



Symbiont Population Structure in *Riftia* and *Oasisia* at 21N: an examination of direct sequencing, host specificity, & the effects of geography on haplotype frequency

Megan Rippy, University of California Santa Cruz

Joe Jones & Bob Vrijenhoek

Summer 2004

Keywords: Vestimentifera, symbiosis, *Riftia*, *Oasisia*, *ITS*, cloning, sequencing

ABSTRACT

Deep-sea hydrothermal vents along the Eastern Pacific Rise are small, chemically enriched, distantly separated habitats. Most of the organisms that inhabit these diverse ecosystems are invertebrates that rely upon chemoautotrophic symbionts for the organic materials they need to survive. Adult vestimentiferans house sulfur-oxidizing symbiont varieties within their trophosome tissues, and recent studies involving 16s DNA show that for *Riftia* and *Oasisia* the microbial strain is the same. With regards to ITS, however, a study of 1 vent site at 21N, reveals three distinct ITS haplotypes, A, D and C (Young, unpublished). When the frequency counts of these haplotypes for different vestimentiferans are analyzed by an AMOVA test, it appears that *Riftia* and *Oasisia* exhibit preference for particular strains. For the 2003 MBARI T557 dive, the AMOVA P value was 0.007495, a marginally significant score. For the 1990 Alvin 2233 dive, the AMOVA P value was small enough to indicate significant differences between *Riftia* and *Oasisia* symbiont composition (0.04327). Using these same dives, as well as the 1990 Alvin 2230 dive, an AMOVA test analyzing the variation in symbiont composition across tubeworm clump was performed. The P value for this test was 0.45584, suggesting that there is no difference in ITS compositions over small distances at 21N.

During the 21N ITS study it was found that 62% of the 26 hosts sampled contained at least two ITS haplotypes, indicative of multiple symbiont infections. This high percentage of double symbionts made the use of direct sequencing to determine symbiont haplotype questionable. When converted from base calls to peak areas and compared graphically to frequencies found through cloning, a correlation between the two measurements was observed. Analysis using correlation coefficients showed that when the entire data set was examined, a t value of 15.80 was obtained, quite large enough to indicate a significant correlation between clone frequency and peak area frequency. This analysis was misleading, however, because a large portion of it hinged upon points where no or only A haplotype was seen. The removal of these data points lowered the t statistic computed to a value of 3.70. This number still indicates a significant relationship between the methods examined, but not an overwhelming one.

INTRODUCTION

The East Pacific Rise (EPR) is a 6,500 km long stretch of seafloor that initially follows the coast of Mexico and then continues due south into some of the deepest parts of the Pacific Ocean. It marks the southern part (27 deg North latitude to 32 deg south latitude) of the Pacific Plate's eastern boarder, and contains 10 individual hydrothermal vent sites. Because these fields can be 100 km or more apart, organisms that exploit these environments, like vestimentiferans, have a patchy habitat distribution (Van Dover et al. 2002 # 4738).

The primary producers of vent communities are chemolithotrophic microbes that oxidize the reduced compounds emitted in vent effluents, like hydrogen sulfide. Any large multi-celled organism living in a vent community relies upon these microbes at some level, whether they consume them, another creature that does, or coexist with them in a kind of symbiosis.

Vestimentiferans are an extreme example of the third case. They have no digestive system of their own and rely upon the organic molecules produced by their symbionts as an energy source (Hurtado, 2002 # 4871). The symbionts benefit from this relationship because the interior of the tubeworm, where they are stored, provides a stable environment that the vent ecosystem does not. Specifically, the bacteria are contained within special cells called bacteriocytes that in turn make up the tissue known as the trophosome (Won, 2003 # 4741). Tubeworms have even evolved a highly developed hemoglobin that transfers hydrogen sulfide along with oxygen and carbon dioxide. This ensures a continual supply of energy to its chemolithotrophic inhabitants, and eventually, itself.

Along the Eastern Pacific Rise there are three main genera of vestimentiferans, *Riftia pachyptila*, *Oasisia alvinae* and *Tevnia jerichoana*. There have never been any *Tevnia* reported above the hydrothermal vent site at 13 N, though *Riftia* span the entire EPR. *Oasisia* exhibits a more complicated distribution because there are believed to be between two and four species of this tubeworm, one of which is only found in southern latitudes (Hurtado et al. 2004 in review). Morphologically speaking, it is easiest to distinguish between these three vestimentiferans based on tube shape, although it can be hard to identify what worms inhabit a clump before it is sampled. This complication is largely due to the fact that *Oasisia* require pre-existing ground structure in order to grow. For this reason they are often found amongst groups of mussels or within large bunches of *Riftia*, which can hide them from view. In general, however, *Riftia* are identified by their smooth white tubes, *Oasisia* by the presence of widely spaced ridges, and *Tevnia*, by ridges that are very close together (Black, 1997 # 3188).

The hydrothermal vent field 21N, the site of interest in this study, contains only *Riftia* and *Oasisia* vestimentiferans. Based on their 16s rDNA genes, the symbionts that inhabit these different hosts are the same. This finding can be justified by the theory that vestimentiferan symbionts are horizontally acquired. This means that each individual tubeworm, at some stage in its life, acquires its symbionts from the environment it is in (Cary, 1993 # 5195). Because *Oasisia* and *Riftia* often share the same clump, their symbiont compositions should be similar. 16s rDNA, however, acquires mutations very slowly because the protein product it codes for is essential to small ribosome subunit function. If vestimentiferans were selective in their acquisition of symbionts, or passively amassing slightly different types due to species succession patterns, an analysis of 16s would fail to shed light these factors because of its high level of functional constraint. For this reason the Internal Transcribed Spacer region (ITS), a non-coding and thus more variable gene, is used as the basis for the molecular biology in this experiment.

The use of ITS in a gene flow study is not novel and was performed for vestimentiferans by Robbie Young and Steve Hallam in 2001. Their research identified five different ITS haplotypes along the entire East Pacific Rise, with only two of these, A and D, occurring at 21N. These haplotypes had two polymorphic sites each, which could read either GG (type A) or AT (type D). The study used direct sequencing to identify haplotypes, though the reliability of this method was questioned due to the occurrence of three odd individual vestimentiferans that contained

more than one of the strains (Young, 2001, unpublished). It is the primary purpose of this experiment to examine direct sequencing through a comparative analysis of direct sequence traces to clone traces, which can only represent one organism. 21N was chosen as a test site because it contained a *Riftia* host with both A and D ITS haplotypes. Where a multiple infection has occurred once, it is quite possible that others have occurred as well, making the re-examination of this site a good starting point in a study of direct sequencing error.

