional Selection in a Marine Invertebrate Gamete Interaction Molecule**

#### **Introduction**

In the process of species divergence, random nucleotide point substitutions can occur as replacement (nonsynonymous) substitutions that result in a new amino acid sequence, or as silent (synonymous) substitutions that do not alter the amino acid sequence. If there is no selection operating (neither directional selection, purifying selection, balancing selection, etc.) there will be many more replacement than silent substitutions at the ratio of 3:1. This ratio favoring amino acid change is a result of the genetic code. All of the nucleotide substitutions at the 2nd position of a codon and nearly all (96%) of the substitutions at the first position, result in an amino acid substitution. Even 30% of the substitutions that occur at the third position will result in a new amino acid (Li and Graur, 1991, p. 14). Yet, with a large majority of proteins examined to date, it is more common to see a large proportion of silent

substitutions as compared to replacement changes (Kimura, 1983; Kreitman and Akashi, 1995; Kumar *et al.*, 1993). This is because nearly all silent changes that occur will confer neither a benefit nor a disadvantage and therefore they are maintained in the population until they are lost or fixed through genetic drift. Although replacement mutations occur at a greater rate, replacement substitutions are often functionally deleterious and therefore do not persist in the population at the same rate that they occur through mutation. If a replacement change confers a benefit, it will be retained and spread through the subpopulation. If the subpopulations are subject to different selective pressures over time, the result of this process is two genes with divergent amino acid sequences and most likely a high proportion of concomitant silent substitutions that hitchhiked along with the beneficial selective sweeps.

The above scenario describes the standard expectations for divergence of genes in different species. For example, Messier and Stewart (1997) examined sequence divergence in the lysozyme gene for 21 different primate species spanning three families, the colobines, the cercopithecines, and the homonoids. In general, the rate of divergence at this locus does not appear to be substantial. Within-group interspecific comparisons show low ratios of nonsynonymous to synonymous

substitutions. According to prediction of the neutral theory of evolution (Kimura, 1983), the proportion of nonsynonymous substitutions per nonsynonymous site ( $d_n$ ) should equal the proportion of synonymous substitutions per expected synonymous site ( $d_s$ ) (Nei and Gojobori, 1986). When  $d_n$  is significantly greater than  $d_s$ , the action of divergent evolution is suggested in comparisons between species and diversity enhancing selection is suggested from comparisons within species. Values of  $d_n:d_s$  significantly less than 1.0 indicate the action of purifying selection eliminating replacement substitutions within a species. In the example of lysozyme, pairwise comparison of  $d_n:d_s$  for the cercopithecines is 0.38 which indicates the absence of directional selection because it is lower than the neutral expectation of 1.0. However, at two points since divergence from an ancestral hominoid lineage, there appears to be instances of accelerated divergence (Messier and Stewart, 1997). This is indicated by significantly high ( $> 5.0$ )  $d_n:d_s$  ratios, suggesting directional selection for rapid divergence at this locus.

In contrast to lysozyme, which shows limited spurts of rapid divergence, for some molecules, directional selection appears to dominate the interspecific divergence. For example, lysin is an abalone sperm molecule that is necessary for penetration of the sperm through the egg vitelline layer. This gamete interaction

molecule has one of the highest measures for interspecific sequence divergence yet reported (Lee *et al.*, 1995). In a statistical analysis of the entire lysin molecule, six out of 190 pairwise comparisons of worldwide abalone species show dn:ds values significantly greater than neutral evolution predictions. Out of 190 comparisons, 29 (15%) had dn:ds values greater than 1.0. In addition, there is a strong trend of decreasing rate of divergence with increasing evolutionary distance. This trend suggests that directional selection for accumulating differences is greatest when species are closely related (perhaps due to character displacement). Alternatively replacement substitutions could reach a saturation point and the proportion of silent substitutions increases.

The apparent rapid divergence of lysin and the nature of external fertilization in marine invertebrates prompted a search for examples of selection operating to increase differences in other marine invertebrate fertilization molecules (Palumbi 1992; Metz and Palumbi, 1996; Vacquier *et al.*, 1995). This type of selection is sometimes called positive selection (Hughes and Nei, 1988; Vacquier *et al.*, 1997; Metz and Palumbi, 1996; Lee *et al.*, 1995). Many marine invertebrates shed their gametes directly into the water column and thus all components of species recognition during reproduction are controlled by the egg-sperm interaction.

Therefore rapid divergence in just the gamete recognition molecules could result in the formation of new species. To evaluate this hypothesis, Metz and Palumbi (1996) examined interspecific sequence divergence in *bindin*, a sea urchin sperm acrosomal protein. *Bindin* is responsible for species-specific binding of sperm to a receptor on the sea urchin egg surface (Vacquier *et al.*, 1995). Metz and Palumbi (1996) concluded that a 39 codon region of the *bindin* molecule was subject to positive selection when *dn:ds* values for the three *Echinometra* species were significantly greater than 1.0 based on a one-tailed t-test ( $t=1.7$ ,  $p<0.05$ ). These three *Echinometra* species are widely distributed throughout the tropical Pacific. All three *Echinometra* species exist sympatrically as well as allopatrically depending on the geographic location (Metz *et al.*, 1994; Kelso, 1970).

