The dissertation of Francis Condon Wham Jr. was reviewed and approved* by the following:

Todd C. LaJeunesse  
Associate Professor of Biology  
The Pennsylvania State University  
Dissertation Advisor

Stephen W. Schaeffer  
Professor of Biology  
Assistant Department Head for Graduate Education  
The Pennsylvania State University  
Chair of Committee

Iliana B. Baums  
Associate Professor of Biology  
The Pennsylvania State University

George Perry  
Assistant Professor of Anthropology and Biology  
The Pennsylvania State University

Tracy Langkilde  
Head of the Department of Biology  
The Pennsylvania State University

*Signatures are on file in the Graduate School
ABSTRACT

In light of modern genetic techniques, there is growing awareness that taxonomic designations do not reflect biological reality. When hidden species boundaries go undetected they cause spurious conclusions about biology. Here I discuss several new analytical approaches useful for finding and making inferences from clusters in genetic data. I then apply these techniques to the results of several recent studies and highlight how critical errors in interpretation might have been avoided by first resolving species boundaries. I then show that these approaches are an important first step in connectivity studies. Finally, I highlight the critical role they have played in justifying the taxonomic status of several new species of Symbiodinium.
# TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. vi
LIST OF TABLES ..................................................................................................................... x
ACKNOWLEDGEMENTS .......................................................................................................... xi

Chapter 1: ON THE ORIGIN OF CLUSTERS IN POPULATION GENETIC DATA .......... 1
References ............................................................................................................................... 21

Chapter 2: EVALUATING THE SPECIES HYPOTHESIS WITH A BAYESIAN
SYMPATRIC STRUCTURE TEST IN ZOANTHUS SANSIBARICUS AND ITS
SYMBIONTS ................................................. 23
Abstract ................................................................................................................................. 23
Introduction ............................................................................................................................ 23
Methods ................................................................................................................................. 26
Results ................................................................................................................................... 32
Discussion ............................................................................................................................ 36
Conclusion ............................................................................................................................ 44
References ............................................................................................................................ 48

Chapter 3: SYMBIODINIUM POPULATION GENETICS: TESTING FOR SPECIES
BOUNDARIES AND ANALYZING SAMPLES WITH MIXED GENOTYPES ............... 54
Abstract ................................................................................................................................. 54
Introduction ............................................................................................................................ 55
Methods ................................................................................................................................. 58
Results ................................................................................................................................... 66
Discussion ............................................................................................................................ 75
Conclusion ............................................................................................................................ 85
References ............................................................................................................................ 87

Chapter 4: THE COALESCENT THEORY OF CLONAL ORGANISMS .................. 96
Abstract ................................................................................................................................. 96
Introduction ............................................................................................................................ 97
Methods ................................................................................................................................. 98
Results .................................................................................................................................. 113
Discussion ........................................................................................................................... 122
Conclusion ........................................................................................................................... 133
References ........................................................................................................................... 135

Chapter 5: MICROSATELLITE LOCI FOR THE HOST-GENERALIST
“ZOOXANTHELLA” SYMBIODINIUM TRENCHI AND OTHER CLADE D
SYMBIODINIUM ........................................... 140
Abstract ............................................................................................................................... 140
References ............................................................................................................................ 146

Chapter 6: MICROSATELLITE LOCI FOR SYMBIODINIUM GOREAUI AND OTHER CLADE C
SYMBIODINIUM ........................................... 151
Abstract ............................................................................................................................... 151
Introduction .......................................................................................................................... 153
Methods ............................................................................................................................... 155
LIST OF FIGURES

Figure 1-1: Figure 1-1) Pairwise genetic distance and Pairwise genetic difference of Symbiodinium genotypes. a) Pairwise genetic distance of Symbiodinium genotypes previously published in Wirshing et al. (2013), arrows point to clonal individuals on the left and unique individuals on the right b) Pairwise genetic difference of Symbiodinium trenchii genotypes previously published in Pettay et al. (2015), arrows point to clonal individuals on the left and unique individuals on the right .........................8

Figure 1-2: Expected results of flipping a fair coin n-times modeled with a continuous beta distribution .................................................. 14

Figure 1-3: Expected frequency of pairwise relationships modeled with the dirichlet distribution .................................................................................. 16

Figure 1-4: Pairwise relationship between individuals in a forward time simulation modeling divergence, sex, clonality and mutation. Histogram colored by the relationship between pairwise comparisons, members of different species shown in red, members of the same species in green, identical individual in the same clone family in blue and members of the same clone family with a single mutation event between them colored in orange. ........................................................................................................18

Figure 2-1: Simplified depiction of experimental design, where the effect of geographic distance is compared to the effect of putative reproductive barriers to assess the species hypothesis ................................................ 29

Figure 2-2: Bayesian sympatric structure test approach to use a region of population equivalence to accept or reject the hypothesis of species .................. 30

Figure 2-3: Results of STRUCTURE analysis on multi-locus genotypes of Z. sansibaricus and their associated clade C symbiodinum. A) k=3 for MLGs of Z. sansibaricus, notice the even probability of all individuals to all clusters b) k=2 for MLGs of Z. sansibaricus, notice the high probability of membership of many individuals to each cluster. C) k=2 individuals plotted by depth at Manza D) k=2 individuals plotted by depth at Tenya, E) k=2 for the clade C symbiodinum at Manza segregated my depth. F) k=2 for the clade C Symbiodinium at Tenyia. G) results of STRUCTURE analysis at k=2 for clade C Symbiodinium subdivided by location ............................................ 33
Figure 2-4: Bayesian analysis for the significance of divergence between sympatric individuals applied to *Z. sansibaricus*. A) the posterior density distribution of $F_{st}$ between individuals assigned to the same cluster across locations. B) the region of population equivalence C) the posterior distribution of $F_{st}$ between sympatric individuals assigned to the same cluster...

Figure 2-5: Bayesian analysis for the significance of divergence between sympatric individuals applied to the clade C *Symbiodinium* occurring in *Z. sansibaricus*. A) the posterior density distribution of $F_{st}$ between individuals assigned to the same cluster across locations. B) the region of population equivalence C) the posterior distribution of $F_{st}$ between sympatric individuals assigned to the same cluster..........................34

Figure 3-1: Flow chart describing the process of acquiring, parsing, and analyzing multilocus data from sample sets potentially containing two or more species.....61

Figure 3-2: Population genetic analyses of simulated data from homogenous and mixed genotype samples. a) Structure plot showing unambiguous multi-locus genotypes (MLGs) that resolve two distinct populations, blue and orange, at k=2. The section with black bars refers to samples with mixed genotypes. b) Analyses of allele frequencies at 3 example loci show bimodal distributions that correspond, and are color coordinated, to evolutionarily distinct populations, or species, with obvious allelic divergence. These bimodal data are then used to parse individual genotypes out of samples with mixtures of two genotypes representing each species; c) Structure plot presented with additional MLGs recovered from mixed samples at k=2. (d and e) These data are then partitioned by species for subsequent population level analysis that examines for genetic structure over various spatial scales..............................67

Figure 3-3: Analysis of MLGs from two *Symbiodinium* species lineages that can co-occur in animals from the Great Barrier Reef and Zamami Island, Japan, in the Okinawan archipelago. The MLGs of *Symbiodinium goreaui* (C1), blue, and C3, yellow, are well resolved from each other in a STRUCTURE plot at k = 2. Several additional MLGs were obtained from corals with mixed populations comprising one genotype from each species by relying on the secondary analysis of relative allele frequencies diagnostic for each species (see Fig. 1 and text for explanation). These later analysis increased the proportion of MLGs for each species, which could then be used for independent population genetic analyses...............................69

Figure 3-4: Histograms for allele size distributions from data generated by Howells et al. (2013). Allele size histograms based on presence absence of allele calls recorded for all samples of *Acropora millepora* on the Great Barrier Reef (n =401 ). The bimodal distribution evident for several loci which indicate the existence of two genetically distinct *Symbiodinium* spp. in
their data set. A few loci (indicated by an asterisk) appear to only
successfully amplify one species based on the lower number of allele calls
over the same number of samples (1.94 and 2.04 alleles per sample compared
to an average of 2.97 alleles per sample across all other loci). Also note that
two loci contained allele sizes that theoretically could not occur at the locus
based on the minimum size calculated after removing the repeat motif (dark
grey shading) ................................................................. 71

Figure 3-5: The partitioning and reanalysis of microsatellite data from Wirshing et
al. (2013). a) Structure plots (k=6, similar to results published previously by
Wirshing et al. (2013), and the optimal k=2) of microsatellite data for
Symbiodinium obtained from the west tropical Atlantic gorgonian, Eunicea
flexuosa and a cryptic sibling species. k=2 indicates the existence of two
Symbiodinium species (blue and yellow) that co-occur at several locations
where host populations were sampled. These likely correspond to the high-
light and low-light adapted lineages proposed by Prada et al. (2014). b) 
Structure analysis of data that were processed further by the removal of
duplicated genotypes found at each location and the recovery of additional
genotypes from mixed samples by parsing alleles into distinct genotypes
[after analysis of allele frequency distributions (see Fig. 4)], which
correspond to each species. c) Bayesian analysis of the MLGs from the most
common species of Symbiodinium resolves weak population genetic
structuring, but support for this is limited by the number of loci assessed and
number of genotypes available from each location. Cluster assignments for
individual genotypes at k values of 1 and 2 are shown. d) Analyses of the
second and rarer species indicates the possibility of a second genetically
distinct entity, yet this dichotomy appears to be driven by the presence of
allele 180 at locus Sym211 in some individuals. Additional k values show no
further genetic structuring for this species .................................................. 73

Figure 3-6: Histograms for allele size distributions from both raw and processed
data generated by Wirshing et al. (2013). a) Allele size histograms based on
presence absence allele calls recorded for all samples (n = 610). b) Allele size
histograms based only on samples where a single and complete unambiguous
MLG was recovered (n =271 ). The bimodal distributions demark separate
size ranges which correspond to and support the supposition that two
genetically distinct Symbiodinium spp. exist in association with E. flexuosa
(sensu lato, Prada et al. 2014). Notice that some individuals assigned to the
yellow cluster have allele sizes associated with the purple cluster. Additional
analysis is needed to resolve this as evidence of introgression or simply mix
samples that only successfully amplified the alleles of one or the other
species ........................................................................................................ 74
Figure 3-7: Summation of allele divergence at all microsatellite loci for two putative species lineages not recognized by Wirshing et al. (2013), but proposed by Prada et al. (2014) predicated on their finding of DNA sequence divergence. a) A Neighbor Joining Tree based on Bruvo Distance comparisons shows a large break in genetic distance between species lineages. b) A histogram of pairwise Bruvo Distance showing intraspecific variation (black) is less than interspecific variation (grey). Identical clones are indicated by the white bar.

Figure 4-1: Performance of G/n, Beta and Psi, corresponding to elements A, B and C respectively. X-axis is the number of samples, y-axis is the estimated value of the summary statistic. Different colors represent different numbers of underlying unique individuals in a population of 5,000 individuals. Values calculated with a sample size of 4,500 can be assumed to be near the true value, if there was no bias we would expect to see a line with a slope of zero for all plots.

Figure 4-2: Performance of Psi and G/n on the Pocillopora damicornis dataset. Dotted line shows the value calculated on the full population set, the blue line is the fitted mean of the values for each sample size along the x-axis. The y-axis is the value calculated for the summary statistics in a) and b). The y-axis is the log error in c) and d).

Figure 4-3: Distributions of E[Ng] values estimated from sample sizes of 30, 50, and 100 samples from the Pocillopora damicornis data from Gorospe and Karl (2013). a) shows the E[Ng] values estimated using N_e from the 1st model from the Haasl and Payseur (2010); each distribution represents an independent subsample from the full dataset. b) shows the E[Ng] values estimated using the census size (N_c=2352) each distribution represents an independent subsample from the full dataset.

Figure 4-4: Psi and DPsi values for the Aspen data, Samples from Fish Lake are divided into understory (S) and overstory (B) collections and Swan Flats (USF) is below. Diploid and Triploid individuals in the same collection are plotted separately on each line and are labeled D and T in the legend.

Figure 4-5: Psi, E[Ng] and DPsi values for the Palasmodium dataset. Locations marked on the y-axis and values marked on the x-axis. Sample periods plotted on the same line and marked as T1 (before transmission decline) and T2 (after transmission decline). Jagged shape of DPsi plot of MLA T1 is due to the low number of Genotype Configurations that are likely under the Psi value observed which produces some rather discreet values to populate the posterior.
LIST OF TABLES

Table 4-1: Psi and DPsi values calculated for the full dataset and E[Ng] and Se values calculated with all 5 Ne models

Table 5-1: Locus name, repeat motif, primer sequences, annealing temperature (T_a) and approximate size range of expected product for nine microsatellite loci for *Symbiodinium* Clade D

Table 5-2: An analysis of 3 *Symbiodinium* Clade D types. We report the allelic size range for each *Symbiodinium* Clade D type, the number of alleles (N_A) and effective number of alleles (A_e), within each population.

Table 6-1: Locus name, repeat motif, primer sequences, annealing temperature (T_a) and approximate size range of expected product for nine microsatellite loci for *Zoanthus sansibaricus*

Table 6-2: Number of alleles, Observed and expected heterozygosity for samples collected at Manza.

Table 6-3: Number of alleles, Observed and expected heterozygosity for samples collected at Teniya.

Table 7-1: Locus name, size of product repeat motif, primer sequences, annealing temperature (T_a) and GenBank accession number for eighteen microsatellite loci for *Symbiodinium* Clade C

Table 7-2: Size ranges of 4 clade C types, N is the number of samples genotyped and Na is the number of alleles observed.
ACKNOWLEDGEMENTS

I would like to thank my advisor Todd LaJeunesse for giving me the tremendous freedom to pursue my own interests. I am also in debt to Zach Fuller who helped me develop the Clonalescent. Without Zach’s help the simulations in this dissertation would still be running even now. I would also like to thank my friend and lab partner Allison Lewis who I enjoyed spending time with in the field and in the office for many years.
Ancora Imparo
Chapter 1
ON THE ORIGIN AND MEANING OF CLUSTERS IN POPULATION GENETIC DATA

Introduction

In this dissertation I seek to demonstrate that the existence of un-even relationships among individuals can be explained by the interaction between the relative rate of sexual recombination and biological forces such as selection, mutation, migration and drift. These un-evenly related individuals can be represented, or thought of, as clusters and the relationship between these clusters can be used to infer the relative rates of the events that form them. This aim is certainly not unique to this work. Rather, I share many of my aims with those who have analyzed population genetic data before me and I will often reach conclusions that will take the same or similar form as the conclusions of others. It is for this reason that I would like to begin here by asking “Are there unifying theories in biology?”

Certainly there are unifying ideas; evolution is one of the most important. I, however, mean “unifying theory” here somewhat differently; more in the sense that physicist have used the idea. That is to ask: is there a single mathematics that allows all of the forces usually thought of as fundamental to evolution to be written in a single form? Could we see the equations for migration equilibrium frequencies, expected heterozygosity, and recombination rates as special cases of a single process? In this essay I do not take the formal proof of this as my goal, rather I’d like to simply suggest that
there is evidence to support this belief and use the discussion as a vehicle to reflect on some ideas that will be fundamental to the conclusions I reach in later chapters.

**Tempering Expectations**

Let us begin here by examining what such a unified theory might look like if it did exist. In physics, it is generally accepted that a handful of forces explain and predict the movement of bodies. That is, the same laws govern the motion of planets as a ball thrown behind my desk. Obviously, a ball thrown behind my desk would be subject to a specific coefficient of friction between it and the floor, as well as variety of other specific effects. In fact, a precise model for a ball thrown behind my desk would be quite complicated and specific because the physical world is rich with complexity. This observation, however, would not be taken as evidence to refute the fundamental laws of motion. Similarly, the rich, complex nature of the biological world would suggest that an exact model of any part would require complicated and specific models. Nevertheless, when these complexities are stripped away there may perhaps be a few unifying features left behind that can be thought of as fundamental.