Besides providing information concerning the accuracy of direct sequencing, the cloning of individuals from 21N provides an opportunity to perform both within host and between clump analyses of ITS variation. If host specificity for particular symbionts is not a factor affecting what haplotypes are observed in a location, a lot can be learned about population structure from the way symbiont haplotype ratio's in tubeworms change over geographic distance. For the cold seep tubeworm *Lamellabrachia*, variation in symbiont composition over areas as small as three meters has been observed (Duhaime, 2003, unpublished). This represents an incredible amount of localized symbiont adaptation. It is possible, however, that for *Riftia* and *Oasisia* some preference for particular symbiont haplotypes does occur. A study performed by DiMeo in 2000 uses rep-PCR fingerprints to argue for this kind of bias (DiMeo, 2000 #4366). Studies of host specificity in vestimentiferans are hard to formulate, however, because this kind of preference cannot really be separated from the possibility that variation due to temporal factors, like *Oasisia*'s late colonization of vent sites, is the real variable driving differences in symbiont composition between hosts. For simplicity's sake this paper will refer to the combined temporal/host specificity variable as host specificity. Of the two hypotheses this is the only one that currently has genetic backing, and thus, for now, it is favored.

MATERIALS AND METHODS

SPECEMINS

The Vestimentiferan specimen hosts were obtained from three different dives, T557, 2230, & 2233, at the East Pacific Rise vent field known as 21N. Prior to their arrival and processing at MBARI, these tubeworms were kept at temperatures below 10 degrees Celsius in order to preserve their tissues and DNA. Upon arrival at the research institute they were separated into

vestmentum, obturculum and trophosome tissues that were frozen. Host identification was based upon the morphological characteristics used in a suite of phylogeographic studies on east pacific rise vestimentiferans (Black, 1997 # 3188; Black, 1998#3223; Hurtado, 2003 #5517)

MOLECULAR METHODS:

Because one objective of this experiment was to validate direct sequencing, as used by Robbie Young and Steve Hallam to identify multiple symbiont strains in the 2233 & 2230 dives, the DNA extractions used for the majority of this molecular work were those prepared by Steve in 2001. These extractions were performed on trophosome tissue containing both host and symbiont DNA, using the Qiagen DNA isolation kit (Qiagen Inc, Valencia, CA). It is unknown how much time these extractions have been stored at -20 deg C versus 4 deg C.

The extractions for the T557 dive and individuals R29 and OCL1 from the 2233 dive were also performed on trophosome tissue using the Qiagen DNA isolation kit. These extractions, however, were performed in 2004 instead of 2001, and thus the DNA itself has had less time to degrade. They were likewise stored at both -20 deg C and 4 deg C.

Symbiont DNA was amplified from the Qiagen extractions, both for direct sequencing and cloning, using sym6 and sym7 primers targeted for the Internal transcribed spacer (ITS) region of the SSU operon. The sym 6 primer attaches 40 base pairs from the 5' end of the 16s rRNA subunit and the sym 7 primer attaches approximately 20 base pairs from the 5' end of the 23S rRNA subunit. Each primer is 15 base pairs long (Young, 2001, to be published).

Sym-6 (5'-GAAGTCGTAACAAGG)

Sym-7 (5'-CAAGGCATCCACCGT)

The PCR reaction mixtures themselves were set up for 25ul, where 1ul of Qiagen DNA extract was added to 2.5ul 10x buffer (Promega Inc. W1), 2.5ul MgCl (2.5uM), 2.5ul of a 2mM dNTP stock, 1ul sym6, 1ul sym7, 0.25ul Taq polymerase (Promega Inc. W1) and 14.25ul nanopure water. When reaction mixtures were created with the intention of cloning the ITS product amplified, 0.25ul ampli-Taq gold was substituted for the Promega Taq, and 2.5ul of ampli-Taq buffer and ampli-taq MgCl were substituted for their Promega equivalents. PCR was performed

using a PerkinElmer machine where the conditions were set to: 25 cycles of 94 deg C/1 min., 52deg C/1min. and 72 deg C/2min. The final extension was performed at 72 deg C for 7 min.

In order to purify the PCR products for direct sequencing, the amplifications were run out on a 1X TAE gel, cut out, spun down at 8000rpm for 10 minutes, and then stored at -20 deg C. The PCR products intended for cloning were treated differently because only fresh PCR products can be taken up by E-Coli as a vector insert. Ligations and transformations were performed before the purified PCR product was frozen at -20 deg C, using the TOPO TA cloning kit (Invitrogen, Carlsbad, California). Transformed E-coli were incubated at 37 deg C in SOC medium for an hour, after which they were plated on agar plates containing kayamycin and rifampicin. The plates were in turn incubated at 37 deg C overnight, and individual colonies picked the next morning. The PCR amplifications performed on those picked colonies had the same proportions of Taq, buffer, MgCl, dNTP's, water and primer, as those discussed above. The Taq, buffer and MgCl used were from Promega, the primers were m13F and m13R, provided by Invitrogen for their TOPO cloning kits, and the PCR reaction conditions were specified by Invitrogen in their m13-check program. The final PCR product was purified using a Millipore filter after being checked for ITS inserts by E-gel.

All sequencing reactions for the ITS DNA in this experiment were performed on an ABI 3100 using Big Dye version 3.0.1 (ABI Biosystems, Inc.). The sequencing traces were edited manually on *Sequencher* version 4.1 (Gene Codes Corp., Ann Arbor, Michigan) and aligned using *MacClade* 4.0. (Sinauer Associates Inc., Sunderland, Massachusetts).

PCR PEAK AREA TEST

In order to examine the relationship between the frequency of ITS haplotypes within an individual vestimentiferan (obtained by cloning) and the relative haplotype amounts for that same vestimentiferan (detected by direct sequencing), it was necessary to develop a quantitative method for measuring haplotype from sequence traces. The method used here is based on *Sequencher* version 4.1's depiction of the data from ABI 3100 as a chromatogram. The formula: (height x width at 1/2 height), a general formula for peak area, was applied to the sequence trace chromatograms at each of two variable sites for ITS. The numerical values for each peak area

seen at that site could then be represented separately as a fraction of total peak area. Because the two sequencing approaches used reflect the same haplotype frequencies, a frequency over frequency plot should produce a line with a slope of 1. The fact that the two variables being compared are measurements, however, means that regression analysis comparing the slope of the data trendline to 1 could not be performed. Instead, correlation coefficients and r^2 values were calculated and then tested for significance through conversion to t scores that could be compared with literature values from the t distribution. In this way it was possible to measure both the strength and significance of the correlation between direct sequencing and cloning. These same sets of values were also computed for the data when it excluded all points where cloning picked up only one haplotype. In such cases direct sequencing and cloning would have to be 100 percent correlated and the ability of direct sequencing to accurately predict symbiont frequencies in a host with multiple infections was not an issue.