To evaluate the ability to generalize the existence of directional selection in marine invertebrate fertilization proteins, we examined deviations from neutral evolution in the *bindin* locus for a group of sea urchins with slightly different characteristics from *Echinometra*. Two of the four species examined here, *Strongylocentrotus franciscanus* and *S. purpuratus*, share a common biogeography yet in contrast to *Echinometra*, their distribution is along a continuous, linear coastline. Also in contrast to *Echinometra*, which is such a closely related genus to

*Strongylocentrotus* in that there are fewer than 2 million years since their divergence (Palumbi and Metz, 1991), *S. franciscanus* and *S. purpuratus* are among the most divergent sea urchin congeners known with 15-20 million years since separation (Hall *et al.*, 1980). The other two species included in the bindin analysis (*Strongylocentrotus droebachiensis* and *Lytechinus variegatus*) do not share a common nor continuous biogeography, yet their shared evolutionary history facilitates sequence alignments and conjectures of protein evolution. The time since divergence for *S. purpuratus* and *L. variegatus* is approximately 30-40 million years (Minor *et al.*, 1991; Smith, 1984). The analysis here seeks to determine whether the bindin locus (excluding the variable region downstream of the conserved core) in these four species exhibits unusually high sequence divergence throughout the entire molecule such as abalone lysin or a portion of the molecule such as *Echinometra* bindin.

### **Materials and Methods**

We examined sequence divergence in two regions of the bindin locus in four species: *S. franciscanus*, *S. purpuratus*, *S. droebachiensis*, and *L. variegatus*. Published cDNA sequences of *S. franciscanus* and *L. variegatus* reported in Minor *et al.* (1991) were used, as well as our own *S. franciscanus* analysis of 268 alleles in

a 91 codon, 5' region of bindin (see Chapter 5). Gao *et al.* (1986) was the source of *S. purpuratus* cDNA bindin sequence and C. Biermann and W. Eanes (State University of New York, Stony Brook) provided unpublished sequence data for *S. droebachiensis*. Alignment of the four bindin sequences was by eye using Seq-App ver 1.98 (Gilbert, 1994) and after Vacquier *et al.* (1995).

We performed our analysis on two regions of bindin. The region designated "bindin 5" spans the beginning of mature bindin to the beginning of the "conserved core" (amino acid sequence TTISA in all four species.) The region examined and designated as the conserved core begins at the end of the 5' bindin region and ends approximately with the sequence MQEEEEEEE in all four species. The variable region (approximately 90 codons) downstream of the conserved core was not included our analysis.

Molecular Evolutionary Genetic Analysis ver. 1.01 (MEGA, Kumar, *et al.*, 1993) was used to calculate nucleotide sequence divergence. Average proportions of replacement substitutions per replacement site (dn) and silent substitutions per silent site (ds) (Nei and Gojobori, 1986), with standard errors (Nei and Jin, 1989), were calculated using the MEGA program (Kumar, *et al.*, 1993). Corrections for multiple hits were calculated using the Jukes-Cantor correction (Jukes and Cantor, 1969). A program available from T. Whittam (Penn. State University) was used to conduct sliding window analyses on the sequence data.

The *S. franciscanus* polymorphic data from our intraspecific analysis (see Chapter 5) was used to compare the number of fixed differences in four other species following the McDonald and Kreitman (1991) test for neutral evolution. We counted replacement and silent sites as described by McDonald and Kreitman (1991) and used a G-test of independence with the Williams correction for continuity (McDonald and Kreitman, 1991; Sokal and Rohlf, 1981).

## **Results**

### ***Sequence Diversity***

Sequence diversity for all four species (*S. franciscanus*, *S. purpuratus*, *S. droebachiensis*, and *L. variegatus*) is moderate in the 5' binding sequences. For the 273 base-pair 5' region of binding, p-distance (calculated with MEGA, Kumar *et al.*, 1993) ranges from 0.07 for the most closely related species pair (*S. purpuratus* and *S. droebachiensis*) to 0.51 for the most divergent pair of *S. droebachiensis* and *L. variegatus*. The p-distance for the other species pairs are 0.27 for *S. franciscanus* and *S. purpuratus*, 0.25 for *S. franciscanus* and *S. droebachiensis*, and 0.50 for *S. franciscanus* and *L. variegatus*, and 0.50 for *S. purpuratus* and *L. variegatus*. The p-values for binding's conserved core are dramatically lower in accordance with the designation of genetic conservation. For a 210 base pair region between base pair

1217 and 1427 (numbers i.e. Minor *et al.*, 1991), there are the following pairwise p-values: *S. franciscanus* and *S. purpuratus*  $p = 0.04$ , of *S. franciscanus* and *L. variegatus*.  $p = 0.16$ , *S. purpuratus* and *L. variegatus purpuratus*  $p = 0.17$ , *S. droebachiensis* and of *S. franciscanus*  $p = 0.03$ , *S. purpuratus* and *S. droebachiensis*  $p = 0.02$ , *S. droebachiensis* and *L. variegatus*  $p = 0.17$ .