**From a null model...**

If we wish to know about the specific coefficient of friction between the ball and the floor the primary means of discovering it would be to calculate the deviation between the observed velocity of the ball and the velocity expected under the simplified model of physics formulated from its fundamental laws. Therefore, even when the assumptions of a simplified model are violated it is the simplified model which provides the key to understanding the parameters of the more complex model. It is for this reason that null genetic models have proven to be broadly useful even when their assumptions are
violated. It seems that if a unified theory could be formulated it would have to arise
naturally from a key aspect of a null genetic model. Here, we consider a key aspect of the
most common null model, the Wright-Fisher model. In particular, we consider the
process of birth, death and replacement.

...to a unified theory

The Wright-Fisher model is a simplified model of a population that consists of a
finite number of individuals (N) which is said to be “large” (we will examine the
meaning of this later). In this population model, the next generation is created by
sampling the current generation with replacement until N individuals have been sampled.
As a result of this sampling with replacement process some individuals have zero
offspring while others have one or more, even though the expectation is that every
individual would be sampled once and give rise to a single offspring. It follows that after
many iterations of this birth-death process any two individuals selected from the
population must share an ancestor at some point in the past. Importantly this results in the
process of ancestry because every time an individual is selected to reproduce twice in a
single generation the resulting offspring share an ancestor. This means that the entire
population can be represented on a single genealogy and hence this can be thought of as
the “genealogical process”. If there is room for a unifying theory in biology the
genealogical process seems like a likely mechanism. Is it possible that many of the most
interesting results in population genetics can be reached by simply framing the
genealogical process appropriately for the question?
Time to Ancestry

“In many important cases the survival for a single generation may be represented by a function simply related to the Poisson series, and in many other cases by the substitution for a generation of the appropriate biological cycle the same formulae will supply an excellent approximation.”

-R. A. Fisher (1930)

To consider this possibility let’s first evaluate a key aspect of the genealogical process: the expected number of generations backwards in time to reach a common ancestor of two randomly chosen individuals (expected time to coalescence). Returning to the sampling with replacement process, the probability of two individuals sharing a common ancestor in the immediately preceding generation is 1/N. That is, if we have two individuals B₁ and B₂ from a population of Bₐ individuals and we know that the ancestor of B₁ was A₁ from the previous generation which consisted of Aₐ individuals, then the probability that the ancestor of B₂ is also A₁ is simply equal to 1/N. It follows then that the expected number of generations we would have to wait until two random individuals shared an ancestor would be N iterations of that 1/N chance, since N(1/N)=1. It is for this reason that N is a convenient unit for keeping track of time in the context of genealogies, because it is the expected time until shared ancestry on a genealogy of two individuals. Therefore the contributions of a single individual to the next generation, or the “survival” of its genes, can be represented as a Poisson distribution with rate = N/N=1. The expected time to a shared ancestor of two random individuals is exponentially distributed with rate = N, which if we are comfortable with expressing our results in terms of N the exponential distribution rate is also 1.
The symmetrical convenience of this result cannot be overstated, but its significance may not be apparent until we consider how we can use this symmetry to understand other events of interest that occur on a genealogy. Before we do this, however, it may be useful to discuss the relationship between the exponential distribution and the gamma and beta distributions. The exponential distribution is simply a special case of the gamma distribution. The exponential distribution describes the waiting time to the first event given a rate $\lambda$. The gamma distribution alternatively describes the waiting time until the $i$th event with notation $\text{gamma}(\lambda, i)$ it follows that the exponential distribution can be written as $\text{gamma}(\lambda, 1)$. The beta distribution then describes two independent gamma distributed variables such that if $\alpha = \text{gamma}(\lambda_1, 1)$ and $\beta = \text{gamma}(\lambda_2, 1)$ then $\text{Beta}(\alpha, \beta)$ represents the distribution of times that event 1 happens first. Now let us consider how these results allow us to frame questions about the identity of individuals in a population. Consider a process that occurs at rate $\lambda_j$ which permanently changes the color of an organism and its subsequent offspring to a new unique color. What is the probability of sampling two differently colored individuals from a population? This is analogous to asking: “What is the probability that event $j$ happens before the two individuals share a common ancestor. Since we have already shown that the time to a common ancestor is equal to $\text{gamma}(N, 1)$, it follows that if we can express $\lambda_j$ in the same time unit then we could use the beta distribution to arrive at the solution.

The first step is to calculate the number of events of type $j$ that happen on the genealogy. If the expected time to a common ancestor is $N$ then the total expected time on the genealogy is $2N$ (1N up the genealogy to the common ancestor and 1N back
down the genealogy to the other individual). Therefore, both the coalescent rate and the event rate of $\lambda_j$ can be expressed in terms of the same time unit, if we allow the coalescent rate to be $N=1$, then the event rate on the same time scale is:

$$[1-1] \quad \theta_j = 2N\lambda_j$$

So we can evaluate $\text{beta}(\theta_j, 1)$ to get the frequency that event $j$ happens first. The mean of the beta distribution is:

$$[1-2] \quad \frac{\alpha}{\alpha + \beta}$$

Thus, the probability of sampling two individuals of a different color is $\theta/(\theta+1)$ mirroring the conclusions of Malacot (1948) regarding the probability of identity under the infinite allele model which is equivalent to evaluating the previous equation in terms of coalescing first which would be $\text{beta}(1, \theta)$ which has an expected value of $1/(1+\theta)$.

**It’s Just a Bunch of Beta Distributions**

Importantly, the equation $\theta=2N\lambda_j$ can be seen as a natural form of expressing a rate in terms of the length of a genealogy of two exchangeable units. Equations of this form have been recapitulated every time a new event needed to be mapped to the genealogical process. For example, the population migration rate is $M = 2Nm$ (Wright 1931), the population mutation rate is $\theta = 2N\mu$ (Kimura and Crow 1964) and the population recombination rate is $P=2Nr$ (McVean et al. 2002). The logic has even been extended to the study of biodiversity, where the fundamental biodiversity constant is $\theta =$
2Jv (Hubbell 2001) where J is the number of exchangeable individuals in a landscape and \( v \) is the speciation rate. Here, I suggest the view that the form \( \theta = 2N\lambda_j \) is fundamental to mapping events to genealogies, where \( \theta \) = the population event rate and \( \lambda_j \) = the per capita, per generation event rate. This view suggests that the population migration rate, the population mutation rate and the population recombination rate are mathematically equivalent. Indeed this view not only implies the solution to the population migration rate, the population mutation rate and the population recombination rate it also implies the solution to yet unstudied genealogical events. The converse, however, can not be shown to be true. I believe this is equivalent to saying that all of these processes are actually special cases of a more general case. If this is correct then perhaps this view can be used to shed light on some patterns observed in data.

**More Betas, More Problems**

Figure 1-1 shows some of the patterns I aim to explain later in this dissertation. Figure 1-1a shows the distribution of pairwise genetic distances between individual *Symbiodinium* genotypes from microsatellite data previously published by Wirshing et al. (2013). Figure 1-1b shows the number of pairwise differences in *Symbiodinium trenchii* from Thailand from microsatellite data previously published by Pettay et al. (2015). *Symbiodinium* are single celled dinoflagellates that reproduces both sexually and asexually. When large numbers of samples are collected it is relatively common to find individuals that are identical to one another. These identical individuals have a genetic distance (Fig. 1-1a) or pairwise distance (Fig. 1-1b) of zero. In Chapter 4 of this dissertation I use the above logic to propose the “population sex rate”. Since the probability of sampling two identical individuals should be equal to the probability of
Figure 1-1) Pairwise genetic distance and Pairwise genetic difference of *Symbiodinium* genotypes. a) Pairwise genetic distance of *Symbiodinium* genotypes previously published in Wirshing et al. (2013), arrows point to clonal individuals on the left and unique individuals on the right b) Pairwise genetic difference of *Symbiodinium trenchii* genotypes previously published in Pettay et al. (2015), arrows point to clonal individuals on the left and unique individuals on the right.
two individuals coalescing before a sex event. I propose the population sex rate is \( \psi = 2N_s \) (Chapter 4 of this dissertation), thus the probability of selecting two identical clones is \( \frac{1}{1 + \psi} \), using equation [1-2], or beta(1, \( \psi \)). Perhaps, even more complex questions can be answered in this way.

In Figure 1-1b, some individuals are likely members of the same clone family but are not identical. That is, they share the same allele at 23 of the 24 loci examined, which is explained by a single mutation event that has occurred on the genealogy of two clones. This means that the first event on the genealogy of some of the individuals that coalesce prior to a sex event is a mutation event. This population is composed of individuals with one of three pairwise relationships, 1) unique individuals that share a sex event prior to coalescing, 2) identical individuals that coalesce prior to a sex event or a mutation event and 3) individuals that have a mutation event prior to a coalescent event. In solving this we will have to expand the beta distribution to accommodate more gamma(\( \lambda_j, 1 \)) processes since we must account for the processes of mutation, sex and coalescence. The dirichlet distribution is the multinomial equivalent of the beta distribution where dir(\( \lambda_1, \lambda_2, \ldots, \lambda_j \)) accommodates any number of events, so beta(\( \lambda, 1 \)) = dir(\( \lambda, 1 \)). Again, we would like our results in units of \( N \) so expanding our analysis to accommodate any \( \lambda_1, \lambda_2, \ldots, \lambda_j \) processes we use equation [1-1] again to get:

\[
[1-2] \quad \text{dir}(\theta_1, \theta_2, \ldots, \theta_j, 1)
\]

Where the final term is 1 because we will express the other values in units of the coalescent rate. In the present case, we can first solve for the proportion of clonemates that will have a mutation event between them:
[1-3] 

\[ E[\text{beta}(24 \theta, 1)] = \frac{24+\theta}{(24+\theta)+1} \]

We can also solve for the proportion of clonemates that do not have a mutation event between them:

[1-4] 

\[ E[\text{beta}(1,24 \theta)] = \frac{1}{1+(24+\theta)} \]

Again using equation [1-2] to solve for both terms, then we use the dirichlet distribution to solve for the probability of two individuals falling into each of the 3 classes:

[1-5] 

\[ \text{dir}(\psi, \frac{24+\theta}{(24+\theta)+1}, \frac{1}{1+(24+\theta)}) \]

This returns a 3-dimensional probability distribution giving the probability of two individuals being in each of the three classes. Here we have to use the beta distribution to solve for the probability of sampling two clonemates that will have a mutation event, then coalesce after that, prior to a sex event since the individuals that have a mutation event first and then a sex event after simply appear as unique individuals. Thus the last two terms in equation [1-5] sum to 1 as in equation [1-2].

**When do clusters exist?**

In the last section I suggested that any number of processes can be mapped on to the genealogical process by first expressing the rates of those processes in terms of the rate of coalescence. This is important because it allows us to infer the rates of processes from data we have collected. Applying inference requires several criteria to be met. In the
last example, we could have solved for the pairwise probability of sampling individuals with sex events, mutation events and coalescent events as their last event with \( \text{dir}(\psi, \theta, 1) \). I mentioned above, however, that we are unable to see all individuals that have mutation events as their last event since some portion occur on individuals that are unique in our dataset so this calculation is not useful. This demonstrates the first of three important aspects of making inferences from clusters about their formative processes: A1) a cluster formative process must have an observable effect on the gathered data in order for us to estimate it. In the present case, observing the state of identity or non-identity is straightforward, however, we may also wish to consider processes that result in cluster formation where the membership of individuals to a particular class can be seen in terms of partial or probabilistic membership. For now we will restrict the discussion to discretely observable states. We will discuss the partial or probabilistic membership case in greater depth in chapter 2 of this dissertation.

**Sample Variance**

The second aspect is: A2) the sample size must be large enough to observe members of both clusters. Assuming that an individual that results from each separate \( \theta_i \) process in \( \text{dir}(\theta_1, \theta_2, \ldots, \theta_j) \) is recognizable if two individuals that contain that event between them are in the dataset then we can set Ewens’ (1972) equation for the number of alleles:

\[
\hat{\pi}_a = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \ldots + \frac{\theta}{\theta+N-1}
\]
equal to two and evaluate $N$ for each value of $\theta_j$. Evaluating this equation in this context provides the expected number of samples to find a member of class $j$. This will result in an expected number of samples for each $\theta_j$ ($E[n_j|\theta_j]$).

**Genealogical Variance**

"Even if we were to census the entire population, our results are still affected by "sampling" in that the particular population sampled is but one of many possible replicates that could have arisen under the same conditions."


The third important aspect of cluster formation I will discuss here is that A3) the rate of the cluster formative process ($\theta_j$) must be large enough that more than 1 event can be expected to occur on the genealogy of the population but not large enough that it is expected to occur between every pair of individuals. This is equivalent to saying that

$$2 \ll E[n_j|\theta_j] \ll N,$$

where $N$ is the effective number of individuals in the population. If a process occurs too often or too infrequently relative to the rate of coalescence then the realized probability of sampling two individuals of that class will not be representative of the actual underlying probability.

To understand this better lets step away from genealogies momentarily. Consider two people with an inclination to wager on the results of flipping a fair coin. Figure 1-2 shows the expected results after 1,2,4 and 8 games in terms of the probability density of the percentage of wins for player A. I encourage readers here to compare Figure 1-2 to Wright’s Figure 6 (1931, p127). It should make sense intuitively that after just 1 game player A has either won or lost 100% of the games. Since the probability of winning a fair coin toss is 0.5 it requires two or more games before we expect to observe both a win
and a loss and it is only after we have observed at least one win and one loss that we can confirm that both the state of winning and losing are possible and calculate a non-zero probability for both states. Laplace addressed this issue in his analysis of “the sunrise problem” (Laplace 1814) and we will use his result to address this very issue with respect to genealogies in chapter 4 of this dissertation.

Returning to a discussion of mapping events onto genealogies, if an event is so rare that we cannot expect it to occur several times on the genealogy of the entire population or the event is so common that it is expected to occur between nearly every pair of individuals we should be concerned about genealogical variance. This means that it is possible for a series of $\theta_1, \theta_2, \ldots, \theta_j$ cluster formative processes to be acting on a genealogy where some processes create clusters which are meaningful for inference while others do not make clusters or if they do the clusters are not meaningful for inference. Eldon and Wakely (2006, 2008 and 2009) considered a case similar to this where a large well mixed population is subdivided into multiple sub-populations that can coalesce rather quickly compared to the rate of migration, even though the absolute rate of migration is high. Eldon and Wakely (2009) show that these sub-population quickly show large significant values for genetic differentiation which would lead a naive observer to conclude that the sup-populations are migration limited. I believe that this is equivalent to my explanation above where the authors have increased the coalescent rate such that the migration rate decreased relative to it so that $2 > E[n_m | m]$ or $E[n_m | m] > N$. Semantically, this means that the genealogy of a subpopulation is so short that at any one time it is either entirely composed of migrant alleles or contains no migrant alleles at
Figure 1-2) Expected results of flipping a fair coin n-times modeled with a continuous beta distribution.
all. So much so that the “migrant” and “native” status of individuals in this model is not meaningful.