Because all sequencing methods involve the use of two primers and the ITS gene has two polymorphic sites of interest, it was necessary to analyze the peak area frequency and clone frequency data for both correlation experiments described above by primer and by site. R^2 values were computed for all of these partial data sets and tested for significance using t scores.

As a subtest within the study just described, an experiment was performed where DNA from a single individual (2233OCL1) was sequenced five separate times from different PCR's. The peak area's for the G/A and G/T ITS sites were calculated, converted to frequencies in terms of G, and graphed as two separate histograms, one representing all of the data for the G/A site and one all the data for the G/T site. Each histogram was further broken down into peak areas that were reported by the sym6 primer and peak areas that were reported by the sym7 primer. The ranges of the results for each primer and each site, as well as for the entire data set, were calculated in order to test for variation within the direct sequence traces of a single individual.

ARLEQUIN

ARLEQUIN (v. 2.000, \ Schneider, 2000 #4656) was used to perform three separate three tiered analyses of molecular variance (AMOVA, \ Excoffier, 1992 # 2038). The first two of these tests partitioned data into between host, between individual/within host, and within individual

groupings, represented by F_{ct} , F_{sc} , and F_{st} statistics, respectively. The third test partitioned data into between dive, between individual/within host and within individual groupings, also represented by F_{ct} , F_{sc} , and F_{st} statistics. These values were tested for significance by the ARLEQUIN Mantel test, which provides P values.

RESULTS

SEQUENCE VARIATION IN VESTIMENTIFERAN ENDOSYMBIONTS

348 ITS DNA fragments (489bp) were examined from a total of 26 vestimentiferans that themselves were acquired in different dives. In total, three separate haplotypes were observed, A, D, & C (figure 1), all three of which concern two variable sites. The first of these sites is a G/A transition and the second, also a transition, is G/T.

62% of the vestimentiferans sampled contained two or more ITS haplotypes, making them heterotypic. Nine of these individuals were *Oasisa* and six were *Riftia*. Of the mixtures seen, 81.25% (13) were D/A, 12.5% (2) were A/C and 6.25% (1) was A/D/C. Haplotype A was the most prominent in that it was present in 88.46% of the tubeworms examined.

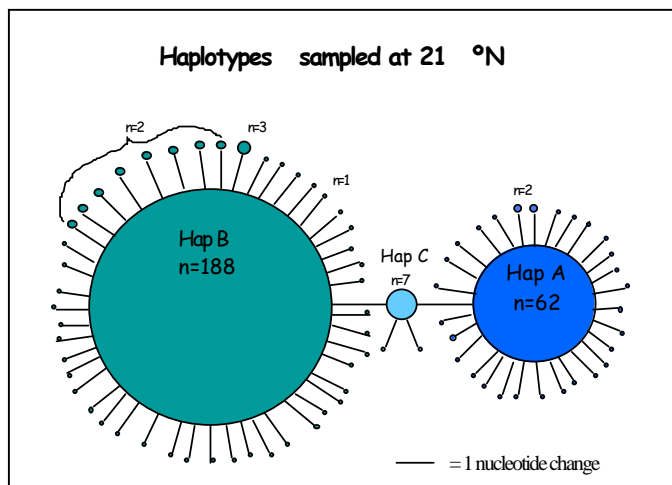


Figure 1: Haplotype Map of 21N vestimentiferan clones. Small balls are singletons, slightly larger balls represent repeat mutations, and the three largest balls show relative amounts of A, D, & C ITS haplotypes. (haplotype B is the same as haplotype A, haplotype A is the same as haplotype D, and haplotype C remains unaltered)

Haplotype D was present in 65.38% (all but nine), and haplotype C was only present in 11.53% (three). The most common homotypic ITS strain was A, seen in 23.07% of vestimentiferans,

only one of which was an *Oasisia*. Strain D was only homotypic 11.54% of the time (in 2 *Riftia* and 1 *Oasisia*), and strain C never stood alone (table 1).

Haplotypes in Cloned Individuals

INDIVIDUAL	TYPE A	TYPE B	TYPE C	INDIVIDUAL	TYPE A	TYPE B	TYPE C
<i>Riftia</i> 18	10	1	0	<i>Oasisia</i> 6	0	10	0
<i>Riftia</i> 19	5	8	0	<i>Oasisia</i> 7	5	5	0
<i>Riftia</i> 16	2	7	0	<i>Oasisia</i> C 1	24	23	0
<i>Riftia</i> 34	0	10	0	<i>Oasisia</i> A2-1	8	0	0
<i>Riftia</i> 32	0	7	0	<i>Oasisia</i> A2-2	2	9	0
<i>Riftia</i> 43	0	10	0	<i>Oasisia</i> A2-3	3	6	0
<i>Riftia</i> 29	0	50	0	<i>Oasisia</i> A2-5	8	1	0
<i>Riftia</i> A2-1	0	3	6	TOTAL	93	246	9
<i>Riftia</i> A2-2	0	6	0				
<i>Riftia</i> A2-3	1	9	0				
<i>Riftia</i> A2-4	9	0	0				
<i>Riftia</i> A2-7	9	0	0				
<i>Riftia</i> A2-13	0	11	2				
<i>Riftia</i> A2-15	1	9	0				
<i>Riftia</i> A2-22	0	10	0				
<i>Oasisia</i> 2	6	11	0				
<i>Oasisia</i> 3	3	11	0				
<i>Oasisia</i> 4	2	4	0				
<i>Oasisia</i> 5	4	6	1				

**Types A, B, & C
come from EPR/ITS
studies made by
Robbie Young &
Steve Hallam**

Table 1: Clones sorted by haplotype and host type (A is written as B, D is written as A, and C is written as C)

DIRECT SEQUENCING VS. CLONING

When clone haplotype frequency, in terms of A, was plotted against peak area frequencies, also in terms of A, a mostly linear relationship was revealed (figure 2).