In *S. franciscanus* a 1,300 base-pair intron marks the separation between the end of the 5' flanking variable region and the conserved core (see Chapter 5). Palumbi and Metz (1996) also report an intron of varying length at this same position in three species of *Echinometra*. Although the published work is based on cDNA sequence, we hypothesize that an intron separates the variable and conserved region in the more closely related species *S. purpuratus*, *S. droebachiensis*, and *L. variegatus* as well.

***dn:ds Analysis.***

We used a dn:ds analysis (i.e. Hughes and Nei, 1988) to evaluate sequence divergence (Table 6-1). The dn:ds statistic evaluates the ratio of nonsynonymous nucleotide substitutions per potential nonsynonymous site (dn), to synonymous substitutions per synonymous site (ds) (Nei and Gojobori, 1986). A dn:ds ratio that is significantly different from 1.0 indicates the action of selection. A value greater than 1.0 signifies selection creating divergence between species. A value significantly less than 1.0 indicates purifying selection acting within species, which results in conserved amino acid sequence between species despite silent nucleotide substitutions (Nei, 1987). Evaluation of dn:ds in all species comparisons in the 5' bindin region ranged from 0.34 to 0.63. These numbers are not significantly different than 1.0, indicating that this test does not detect the presence of selection for divergence between species. A sliding window analysis (e.g. Metz and Palumbi, 1996; Ina, 1996) of a variety of window sizes (20, 25, 30, 40 codons) did not reveal any region containing significant dn:ds ratios (i.e. > 1.0). For the sliding window analysis comparisons of *S. franciscanus* to *S. purpuratus*, the dn:ds for all windows ranges from 0.07 to 2.04. Specific regions identified by the sliding window analysis are significantly less than 1.0 suggesting genetic conservation. The dn:ds value of 2.04 corresponds to a slight, but non significant, peak in the dn:ds values at the end of the 5' region. This increase in dn:ds could indicate a region of the molecule with

functional importance, but more likely it is the result of random clustering of a few replacement substitutions.

We also conducted a more focused dn:ds analysis on the first 38 codons of the 5' bindin region in the four species primarily because this region corresponds to a region within *S. franciscanus* that displays the greatest concentration of replacement polymorphisms (see Chapter 5). We were interested to learn if there would be a similar concentration of fixed replacement substitutions revealed by an interspecific comparison. All measures for dn:ds in the first 38 codons of bindin are lower than for the whole of 5' bindin with a range of 0 to 0.41 (Table 6-1). In the comparison of *S. purpuratus* and *S. droebachiensis*, this region of bindin has no replacement substitutions resulting in a dn:ds value of zero. Similar to the larger region examined, the average dn:ds values are not significantly different from 1.0. Although this region has a concentration of replacement polymorphisms within *S. franciscanus*, the lower dn:ds values observed between species indicate either fewer replacement substitutions and/or more synonymous substitutions in the region compared to the whole of 5' bindin. Finally, as a comparison to the variable 5' bindin region, we conducted a dn:ds analysis on the conserved core for the four species. The dn:ds values for the conserved core are significantly below 1.0 (Table 6-1) and support the suggestion that purifying selection is maintaining structure and thus function of this region (Minor *et al.*, 1991; Vacquier *et al.*, 1995).

***Sequence alignment and McDonald-Kreitman test for neutral evolution***

Figure 6-1 in shows amino acid alignment of all four species as originally published by Vacquier *et al.* (1995). In the 5' region there are 34 (12%) amino acid codons that are identical between all four species suggesting the action of purifying selection within species results in genetic conservation between species. Table 6-2 shows aligned nucleotide sequence for the first 38 codons of the 5' bindin region for all four species. For all species, we used the McDonald-Kreitman test to evaluate directional selection in this first section (38 codons) of 5' bindin. This region begins at the beginning of mature bindin and corresponds to base pairs 944 to 1058 in Minor *et al.* (1991). This region of bindin represents at least one-third of the entire bindin 5' region. The original use of the McDonald-Kreitman test identified accelerated species diversification between the alcohol dehydrogenase (Adh) locus of *Drosophila sp.* signified by an excess of fixed replacement substitutions compared to silent substitutions between species. The test uses polymorphic data from within at least one species to compare the ratios of replacement to silent substitutions within versus between species. An unequal ratio, and rejection of the null hypothesis of neutral evolution, indicates the action of selection at the locus examined.