**But does it work?**

Let's consider a population of 2,000 individuals with 2 groups of individuals (A and B) with a reproductive barrier between them. For simplicity let us assume that the reproductive barrier is strong, but not complete and equal in both directions so that if we represent the barriers as $\alpha_A$ and $\beta_B$ we can say $\alpha_A = \beta_B$, so we can simplify the effects of the reproductive barrier because $\alpha_A / (\alpha_A + \beta_B) = \beta_B / (\beta_B + \alpha_A) = \frac{1}{2}$. We will consider the case where both A and B individuals reproduce sexually with probability $s=0.1$ every generation and the mutation rate is $\mu=5 \times 10^{-6}$ per generation per locus. Here we want to evaluate the probability of sampling 2 individuals and their relationship as being one of four states: 1) unique members of the same group, 2) members of different groups, 3) identical individuals or 4) members of the same clone family different by a single mutation at 1 of 10 loci. The resulting parameterization of the Dirichlet distribution in that order is:

$$dir\left(\frac{1}{2} 2Ns, \frac{1}{2} 2Ns, \frac{1}{1+20N\mu}, \frac{20N\mu}{20N\mu+1}\right)$$

Figure 1-3 shows four probability density distributions each giving the probability of sampling two individuals from this population and those individuals having one of the above four relationships. Notice that we have specified that the sample is of size two and so the variance in Figure 1-3 is not the sample variance rather it is the variance that was discussed in the last section, the genealogical variance. Another way of thinking about
Figure 1-3) Expected frequency of pairwise relationships modeled with the Dirichlet distribution for a) members of the same group b) members of different groups c) identical clonemates and d) clone mates with a single mutation.
this is that Figure 1-3 shows, for each of these processes, the range of values the population’s genealogy would take over time. That is, if you sampled the population many times, over time periods that were sufficiently spread apart, Figure 1-3 shows the range of values you would observe. With reference to (A3), notice the width of the variance of identical individuals (Fig. 1-3c) and individuals with one mutation difference (Fig. 1-3d). Even though both of these have a positive expected value both can take on a number of very small values such that $E[n_j| \theta_j] > N$ which means that at any one time the population may have zero individuals of this type and additional sampling of this population at the same time point would only reinforce an incorrect conclusion about its rate.

Figure 1-4 approaches the same problem in an entirely different direction. Here I have used an individual based forward time simulation in the program *rmetasim* (Strand 2002) with the above parameters to simulate the previously described processes. This program creates a data matrix of individuals with multi-locus genotypes and then iterates a Wright-Fisher model with mutation. Importantly, I began the simulation with the two groups A and B having the same alleles and so we only begin to have resolution after 4000 generations. Pairwise comparisons between individuals have been colored to match the four classes discussed above and appearing in Figure 1-3. Notice that the forward time simulation converges on pairwise frequency values that match those predicted by the equation [1-7] shown in figure 1-3. Also notice the relative frequency of individuals with a difference of just one mutation (appearing as orange but also marked by an arrow). It is important to notice that under the above conditions these individuals either do not exist or
Figure 1-4) Pairwise relationship between individuals in a forward time simulation modeling divergence, sex, clonality and mutation. Histogram colored by the relationship between pairwise comparisons, members of different species shown in red, members of the same species in green, identical individual in the same clone family in blue and members of the same clone family with a single mutation event between them colored in orange.
are likely over-represented in the population and it is only by averaging over time that we would be able to accurately measure their expected frequency.

**Conclusions**

In this discussion I have highlighted some of the key aspects of the coalescent model, which I will use extensively in the subsequent chapters. I wish to mention here that many of the results I’ve discussed are well known. The application of the beta distribution to model genealogical variance among populations is widely used in forensic DNA analysis (Balding and Nicholas 1995). The approach I take here in using the beta and dirichlet distributions to approximate pairwise frequencies is, I think, implied by Wakeley (2009), although the author uses a transition state markov process to accomplish the same types of analysis in many cases. I also believe that the idea that the genealogical process of coalescence is fundamental or unifying has been implied by (Wakeley 2009, Hossjer 2014) or even stated by others (Nordborg and Krone 2002, Sjodin et al. 2005). Thus, I make no claim to the idea, rather I simply find the argument compelling. Finally, I wish to say that I am aware of a number of faults in the methods and logic presented above. In particular, I am aware that I have substituted continuous approximations in a number of discrete instances and I am also aware of the difference between the Wright-Fisher model and the approximations I have made when the population size is particularly small. I hope that anyone reading the works presented here will take the time to understand the limitations and assumptions of the methods, in many cases Wakeley (2009) will provide an excellent starting point for additional reading. Finally, I hope that future readers will be kind when finding flaws in either my logic or my calculations, undoubtedly there are many. In the words of George Box:
“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

-George E. P. Box (1984)
References


Chapter 2

EVALUATING THE SPECIES HYPOTHESIS WITH A BAYESIAN SYMPATRIC STRUCTURE TEST IN ZOANTHUS SANSBARICUS AND ITS SYMBIONTS

Abstract

There is growing awareness that taxonomic designations do not reflect biological reality. The designation of “species”, however, can be formalized as a testable and falsifiable hypothesis. In the present manuscript we discuss an experimental design for testing the designation of “species” on a set of individuals using multi-locus genetic data and Bayesian clustering. We focus on how to evaluate if the ideal number of clusters is one or two and suggest a Bayesian analysis of the resulting clusters to evaluate the two competing hypotheses. We then argue that this approach can be applied universally across all sexually recombining organisms where the biological species concept is applicable to define species boundaries. Under this experimental design the null hypothesis of a single species can be accepted or rejected among the individuals in the experiment. The results can designate those individuals as the same or different species with respect to the other individuals in the experiment; even if the results alone do not warrant formal taxonomic descriptions.

Introduction

Species delimitation is one of the most fundamental and original challenges in the field of biology. Despite the importance of the species designation there exists no single criteria for defining a group of organisms as a species. Rather, there are separate criteria for the many hierarchical groups of organisms; these criteria reflect the unique history of species designation practices of the subfield of biology for which the studied organism fits. The lack of a unified framework and prevalence of many competing species concepts
among biologist is commonly referred to as “the species problem” (Coyne and Orr 2004). For this reason the degree to which current taxonomic designations reflect underlying patterns of genetic exchange among organisms is highly variable among taxa.

Recently, a number of methods for testing the assumptions of the BSC with genetic data have emerged. Many of these methods differed by the type of data they are designed to analyze, i.e. multi-locus sequence data (Yang and Rannala 2009), population genetic allelic data (Hebert et al. 2003, Hebert et al. 2004, DeSalle et al. 2005, Pereyra et al. 2009) and next-generation single nucleotide polymorphism data (Pante et al. 2015) and still others that integrate data from multiple potential sources (Camargo et al. 2011, Hay and Pinho 2012). In the present manuscript, we focus on an increasingly common approach that employs unsupervised clustering algorithms to resolve hierarchical patterns of genetic structure with multi-locus genetic data following an approach similar to Hausdorf and Hennig (2010). This approach has been used successfully to resolve species boundaries in a large number of taxa (Pereyra et al. 2009, Pinzon and LaJeunesse et al. 2011, Ladner and Palumbi 2012, Pinzon et al. 2013, Cowart et al. 2013, Boulay et al. 2014, LaJeunesse et al. 2014, Thornhill et al. 2014, Warner et al. 2015, Wham and LaJeunesse 2016). While the sampling design and source of genetic data vary widely among these studies, all of them employ the same basic logic to resolve species: populations do not contain reproductive barriers, so if sympatric individuals form strong clusters with geographically distant individuals then those clusters can be interpreted as species.

In the present manuscript we discuss the experimental design for using this method (which we call a Sympatric Structure Test) to test the designation of “species” on
a set of individuals. We then identify some areas of weakness in the most commonly used approach. Specifically, we focus on how to evaluate if the ideal number of clusters is one or two and suggest a Bayesian analysis of the resulting clusters to evaluate the two competing hypotheses. We then argue that this approach can be applied universally across all sexually recombining organisms where the BSC is applicable to define species boundaries. Under this experimental design the null hypothesis of a single species can be accepted or rejected among the individuals in the experiment. The results can designate those individuals as the same or different species with respect to the other individuals in the experiment; even if the results alone do not warrant formal taxonomic descriptions.

To demonstrate this approach, we apply a Bayesian Sympatric Structure Test (BSST) to the hypothesis of species for a collection of *Zoanthus sansibaricus* and their unicellular eukaryotic symbionts. *Zoanthus sansibaricus* is zoanthid widely distributed from shallow and deep reef environments ranging geographically from the Indian Ocean (Reimer et al. 2004) to the Galapagos Islands (Reimer and Hickman 2009) and tropical to sub-tropical latitudes. It forms an obligate symbiotic relationship with dinoflagellate micro-algae in the genus *Symbiodinium*. The order *Zoantharia* and the genus *Symbiodinium* have been the subject of recent taxonomic revision. Within Zoantharia, a number of new genera (Sinniger et al. 2010) and species have recently been described (Reimer et al. 2006) while other closely related species have been collapsed (Reimer et al. 2004). Similarly, new *Symbiodinium* species have been recently described (LaJeunesse et al. 2012 & 2014, Hume et al. 2015; Lee et al. 2015) and there is evidence suggesting that hundreds of currently recognized ITS types are probably species and in need of formal characterization (Sampayo et al. 2009, Thornhill et al. 2014).
sansibaricus occurs at a wide but discontinuous depth range, common in both the intertidal and depths as great as 35 meters, with a sharp drop in abundance at intermediate depths of 4-7m (Kamezaki et al. 2013). This discontinuous distribution of the host coincides with the prevalence of minor variations in ITS sequences of its clade C Symbiodinium (Reimer et al. 2011). The uncertainty of species continuity across the host and symbiont populations along a broad depth gradient provides an opportunity to apply the SST to examine the potential for species boundaries occurring in distantly related organisms with different life histories.

Methods

Specimen Collection and Microsatellite genotyping

Z. sansibaricus was collected from 2 sites on opposite sides of Okinawa-Jima Island, Japan and one site in Kagoshima prefecture. 24 samples were taken from the intertidal zone and 33 samples were taken from colonies that occurred between 7 and 35m at the west site (Manza) on Okinawa-Jima Island which faces the East China Sea. 29 samples were taken from the intertidal and 7 samples were taken between 7 and 35m at the east site (Teniya) on Okinawa-Jima Island, which faces the Pacific Ocean. 10 samples were taken from the intertidal Sakurajima in Kagoshima prefecture where Z. sansibaricus is known to occur primarily in the intertidal (pers. communication James Reimer). All samples were collected as small clusters of 1-30 polyps from colonies that were separated by 5-100m. These samples were individually bagged and labeled and then transferred to 99% ethanol for preservation. DNA was then extracted using a Qiagen tissue DNA isolation kit from a single polyp from each sample.
We obtained multi-locus genotypes for both *Z.sansibaricus* and its associated clade C *Symbiodinium* using microsatellite primers that were designed to be specific to each organism. To target the nDNA of the host we used the primers Zsan_71, M13-Zsan_67, M13-Zsan_30, M13-Zsan_104, M13-Zsan_102 and M13-Zsan_130, originally published by Wham *et al.* 2013. To target the nDNA of the clade C *Symbiodinium* occurring in each sample we used the primers Sgr_21, SgrSpl_25, SgrSpl_30, Spl_33, Sgr_40, SgrSpl_78 and S_99, originally published by Wham *et al.* 2014. All primers with the “M13” designation contained an additional M13 sequence (Schuelke 2000) on the forward primer to attach a florescent label to the final PCR product following the methods described by Schuelke (2000). All other primers were directly labeled with either a FAM, HEX, NED or PET florescent tag on the forward primer. Polymerase chain reactions were carried out following the protocols previously described in Wham *et al.* (2013 & 2014). The PCR products were then screened by gel electrophoresis (Penn State Genomics Core Facility, University Park, PA). Alleles were scored using the genetic software Geneious (v. 6.1.6). Both the host and symbiont were scored as diploid organisms following the ploidy interpretation of Clade C *Symbiodinium* discussed in Wham *et al.* (2014), Thornhill *et al.* (2014) and Wham *et al.* (2015).

**Experimental Design for Bayesian Sympatric Species Test**

Studies that have employed unsupervised clustering to resolve species boundaries (Pereyra *et al.* 2009, Pinzon and LaJeunesse *et al.* 2011, Ladner and Palumbi 2012, Pinzon *et al.* 2013, Cowart *et al.* 2013, Boulay *et al.* 2014, LaJeunesse *et al.* 2014, Thornhill *et al.* 2014, Warner *et al.* 2015, Wham *et al.* 2015) vary widely in the number of taxa considered, the number of locations sampled and the number of individuals
employed in their analysis. We begin by describing the Sympatric Species test in its simplest form, two sample locations with two putative species that are sympatric in at least one of the 2 locations (Fig. 2-1). The null hypothesis that putative species A and putative species B are a single species with respect to putative species A can then be tested by acquiring sympatric samples of both putative species A and B from a single location and samples of species A from a second location. Here “a second location” means a site that exceeds the range at which samples of species A and B were collected in sympatry. Ideally the distance among sites exceeds the average distance of sympatric individuals by several orders of magnitude such that the sites can be considered independent. The selection of the second site can influence the inferences made from the analysis of a BSST; we discuss these implications in depth in the discussion section “A Second Location”. Genetic data is then generated from unlinked nDNA to produce multi-locus genotypes (MLGs) for each individual; the approach can accommodate a number of different sources of genetic data, from sequence-based haplotypes to microsatellites to single nucleotide polymorphisms. The MLGs are then analyzed with an unsupervised clustering algorithm at a series of different k-values (k = number of clusters). In the present analysis we use STRUCTURE under the no admixture model with alleles uncorrelated, however the present approach can be extended to other unsupervised clustering algorithms. After performing the cluster analysis, commonly, the ideal number of clusters is interpreted by evaluating the change in log-likelihood of the data at each progressive k-value. When evaluating the data to determine if there are two sympatric species when sympatric individuals form strong clusters with geographically distant individuals at k=2 those clusters should be interpreted as species. When no such
Figure 2-1) Simplified depiction of experimental design, where the effect of geographic distance is compared to the effect of putative reproductive barriers to assess the species hypothesis.
Figure 2-2) Bayesian approach designating a region of population equivalence to accept or reject the hypothesis of species
clusters are observed, or clusters occur between locations, individuals should be interpreted as conspecific.

Here we apply an additional statistical approach to evaluate the two competing hypotheses \( k = 1 \) and \( k = 2 \). We begin by evaluating the data at \( k = 2 \), first assigning all individuals to their cluster of majority membership (i.e. >50% probability of belonging to a cluster). We then estimate the posterior distribution of \( F_{st} \) (Fig. 2-2a) between the two collection sites for all individuals of cluster 1 by bootstrapping over loci (Weir and Cockerham 1984). Alternatively, the posterior distribution of \( F_{st} \) could be estimated for every Markov chain step after burn-in and then a final estimate could be produced by integrating over all of these posterior distributions. The region between zero and the upper bound of the 95% highest density credibility interval (HDCI) is then retained as a “region of population equivalence”, referred to from here forward as the ROPE following the Bayesian approach to hypothesis testing of Kruschke (2013) (Fig 2-2b). The posterior distribution of \( F_{st} \) between the sympatric samples of cluster 1 and 2 is then calculated and compared to the ROPE (Fig. 2-2c). The resulting comparison will have one of three outcomes which suggest the following interpretations: 1) more than 95% of the density of the posterior distribution for \( F_{st} \) is outside of the ROPE, the null hypothesis of a single species identity for putative species A and B (Fig. 2-2c1) is rejected, 2) less than 5% of the density of the posterior distribution for \( F_{st} \) is outside of the ROPE, that is, more than 95% of the density of the posterior distribution for \( F_{st} \) is inside of the ROPE; the null hypothesis that putative species A and B are the same species (Fig. 2-2c2) should be accepted, 3) between 5% and 95% of the density of the posterior distribution for \( F_{st} \) is
outside of the ROPE; additional data is necessary to conclude the relationship between A and B, (Fig. 2-2c3).