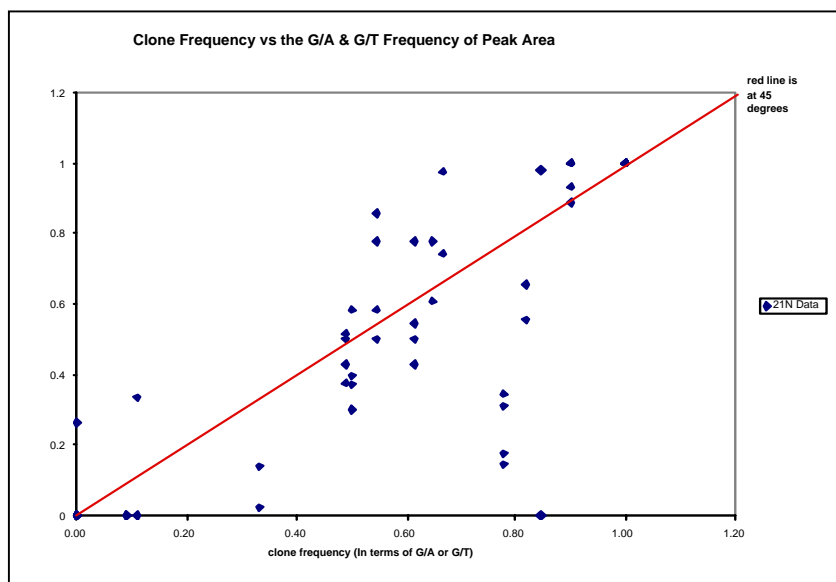


Figure 2: Graph of clone frequency versus peak area frequency (in terms of haplotype A)

Specifically, when the actual points plotted were separated out into categories of; sym6 primer & G/A site, sym7 primer & G/A site, sym6 primer & G/T site, and sym7 primer & G/T site, data trends with r^2 values of 0.81, 0.77, 0.89, and 0.73, respectively, were observed. This indicates that no individual parameter examined had a linear correlation of less than 85%. The corresponding t values were all large enough for the hypothesis of correlation between direct sequencing and cloning to be accepted (table 2). When the data from the sym6 and 7 primers was examined as a whole, r^2 values of 0.80 and 0.74 were calculated, marking the sym7 primer as less reliable. The r^2 values for the G/A and G/T sites, 0.80 and 0.82 respectively, were similar to the value for the sym6 primer. The overall data set, however, had a lower r^2 value (0.76), indicating that the correlations seen in the individual parameters are offset, leading to a larger range of values when they are observed together. Even so, this r^2 value, along with those for the sym6 primer, sym7 primer, G/A site and G/T site, converts to a t score of unquestionable significance (table 2).

Data set examined	R^2 value	T score	T from text	Accept/reject null
Sym6 g/a	0.81	7.84	1.75	Accept
Sym7 g/a	0.77	7.49	1.75	Accept
Sym 6 g/t	0.80	8.70	1.71	Accept
Sym7 g/t	0.73	7.98	1.71	Accept
Total g/a	0.80	11.06	1.70	Accept
Total g/t	0.82	12.19	1.68	Accept
Total sym6	0.80	12.51	1.68	Accept
Total sym 7	0.74	10.57	1.68	Accept
All data	0.76	15.80	1.67	Accept

Table 2: unmodified data set r^2 values, t scores, and significance analyzed both by primer and by site

When the total data set used to derive the calculations reported above was modified to exclude all individuals where cloning picked up only one haplotype, the r^2 values calculated decreased. For the individual categories of; sym6 primer G/A site, sym7 primer G/A site, sym6 primer G/T site, and sym7 primer G/T site, r^2 values of 0.16, 0.13, 0.61 & 0.47, respectively, were calculated. This makes the G/T values for both primers approximately 3/4ths the magnitude of

those calculated using the entire data set and the G/A values only 1/3 of those observed for G/T (figure 3 & 4).

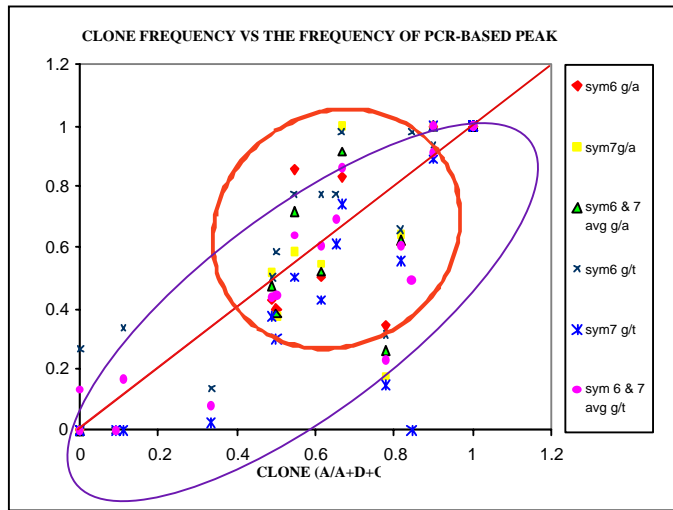


Figure 3: Total data set where the G/A site is circled in orange and the G/T site in purple

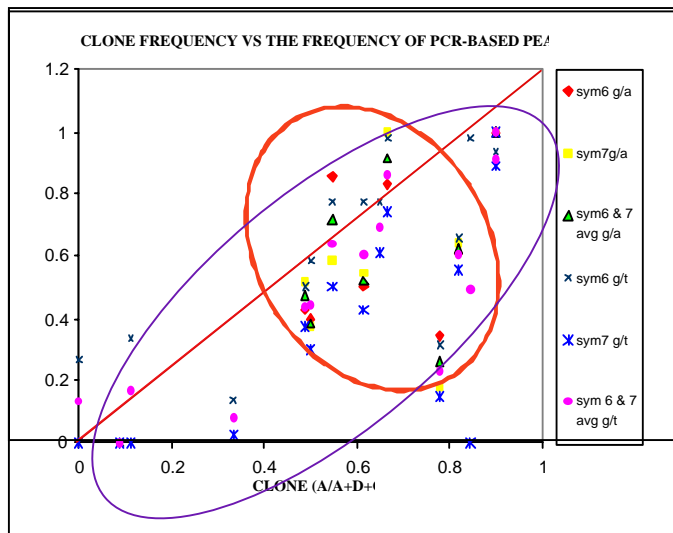


Figure 4: Data set excluding clones from individuals with only one symbiont type. The G/A site is circled in orange and the G/T in purple.