The bottom portion of Table 6-2 shows summary numbers and ratios for three species compared to *S. franciscanus*. The ratios of replacement to silent substitutions are not equal, and in all cases the fixed ratio is much lower than the polymorphic ratio. For example, for the first 117 bases of *S. franciscanus* bindin, the total number of replacement polymorphisms is 8 and the total number of silent polymorphisms is 1 yielding a ratio of 8:1. Comparing fixed differences between *S. franciscanus* and *S. purpuratus* results in a replacement to silent substitution ratio of 2.3:1. Although the ratios are unequal, there is no statistical difference in any of the ratios. If the polymorphic ratio had been significantly greater than the fixed ratio, the process of directional selection acting within *S. franciscanus* bindin would be suggested (see Chapter 5). If the polymorphic ratio had been significantly less than the fixed ratio, a selective force for species divergence, similar to what is seen at the Adh (McDonald and Kreitman, 1991) would be suggested. Without statistical significance, it is not possible to reject the null hypothesis of neutral evolution as the force creating sequence differences in *S. franciscanus* compared to the other three species.

***Comparison to selection operating in Echinometra sp.***

Based on dn:ds analysis and the McDonald-Kreitman test, interspecific directional selection is not indicated in 5' bindin region of the three

Strongylocentrotid and *L. variegatus* species examined here. Yet, Metz and Palumbi (1996) report selection in this 5' region of bindin between three species of *Echinometra*. There are at least three possibilities to explain this discrepancy in bindin evolution for the different species: 1) the results presented here are incorrect, 2) the two evolutionary systems in the varying species are fundamentally different, or 3) the signal of significant clustering artifact of random clustering of replacement changes seen in *Echinometra* is not a correct indication of accelerated species divergence. The first two options including the potential differences in the species that would result in alternate patterns of selection and sequence variation are considered thoroughly in the Discussion section. To test the third possibility, the sequence data of *Echinometra sp.* 5' bindin (Metz and Palumbi, 1996) were re-evaluated to determine the probability of obtaining a false signal of significant clustering of replacement substitutions. The size of the original *Echinometra sp.* data set of the bindin 5' region is 384 nucleotides (128 codons) and includes sequence information of 16 *E. oblonga*, 19 *E. mathaei*, and 8 *E.* "species A" alleles (43 alleles total). To perform a dn:ds analysis, it is necessary for each sequence to be equal in length (Kumar, *et al.*, 1993). The *Echinometra* sequence data show insertions and deletions (indels) both within and between species that range from one to ten codons in length. Because dn:ds represents an average over all possible pairwise comparisons, it is possible to remove indels only from those pairwise

comparisons that would have different sequence lengths. However, it is a more conservative approach to eliminate all positions aligned with an indel in all of the sequences and not just specific pairs of sequences. The result after removal of indels was a 228 base-pair (76 codon) data set that had slightly fewer replacement and silent amino acid substitutions than the original *Metz and Palumbi (1996)* data set. Subsequently the test of selection conducted here is more conservative than if the indels had been included.

Using this truncated data set, we then created 100 theoretical data sets identical to the *Echinometra* data (Metz and Palumbi, 1996) yet with a randomized order of codon position. In other words, where the original data had codon positions in a sequential order 1 to 228, here each theoretical data set contains exactly the same codons, yet the arrangement of the codons was randomized (for example, 76, 3, 211, 15, 19 etc.). Each codon was represented only once and thus the overall dn and ds values remain constant yet the clustering of replacement and silent substitutions varies.

The purpose of this simulation is not to mimic the process of evolution. Clearly it is not common for molecules to evolve by the rearrangement of codon order. The goal was to determine the probability of finding a significant cluster of replacement nucleotide substitutions given an overall value of nonsynonymous and synonymous substitutions equal to what Metz and Palumbi (1996) observed. In other words,

given evolution results in a particular level of replacement and silent substitutions, we determined the likelihood that this signal would be spread out over the whole region or clustered into "hotspots." We also evaluated how the method of determining the degree of clustering (i.e. codon window size) would influence the likelihood of observing a significant cluster of nonsynonymous substitutions.

We conducted sliding window analysis on each of the 100 randomized data sets using a 20-, 25, or 30-codon window size. Figure 6-2a in a represents a typical distribution of replacement and silent substitutions for the three window sizes. Out of 100 data sets, using any one of these three window sizes, the probability of obtaining a data set with at least one codon window with a dn:ds ratio significantly different than 1.0 using a two-tailed t-test of significance ranges from 3% to 14% (Table 6-3, see Kumar *et al.*, 1993). For example, out of 100 data sets, evaluated in 25 codon window segments, 14 of these data sets had at least one window where dn was significantly greater than ds using a one-tailed t-test ( $t > 1.96$ ,  $p < 0.05$ ). However, if the test evaluates the hypothesis of directional selection for divergence with a one-tailed test as used in Metz and Palumbi (1996), the range of the probability increases. With a one-tailed test there is between a 10-25% chance of obtaining a data set with a significant region where dn:ds is greater than 1.0. For example, Figure 6-2b shows a region where using a 25 codon window, dn:ds exceeds 20:1. Finally, there are 18 data sets where dn:ds is significant based on at

least one but not all three window sizes. Therefore, the probability of finding a significant region of replacement substitutions clustering increases as the number of window sizes evaluated increases. Thus, the size of the window and the number of different window sizes used as well as the statistical power used (one-tailed versus two-tailed) influences the likelihood of observing a region of the bindin locus where  $d_n:d_s$  is statistically greater than 1.0.