**Results**

**BSST on *Z. sansibaricus* and its Clade C Symbionts**

We applied this approach to test if intertidal *Z. sansibaricus* and/or their clade C symbionts were different species than their subtidal counterparts. The STRUCTURE analysis revealed two different patterns of genetic relationship for the host and symbiont (Fig 2-3a,b & g). In *Z. sansibaricus*, the log likelihood of the data evaluated at k-values from 1 to 8 showed no strong pattern of increase, suggesting that k=1 was the best model for this data. In figure 3 we show the results of k=2 and k=3. While there are a number of individuals that are given a high probability of assignment to clusters at k=2, this pattern did not correspond to a significant increase in the log likelihood of the data nor did the pattern of individuals assigned to each cluster correspond to sample depth or location, although the most common cluster membership was different at each location (Fig. 2-3c & d). This pattern was not observed at k=3 (Fig. 2-3a) or at higher k-values. In contrast, in the clade C *Symbiodinium* found within these individuals showed a large difference in log likelihood between k=1 and k=2 with only a marginal increase in log likelihood in successive k-values. This suggests that k=2 is the best model for these data (Fig. 2-3g). At k=2, all individuals showed a high probability of membership to only 1 of 2 clusters, however this pattern showed no strong relationship between site of collection. Rather, these clusters corresponded to the position of the host being above or below the discontinuous break in its distribution along the depth gradient. That is, all individuals
Figure 2-3) Results of STRUCTURE analysis on multi-locus genotypes of *Z. sansibaricus* and their associated clade C *symbiodinium*. A) k=3 for MLGs of *Z. sansibaricus*, notice the even probability of all individuals to all clusters b) k=2 for MLGs of *Z. sansibaricus*, notice the high probability of membership of many individuals to each cluster. C) k=2 individuals plotted by depth at Manza D) k=2 individuals plotted by depth at Teniya, E) k=2 fro the clade *C. symbiodinium* at Manza segregated my depth. F) k=2 for the clade C *Symbiodinium* at Teniya. G) results of STRUCTURE analysis at k=2 for clade C *Symbiodinium* subdivided by location.
Figure 2-4) Bayesian analysis for the significance of divergence between sympatric individuals applied to Z. sansibaricus. A) the posterior density distribution of $F_{st}$ between individuals assigned to the same cluster across locations. B) the region of population equivalence C) the posterior distribution of $F_{st}$ between sympatric individuals assigned to the same cluster.
Figure 2-5) Bayesian analysis for the significance of divergence between sympatric individuals applied to the clade C Symbiodinium occurring in Z. sansibaricus. A) the posterior density distribution of $F_{st}$ between individuals assigned to the same cluster across locations. B) the region of population equivalence C) the posterior distribution of $F_{st}$ between sympatric individuals assigned to the same cluster.
that were sampled from the intertidal and the shallow sub-tidal clustered together respectively, regardless of the site of collection (Fig. 2-3g).

To interpret the significance of these clusters we calculated the upper bound of the 95% credibility interval (CI) of $F_{st}$ between individuals at separate locations assigned to the same cluster. In *Z. sansibaricus*, this resulted in an upper bound of 0.078 and 0.092 for the ROPE of the 2 clusters respectively (Fig. 2-4a&b - 1&2). We then calculated the posterior density distribution for $F_{st}$ between clusters among sympatric individuals belonging to different clusters. At Manza, the upper bound of the 95% CI of $F_{st}$ between individuals assigned to each cluster was 0.022. At Teniya, the upper bound of the 95% CI of $F_{st}$ between individuals assigned to each cluster was 0.092, both estimates falling entirely within the ROPE (Fig. 2-4c - 1&2). The upper bound of the 95% CI of the posterior density distribution of $F_{st}$ between clade C *Symbiodinium* found within Z. *sansibaricus* at separate locations assigned to the same cluster was calculated to be 0.030 for the cluster corresponding to the intertidal individuals and 0.221 for the cluster corresponding to the sub-tidal individuals (Fig. 2-5a&b - 1&2). We then calculated the posterior density distribution of $F_{st}$ between sympatric individuals belonging to different clusters. At Manza, the 95% HDCI for $F_{st}$ between clusters was 0.360 – 0.448 and at Teniya $F_{st}$ between clusters was 0.446 – 0.534, both distributions being entirely outside of the ROPE (Fig. 2-5c - 1&2).

**Discussion**

**What is a population and what is a species?**

In light of modern genetic techniques, there is growing awareness that many formal taxonomic designations do not reflect biological reality, particularly in marine
systems where cryptic species seem to be common (Knowlton 1993) and the history of

taxonomic description has been particularly inadequate (Pante et al. 2014). In recognition

of this, there is an emerging paradigm in biology that recognizes that the designation of

“species” can be formalized as a testable and falsifiable hypothesis (Knowles and

Carstens 2007, Ence and Carstens 2011, Hausdorf and Hennig 2010, Yang and Rannala

2010, Hay and Pinho 2011, Pante et al. 2014). Further, experiments to test or falsify the

species hypothesis have utility beyond their use in formally designating a binomial.

Indeed, it may be useful or even necessary to test if a set of individuals is conspecific
even if the results are not immediately used to name the species that are resolved (Pinzon

and LaJeunesse et al. 2011, Pinzon et al. 2013, Thornhill et al. 2014, Warner et al. 2015),

although the results of such a test might be later integrated with other data to provide

adequate criteria for naming species (LaJeunesse et al 2014). The unique history and

tradition of species designation in each subfield of biology may always require different
criteria for formal species recognition, causing the “species problem” to be unsolvable as

it applies to formally designating taxonomic ranks. The “Species problem” may,

however, be solvable as it applies to a discrete set of individuals by testing the Biological

Species Concept (BSC) with genetic data. Indeed, in some taxonomic groups a test for

conspecificity of all individuals may be a necessary first step for any biological study to

avoid spurious conclusions (Wham et al. 2015).

He we make the fundamental assumption that biological diversity is structured

hierarchically. Specifically we assume that collections of individuals exist in groups that

generically recombine frequently at a relatively constant rate; these groups are

populations. We also assume that the geographic range of an organism potentially
comprises many populations, which are all connected by continuous genetic exchange. These populations do not necessarily exist as discreet groups (i.e. as islands), although they may be discreet in some cases. Individuals in the same population can be thought of as the group of co-occurring individuals that are equally likely to mate with each other.

Separate populations then arise when groups of individuals genetically recombine more frequently within their group than between groups. Populations can therefore exist on a continuum of relatedness relative to their per-generation, per-individual rate of genetic exchange, i.e. $2N_e m$ (Wright 1931), where $N_e$ is the effective population size and $m$ is the migration rate. All of these connected populations together comprise a species. These species are related to other such groups (species) and may exist in separate locations (allopatric) or co-occur (sympatric).

In this model, a collection of individuals can be identified as members of the same population, different populations or different species. Here we assume that in general, it has been accepted that species are defined by their potential for genetic exchange, under the biological species concept (Mayr 1942). The issue remains on how to assess if groups of individuals engage in genetic exchange. This is made difficult by the fact that patterns of infrequent genetic exchange, expected between distinct populations, are also consistent with the patterns expected between groups that have recently ceased exchanging genetic material. A key aspect of the present debate is if the observable qualities (genetic divergence) of recently diverged species are qualitatively or quantitatively different than the observable qualities of populations (Hey and Pinho 2011). Some have suggested that recently diverged species and populations do not qualitatively differ (Levin 1979, Hendry et al. 2000). There is evidence, however, that
particular genes might be more active in the speciation process (Nosil and Schluter 2011), but assessing these genes require prior knowledge about divergence patterns and so they do not yet offer diagnostic utility.

We suggest, as others have (Shluter 2010), that the qualities of population and species divergence are qualitatively different in that population-forming boundaries are extrinsic and species forming boundaries are intrinsic. Further, the observable quantities of divergence between newly formed species and populations are not necessarily different. If the state of “species” was simply the result of more of the same kind of divergence observable between populations then we might expect to observe a threshold at which the amount of this difference might indicate population or species status. This hypothesis was, however, investigated by Hey and Pinho (2011) who showed that there is no correspondence between the inferred values of $2N_e m$ and the status of population or species in a meta-study that summarized the results of 97 reports from a broad range of taxa. This is likely explained by the fact that populations can maintain a broad range of divergence levels as stable states (Wright 1931). Haploid populations that engage in genetic exchange at a rate $2NM > 1/2$ will share similar gene frequencies but when $2NM < 1/2$ the populations will maintain independent gene frequencies (Wright 1931), thus divergence in populations is related to the frequency of genetic exchange, $2NM$.

Divergence between species on the other hand is often thought of as relating to the time since genetic exchange ended ($T$), the expected divergence increasing with time. It is therefore possible to observe equal genetic divergence between two populations that have been engaging in a low level of genetic exchange for a long period of time and species that have been completely separate for a relatively short period of time (Wright 1931,
Hey and Pinho (2011). Hey and Pinho (2011) demonstrated, however, that if both 2NM and T are considered simultaneously a threshold might be adopted to differentiate species from populations. This threshold is still subject to misclassifying species as populations and populations as species because species and populations very widely in their total number of individuals (Hey and Pinho 2011) and so making equivalent comparisons across all taxa may be impossible. Nevertheless this approach may be very useful when evaluating the species or population status of allopatric groups.

Many of the limitations considered by Hey and Pinho’s (2011) approach can be avoided when evaluating putative species that are sympatric. Because members of the same species share the potential for genetic exchange, when genetic divergence is observed in sympatric individuals it is necessary to evoke a reproductive barrier to explain the maintenance of that divergence. Here we submit that the formative process of populations does not involve intrinsic reproductive barriers only extrinsic barriers to dispersal. That is, populations form because individuals lack the opportunity to reproduce, not the ability. Intrinsic barriers to genetic exchange, in contrast, are exclusive features of species formation, even if these intrinsic barriers are porous. Here we suggest that the designation of species can be resolved from the designation of population by demonstrating the existence of an intrinsic barrier to genetic exchange.

**Why specify a region of population equivalence?**

The logic presented above is not unique to this manuscript. In fact many species delimitation methods seek to assess some measure of genetic divergence (see Sites and Marshall 2003 for a review of methods and Hey and Pinho 2011 for a discussion). However these methods almost universally test if the measure of genetic divergence is
non-zero, which, given enough samples, will always be true and is not necessarily indicative of species, as pointed out by Hey (2009). Threshold based methods of species delimitation seek to avoid this problem by defining a minimum non-zero level of divergence to define the boundary between populations and species (Wiens and Servedio 2000, Tobias et al. 2010 and Hey and Pinho 2011). These methods, however, assume that there is some divergence metric with a universal boundary between populations and species. Divergence patterns, however, are unique for every organism, varying based on life history, population size and many other factors. For this reason, threshold based techniques can only hope to minimize their error rate and concede some acceptable rate of misclassification.

The approach presented here does not assume a universal degree of divergence that is indicative of species status. Rather, we simply assume that the minimum species divergence that we wish to detect is greater than population level divergence within highly related groups. The designation of a “Region of Population Equivalence” is therefore two fold.

First, to test if sympatric clusters are significantly more diverged than individuals of the same cluster from different populations. In other words, traditionally $F_{st}$ has an expected value of zero when it is calculated on groups of individuals from the same panmictic population. Given a finite sample of individuals from the same population and a finite number of loci, $F_{st}$ will always be more than zero. For this reason, traditionally a null distribution is computed for $F_{st}$ to test the significance of the estimate of $F_{st}$ as being greater than zero (Weir and Cockerham 1984), which is often described by a p-value. This approach is useful for describing the presence of statistically significant structure but
in this case the experiment aims to designate a range of genetic distance from zero to some value that is significantly greater than zero, yet still corresponds to the divergence expected between populations. By first estimating a population level, non-zero value for $F_{st}$ we are providing this benchmark.

The second aim in designating a ROPE is to insure that there is sufficient power to resolve population level divergence prior to accepting the null hypothesis of a single species. We assume that the minimum level of divergence between two species that we would like to detect is equal to the amount of divergence between two populations of the same species. If there are intrinsic barriers to reproduction that cause genetic divergence below this level we are still comfortable with considering those individuals as the same species. When 95% or more of the density of the posterior distribution for $F_{st}$ is within the ROPE we can not only infer that the genetic distance between putative species A and B is less than or equal to the two populations of putative species A, we can also infer that the experiment had enough power to resolve divergence at least as great as expected between populations of putative species A. This is an important feature of this approach, consider that if $F_{st}$ was calculated between putative species A and B without designating the ROPE, if $F_{st}$ could not be shown to be greater than zero the null hypothesis that species A and B are the same species might naively be accepted when the experiment simply lacked sufficient power to resolve the two putative species.

In other words, this approach provides a 3-way decision boundary for resolving the relationship between A and B. If we have observed at least one example of a population of A that is more divergent than A and B we are comfortable with treating A and B as the same species given the present data. Alternatively, if A and B are more
divergent than at least one population of A it is reasonable to treat them as the separate species. Finally, if the divergence between A and B can not be shown to be significantly greater or less than a second population of A any further analysis of A and B which assumes a specific relationship between A and B should be halted until more data is gathered. This approach is not subject to the weaknesses of other distance based methods which tend to find more species as sample size increases (Hey 2009). Additionally, this method is not subject to the weaknesses of threshold approaches which can not accommodate the unique nature of every organismal lineage.

A Second Location?

Perhaps an obvious criticism of this technique is the subjective nature of choosing a second location to sample individuals. Indeed we should expect a wide range of divergence among populations and it follows that by choosing a distant population the width of the ROPE will increase resulting in a greater tendency to accept the null hypothesis of single species for the sympatric individuals. This is indeed the case, but we think that in the most general sense species are expected to have more evolutionary independence than populations, so we think that any second population is suitable. Importantly this results in a relative basis of species distinction, where species are delimited or not delimited based on divergence measures relative to population level divergence. This means that when this approach is extended to many pairwise comparisons it may be possible (although unusual) for some pairwise comparisons to yield the distinction of species while others do not, violating the logical expectation of transitivity. Such a result should be restricted to cases where speciation is ongoing or
incomplete and so we think that this result is an accurate one if and when it occurs because species do not necessarily follow a transitive property (Martins et al. 2013).

Perhaps the biggest danger is in selecting a second location which is composed entirely of a third species. If this happens the experiment cannot provide meaningful inference about the status of population or species among the sympatric individuals. For this reason, we suggest that researchers generally attempt to collect individuals from a similar, but independent, nearby location. The primary down side of collecting from closely related sites is that the ROPE will be narrower requiring more samples in order to accept the null. However, as we show here, we were able to accept the hypothesis of a single species with a relatively modest number of samples in *Z. sansibaricus*. Potentially other genetic markers may also provide supporting evidence of the conspecific status of individuals sampled from separate locations. For example, Thornhill et al. (2014) showed that Symbiodinium C7a and C7 were different species where they overlapped in Curacao and then inferred that all C7 individuals were conspecific within each genetic type across other sample locations because they shared sequence haplotypes or showed chloroplast haplotype sequence similarity (Thornhill et al. 2014). Wham et al. (2015) also showed that allele identity can provide useful insights into the continuity of species across locations. Ultimately, researchers will have to weigh the risk of sampling a third species at a second site with specific knowledge of diversity and dispersal in their organism of interest.