When this relationship is examined in terms of total G/A and total G/T r^2 scores, the modified data can be used to create a correlation ratio of 0.15 : 0.61, while the original data's ratio is 0.80 : 0.82. Not only is the G/A : G/T ratio for the modified data 23 times smaller than the ratio for the raw data, but the majority of this size discrepancy is accounted for by the low correlation between direct sequencing and clone frequency at the G/A site. When t scores for the total G/A and the total G/T data are calculated, direct sequencing is shown to be insignificantly correlated

with cloning for the G/A site and significantly correlated with cloning in regards to the G/T site (table 3).

Data set examined	R ² value	T score	T from text	Accept/reject null
Sym6 g/a	0.16	1.05	1.89	Reject
Sym7 g/a	0.13	0.95	1.89	Reject
Sym 6 g/t	0.62	4.59	1.76	Accept
Sym7 g/t	0.47	3.37	1.76	Accept
Total g/a	0.15	1.58	1.75	Reject
Total g/t	0.61	4.10	1.70	Accept
Total sym6	0.50	4.57	1.72	Accept
Total sym 7	0.35	3.39	1.72	Accept
All data	0.38	3.70	1.71	Accept

Table 3: modified data set r² values, t scores, and significance analyzed both by primer and by site (modified refers to the absence of data from individuals containing symbionts of only one haplotype)

When the modified data set is examined on a primer to primer basis, it is seen that the r² value for sym7 (0.35) is smaller than the r² value for sym6 (0.50). Although both of these r² values are lower than the corresponding ones from the full data set, they show the same trend; direct sequencing appears to be more correlated with cloning for the sym6 primer (figure 5).

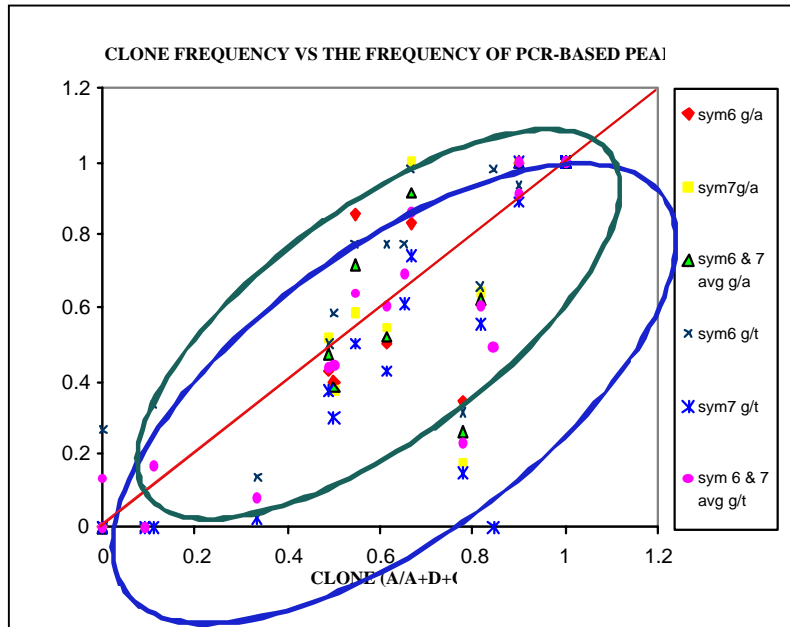


Figure 5: Graph of clone frequency versus peak area frequency in terms of haplotype A. The blue oval represents the sym7 primer and the green oval the sym6 primer.

It is important to note, however, that this trend does not indicate there is no correlation between direct sequencing and cloning for the sym 7 primer. The t scores for both primers, as well as the r^2 score of 0.38 calculated for the entire data set, indicate a significant correlation between these two methods (table 3).

DIRECT SEQUENCING VS CLONING (FOR 1 INDIVIDUAL)

Through the examination of multiple direct sequence traces from one individual it was found that different haplotype frequencies were in each trace, which should not have been the case. The ratio of A:D in the 2233OCL1 individual chosen, as observed through cloning, was 23:24, or 0.49. This makes the average values at the G/A site of 0.37 (sym6 primer) and 0.47 (sym 7 primer), low. Here the sym 7 primer was more accurate (figure 6). This changes for the G/T site, where the average value for the sym6 primer was 0.64 and the average value for the sym7 primer was 0.3080 (figure 7). This relationship becomes clearer when shown as percent divergence from the expected clone frequency of 0.49. This makes the sym6 primer at the G/A 24.36% divergent, the sym 7 primer at the G/A site 4.08% divergent, the sym6 primer at the G/T site 30.14% divergent, and the sym7 primer at the G/T site 37.01% divergent. Clearly sym 7 is more accurate for the G/A site and less for the G/T site. Furthermore, these percents show that overall the G/A site produced more accurate results than the G/T site.

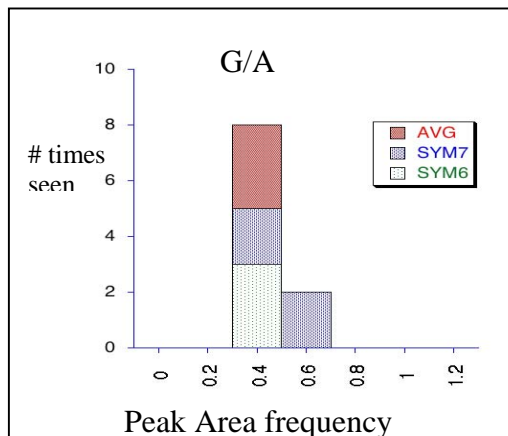


Figure 6: A histogram depicting peak areas for multiple direct sequencing reactions from one individual. (this graph is formatted in terms of the G/A site for both the sym6 & sym7 primers)

Precision values for the primers at the two polymorphic sites were examined by looking at total range. The largest of these ranges was 0.5, a 50% difference in reported frequency seen for sym6 at the G/T site. The smallest range was 0.095, a 9.5% difference in reported frequency seen for sym6 at the G/A site. The other ranges were 0.13 and 0.16, for the sym6 primer at the

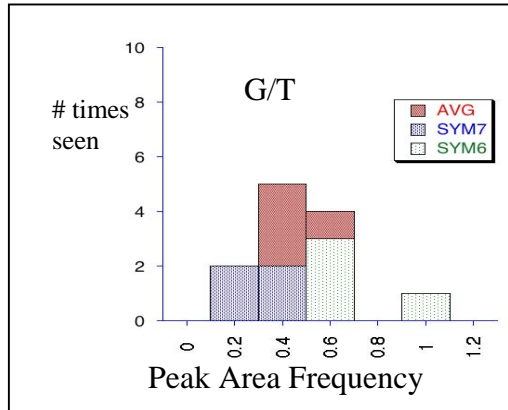


Figure 7: Figure 6: A histogram depicting peak areas for multiple direct sequencing reactions from one individual. (this graph is formatted in terms of the G/T site for both the sym6 & sym7 primers)

G/A site & the 7 primer at the G/T site, respectively. These values indicate that overall the G/A site was more precise, and that within this site the sym6 primer was more precise than the sym7. Although the G/T site was less precise overall, the sym7 primer at this site produced more repeatable results than the sym7 primer at the other. It was also more precise than the sym6 primer at its same site, a result not seen at the G/A location (figure 6 & 7).