## **Discussion**

All measures suggest an absence of selection for diversification in the bindin 5' locus for all four species of urchins examined in this study. Based on the  $d_n:d_s$  and sliding window analyses, it is not possible to reject the null hypothesis of neutral evolution creating sequence differences observed in the four species. The McDonald-Kreitman test only allows a conclusion about the lack of directional selection in *S. franciscanus* compared to the other three species. To use the McDonald-Kreitman test to explore evolution in the other three species, it is necessary to have intraspecific polymorphic data for those species. The existence of amino acid identity in approximately 10% of the 5' bindin locus in the four species suggests that functional constraints on specific amino acids results in genetic

conservation in portions of bindin. Thus, in *L. variegatus* and three species of *Strongylocentrotus*, neutral and purifying selection acting at the species level result in nucleotide and amino acid sequence differences observed between species.

The methods used here did not detect selection operating in the 5' bindin locus in *Strongylocentrotus sp.* or *L. variegatus*. In contrast, Metz and Palumbi (1996) report that selection in three *Echinometra* species results in rapid divergence of a portion of the same 5' bindin region. There are three possible ways to reconcile the disparate accounts of evolution in bindin. First, it is possible that the methods used here, although similar to the tests employed by Metz and Palumbi (1996), are not sensitive enough to detect selection. This would suggest that the results regarding *Strongylocentrotus* and *L. variegatus* are inconclusive and/or incorrect.

Unfortunately, our analysis is limited to the current understanding of selection and sequence variation and we will have to rely on future work to identify possible misunderstandings. As a second possibility, the results reported here could be correct yet there are differences between *Strongylocentrotus* and *Echinometra* such that selection can create divergence in *Echinometra* and not *Strongylocentrotus*.

Third, it is possible that the signal of selection detected in *Echinometra* is a false signal created by a random clustering of replacement substitutions. To address the

second possibility, we will focus the discussion on some major differences between *Echinometra* and *Strongylocentrotus*.

***Differences in Echinometra sp. and Strongylocentrotus sp. could result in different portraits of selection***

If the results here and in Metz and Palumbi (1996) are correct, different selective pressures create interspecific sequence divergence in *Echinometra* and *Strongylocentrotus* sea urchin species. Because of differences in evolutionary history, fertilization, and species distribution, it is possible that selection shapes sequence divergence in *Echinometra* *bindin* and not *Strongylocentrotus*.

Although there is no direct evidence, it is possible that differences in geographic distribution could result in substantially different population sizes for the two genera. Although the life history characteristics that determine population size of *Strongylocentrotus* and *Echinometra* are highly similar (high fecundity, long planktonic larval stage, etc.) the differences in geographic distribution (continuous versus sporadic) could result in different effective population sizes. For example, biogeography for the two genera are quite different. *S. franciscanus* and *S. purpuratus* inhabit a continuous coastline (Mooris, *et al.*, 1980). *S. droebachiensis*

mostly inhabits a continuous coastline however there are a few large interruptions in species distribution in the Arctic region (Palumbi and Kessing, 1991; Jensen, 1974). The three *Echinometra* species also share common ranges and habitats yet the distribution is not continuous. Frequently, there are large stretches of Pacific Ocean isolating populations of *Echinometra*. For example, Ohta (1990) discusses how small population size can conceivably increase the replacement substitution rate relative to the synonymous rate 40-50% in *Drosophila obscura* Adh genes. The results of a large intraspecific analysis of population structure (see Chapter 4) indicates that the *S. franciscanus* bindin locus has genetic exchange throughout the entire species range suggesting an extremely large effective population size for the species. With potentially smaller population sizes, *Echinometra* sp. bindin could respond more readily to selective forces promoting divergence.

Evolutionary history is another major difference between *Echinometra* and *Strongylocentrotus* that could result in different patterns of selection. *Echinometra* is a recently diverged genus (0.5 - 2 million years ago; Palumbi and Metz, 1991). On the other hand, *Strongylocentrotus* is a relatively ancient lineage with 15-20 million years separating *S. franciscanus* and *S. purpuratus* (Hall, *et al.*, 1980). It is possible that *Echinometra* is so recently diverged that it is in the process of

speciation and experiencing a period of accelerated sequence divergence. Perhaps in the evolutionary past during *Strongylocentrotus* speciation, there was a similar period of accelerated divergence that has been masked by an accumulation of silent substitutions.