**Conclusions**

Here we show that the clade C *Symbiodinium* associated with *Z. sansibaricus* are different species and these species are depth segregated, each primarily occurring above
or below a major break in the abundance of the host along a depth gradient (Fig. 2-3 & 2-5). Kamzaki et al. (2013) also observed differences in the clade C symbionts across depth at phylogenetic marks, interpreting, as we have, that this corresponds to differences in symbiotic species association across depth. We also concluded that all host individuals are members of the same species because the within location $F_{st}$ values fell fully within the ROPE (Fig. 2-4c). This interestingly suggests that both shallow and deep colonies arise from the same gene pool which support the conclusions of Kamzaki et al. (2013) based on the analysis of phylogenetic markers. *Z. sansibaricus* is thought to be an external broadcast spawner with horizontal symbiont transmission (Kamzaki et al. 2013). Evidently then, larvae establish the appropriate symbiotic association based on the location of their settlement by up taking symbionts which are both compatible with the host and with the environment from neighboring colonies. The discontinuous distribution of the host across depth may be explained by competition with Acroporids and other stony corals and/or predation by butterflyfish. However, the correspondence of break in host abundance to the timing of the switch in the dominant symbiont species suggests a potential explanation involving the niche boundaries of its compatible symbiotic partners (Iglesias-Prieto et al. 2004). Alternatively, this discontinuous distribution may be caused by to the niche space of the host alone; potentially then, this host has evolved the flexibility to associate with these two specific symbionts to accommodate its wide depth range. Potentially then, this host has evolved the flexibility to associate with these two specific symbionts to accommodate its wide depth range. Resolving these explanations will be the subject of future work in this system.
In our analysis of the host we observed a large number of individuals that showed a high percent assignment to only one of the two clusters (Fig. 2-3) even though we did not observe a large change in log likelihood between k=1 and k=2. This highlights a potential limitation of just considering the correspondence of clusters to geographic location at successive k-values to recognize species (Pinzon and LaJeunesse et al. 2011, Pinzon et al. 2013, Thornhill et al. 2014, Warner et al. 2015, Wham et al. 2015), although most of these studies also employed independent lines of evidence to support their species diagnoses and all of these studies considered the change in log likelihood between k-values. In the present case, the change in log likelihood suggested that the clusters observed at k=2 were artifacts, perhaps due to the relatively small number of loci used (6). It is worth noting that the Bayesian approach we describe here, however, allowed us to evaluate objectively between k=1 and k=2. Rather than having to make a subjective judgment as to whether or not the change in log likelihood between k=1 and k=2 was sufficient to consider the assignments at k=2 significant. We also observed a slight difference in the frequency of individuals assigned to each cluster within each location at k=2 (Fig. 2-3). We interpret this as potentially arising from weak genetic structure between these locations.

Our analysis demonstrates how the SST can be used to accept or reject the hypothesis of species in an objective way. Objective here meaning that the approach can be applied to any set of organisms for which the BSC is applicable and multiple independent applications of the approach should arrive at the same conclusion. Perhaps the most significant limitation of the approach is that it is restricted to delimiting sympatricly co-distributed species, still leaving the question of how to define the many
groups that do not co-occur with the individuals that they need to be resolved from. In these cases, we believe that the application of threshold based techniques (Hey and Pinho 2011) when integrated with other lines of evidence can provide strong and useful insights. Indeed, one of the critical limitations of all threshold based techniques is that the data used to interpret the ideal boundary between populations and species is itself mislabeled, species marked as populations and populations marked as species. Yet many of the species and populations used to calibrate these threshold methods comprise groups with sympatric distributions where the present analysis might be used to provide population or species labels that better reflect biological reality. While we do not believe that any threshold method could ever be error free, we anticipate that their accuracy when applied to allopatric groups could be improve greatly if trained on correctly labeled data from sympatric defined groups (as we described here). For this reason taxonomy may never reach ground truth, but its best days are almost certainly still ahead.
References


LaJeunesse, T. C., Parkinson, J. E., & Reimer, J. D. (2012). A genetics based description of *Symbiodinium minutum* sp. nov. and *S. psygmophilum* sp. nov. (Dinophyceae), two dinoflagellates symbiotic with cnidaria. *Journal of Phycology*, 48(6), 1380-1391.

genus *Symbiodinium* (Dinophyceae) Clade D are different species. *Phycologia*, 53(4), 305-319.


common species and few endemics among Indo-Pacific cauliflower corals (*Pocillopora, Scleractinia*). *Journal of Biogeography, 40*(8), 1595-1608.


Sinniger, F., Reimer, J. D., & Pawlowski, J. (2010). The Parazoanthidae (Hexacorallia: 
Zoantharia) DNA taxonomy: description of two new genera. *Marine Biodiversity, 40*(1), 
57-70.


Lineages Dominate The Adaptive Radiation Of Reef Coral Endosymbionts. *Evolution, 
68*(2), 352-367.


the widespread coral *Seriatopora hystrix* masks spatial genetic patterns of connectivity. 
*Molecular ecology.*

structure. *Evolution, 1358-1370.*

differences between species. *Proceedings of the Royal Society of London B: Biological 
Sciences, 267*(1444), 631-636.

microsatellite loci for the Indo-Pacific-wide zoanthid, *Zoanthus sansibaricus.* *Marine 
Biodiversity, 43*(3), 247-250.


Chapter 3

SYMBIODINIUM POPULATION GENETICS: TESTING FOR SPECIES BOUNDARIES AND ANALYZING SAMPLES WITH MIXED GENOTYPES

Abstract

Population genetic markers are increasingly being used to study the diversity, ecology, and evolution of Symbiodinium, a group of eukaryotic microbes that are often mutualistic with reef-building corals. Population genetic markers can resolve individual clones, or strains, from samples of host tissue, however samples may comprise different species that may confound interpretations of gene flow and genetic structure. Here we propose a method for resolving species from population genetic data using tests of genetic recombination. Assigning individuals to genetically recombining populations prior to further analyses avoids critical errors in the interpretation of gene flow and dispersal. To demonstrate the effectiveness of the approach we first apply this method to a simulated dataset. We then use the method to resolve two species of host-generalist Symbiodinium that commonly co-occur in reef-building corals collected from Indo-West Pacific reefs. We demonstrate that the method is robust even when some hosts contain genotypes of two distinct species. Finally, we examine population genetic datasets from two recently published papers in Molecular Ecology. We show that each strongly supports a two species interpretation, which significantly changes the original conclusions presented in these studies. When combined with available phylogenetic and ecological evidence, the use of population genetic data offers a robust method for unambiguously delimiting morphologically cryptic species.
Introduction

Population genetic data are increasingly generated in ecological studies of the dinoflagellate genus, *Symbiodinium*, a diverse group of dinoflagellates often associated with reef-building corals and other Cnidaria (Andras et al. 2011; Pettay and LaJeunesse 2013; Baums et al. 2014; Thornhill et al. 2014). The analysis of these data generally requires alleles to be assigned to individuals and assumes that all samples analyzed together are members of the same species. This allows the data to be organized into multi-locus genotypes (MLGs). These MLGs can then be analyzed under the assumptions of population genetic models. A number of factors must be considered, however, before meeting these basic criteria for analysis.

The genus *Symbiodinium* is widely recognized as ecologically and genetically diverse, however there are relatively few described species (Trench and Blank 1987, Hansen and Daugbjerg 2009, LaJeunesse et al. 2012, LaJeunesse et al. 2014), making initial species identification of samples problematic. Ribosomal DNA (rDNA) has often been used to provisionally resolve ‘species’ diversity (LaJeunesse 2001) because *Symbiodinium* with fixed differences in rDNA exhibit distinct ecological, biogeographic, and physiological attributes (Rodriguez-Lanetty et al. 2004; Sampayo et al. 2007; Frade et al. 2008; LaJeunesse et al. 2010). However, sometimes rDNA can fail to resolve ecologically distinct *Symbiodinium*. At least a few studies have resolved *Symbiodinium* with the same ITS or LSU sequence with higher resolution genetic markers into multiple distinct entities, or species (Santos et al. 2004; Finney et al. 2010; LaJeunesse and Thornhill 2011; Thornhill et al. 2014). Consequently, no single genetic marker currently
resolves species in this group and a hierarchical multi-locus approach is generally required to obtain species level resolution (LaJeunesse et al. 2012; 2014).

Host tissues often contain millions of *Symbiodinium* cells per square centimeter (Fitt et al. 2000; Thornhill et al. 2011); all of which contribute to the pool of DNA isolated from a sample. In most cases multi-locus genotyping is possible because samples contain a clonal population of cells that are either haploid (Pettay et al. 2011; Baums et al. 2014) or, in some cases, diploid through genome duplication (LaJeunesse et al. 2014; Wham et al. 2014). In several studies the number of samples with mixed genotypes was thought to be high, which deterred the characterization of MLGs for population genetic inference (Howells et al. 2009, 2013, Wirshing et al 2013). When mixtures occur, the origin of mixed genotypes may be two or more resident individuals (i.e clones) of the same species (Pettay et al. 2011; Baums et al. 2014), genotypes from multiple species (LaJeunesse et al. 2014), or the mixture of genotypes from multiple individual clones and species (LaJeunesse et al 2014). It is important to distinguish between these potential explanations when making inferences from mixed samples, however, these competing explanations have rarely been considered.

Here we employ a hierarchical approach to analyzing *Symbiodinium* population genetic data generated from host samples. Our approach aims to assess the occurrence of separate species in a dataset and then designate the origin of mixed samples. We start by resolving putative species as genetic groups that form clusters across a geographic range by analyzing correlations in allelic identity among individuals. We then parse MLGs into separate species and identify alleles diagnostic for these species; in many cases allowing mixed samples that contain individuals of two species to
be separated into separate MLGs unambiguously. To test the accuracy and utility of this approach we applied these techniques to a simulated microsatellite dataset, which we modeled as two reproductively isolated species, diverging by the processes of mutation and drift. We then analyzed a *Symbiodinium* dataset chosen to test our ability to discern the presence of multiple *Symbiodinium* spp. and to demonstrate the potential flaws in analysis that would have resulted had the presence of multiple species gone undetected.

Having established the consequences of inaccurately delineating species prior to population level analysis and how differences in allelic identity (size ranges in this case) can be used to recognize and assign individuals to species (and in some cases partition mixed samples into individual MLGs), we examine two papers recently published in *Molecular Ecology* to show how patterns observed in these studies are attributable to the presence of previously unrecognized species. We first examined microsatellite data from *Symbiodinium* associated with *Acropora millepora* colonies living in different reef habitats published by Howells *et al.* (2013). We then examined multi-locus data from *Symbiodinium* occurring in the common gorgonian morphospecies, *Euniciea flexuosa*, from the northwestern tropical Atlantic published by Wirshing *et al.* (2013). In both cases we find strong evidence that supports a two species interpretation of data that was originally analyzed as coming from a single species. We then offer a more parsimonious interpretations of these data based on this new knowledge.
Methods

Forward time and Coalescent Based Model for Simulated Data

We simulated microsatellite variation assuming: i) divergence by genetic drift in reproductively isolated groups ii) the appropriate allelic diversity within groups of reproducing individuals based on coalescent expectations of allele frequency and number given a population size (Kimura and Crow 1964, Ewens 1972), iii) linkage equilibrium between loci within individuals in the population separated by at least one sex event iv) the occurrence of identical genotypes arising from asexual reproduction v) multiple individuals occurring in a sample.

We created a dataset that reflected the above attributes in several steps that involved running a forward time simulation and then permuting the allelic configurations of individuals to reflect iii, iv and v. We used the program rMETASIM (Strand 2002) to implement the individual based forward time simulation to account for attributes i and ii. The program simulated two asexual populations of 3,000 individuals ($N_e$) with no genetic exchange at ten loci, representing two species. The simulation began with all individuals having the same allele size at all loci. The simulation, implemented a size-step mutation model every generation with a per locus mutation rate ($\mu$) of $4.5 \times 10^{-4}$ (Whittaker et al. 2003). The simulation ran for 6,000 generations, at the end of the simulation 100 individuals were sampled from each species, MLGs were saved for these individuals in two separate data files. Then, to transform these asexually created MLGs into MLGs expected from a sexual population, allele scores were then sampled with replacement within each locus, creating MLGs in linkage equilibrium, accomplishing iii while maintaining i and ii. We then used Ewens sampling distribution (Ewens 1972) to give the
number of unique genotypes (U) and their respective number of occurrences \( (K_i) \) expected in a sample of 100 \( (n) \) individuals from a population of 3,000 and an effective sex rate \( (S_e) \) of 0.01, i.e. a population sex rate \( (\psi) \) of 60. The first U individuals were retained in the MLG data file with each individual being repeated \( K_i \) times creating a new dataset that incorporates iv while maintaining i, ii, and iii. We then pooled the resulting two data files into a single data file, as if both ‘species’ had been sampled from the same location. We created mixtures of individuals to account for v by drawing from a Poisson distribution with \( p=0.7 \) and merging the resulting number of individuals’ allele calls into one line until the sum of all draws equaled the number of samples. Alleles of equal size were condensed to a single call to reflect the way someone naive of the nature of the mixed sample would have scored the genotype. The frequency of unmixed samples in the final data set was therefore 73.7% (56 of 76), the frequency of mixtures of exactly two genotypes arising from two species was 11.8% (9 of 76) and the remaining 14.5% (11 of 76) samples were comprised of at least two individuals of the same species. Finally, the species and individual designations were hidden from the upstream analysis so that the method would be implemented on the samples blind of their original designation.

**Identifying Species Boundaries with Population Genetic Data**

We retrieved all simulated samples with a single allele size call at all loci and placed these unambiguous MLGs into a single datafile. All samples with identical allele scores across all loci were interpreted as asexual clones when they occurred in the same population, repeated clones were removed from the dataset. We then implemented a multi-step process to determine if more than one species was present in the data; our general approach is outlined in figure 1. We ran a STRUCTURE (Pritchard et al. 2000)
analysis of all unique individuals with a burn-in of 100,000 and 1,000,000 MCMC repetitions with no location prior, under the no admixture model with allele frequencies uncorrelated. This analysis was run three times for each k-value across a range of k-values. We followed the hierarchical clustering of samples at increasing k-values. When the hierarchical partitioning of data returned clusters that did not follow geographic patterns (Fig. 3-1A) and the probability of cluster assignment for each individual was high (Fig. 3-1B), we retained the assignment of individuals for further analysis. We then analyzed the correlation between allele sizes and STRUCTURE assignment across all samples (Fig. 3-1C). We plotted a histogram of allele sizes for each locus coloring histogram bars according to their cluster assignment from the STRUCTURE analysis (e.g. Fig. 3-2b). We then calculated the pairwise Bruvo distance, a genetic distance calculation that considers both the sharing of alleles as well as the size differences between loci that are different between individuals (Bruvo et al. 2004), between each individual and plotted the resulting histogram, labeling the histogram bars by comparisons made within and between STRUCTURE assigned clusters. When clusters corresponded to differences in allele size distribution and pairwise Bruvo distance, the probability of assignment from the STRUCTURE analysis was interpreted as the probability of assignment to a species.

We then used a Bayesian model to recover MLGs from mixed samples. Our approach uses Bayes formula to interpret the posterior probability of species assignment for each of k possible species given the observance of any allele in the dataset:
Figure 3-1) Flow chart describing the process of acquiring, parsing, and analyzing multi-locus data from sample sets potentially containing two or more species.

Dataset lacks the statistical power to clearly resolve genetic clusters explained by the presence of two, or more, species; or when data are evaluated at K-values that are higher than the natural underlying genetic partitions (i.e. created by the presence of multiple species, or populations.)

Sample Collection and DNA extraction.

Select, or create, appropriate MLG markers and proceed to genotyping samples.

Remove identical MLGs that occur in the same location.

**STRUCTURE** analysis

A) Do samples partition by geography?

1) Retain **STRUCTURE** out-file at k-value where clustering does not correspond to geographic patterning.

2) Produce histograms for relationship between cluster assignment and allele identities.

B) Is the average cluster assignment greater than 0.95?

Data either comprise a single species with genetically distinct populations, or multiple allopatric species. Additional evidence is needed to resolve these alternate hypotheses.

C) Does cluster assignment relate to allele size or identity?

Data are consistent with a two species interpretation, clusters/species should be separated and further evaluated for geographic structure. Post hoc analyses of alternate genetic, morphological or ecological traits should be assessed to determine support for species designations.
\[
p(S_k, x_j) = \frac{p(x_j | S_k) p(S_k)}{p(x_j)}
\]

Where \( p(x_j | S_k) \) is the probability of observing allele \( x_j \) in species \( S_k \), \( p(S_k) \) is the prior probability of a MLG being species \( S_k \) (here evaluated as the percentage of individuals assigned to that species in the unambiguous dataset), \( p(x_j) \) is the probability of observing the allele in the unambiguous dataset as a whole. We then evaluated the ambiguous (i.e. mixed) MLG samples by iteratively proposing all combinatorial arrangements of the mixed genotype consistent with a single MLG and recording the probability of each allele in the proposed species. When a single MLG with a probability \( >0.95 \) for all alleles for a given species could be constructed from a mixed sample, that MLG was interpreted as a new MLG assigned to that species. When a mixed sample returned multiple arrangements of alleles with a probability \( >0.95 \) for a given species that sample was interpreted as a mixture of that species. When the model returned MLGs with probabilities \( <0.95 \) for at least one locus for all species the sample was interpreted as a mixture of indeterminable origin since the alleles in the samples were not sufficiently diverged to assign their origin. The model could then return zero to \( k \) new MLGs for every mixed sample evaluated.