ANALYSIS OF MOLECULAR VARIANCE (AMOVA)

For the three-tiered AMOVA analyzing the relationship between host vestimentiferan and symbiont haplotype, the percentages of variation due to; between host group, between individual/within host group, and within individual differences, were not the same for the two dives examined. In the T557 dive, 16.54 percent of the variation observed was due to differences between host type, 47.55 percent was due to differences between individuals/within host type, and 35.91 percent was due to differences within individuals. The between host type parameter for the 2233 dive, however, was almost twice that of the T557 dive at 30.14 percent, while its between individuals/within host type parameter was about 7 times smaller (6.42%). Like the between host group comparison, the 63.44% within individuals parameter for the 2233 dive was almost twice that of the T557 dive. These variation percentages, when converted by the

Mantel test to a P score, were all at least marginally significant, indicating that the null hypothesis (an assumption that there is no correlation between the data analyzed) can be rejected.

2233		
Source of variation	Percentage of variation	Probability
Between host type	30.14	0.043
Between individuals within host type	6.42	0.055
Within individuals	63.44	0.000

T557		
Source of variation	Percentage of variation	Probability
Between host type	16.54	0.075
Between individuals within host type	47.55	0.000
Within individuals	35.91	0.000

Table 4: Displays the results of the AMOVA tests analyzing host specificity (blue P values are significant and black P values are marginally significant)

It is important to note, however, that the between host variation for the T557 dive is greater than 0.05, and thus in traditional statistics would not be used to state that hosts discriminate in their symbiont acquisition. The 2233 dive exhibits a similar result for the comparison between individuals within host type, though this 0.05455 value is even closer to the 0.05 cut off point. Its P value for the between host type comparison parameter (0.04327), however, is undeniably significant and indicative of host specificity. All of the other parameters both for the 2233 and the T557 dives had significant values of 0 indicating variation within individuals and, for the T557 dive, between individual/ within a host type (table 4).

For the three tiered AMOVA analyzing the relationship between dive at 21N and symbiont haplotype, the between dive component of the test was effectively zero, though the test reported a score of -1.78% in order to make the other percents sum to 100. The between individuals within a dive parameter made up 42.85% of the variation observed and the within individual parameter accounted for 58.93 percent. These two values had significant P scores of 0, indicating that individual vestimentiferans have significantly different amounts of particular

symbiont types and that individuals within a single dive have significantly different ratios of those types. The P value for the between dive parameter was 0.45584, which definitively indicates that over the geographic distance separating the dives examined there was no variation in total clump symbiont composition (table 5).

AMOVA (dives 2233, 2230 & 557)

Source of Variation	Percentage of variation	Probability
Between dives	-1.78 (effectively 0)	0.45584
Between individuals within a dive	42.85	0.00000
Within individuals	58.93	0.00000

Figure 5: AMOVA results for the between 2233, 2230 and T557 dives (significant values are bold blue and insignificant values are in black)

DISCUSSION

In the study performed by Robbie Young and Steve Hallam in 2001, two haplotypes, A and D were observed at 21N (Young, 2001 unpublished). The use of cloning for this 2004 study revealed nine individuals with the C haplotype, 39 symbionts that were type A with a singleton, 27 symbionts that were type D with a singleton, and 2 symbionts that were type C with a singleton (figure 1). It is entirely possible that some of these singletons were due to Taq error caused by the lack of exonuclease activity in Promega Taq, though a miss-incorporation rate of 1×10^{-4} could not have accounted for all of the singletons observed. It is fairly certain that several of these single base pair mutations are real because they occurred more than once. Repeat singletons occurred nine times off of the A haplotype, zero times off of the C haplotype, which is not surprising because there were only nine C clones observed in total, and two times off of the D haplotype (figure 1). Any way you look at it cloning allows scientists to see more of the haplotype variation within a particular location or individual. Because cloning requires a second

PCR reaction to amplify gene inserts from their bacterial vector, however, it is more likely to be subject to Taq error. This needs to be kept in mind when analyzing the variation observed.

There are other problems with cloning besides Taq error that need to be considered in any discussion of molecular method choice. First of all, trying to avoid PCR bias by cloning does not work because, as mentioned above, cloning requires PCR. If PCR is so biased for one haplotype of a gene that it is not copied at all, neither cloning nor direct sequencing will be able to detect that haplotype. Cloning, however, does have the advantage of detecting things that occur at low frequency. This means that a haplotype PCR does not amplify well might be picked up through cloning, but not direct sequencing. Cloning, however, is really a probability problem. When you have certain haplotype frequencies that translate into bacterial colonies on a plate it is entirely possible to randomly pick colonies and never see the less frequent haplotype. It is also possible to over represent this less frequent strain simply by chance. For this reason sample size is very important. In this experiment ten colonies were picked as the standard sample size, though it is hard to show whether or not this was adequately large to correctly represent haplotype frequency.