Another marked difference between *Echinometra* and *Strongylocentrotus* is the receptivity of egg and sperm to heterospecific fertilization. Within the *Echinometra* genus, there are very strong, reciprocal barriers to interspecific cross-fertilization (Metz *et al.*, 1994). For *Strongylocentrotus*, however, it is relatively more likely to obtain hybrid zygotes (Minor *et al.*, 1991; Vacquier *et al.* 1995). Barriers to cross-fertilization do exist in *Strongylocentrotus* but can be overcome by high sperm concentration and these barriers are not reciprocal. For example, regardless of sperm concentration, *S. purpuratus* sperm are not effective at fertilizing *S. franciscanus* eggs. However, the reciprocal cross of *S. franciscanus* sperm fertilizing *S. purpuratus* eggs produces hybrid zygotes at sufficiently high sperm concentrations (Minor *et al.*, 1991). Perhaps as a recently speciated genus, there is an adaptive advantage to maintaining the strong, reciprocal block to interspecific fertilization through enhancement of selection for rapid interspecific divergence. However, in the time since *Strongylocentrotus* diverged, the selective

pressure to keep heterospecific egg and sperm separate has relaxed as other mechanisms have evolved to avoid gamete waste.

***Implications for proposed mechanisms of diversifying selection***

If it is true that divergent evolution exists in *Echinometra sp.* and not *Strongylocentrotus sp.*, the existence of selection operating in one genus and not another would shed some light on several of the previously proposed mechanisms of positive selection. For example, Lee *et al.* (1995) suggested diversifying selection as a means for free-spawning gametes to avoid microbial pathogens. It is also unlikely that pathogens have adapted to target a widespread genus such as *Echinometra*, but not *Strongylocentrotus*. Thus, if selection shapes sequence divergence in *Echinometra* and not *Strongylocentrotus*, pathogen avoidance is probably not the driving force for sequence divergence.

There are two mechanisms that are still possible as a means of creating sequence divergence even if it is specific to a particular genus and not another. For example, it is possible for assortative mating to be limited to only one genus. Although the adaptive advantage is unclear, mate matching could exist in *Echinometra* bindin and receptor alleles at the species level. If there were an advantage to more than one

bindin type within species, selection favoring intraspecific diversity could also accelerate interspecific species divergence. Again, there is no statistical or experimental support for the necessary selective force resulting in sequence diversity within *Echinometra* species. In addition, this mechanism would require that the bindin and receptor loci are linked to other traits — another conjecture that requires empirical support. Therefore, although mate matching remains a possible mechanism that can result in intra- and interspecific sequence divergence, there is currently no empirical support for this mechanism.

A process of reinforcement that accelerates reproductive isolation could potentially operate in *Echinometra* and not *Strongylocentrotus*. As a very recently diverged genus, *Echinometra* may be completing the process of speciation in which the selective pressure to reduce hybridization is stronger than in *Strongylocentrotus*. This mechanism of reinforcement would favor rapid divergence of gamete recognition molecules between *Echinometra* species. This process would only predict divergence between, and in contrast to the other mechanisms discussed, would not require sequence divergence within species. *Strongylocentrotus* perhaps experienced a similar selective pressure during initial species divergence, however with 10-20 million years since separation, the accumulation of synonymous

substitutions has since diluted the interspecific signal of selection. Therefore, if directional selection shapes sequence divergence in *Echinometra* and not *Strongylocentrotus*, these two mechanisms (assortative mating and reinforcement) remain viable mechanisms driving sequence divergence.

***Patterns in Echinometra and Strongylocentrotus suggest neutral evolution***

Having explored how sequence diversity in *Echinometra* and *Strongylocentrotus* may be the result of different selective pressures, it is also necessary to explore the possibility that the two systems are similar. If the two systems are actually similar, and the results represented here are correct, then the signal of concentrated amino acid substitutions in *Echinometra* may not be an indication of accelerated sequence divergence. Random clustering of neutral mutations may create a false signal that  $d_n$  is significantly greater than  $d_s$ .

The sliding window analysis is not a statistical test but rather a tool to identify regions of unusually high replacement substitutions. There is no error in the use of this tool and at times the sliding window analysis can identify subsets of larger molecules with high proportion of replacement substitutions that correlate with functional regions of a protein. For example, the 57 amino acids in MHCs that

show  $d_n:d_s$  values statistically greater than 1.0 correlate with the functional antigen recognition site (ARS) of the molecule (Hughes and Nei, 1988). However, our evaluation of 100 binding data sets with randomized order of codon position shows that a sliding window analysis has a high probability of identifying a random cluster of replacement substitutions as evidence for the rejection of the neutral theory of evolution in that region. With an *a priori* choice of a single random window size (we chose 20, 25, or 30), the probability of finding a significant cluster of amino acid substitutions can range as high as 10%-25% (Table 6-3). For example, with the 20 codon window analysis, one quarter of the time the data suggest rejection of the null hypothesis of neutral evolution. Clearly no region of the randomized data set can be correlated with sequence function, yet the data would support this conclusion. Additionally, evaluation of more than one window size (i.e. 20, 25, and 30) further increases the probability of finding a significant cluster of amino acid substitutions between species to 29%. Therefore, the use of a sliding window analysis is perhaps a tenuous method to detect regions of elevated sequence divergence.