Most of the analyses described above can be done with the R script “Bayes Allele.R” available at the authors website: www.thecoalescent.com/bayesallele. This script will calculate the probability of assignment to clusters given the input and output of the program STRUCTURE. The script will then partition samples into their structure assigned species, as well as, provide histograms for each locus showing the allele
frequencies of samples that have been assigned to those species. Finally, it will provide conditional probability tables useful for recovering mixed MLGs.

**Symbiodinium goreaui (C1) and C3 genotyping**

To empirically test our species designation approach we collected genetic samples of *Symbiodinium goreaui* (C1) from culture (rt-152 and rt-113, The Robert K. Trench Collection) as well as *Symbiodinium goreaui* (C1) and C3 from Zamami Island, Japan (n=32), and the Great Barrier Reef Australia (n=59) which had been previously collected and identified as *Symbiodinium goreaui* (C1) or C3 or mixtures of the two species by ITS2-DGGE (LaJeunesse et al. 2003, 2004) [Genbank accessions AF380551 *S. goreaui* (C1) and AF380552 *Symbiodinium C3* for ITS 1 and GU111864 *S. goreaui* (C1) and GU111863 *Symbiodinium C3* for ITS2]. These samples were genotyped at 6 diploid microsatellite loci (SgrSpl_13, SgrSpl_22, SgrSpl_25, SgrSpl_26, SgrSpl_30 and Sgr_40) following the methods described in Wham *et al.* (2013).

We chose these two *Symbiodinium* species for several reasons. They are both members of Clade C and conserved primers amplify the same microsatellite loci (Wham *et al.* 2013). They also exhibit widespread geographic distributions in the western Pacific (LaJeunesse *et al.* 2003, 2004) where they associate with the same host species (LaJeunesse *et al.* 2003, 2004; Ulstrup and van Oppen 2003; van Oppen 2005; Abrego *et al.* 2008); yet, they appear to have distinct ecological niches (Ulstrup and van Oppen 2003).

**Background on the Symbiodinium C3 from Howells *et al.* (2013)**

Howells *et al.* (2013) genotyped the *Symbiodinium* clade C symbionts of 401 coral tissue samples from *Acropora millepora*. These samples originated from 7
sites in the central Great Barrier Reef ranging from 0.4-13kms apart. In order to investigate temporal stability as well as regional connectivity samples were gathered over a 12 year period, with sampling occurring in 1997, 2004 and 2009. Single-strand confirmation polymorphism (van Oppen et al. 2005) of the internal transcribed spacer 1 (ITS1) was used to genotype the samples. All samples were identified with this method as *Symbiodinium C2 sensu* Ulstrup and van Oppen (2003) here after referred to as C3 sensu LaJeunesse (2002). The authors genotyped all samples with 8 polymorphic microsatellite loci (Bay et al. 2009) by fragment analysis. Fragment analysis of these samples returned a high frequency of samples with multiple peaks; the authors interpreted this as evidence of multiple individuals occurring in each sample. For this reason all of the alleles in each sample were recorded and analyzed by a presence-absence method. The resulting presence-absence matrix contained a row for each sample (n=401) and a column for each allele that was observed at each locus. The matrix was populated with 0’s and 1’s, with 1’s representing the presence of a given allele in a sample; deposited in Dryad by Howells et al. at (doi:10.5061/dryad.5dh0j).

**Background on the Symbiodinium B1 Data from Wirshing et al. (2013)**

Wirshing et al. (2013) collected samples of *Euniciea flexuosa* from fourteen sites in the Caribbean. Sampling was primarily focused along the Florida reef track with nine of the fourteen sites in the Florida Keys (6 sites, n=42-82) and Biscayne Bay (3 sites, n=54-76). The remaining five sites were spread between three sites in the Bocos del Toro Province, Panama (n=9-25); one site from Punta Cana, Dominican Republic (n=22); and one site in Saba Bank(n=12), a small island 550km east of the Dominican Republic. The ITS-2 region of the rDNA of the symbiont was analyzed by denaturing gradient gel
electrophoresis; all samples were determined to be *Symbiodinium* B1. The authors then developed four microsatellites markers for the host, *E. flexuosa*, and five microsatellite markers for the symbiont. All samples were genotyped with these markers.

The resulting allele calls were divided into three datasets, our analysis is based on these three datasets which were deposited at Dryad (datadryad.org) by Wirshing et al. (2013) under (doi:10.5061/dryad.82ms3). The allele calls for *Symbiodinium* B1 were placed in two data files, a presence absence matrix (totalmatrix_structure) and the unambiguous MLGs, samples with only one allele scored at each locus, (singlehaps_structure). The presence-absence matrix contains a row for each sample (n=610) and a column for each allele that was observed at each locus. The matrix was populated with 0’s and 1’s, with 1’s representing the presence of a given allele in a sample. The rows of totalmatrix_structure were not labeled with the alleles they corresponded to; we assigned rows to allele sizes by a process described later in *Assigning Allele Sizes to data from Wirshing et al. (2013).* Samples that contained only one allele call per locus (n= 326 or 53% of all samples) were also placed into a second dataset with their allele sizes at each locus.

*Assigning Allele Sizes to data from Wirshing et al. (2013)*

The data file totalmatrix_structure did not have the rows labeled with the allele size. In order to retrieve the allele sizes for these rows we cross-referenced the size calls made in the singlehaps_structure file with their presence absence calls in totalmatrix_structure. This allowed us to label all alleles sizes that appeared in the presence absence matrix that also appeared in the MLG dataset. Several rare alleles (alleles that did not appear in the 326 individual singlehaps_structure MLG dataset) in the
presence absence matrix could not be labeled in this way because they only appeared in mixed samples. Samples with these rare alleles were removed from our allele size analysis since their size was unknown to us.

Results

Simulated Data

The simulated dataset contained a total of 76 samples, of these samples 56 samples contained a single allele call at all ten loci. The identity of these 56 samples was then evaluated to remove clones resulting in 33 unique genotypes and 23 repeated clones. We performed a STRUCTURE analysis on these 33 unique genotypes, at k=2 we observed clusters with high per individual probability of assignment, the Evanno method returned k=2 as the most likely value for k (Fig. 3-2a). Under a single species hypothesis these data are expected to form a single cluster; the STRUCTURE analysis however returned high assignment probabilities for individuals at k=2 so the assignment probability for individuals were retained for allele size and Bruvo distance analysis. We therefore made a histogram of allele sizes colored by cluster assignment (Fig. 3-2b).

Accepting the cluster assignments from the STRUCTURE analysis as the probabilities of assignment for each individual to each species, we then allocated the mixed samples to species and individual assignments when a single unambiguous MLG could be composed from a mixed sample that had > 95% probability of membership to one of the two species based on the conditional probability of species assignment of alleles (Fig. 3-2c). Through this process, we recovered an additional 23 unambiguous MLGs, 11 of these MLGs were new unique genotypes and, with their inclusion, increased the sample size of MLGs in the full dataset to 44, a 25% increase. When these
Figure 3-2) Population genetic analyses of simulated data from homogenous and mixed genotype samples. a) Structure plot showing unambiguous multi-locus genotypes (MLGs) that resolve two distinct populations, blue and orange, at k=2. The section with black bars refers to samples with mixed genotypes. b) Analyses of allele frequencies at 3 example loci show bimodal distributions that correspond, and are color coordinated, to evolutionarily distinct populations, or species, with obvious allelic divergence. These bimodal data are then used to parse individual genotypes out of samples with mixtures of two genotypes representing each species; c) Structure plot presented with additional MLGs recovered from mixed samples at k=2. (d and e) These data are then partitioned by species for subsequent population level analysis that examines for genetic structure over various spatial scales.
assignments were compared to the original data file we found that all species assignments agreed with the actual species designations and that all MLGs inferred from parsing mixed samples reflected the actual MLG of the underlying samples. An additional STRUCTURE analysis with the newly partitioned mixed samples showed a high probability of membership of recovered genotypes to their respective species clusters with no evidence of structure at higher clustering within each group (Fig. 3-2d and e).

**Pacific populations of Symbiodinium goreaui (C1) and C3.**

A total of 78 samples were scored that had been previously identified as Symbiodinium goreaui (C1), C3, or mixtures of both species. The resulting dataset contained 65 samples (53 samples of S. goreaui (C1) and 12 samples of C3) with unambiguous multi-locus genotypes and 13 mixed samples. A structure analysis was performed on these 65 samples which returned high probability of assignment for all samples at K=2, the k-value selected by the Evanno method as the most likely k-value (Fig. 3-3a). These assignments did not follow the geographic partitioning of the data, rather, the assignment of samples to clusters matched their ITS2 sequences observed in DGGE rDNA fingerprinting analyses. The correlation between structure assignment and allele sizes was then evaluated which showed a strong relationship between allele size distributions and structure assignment (Fig. 3-3a). Mixed samples were then allocated to species assignments when a single unambiguous MLG could be composed from a mixed sample that had > 95% probability of membership to one of the two species based on the conditional probability of species assignment of alleles. These newly recovered MLGs were then added to the dataset and showed allele size and composition similar to the
Figure 3-3) Analysis of MLGs from two *Symbiodinium* species lineages that can co-occur in animals from the Great Barrier Reef and Zamami Island, Japan, in the Okinawan archipelago. The MLGs of *Symbiodinium goreaui* (C1), blue, and C3, yellow, are well resolved from each other in a STRUCTURE plot at $k = 2$. Several additional MLGs were obtained from corals with mixed populations comprising one genotype from each species by relying on the secondary analysis of relative allele frequencies diagnostic for each species (see Fig. 1 and text for explanation). These later analysis increased the proportion of MLGs for each species, which could then be used for independent population genetic analyses.
unambiguous MLGs from the non-mixed samples (Fig. 3-3b). We then performed a STRUCTURE analysis on these two species separately and selected k=1 as the most likely k-value for both species (Fig. 3-3b) since we only observed marginal increases in the likelihood at higher k-values.

Analysis of data from Howells et al. (2013).

The *Symbiodinium* C3 (sub-clade C2 *sensu* van Oppen et al. 2001) data from 8 loci (n=401) contained zero samples that could be interpreted as unambiguous haploid genotypes. In our analysis we interpreted these data as arising from diploid individuals, consistent with the observations of Wham et al. (2014) and the interpretations of population genetic data from other related Clade C *Symbiodinium* by Thorhill *et al.* (2014) and with the genotypes of the monoclonal clade C cultures examined in the present study. Even under this interpretation unambiguous MLGs could not be interpreted from the raw data since more than 2 alleles were observed at most loci. This pattern, however, was not consistent across all loci. Locus C1.05 and C3.02 returned 1.94 and 2.04 alleles per sample respectively, while the other six loci averaged 2.97 alleles per sample. Because our model for partitioning samples that comprise mixed species requires at least a few individuals of each species appearing as unambiguous MLGs in the dataset we were unable to proceed with a STRUCTURE analysis and subsequent division of mixed samples. We however continued with our analytical approach by plotting a histogram of the allele size distributions observed by Howells et al. (2013) (Fig. 3-4), assuming that the presence of a second species could be evaluated by the continuity of allele sizes. In all 6 loci that contained a high frequency of alleles per sample we also observed discontinuous allele sizes (Fig. 3-4). We then noted the
Figure 3-4) Histograms for allele size distributions from data generated by Howells et al. (2013). Allele size histograms based on presence absence of allele calls recorded for all samples of *Acropora millepora* on the Great Barrier Reef (n =401 ). The bimodal distribution evident for several loci which indicate the existence of two genetically distinct *Symbiodinium* spp. in their data set. A few loci (indicated by an asterisk) appear to only successfully amplify one species based on the lower number of allele calls over the same number of samples (1.94 and 2.04 alleles per sample compared to an average of 2.97 alleles per sample across all other loci). Also note that two loci contained allele sizes that theoretically could not occur at the locus based on the minimum size calculated after removing the repeat motif (dark grey shading).
minimum allele size for each locus by subtracting the entire repetitive section of each locus from the reference sequence on GenBank (GenBank GQ254810-16). In several cases the discontinuous allele sizes span a range that was less than the size that would be expected if the entire repeat region was lost.

*Patterns found in the reanalysis of data from Wirshing et al. (2013).*

The *Symbiodinium B1* data from 5 microsatellite loci (n=326) had already been separated from mixed samples (n=610) into the data file “singlehaps_structure”. We then removed individuals with more than one missing allele prior to further analysis reducing the total number from 326 to 271 samples. We observed the greatest likelihood value for the data at k=6, similar to Wirshing *et al.* (2013) who identified k=7 (Fig. 3-5a). However, the most significant change in likelihood and the k-value suggested by the Evanno method was k=2 (Fig. 3-5a). When we removed all repeated genotypes within a sample site, allowing identical MLGs across sites, the genetic data again partitioned into 2 clusters (Fig 3-5b). Like Wirshing *et al.* (2013) we detected a non-hierarchical pattern of clustering as k-values increased (Fig. 3-5a and b, Also see Wirshing *et al.* 2013 supplemental Fig. S1A), that is, the clusters did not follow the geographic partitioning of the samples. This non-hierarchical clustering was first seen at k=2 and supported by the Evanno method as the most significant difference in data likelihood between successive k-values. In a histogram of allele sizes based on individuals assigned to one of two genetic clusters, all five loci showed differences in allele sizes corresponding to clusters 1 (yellow) and 2 (blue), respectively (Fig. 3-6). As expected by the stochastic process of neutral genetic drift, some loci had more pronounced differences in size-range differences than others. The difference in average allele size between samples assigned to the two
Figure 3-5) The partitioning and reanalysis of microsatellite data from Wirshing \textit{et al}. (2013). a) Structure plots (k=6, similar to results published previously by Wirshing \textit{et al}. (2013), and the optimal k=2) of microsatellite data for \textit{Symbiodinium} obtained from the west tropical Atlantic gorgonian, \textit{Eunicea flexuosa} and a cryptic sibling species. k=2 indicates the existence of two \textit{Symbiodinium} species (blue and yellow) that co-occur at several locations where host populations were sampled. These likely correspond to the high-light and low-light adapted lineages proposed by Prada \textit{et al}. (2014). b) Structure analysis of data that were processed further by the removal of duplicated genotypes found at each location and the recovery of additional genotypes from mixed samples by parsing alleles into distinct genotypes [after analysis of allele frequency distributions (see Fig. 4)], which correspond to each species. c) Bayesian analysis of the MLGs from the most common species of \textit{Symbiodinium} resolves weak population genetic structuring, but support for this is limited by the number of loci assessed and number of genotypes available from each location. Cluster assignments for individual genotypes at k values of 1 and 2 are shown. d) Analyses of the second and rarer species indicates the possibility of a second genetically distinct entity, yet this dichotomy appears to be driven by the presence of allele 180 at locus Sym211 in some individuals. Additional k values show no further genetic structuring for this species.
Figure 3-6) Histograms for allele size distributions from both raw and processed data generated by Wirshing et al. (2013). a) Allele size histograms based on presence absence allele calls recorded for all samples (n = 610). b) Allele size histograms based only on samples where a single and complete unambiguous MLG was recovered (n = 271). The bimodal distributions demark separate size ranges which correspond to and support the supposition that two genetically distinct *Symbiodinium* spp. exist in association with *E. flexuosa* (sensu lato, Prada et al. 2014). Notice that some individuals assigned to the yellow cluster have allele sizes associated with the purple cluster. Additional analysis is needed to resolve this as evidence of introgression or simply mix samples that only successfully amplified the alleles of one or the other species.
clusters ranged from 20 bases to 75 bases with allele size differences present in all 5 loci (Fig. 3-6). We also made a histogram of pairwise Bruvo distances between samples (Fig. 3-7a) and a hierarchical clustering network of pairwise Bruvo distances (Fig. 3-7b) to investigate the relationship between allele sizes at all 5 loci and the STRUCTURE assigned cluster of each sample. We then partitioned individual assignments representing each species from mixed samples. We recovered an additional 46 unambiguous MLGs, 27 of these MLGs were new unique genotypes. These 27 new MLGs comprised of 11 MLGs assigned to cluster 1 and 16 MLGs assigned to cluster 2 (Fig. 3-5c).