Although cloning is not perfect it is absolutely necessary in any study involving hosts that acquire their symbionts from the environment, because of the potential for acquisition of multiple strains. Little is currently known about what causes tubeworm hosts to uptake symbionts, when they do it, and if they do it more than once, but these symbionts are never found in larval tubeworms so they must be horizontally acquired (Jones, 1988 # 3422; Southward, 1988 # 2988; Feldman, 1997 # 3185). Because this 21 N study indicates that tubeworms are a mixed bag of bacteria at least 62% of the time, direct sequencing under-represents the diversity there. In particular, because the observed frequencies of haplotype D and C were low, it would have been easy to dismiss the peaks distinguishing them from haplotype A as noise. This qualitative observation is supported by quantitative numbers (peak areas and clone frequencies), which suggest that only 76 percent of the symbiont assemblages observed at 21 N can be explained by a linear correlation between clone frequency and direct sequencing. Even this particular correlation coefficient is questionably high, however, because approximately 1/3 of the data it uses concerns vestimentiferans where clone frequency information and direct sequencing agree there is only one haplotype present. It is well documented that direct

sequencing performs well when it is used for organisms containing only one genetic trace for a particular gene. The question proposed here really concerns individuals with multiple infections, and thus it seems appropriate to analyze a data set composed of this kind of individual. Because 2/3rds of the vestimentiferan population at 21N exhibits double or triple symbiont conditions, not only is it possible to say that direct sequencing only represents accurate haplotype frequencies 38 percent of the time for this sort of individual, but also that it is misrepresenting the majority of the population. Direct sequencing should not be used as a method for haplotype identification for vestimentiferan symbionts.

The list of information concerning problems that arise when direct sequencing is used in vestimentiferan symbiont haplotype enumeration just goes on. First and foremost, it is undeniable that there is a discrepancy between what the sym6 and sym7 primers used in this method incorporate at both the G/A and G/T polymorphic sites. This may be due to some sort of primer bias, where for some reason the sym6 primer has less trouble amplifying all haplotypes and sym7 preferentially amplifies D. This seems odd because the two haplotype strains only differ in two places. There may, however, be a directional component to it, where the sequence of bases before or after the nucleotide of interest affect the way the primer adds big dye during sequencing. It is also possible that these primers contain m13F and R primers mixed within them and that these primer mixtures are somehow biasing ITS sequencing. The directional component, discussed first, may also play a part in a second reason not to use direct sequencing, the fact that cloning and this method are less correlated for the G/A site than the G/T site. This particular problem was significant enough to be picked up by a t score, indicating that the probability of correlation at the G/A site was less than 5 percent. Obviously direct sequencing can no more be trusted to detect the same haplotype frequencies across polymorphic sites than it can be trusted to determine equal frequencies with both of its primers. When a method is biased in so many different ways it becomes hard to use with any expectation of accuracy.

A third reason not to trust direct sequencing where multiple infections are probable is that, for one individual vestimentiferan, different primers and different sites reported disparate haplotype frequencies. This test however, indicated that direct sequencing was biased in different directions than those discussed above. One reason this may have been the case concerns sample size. The individual study was only done with four trials only three of which could be used for

the G/A site, where there appeared to be less variation (figure 6). Furthermore, in the sym6 category for the G/T site there was one sequence that only showed an A haplotype, where the others showed D and A (figure 7). This one very variable reading, although a real direct sequencing error, may have over-represented the general error range at that site with regards to the small sample size available. Support for this is seen both by the fact that the sym7 primer for that site is more accurate than the sym 7 primer for the G/A site and the results when this individual is removed entirely, which show the average value for G/A becoming less accurate than the value for G/T. Above and beyond this issue of the direction sequencing inconsistencies take, it is important to note that they exist. The ranges for the data examined both site to site and primer to primer are large, all of which shows that direct sequencing does not repeatedly produce identical sequence traces. Any result that is not repeatable must be questioned.

Clone frequency information, far beyond its use in the studies above to examine the reliability of direct sequencing, can provide valuable information regarding the field of symbiont population genetics. The between host AMOVA values based on clone data for both the 2233 and the T557 dives indicate that for some reason *Riftia* and *Oasisia* are acquiring different symbiont compositions. Observations of raw data suggest that *Riftia* accumulates little of the A haplotype relative to *Oasisia*, while *Oasisia* is more likely to exhibit high levels of two haplotypes. This trend changed for the 2230 dive, which only obtained *Riftia* samples. Here, one *Riftia* individual contained only A, and two had mixed compositions of A and D (figure 2). It appears that without *Oasisia* present *Riftia* acquire a more diverse variety of haplotypes, though this may prove to be untrue as further data is collected.

Because the biological meaning of the differences between haplotype A, C and D is unknown, it is hard to ask the pertinent question; why would a vestimentiferan select a particular symbiont strain over another? It is possible that changes in ITS are correlated with changes in another gene that make particular symbionts better for tubeworms that settle in areas with specific kinds of substrate, but it is equally plausible to state that symbionts are actually selecting hosts based on their own sets of biological preferences. The significant (2233) and marginally significant (T557) values obtained for *Riftia* and *Oasisia* between host AMOVA studies, though they cannot answer this kind of question, do suggest that such questioning is necessary. Hypotheses cannot be formed if a point of interest has not been identified.

Beyond the study of between host P values, analytical comparisons between the variance scores for the 557 and 2233 AMOVA tests provide valuable information concerning differences between the dives. For the T557 dive the largest portion of observed variation was due to the differences between individual worms within a particular host type. This may be due to the fact that the majority of C observed at 21 North was in the *Riftia* of this dive. The *Riftia* that didn't have A & C were solely A or solely D. The fact that these very different symbiont assemblages were present in relatively equal numbers likely led the between individual within host parameter to be large. Also contributing to this factor for the dive would be the fact that the *Oasisia* sampled were either all D or mostly A with some D. These symbiont groupings were also present in relatively equal amounts and would have made a clean split regarding between individual variation for this dive (figure 3). The second largest percentage of variation for T557 was the within individual parameter, which reflects the fact that symbiont haplotypes were not consistently 50/50 within an individual host. This may be a result of uneven bacterial composition in the environment, or, as indicated by the fact that some individuals only contained rare haplotypes, it may be a result of host specificity.

Unlike the T557 dive, the majority of the variation within the 2233 dive was attributed to the within individual parameter. Only a small portion suggested that variation between individuals within a host type was important, likely a result of the fact that all of the *Riftia* for the dive contained only the A haplotype and 2/3rds of the *Oasisia* were AD splits. The high value for the within individual parameter seen for this dive is also mainly due to *Riftia* individuals, which because they were only A, exhibited extremely high values for 1 haplotype and low values for the other two.

The second AMOVA test performed using this data, the one that examined between dive variation in the attempt to quantify differences in vestimentiferan symbiont composition over short distances, assumed each dive only sampled one tubeworm clump. For the 2230 and 2233 dives this is not really known, so any conclusions drawn from this data must be used cautiously. It does seem fairly apparent from the high P value for the between dive parameter, however, that the dives are not different. Because sampling of a tubeworm bush destroys the bush it is highly unlikely that the 2230 and 2233 dives re-sampled the same area. If the vestimentiferans for each

dive were not from the same location, they probably came from a larger geographic range than this test assumes they did. If this was the case, then it is still accurate to say that over the distances sampled at 21N, tubeworm symbiont composition remained the same.