Finally, the presence of many relatively large insertions and deletions (ranging from 2-11 codons in length) both within and between species, is another suggestion

that neutral and not divergent evolution may dominate *Echinometra* evolution. Alleles with alternate insertions could have alternate function, yet the ability to tolerate relatively large changes to nucleotide sequence suggests the molecule is not rigidly constrained at the species level. Although there are no data to support this supposition, it is difficult to imagine a widely varying molecule that performs a crucial function.

If sequence diversity in *Echinometra* represents a false signal of selection, the divergence in *Echinometra* and *Strongylocentrotus* is the result of neutral evolution. Subsequently, abalone lysin would be the only example of significant interspecific divergence reported for a marine invertebrate gamete interaction molecule. It is a convincing argument for directional selection for divergence that the dn:ds values for lysin are among the highest reported and furthermore these high values result from analysis of the entire molecule and not a sub-region (Vacquier and Lee, 1993). It is unclear why lysin and not other fertilization molecules such as bindin would diverge rapidly. It is possible that functional differences account for the disparate divergence. Where bindin is responsible for species-specific binding of sea urchin sperm to eggs (Vacquier *et al.*, 1995), lysin solubilizes the egg vitelline envelope

(Vacquier and Lee, 1993). It is not clear if this difference in specialized functions would also be linked to differences in potential for sequence divergence.

Where the lysin locus may be the only example of accelerated divergence between species of a marine invertebrate, there are currently no examples of diversity enhancing selection within a marine species. In fact, the major histocompatibility locus (MHC) in humans and mice (i.e. Hughes and Nei, 1988; Hughes and Nei, 1989) and self-incompatibility (S-alleles) in plants (Clarke and Kao, 1991) are the only two systems displaying intraspecific signal for allelic diversity. In both of these cases, overdominant selection increases allelic diversity at a specific locus. An extensive examination of 268 alleles in *S. franciscanus* revealed moderately high polymorphism yet neutral evolution (see Chapter 5). Likewise for each of the three *Echinometra* species examined, there is no statistical support for positive selection at the species level (Metz and Palumbi, 1996). And despite lysin's strong signal of interspecific divergence (Lee *et al.*, 1995), there is no evidence of intraspecific allelic diversity (E. Metz, unpublished data).

There are many accounts of accelerated sequence divergence of particular genes between species (i.e. Endo *et al.*, 1996; Messier and Stewart, 1997; Lee *et al.*, 1993). Interspecifically, the term positive selection is used to describe a process

that results in the divergence of a specific locus between two species. Intense directional selection for divergence "character displacement" is an expected outcome when two recent and related species come into secondary contact. Thus, the process termed positive selection between species may be no different than the expected action of natural selection. The process becomes interesting is when unique selective pressures on that locus between the different species results in relatively rapid divergence and a higher than expected amount of amino acid substitutions as is seen in lysin (Lee *et al.*, 1995). For some of the pairwise comparisons of abalone species, the sequences are not very divergent and show no statistical reasons to reject the null hypothesis of neutral evolution. However, for six out of 190 pairwise comparisons,  $d_n:d_s$  is statistically greater than the neutral expectation and results in more rapid divergence than neutral evolution would create, thus implying an increased selection for species-specific function at that locus.

Thus there is an important distinction between a selective force that results in accelerated sequence divergence between species compared to a selective force that results in allelic diversity within species. Unfortunately, throughout the literature the term "positive selection" has been used to describe both of these processes. In

the sense that both of these processes are the antithesis of purifying (or negative) selection, it is correct to use the term positive selection. Yet, it is misleading, if not incorrect, to equate the two processes. For example, it is incorrect to compare selection promoting interspecific sequence divergence (such as lysin) with intraspecific diversity-enhancing selection as seen in MHCs. The former is accelerated directional selection, while the latter is the result of overdominant selection creating extreme allelic diversity. It would be helpful if a new standard could be adopted. For example, when discussing the rapid divergence of a locus between species it would be informative to use the term directional selection and not positive selection. If the term positive selection is used at all, it should be confined to describing selection for intraspecific sequence variation. Yet, perhaps the most appropriate term to describe the intraspecific process is diversity enhancing selection. Regardless which term is used, selection for sequence divergence, either within or between species, does not appear to be a general property of marine invertebrate fertilization molecules. Perhaps before searching for this phenomenon in other marine invertebrates, it would be worthwhile to learn more about the particular operation of sequence divergence in abalone lysin by understanding more of its function and interaction with the abalone egg.