After allocating samples to individual and cluster designations we examined the degree of genetic structure across sampling locations within each cluster separately. Using the Evanno method applied to the results of the program STRUCTURE the most likely clustering of the two species was K=2 for cluster 1 and K=1 for cluster 2 (Fig. 3-5c). While the Evanno method suggested K=2 as the most likely k-value for cluster 1 the clustering of individuals into these two groups did not strongly follow the geographic partitioning of the data nor did the cluster assignment correlate with allele size differences (e.g. Fig. 3-6).

Discussion

Population genetics offer valuable tools for estimating migration rates and connectivity among populations, including eukaryotic microbes. However, errors of interpretation can arise when unrecognized species occur in datasets and go undetected
prior to downstream statistical analyses (Pante et al. 2015). Genetic differentiation may occur independent of morphological differences. Therefore, when using population

Figure 3-7) Summation of allele divergence at all microsatellite loci for two putative species lineages not recognized by Wirshing et al. (2013), but proposed by Prada et al. (2014) predicated on their finding of DNA sequence divergence. a) A Neighbor Joining Tree based on Bruvo Distance comparisons shows a large break in genetic distance between species lineages b) a histogram of pairwise Bruvo Distance showing intraspecific variation (black) is less than interspecific variation (grey). Identical clones are indicated by the white bar.
genetic approaches, it is important that individual samples be correctly resolved into species whose populations share the potential for reproductive genetic exchange.

We reason that multi-locus genotype (MLG) data, especially among taxonomically problematic organisms, can be used to directly test for patterns of sexual reproduction between co-occurring (sympatric) individuals obtained from one, or several, locations. In cases of closely-related species, population-level allelic data can be used to examine reproductive isolation between populations in sympathy based on linkage disequilibrium and heterozygote deficiencies (Hebert et al. 2003, Hebert et al. 2004, DeSalle et al. 2005, Percyra et al. 2009). In cases where morphology and traditional genetic techniques produced discordant (Pinzon and LaJeunesse et al. 2011, Pinzon et al. 2013), or ambiguous, results (Ross et al. 2009, Ladner and Palumbi 2012, LaJeunesse et al. 2014, Thornhill et al. 2014), this method offers the best evidence for delimiting distinct biological entities because sexual reproduction is the fundamental component of most species concepts in eukaryotes (Mayr 1942, Mallet 1995, Mayr 1995, De Queiroz 2005, Coyne and Orr 2004).

Our simulation of the divergence between two populations shows how allele sizes and frequencies become highly distinct for isolated populations over evolutionary time (Fig. 3-2). These artificial data illustrate how the genetic isolation of sympatric populations has, over time, an observable signature of genetic divergence. Subsequent analysis of dispersal and genetic connectivity could only be correctly conducted when this pattern was recognized and individuals were partitioned into their respective species; (Fig. 3-2d and e).
The effectiveness of this method depends in part on sufficient divergence of allele frequencies between reproductively isolated groups (Fig. 3-2b). Below we discuss several case studies, one new and two recently published datasets, where we have applied this screening to more accurately analyze and interpret findings, and then discuss the disparity in old vs. new conclusions created by each reanalysis.

*Population genetics of* S. goreaui (C1) *and Symbiodinium C3 (Pacific)*

Having demonstrated the expected patterns in data that contain two species in sympathy with simulations, we presumed that the same patterns would be observable in microsatellite data generated from samples that comprised two species of *Symbiodinium*. While only two described species exists for Clade C, *Symbiodinium goreaui* (Trench and Blank 1987) and *S. thermophilum* (Hume et al. 2015), there is considerable phylogenetic and ecological evidence that point to the existence of numerous species within this group (Sampayo et al. 2009; Thornhill et al. 2014). Albeit small, fixed differences in rDNA sequence differentiate *Symbiodinium goreaui* (C1) from *Symbiodinium C3* where they co-occur in the Pacific. Analysis of microsatellite evidence verifies that no genetic exchange occurs between them (Fig 3-2), and hence they are separate species (LaJeunesse et al. 2003; Thornhill et al. 2014), with clear ecological (LaJeunesse et al. 2003, 2004) and physiological differences (Berkelmans and van Oppen 2006; Abrego et al. 2009).

Consistent with estimations of evolutionary divergence (Thornhill et al. 2014), our analysis of MLGs from *S. goreaui* and *C3* (Pacific) showed that the microsatellite allele size ranges for each lineage were significantly different and/or several samples where mixtures existed we were able to separate distinct MLGs representative of each species. If population genetic data from these two species were mistakenly combined, this
would significantly affect $F_{st}$ summary statistics (used to make inferences about gene flow). Indeed, changes in the relative frequency of each species would produce significant differences in statistical calculations. For example, $C3$ was sampled more often in Japan than in Australia (Fig. 3-3). This difference in relative abundance results in an $F_{st}$ estimate of 0.242 ($p=0.001$) between Japan and Australia. However, when the data are correctly partitioned, $F_{st}$ values are relatively low for $S. goreai$ (0.063, $p=0.001$) and statistically insignificant for $Symbiodinium C3$ (0.021, $p=0.142$). While one analysis suggests very strong genetic structure, the correct analysis indicates a high degree of connectivity between two distant locations (Fig. 3-3).

**Genetic structure at scales of 1-10 Kms in Symbiodinium?**

One of the primary findings of Howells et al. (2013) was that 19% of the total genetic variation among $Symbiodinium$ found in $Acropora millepora$ populations was explained by site of sample collection. The windward or leeward sides of islands had a greater effect on the clustering populations of $Symbiodinium$ in $Acropora millepora$ than geographic proximity (Howells *et al.* 2013). This was interpreted as evidence of significant genetic structure occurring in $Symbiodinium$ over relatively small geographic distances. The authors concluded that this genetic structure was due to limited dispersal ability, or “limited oceanographic transport”. They also observed a very high frequency of samples with multiple alleles, which were interpreted as arising from a mixed population of individuals of the same symbiont species in each sample. With this assumption, the authors justified that their data could not be integrated into subsequent genetic analyses.
Our re-analyses of these data indicate that the co-occurrence of two species, rather than mixed individuals of the same species, explains the high frequency of alleles per sample. Indeed, most loci (6 of 8) contained bimodal allele size distributions similar to those observed in the simulated data and the data presented from *S. goreaui* (C1) and *Symbiodinium C3* (Fig. 3-4). In several cases, allele sizes are too small to be explained by the loss of the microsatellite repeat motif alone, indicating major alterations (sequence deletions) occurred in the flanking regions (e.g. allele sizes between 131 and 149 at locus C1.07). The bimodal distribution of allele sizes across most loci indicate that allele calls from a second distinct species and/or the products of non-specific binding (false alleles scored from non-specific PCR products) occur in this dataset.

The two loci (C1.05 and C3.02) that do not show a bimodal distribution have a significantly lower number of alleles per sample (1.94 and 2.04 alleles per sample respectively) than the other six loci (averaging 2.97 alleles per sample). Indeed, 87% (348 of 401) of the samples could be interpreted as unambiguous MLGs at these two loci indicating that they arise not only from a single species but also from a single genotype of that species in most cases. We believe the markers for these loci are specific to their target species, unable to amplify the second species. Using the number of alleles observed per sample at these loci compared to the number of alleles observed per sample in the other six loci to extrapolate, approximately 27% of the dataset may have arisen from a second species.

When multiple species are erroneously counted as one genetically cohesive entity, differences in the relative frequencies of each species collected at each location alone can explain patterns of genetic structure over extremely small distances (see previous
discussion of *S. goreai* (C1) and *Symbiodinium* C3). With the recognition that there are at least two species of *Symbiodinium* represented in their datasets the genetic structure observed between locations is better explained by differences in the relative proportions of each species collected rather than barriers to dispersal, the lack of gene flow, or selection (local adaptation). The reinterpretation that species with distinct niches are more prevalent in some environments and less common in others, dramatically changes the results and interpretations of the population genetic analyses; and hence the conclusions of Howells *et al.* (2013)

*Vectored dispersal from Panama to Florida?*

The existence of a second unrecognized species appears to have confounded the analyses, interpretations and conclusions presented by Wirshing *et al.* (2013). From their analyses, the authors hypothesized that symbionts from Panama had been recently introduced (vectored) into host populations living in the Florida Keys. However the recent findings of Prada *et al.* (2014) identified two phylogenetically distinct lineages of *Symbiodinium* one of which associated specifically with a newly recognized sibling species of *Eunicia flexuosa*. This cryptic species, which is easily confused with *E. flexiosa*, is generally found in low-light habitats corresponding with greater water depths or in environments with high water turbidity such as Panama (Prada *et al.* 2013). Indeed when the allele calls from published raw data were analyzed in detail, gaps in allele size ranges were discovered at most loci, indicating that the sample set comprised separate symbiont species with distinct gene pools (Fig. 3-6), and further substantiating the existence of a second *Symbiodinium* species (Prada *et al.* 2014). Members of each species possess the same intra-genomically dominant ITS2 sequence (sequence B1). This could
explain why data from Wirshing et al. (2013) was analyzed as if representing a single species.

This alternative explanation accounts for why K=2 best describes the number of populations to which these data are partitioned (based on the ΔK method; Evanno et al. 2005). Locations that were comprised of mostly one species are the locations where Wirshing et al. (2013) observed linkage equilibrium. Consequently, significant values of disequilibrium were reported in populations with both species present. This linkage disequilibrium was significantly reduced when Wirshing et al. (2013) divided samples into separate populations which, as we suggest here, had the effect of separating samples into species (Fig. 3-5). Finally, the five “vectored individuals” from Panama to the Florida Keys belong to the low-light adapted Symbiodinium sp. (Prada et al. 2014). This second species was sampled more rarely, but nevertheless occurred at most Caribbean locations (Fig. 3-5b). Our genetic analysis on the populations of each species shows further, albeit weak genetic structuring (Fig. 3-5c). This additional genetic structuring within each species appears to have contributed to the k=7 adopted by Wirshing et al. (2013). Analyzing these data at k=7, however, may have obscured the relationship between the five individuals Wirshing et al. (2013) focused on and the numerous other closely related individuals in the dataset that we here interpret as a separate species (Fig. 3-5b&c).

**When population genetic data have limitations**

The recognition of species boundaries depends on sufficient divergence of allele frequencies between reproductively isolated groups (Fig. 3-1b). In our simulated example, as well as the examples from empirical datasets, large differences were
observed in allele frequencies. This resulted in allele size ranges that were diagnostic of each species. However, this method is effective in recognizing less diverged groups that even share allele sizes at different frequencies, as in the *Symbiodinium* Clade D species described by LaJeunesse et al. (2014). It is worth noting, however, that the likelihood of type II error (*i.e.* failing to recognize two reproductively isolated units) goes up as the degree of genetic divergence goes down; the approach may also be prone to type II error when there are very few individuals of the second species in the dataset. This can be addressed by including additional samples and/or additional loci if researchers suspect that a second species is present.

The approach outlined in Figure 3-1 is less prone to type I error, (*i.e.* resolving two species that are actually just one species). Following the outline of analysis in Figure 3-1, a dataset could be recognized as arising from a single species in one of the following ways: 1) at K=2 the existence of low or split assignment probabilities among individuals in each cluster; or 2) the assignment probabilities for individuals in each cluster is high, but there is no difference in allele identity between clusters. Finally, there could be a single species with population structure, in which case there would be high cluster assignment probabilities for individuals and allele identities might correspond to these clusters. The clusters, however, would also correspond to sample locations (Fig. 3-1a). In this case other lines of evidence must be used to decide if these groups are populations of the same species or different species.

The ability to resolve MLGs from mixed species samples also depends on the degree of divergence. As the degree of genetic divergence goes down there is less certainty in the assignment of alleles at each locus to a single species resulting in a higher
rate of type II error (i.e. failing to recover an MLG in mixed sample). Conversely, the rate of type I error (i.e. accepting a new MLG that is actually a mixture of individuals) is relatively insensitive to the degree of divergence, being primarily related to the frequency of mixtures of the same species and errors in genotyping. Researchers should note that the inclusion of additional, real MLGs only marginally improves the inferences that can be made from the data but the inclusion of false MLGs could cause spurious conclusions. It is for this reason our approach is designed to favor type II errors over type I errors. That is, we favor leaving out MLGs from mixed samples over including them when there is clear uncertainty. One of the strengths of the Bayesian model is that the statistical significance of this uncertainty is easy to assess. In this way, the approach’s utility scales with the genetic distance of the clusters it is attempting to resolve but can be robustly applied to clusters that span any degree of divergence.

**Best practices in the analysis of population genetic data from *Symbiodinium***

During the analysis of population genetic data from *Symbiodinium*, we discourage retaining all allele calls and then analyzing presence/absence of alleles within populations (e.g. Howells et al. 2009, 2013, Wirshing et al 2013). While this method may seem conservative in that it retains all data, it presumes that all alleles belong to the same gene pool. For this reason community level processes (mixtures of species affected by environmental conditions) cannot be resolved from population level processes (dispersal, genetic connectivity, selection). Relying on unambiguous MLG data increases statistical power and allows for the use of the Hardy-Weinberg and linkage equilibrium models in population genetic analyses. This is still preferable despite the loss of samples with
mixtures of genotypes that cannot be partitioned into distinct MLGs, which ranges as much as 50% but typically between 10-20% of the raw dataset.

**Conclusion**

Through our analyses, we show that observing the occurrence of highly structured gene pools in sympathy should be interpreted as arising from reproductively isolated populations where individuals reproduce within their species to the exclusion of the other; the central concept of species under the biological species concept (Mayr 1942, 1995). As we have shown, this pattern can be used to recognize species in datasets even when phylogenetic markers may fail to adequately resolve them. The availability of population genetic data for *Symbiodinium* presents opportunities to investigate barriers to dispersal, clues to whether populations have undergone genetic bottle necks because of severe selection or recent introductions to new locations, and sufficient evidence to recognize the existence of reproductive isolation. Applied correctly, these data will help to expand our knowledge of *Symbiodinium* ecology and evolution. Moreover, when combined with phylogenetic and ecological data, population genetic markers offer a promising means of defining morphologically cryptic species of microbial eukaryote (LaJeunesse et al. 2014).
Data Accessibility

All multi-locus genotypes produced from simulations as well as samples of *Symbiodinium goreau* and C3 are archived on the Dryad Digital Repository doi:10.5061/dryad.66f8c. Scripts for the analysis and partitioning of species described in the method section are also available for download from the authors site thecoalescent.com/bayesallele
References


Frade PR, Englebert N, Faria J, Visser PM, Bak, RPM (2008). Distribution and photobiology of *Symbiodinium* types in different light environments for three colour morphs of the coral *Madracis pharensis*: is there more to it than total irradiance?. *Coral Reefs*, 27 (4), 913-925.


Lajeunesse TC, Parkinson JE, Reimer JD (2012). A genetics-based description of *Symbiodinium minutum* sp. nov. and *S. psygmophilum* sp. nov. (Dinophyceae), two dinoflagellates symbiotic with cnidaria. *Journal of Phycology*, 48 (6), 1380-1391.