The uniformity of symbiont distribution between vestimentiferan patches at 21N is surprising in light of the changes in symbiont composition found within the cold seep tubeworm *Lamellabrachia* over distances as short as 3 meters. It may be the case that hydrothermal vent symbiont distributions are much more uniform than those at cold seeps, though between vent studies (covering larger distances) would have to be performed to test this hypothesis.

Though it provides very valuable information concerning the between dive parameter, the AMOVA test just discussed is not as complete as it could be. It collapses the between host within a dive and between individual within a host variables into a between individual within a dive parameter, thus making it difficult to tell what a significant P score for this parameter means. In order to break this component down into its parts, it would be necessary to construct a four-tiered AMOVA test, which there is currently no format for. Just by observing the data, however, it appears that *Oasisia* across dives are relatively the same, while *Riftia* vary considerably. What affect this actually has on the between host within a dive or between individual within a host variables remains to be seen, though it is likely that the between host factor will make up a large percentage of the variation.

CONCLUSIONS/RECOMMENDATIONS

In light of the small percentage of symbiont haplotype frequencies that can accurately be detected by direct sequencing I would advocate cloning in any further vestimentiferan study. For this particular experiment I believe it would be beneficial to sequence ten clones from ten individuals at each of the nine remaining sites along the Eastern Pacific Rise. The use of cloning would allow the observation of more genetic variation, with increased reliability for individuals that contain multiple symbionts. The extensive examination of the EPR is advised because the samples are already available, the examination of 21N only tested for localized geographic variation, and *Tevnia* can only be sampled below 13N. The inclusion of *Tevnia* in this study would also allow for further, more rigorous, host specificity tests. Any site for which all three

vestimentiferans are available would be particularly relevant in this particular line of questioning.

On a completely different track, I would advise that this study take on another gene of interest to be used in consort with ITS. The 2001 study by Steve Hallam and Robbie Young used *fliC*, which is slightly more variable than ITS, and could thus add detail to any study of symbiont population genetics performed using this gene (Young, 2001 unpublished). An alternative to *fliC* would be *RuBisCo*, which was used in 2003 for the cold seep tubeworm *Lamellabrachia* (Duhaime, 2003 unpublished). Because variation over short distances for hydrothermal vent symbionts seems so different from cold seep symbionts, it might be advisable to gather data on *RuBisCo* for hydrothermal vents so that a true comparison study can be made.

ACKNOWLEDGEMENTS

I would like to thank Joe Jones for his help answering my endless questions as well as for his work extracting the DNA's used for the T557 dive. Robbie Young offered his expertise in statistical methodology, Shannon Johnson dissected the majority of the tubeworms, and Bob Vrijenhoek offered valuable advice both with statistics and population genetics. Without these people, the pilots of the Alvin and Tiberon as well as the David and Lucile Packard Foundation, this study would not have been possible. I would also like to send a very special thank you out to George Matsumoto, without whom this internship would not have been possible. Thank you all very much.

REFERENCES CITED

- Black, M. W., K. M. Halanych, P. A. Y. Maas, W. R. Hoeh, J. Hashimoto, D. Desbruyeres, R. A. Lutz, and R. C. Vrijenhoek. 1997. Molecular systematics of vestimentiferan tube worms from hydrothermal vents and cold-water seeps. *Biology*. **130**:141-149
- Black, M. W., A. Trivedi, P. Maas, R. A. Lutz, and R. C. Vrijenhoek. 1998. Population genetics and biogeography of vestimentiferan tube worms. *Deep Sea Res. II* **45**:365-382.
- Cary, S. C., and S. J. Giovannoni. 1993. Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. *pnas* **90**:5695-5699.
- DiMeo, C. A., A. E. Wo;bir, W. E. Holben, R. A. Feldman, R. C. Vrijenhoek, and S. C. Cary. 2000. Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. *Aem*. **66**:651-658.
- Duhaime, M., J. Jones, and R. C. Vrijenhoek. 2003 unpublished. Phylogenetics of Vestimentiferan Symbionts from Guaymas Basin Using the 16s and RuBisCO Genes.
- Feldman, R. A., M. B. Black, C. S. Cary, R. A. Lutz, and R. C. Vrijenhoek. 1997. Molecular phylogenetics of bacterial endosymbionts and their vestimentiferan hosts. *mmbb* **6**:268-277.
- Hurtado, L. A., M. Mateos, R. A. Lutz, and R. C. Vrijenhoek. 2002, Molecular evidence for multiple species of *Oasisia* (Annelida: Siboglinidae) at eastern Pacific hydrothermal vents. *Cahiers de Biologie Marine*. **34**:277-380.
- Hurtado, L. A., M. Mateos, R. A. Lutz, and R. C. Vrijenhoek. 2003, Coupling of bacterial endosymbiont and hostmitochondrial genomes in the hydrothermal vent clam *Calypptogena magnifica*. *Applied Environmental Microbiology*. **69**:2058-2064.
- Hurtado, L. A., R. A. Lutz, and R. C. Vrijenhoek. 2004 in review. Distinct patterns of genetic differentiation among annelids of eastern pacific hydrothermal vents. *Molecular Ecology*: submitted 4/16/04
- Jones, M. L., and S. L. Gardiner. 1988. Evidence for a transient digestive tract in Vestimentifera. *pbsw* **101**:423-433.
- Southward, E. C. 1988. Development of the gut and segmentation of newly settled stages of *Ridgeia* (Vestimentifera): implications for relationship between Vestimentifera and Pogonophora. *jmba* **68**:465-487.
- Van Dover, C. L. 2000. The Ecology of Deep-Sea Hydrothermal Vents. *Princeton University Press*, Princeton.

Won, Y., S. J. Hallam, G. D. O'Mullan, and R. C. Vrijenhoek. 2003a. Cytonuclear disequilibrium in a hybrid zone involving deep-sea hydrothermal vent mussels of the genus *Bathymodiolus*. *Molecular Ecology* **12**:3185-3190.

Young, R., S. Hallam, and R. C. Vrijenhoek. 2001 unpublished. Hierarchical Population Structure of Bacterial Endosymbionts in Co-Distributed Species of Hydrothermal Vent Tubeworms.