## References

- Clarke, A. G., and Kao, T.-H. (1991). Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of *Solanaceae*. PNAS USA 88: 9823-9827.
- Edmands, S., Moberg, P. E., and Burton, R. S. (1996). Allozyme and mitochondrial DNA evidence of population subdivision in the purple sea urchin *Strongylocentrotus purpuratus*. Marine Biology 126:(3) 443-450.
- Endo, T., Ikeo, K., and Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. 13:(5) 685-690.
- Gao, B., Klein, L. E., Britten, R. J., and Davidson, E. H. (1986). Sequence of mRNA coding for bindin, a species-specific sea urchin sperm protein required for fertilization. Proc. Natl. Acad. Sci. USA 83: 8634-8638.
- Gilbert, D. (1994). SeqApp. Multiple sequence alignment program (1.9a157+). Indiana State University.
- Hall, T. J., Grula, J. W., Davidson, E. H., and Britten, R. J. (1980). Evolution of sea urchin nonrepetitive DNA. J. Mol. Evol. 16: 95-110.
- Hughes, A. L., and Nei, M. (1989). Major histocompatibility complex class II loci: evidence for overdominant selection. PNAS USA 86: 958-962.
- Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:(8 September) 167-168.

- Ina, Y. (1996). Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. *J. Genet.* 75: 91-115.
- Jensen, M. (1974). The Strongylocentrotidae (Echinoidea) L. a morphological and phylogenetic study. *Sarsia* 57: 113-148.
- Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian protein metabolism*, H. N. Munro, ed. New York: Academic Press, pp. 21-132.
- Kelso, D. (1970). A comparative morphological and ecological study of two species of sea urchins, genus *Echinometra*, in Hawaii. In Department of Zoology Honolulu: University of Hawaii.
- Kimura, M. (1983). *The neutral theory of molecular evolution* Cambridge: Cambridge University Press.
- Kreitman, M., and Akashi, H. (1995). Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* 26: 403-422.
- Kumar, S., Tamura, K., and Nei, M. (1993). MEGA: Molecular Evolutionary Genetics Analysis (version 1.01). University Park, PA: The Pennsylvania State University.
- Lee, Y.-H., Ota, T., and Vacquier, V. D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 12:(2) 231-238.
- Li, W.-H., and Graur, D. (1991). *Fundamentals of molecular evolution* Sunderland, MA: Sinauer Assoc.
- McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:(20 June) 652-654.

Messier, W., and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature* 385:(9 January) 151-154.

Metz, E. C., Kane, R. E., Yanagimachi, H., and Palumbi, S. R. (1994). Fertilization between closely related sea urchins is blocked by incompatibilities during sperm-egg attachment and early stages of fusion. *Biol. Bull.* 187: 23-34.

Metz, E. C., and Palumbi, S. R. (1996). Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* 13:(2) 397-406.

- Minor, J. E., Fromson, D. R., Britten, R. J., and Davidson, E. H. (1991). Comparison of the bindin proteins of *Strongylocentrotus franciscanus*, *S. purpuratus*, and *Lytechinus variegatus*: Sequences involved in the species specificity of fertilization. *Mol. Biol. Evol.* 8:(6) 781-795.
- Morris, R. H., Abbott, D. P., and Haderlie, E. C. (1980). *Intertidal invertebrates of California* Stanford, CA: Stanford University Press,.
- Nei, M. (1987). *Molecular evolutionary genetics* New York: Columbia University Press.
- Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:(5) 418-26.
- Nei, M., and Jin, L. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* 6: 290-300.
- Palumbi, S. R. (1991). Large mitochondrial DNA differences between morphologically similar Penaeid shrimp. *Molecular Marine Biology and Biotechnology* 1:(1) 27-34.
- Palumbi, S. R. (1992). Marine speciation on a small planet. *TREE* 7:(4) 114-117.
- Palumbi, S. R., and Metz, E. C. (1991). Strong reproductive isolation between closely related tropical sea urchins (genus *Echinometra*). *Mol. Biol. Evol.* 8:(2) 227-239.
- Smith, A. E. (1984). *Echinoid Paleobiology* London: George Allen and Unwin
- Sokal, R. R., and Rohlf, F. J. (1981). *Biometry: The principles and practice of statistics in biological research*, 2nd Edition New York: W.H. Freeman and Co.
- Vacquier, V. D., and Lee, Y.-H. (1993). Abalone sperm lysin: unusual mode of evolution of a gamete recognition protein. *Zygote* 1:(August) 181-196.

Vacquier, V. D., Swanson, W. J., and Hellberg, M. E. (1995). What have we learned about sea urchin sperm binding? *Develop., Growth and Differ.* 37: 1-10.

Vacquier, V. D., Swanson, W. J., and Lee, Y.-H. (1997). Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure for their divergence. *J. Mol. Evol.* 44 (*Suppl. 1*): S15-S22.