LaJeunesse TC, Wham DC, Pettay DT, Parkinson JE, Keshavmurthy S, Chen CA (2014) Ecologically differentiated, stress tolerant endosymbionts in the dinoflagellate genus *Symbiodinium* Clade D are different species. *Phycologia*


Wham DC, Carmichael M, LaJeunesse T C (2014). Microsatellite loci for *Symbiodinium goreauii* and other Clade C *Symbiodinium*. *Conservation Genetics Resources*, 6(1), 127-129.


Chapter 4

THE COALESCENT THEORY OF CLONAL ORGANISMS

Abstract

The rate of sexual recombination describes the relative frequency of sexual and asexual reproduction in organisms with mixed modes of reproduction. This rate is influenced by both intrinsic biological factors and external forces acting on the population, affecting the tempo and mode of evolution in lineages that rely partially on clonal propagation as well as sexual recombination for maintaining or growing a population. The predominance of one mode over the other is of interest in many contexts to the fields of evolutionary, agricultural and medical biology. Here we propose new methods, based on the coalescent model, to estimate a population’s sexual recombination rate ($\psi$) and to describe the distribution of genotypes ($D^\psi$) from multi-locus genetic data. Simulations are used to explore the accuracy, precision and the potential for biases of $\psi$ and $D^\psi$ at various levels of sexual recombination, genotypic diversity and sample size. These simulations demonstrate the advantages of using $\psi$ and $D^\psi$ in describing data containing various amounts of clonality over commonly used summary statistics. We then apply this method to previously published multi-locus genetic data from a reef-coral (Pocillopora damicornis, a reef coral), a highly clonal tree (Populas tremuloides, an Aspen), and a partially clonal human pathogen (Plasmodium falciparum, the causative agent of malaria). Using this approach, we investigated elements of these datasets that could not be addressed with previous analytical techniques. We demonstrate that a coalescent-based model exceeds the performance of other models at inferring population
level parameters, particularly when sample sizes are low. Finally, we highlight how these new analysis lead to more accurate biological inference.

**Introduction**

The life-history of many organisms includes clonal reproduction. This reproductive strategy is particularly common in protists, plants and invertebrates. Many of these taxa are of major ecological, economical and medical importance including reef-building corals, agricultural pests and parasitic human pathogens. The frequency of sexual reproduction in these organisms can have a large effect on their ability to respond to environmental change by influencing the way new mutations spread across a population. The process of sexual recombination facilitates the combination and spread of favorable mutations rapidly creating new genotypic variability. Alternatively, clonal reproduction allows for the rapid proliferation of successful genotypes allowing for genetic stability over multiple loci. Organisms with mixed modes of reproduction may be able to shift between strategies in order to best respond to environmental pressures. The relative contribution of one reproductive strategy to the other may be constrained biologically in some organisms. These less flexible organisms may not be able to fully take advantage of alternative reproductive strategies. For this reason, knowledge about the rate of sexual recombination can have profound consequences in best practices for conservation, drug resistance and pesticide/herbicide resistance.

Estimating the rate of sexual reproduction in partially clonal organisms from genetic data has been historically problematic. An established approach employs summary statistics such as linkage-disequilibrium (LD) and the inbreeding coefficient \( F_{is} \) to detect sex frequency (Meeus and Balloux 2004). These statistics are often
informative in detecting the presence or absence of sexual recombination. They, however, preform poorly in providing estimates of sexual recombination rates, and can be strongly influenced by other factors such as allelic diversity. Other approaches use the ratio of unique genotypes to the number of samples (G/N) (Ellstrand and Roose 1987) or the ratio of genotypic diversity in a dataset to the expected under Hardy-Weinberg (Stoddart and Taylor 1988). While these approaches provide the statistical ability to compare the contribution of sexual reproduction across multiple populations, however they do not provide the actual rate of sexual reproduction because they do arise from an underlying genetic model. Further these methods of describing clonal rates suffer from significant biases when ramet distributions are skewed among genets (Arnaud-Haond et al 2007) and when clonality is high (Gorospe et al. 2015).

While these approaches do not allow for the direct assessment of the rate of sexual recombination they take advantage of the signature of asexual reproduction in genetic data. This signature is generally found in significant values of LD and F<sub>is</sub> (Meeus and Balloux 2004). Individuals in asexual populations may have abundance greater than 1/n and therefore may be sampled more than once, which strongly contributes to significant values of LD and F<sub>is</sub>.

Statistical significance of these summary statistics primarily comes from repeated genotypes in a sample. For this reason, studies involving clonal rates often simply report the ratio of unique genotypes to the number of samples or the ratio of observed genotypic diversity in a dataset to the expected genotypic diversity under Hardy-Weinberg. This ratio provides a basic description of the relative contribution of sexual and asexual
reproduction to the sample. The inability to relate this ratio to an actual rate of sexual reproduction is, however, a serious limitation.

Coalescent theory provides an evolutionary model appropriate for inferring the per generation per individual rate of an event that is detectable in a genealogical sample. The coalescent process provides a description of the genealogical history of a set of samples from a large Wright-Fisher (Fisher 1930, Wright 1931) or Moran (1958, 1962) population of constant size with no selection or migration (Kingman 1982). Events such as mutation or recombination with constant rates per individual can be mapped onto this genealogy. Under the neutral coalescent, the shape of the genealogy and the placement of events on that genealogy is a function of both the effective population size, \( N_e \) and the per individual per generation rate of the event (\( \lambda \)). If there is no information available for \( N_e \) the model can only estimate the population event rate (E) rather than the event rate itself (\( \lambda \)). This is because large populations with low mutation rates produce similar coalescent trees to small populations with high mutation rates. The population event rate is equal to \( 4N_e \lambda \) in diploids and \( 2N_e \lambda \) in haploids (Wright 1931, Kimura 1964, Tajima 1984, Hubbel 2001, McVean et al. 2002). Classically, mutation is the event that is being mapped to a genealogy and in this case the population mutation rate (\( \theta \)) is determined by the effective population size and the per individual per generation mutation rate (\( \mu \)), often written as \( \theta = 2N_e \mu \) (Tajima 1984).

In this study we extend the coalescent to infer the relative contribution of sexual and asexual reproduction to a set of MLGs. Additionally, we propose an Markov chain Monte Carlo approach to test hypothesis regarding the rate of sexual recombination in a population including the potential for data to arise from completely sexual and asexual
reproduction. We also propose an additional statistic ($D^\psi$) to describe the distribution of genotype frequency compared to their expected neutral distribution, allowing for model based assessment of deviations from this distribution. We then demonstrate the accuracy of this method at various levels of sexual recombination, diversity, sample size, and locus number. Finally we apply this method to several empirical datasets from symbiotic dinoflagellates, parasitic protozoa, pathogenic fungi and agricultural pests.

We then examine the performance of ($\psi$) over various underlying sex rates, comparing its performance to the commonly used statistic G/N (Ellstrand and Roose 1987). We then move on to develop a second summary statistic $D^\psi$ to test MLG data for deviations from the neutral model and discuss the potential biological significance of these deviations. Finally we introduce three previously published datasets that these models can be applied to.

**Methods**

We began by developing a coalescent model for multi-locus genotypes (MLGs) from a population of individuals undergoing both sexual and asexual modes of reproduction. Based on this model we propose the population sex rate ($\psi$) as a summary statistic for describing clonality in natural populations.

**Applying the Coalescent to Multi-Locus Data**

The genotype of an individual at multiple (often unlinked) loci collected together creates a multi-locus genotype (MLG). These MLGs may comprise genotypes from any type of polymorphic genetic marker, including allozymes, microsatellites, single nucleotide polymorphisms or next-generation sequence data. The field of population genetics has a long history of studying genetic patterns through the analysis of collections
of MLGs from populations. The most common analytical approaches, however, model patterns of variation within loci (Wakeley et al. 2010) or patterns between pairs of loci (McVean et al. 2002). Here we develop a model-based analyses of full genotypes, inferring the rate of the creation of new MLGs as the rate of sexual recombination in the genealogy of a sample of individuals from a population. The approach follows the basic logic of the coalescent under an infinite allele model (Ewens 1974) but with some additional assumptions. We outline these assumptions in the following sections:

**Assumptions: Data**

MLG datasets may contain any number of loci greater than two and these loci may vary in their degree of polymorphism. Here we assume that a sufficient number of polymorphic loci have been collected so that the probability of identity of two MLGs without identity by descent is negligible. That is, assuming sexual reproduction was the only means of reproduction in the population, the expected number of unique genotypes ($G_e$) is effectively equal to the number of samples ($n$). This assumption is frequently meet by modern polymorphic markers (Baums et al. 2006, Mock et al. 2008, Gorospe et al. 2015) and next-generation sequencing data with a large number of bi- or tri-allelic polymorphisms (Nkhoma et al. 2013). Thus, when sufficient data is collected the MLGs of individuals produced from sexual reproduction are always recognizable as new unique genotypes in the population. Conversely, the MLGs of individuals produced from asexual reproduction are identical (with the exception of mutation, discussed in the following section).

**Assumptions: Mutation Rate is Low**
We assume that the mutation rate is low relative to the rate of sex. The models developed here specifically assume that the mutation rate ($\mu$) is low enough that $Se/(\mu+Se) \sim 1$. This insures that the occurrence of clone mates with mutation events between them can not be confused with distinct MLGs. In practice, however, the mutation rate only needs to be low enough that there is a gap between the highest observed pair-wise distance (PWD) between clones and the lowest PWD observed between sexual genotypes (Meirmans and Van Tienderen 2004). This allows for the possibility of non-identical clone mates (ramets) so long as these clone families can be unambiguously resolved from unique sexual individuals (genets). A number of studies have focused on methods of genet assignment (Douhovnikoff and Dodd 2003, Meirmans and Van Tienderen 2004, Arnaud-Haond et al. 2007) and a number of programs are available to perform genet assignment including GenoType and GenoDive (Meirmans and Van Tienderen 2004) and GENCLONE (Arnaud-Haond et al. 2007). It is also important to note that asexual populations can produce MLGs with high mean PWDs so it is important to tests if datasets with large PWDs can be assumed to originate from sex events. One way of doing this is to employ four-gamete tests (Hudson and Kaplan 1985) to assess the number of events in a dataset that are incompatible with an asexual model of the data.

**Assumptions: Individuals are part of a population and Sampling defines populations**

We further assume that the MLGs in a sample set are related to one another by a coalescent genealogy, extending the assumptions of the coalescent model (Kingman 1982).
This analysis is flexible in that it can be used to analyse samples that were
gathered in a number of different ways, however the method of sampling defines the
extent of what is considered a population. Thus the extent of sampling is reflected in the
resulting inferred rate of sex. For instance, if clone mates are spatially correlated
(identical clones are more likely to occur near one another) samples taken at regular 1m
intervals from the same population would return a lower rate of sexual reproduction than
samples taken at 10m intervals from the same population. Such a result correctly reflects
the difference in the rate of sexual recombination at different scales. When used correctly
this allows for the precise comparison of datasets, but when differences in sampling are
not accounted for the resulting inferences may be spurious. It is therefore important to
note the sampling method when comparing across collections.

The Population Sex Rate and the Genotype Frequency Spectrum

Under these assumptions the coalescent can be extended to the MLGs themselves
in order to infer the population sex rate of the individuals that gave rise to that set of
MLGs. In this view, generations are defined by reproductive events that can be either
sexual or asexual. Therefore the effective sex rate ($S_e$) and the effective clonal rate ($C_e$)
are related by

\[ 1 = S_e + C_e \]

[4-1]

And the population sex rate ($\psi$) is

\[ \psi = 2N_eS_e \]

[4-2]
Where $N_e$ is the coalescent effective population size which is $N_e = N / \sigma^2$ (Kingman 1982) or $N_e = N / c$ (Sjodin 2005), where $N$ is the number of individuals in the population and $\sigma^2$ is the variance in offspring number (Kingman 1982) and $c$ is a similarly derived constant (Sjodin 2005). Thus, $N_e$ here functions to rescale the population model as a Wright-Fisher model (Sjodin 2005). A sample of MLGs can then be followed backwards towards their most recent common ancestor (MRCA) on a genealogy until samples coalesce or reach a sex event. The ability to describe MLGs on a single genealogy ends at the first sex event and for this reason the infinite alleles model (Ewens 1972) is an appropriate model for these data since we must infer the total rate of the event from data that can only be followed up until the occurrence of the first event. Following the conclusions of Malecot (1948) and Kimura and Crow (1964) regarding identity by decent in an infinite allele model, the probability of any two individuals belonging to the same clone family or genet ($F$) is:

$$[4-3] \quad F = \frac{(1 - S_e)^2}{N_e S_e (2 - S_e) + (1 - S_e)^2}$$

That is, the probability of two individuals belonging to the same clone family or genet is equal to the probability of no sex event occurring on the genealogy between them. It follows from equation [4-3] that the effective sex rate is:

$$[4-4] \quad S_e = \frac{1 - (N_e F)}{\sqrt{N_e F [N_e - 1] + 1}}$$
Sex events create new genotypes in the population and asexual events create copies of those genotypes. The expected number of unique genotypes in the population (E[Ng]) is therefore equal to the sum of the probability of sexual reproduction occurring in the genealogy between all individuals in the population:

\[ E[Ng | N, \psi] = \frac{\psi}{\psi + 1} + \frac{\psi}{\psi + 2} + \cdots + \frac{\psi}{\psi + N - 1} \]

A collection of \( n \) MLGs from a population is then comprised of \( G_o \) unique individuals. Each of these unique genotypes could be represented in the sample with a frequency of 1 to \( n \) times.

Here we define the list of the number of occurrences of each unique genotype as the Genotype Frequency Spectrum (GFS), where the sum of the GFS is \( n \) and the length of the GFS is equal to \( G_o \). Following the tradition of representing allelic configurations in Ewens sampling formula (Ewens 1972), the genotype configuration (GC) can be described by a list \( GC=(a_1, a_2, \ldots, a_n) \) where \( a_j \) represents the number of genotypes that are observed \( j \) times (see Wakely 2009 for additional details on this notation). Under this model the probability of a sample of \( n \) MLGs contains \( G_o \) unique individuals with a given GC is

\[ P\{n, G_o, GC\} = \frac{n! \psi^{G_o}}{\psi(n)} \prod_{j=1}^{n} \frac{1}{j^{a_j/a_j!}} \]

following Ewens (1972).
Testing Deviations from the Neutral Model

Tajma described a statistic for testing for, and summarizing, deviations from the neutral model in infinite site data based on the shape of site frequency spectrums; Tajma’s D (Tajma 1989). The direction of these deviations can suggest potential biological forces that might be affecting the population. Similarly, under this model, different configurations of the GFS given n and $\psi$ have different probabilities. The direction and significance of deviations from the maximum likelihood GFS can be used to make inferences about what biological forces might explain these deviations. Here we describe the statistic $D^\psi$ which is conceptually similar to Tajma’s D but is useful in describing the present data based on the infinite allele model. We define $D^\psi$ as

$[4-7]$

$$D^\psi = \frac{(F_o - E[F|n,\psi])}{E[F|n,\psi]}$$

Where, $F_o$ is the observed probability of any two individuals belonging to the same clone family or genet calculated as:

$[4-8]$

$$F_o = \sum_{j=1}^{n} a_j \binom{j}{2}$$

and $E[F|n,\psi]$ is the expected probability of any two individuals belonging to the same clone family or genet given n and $\psi$ it can be calculated as:

$[4-9]$

$$E[F|n,\psi] = \sum_{j=1}^{n} a_j \binom{j}{2}$$
Where $E[F| n, \psi]$ can also be estimated by simulating GFS with the given $n$ and $\psi$ values and calculating $F_o$ on the simulated data to obtain the posterior distribution of $[F| n, \psi]$. We can then assign significance to $D^\psi$ by evaluating equation [4-7] with the value of $F_o$ obtained from the data and the posterior distribution of $[F| n, \psi]$ to yield the posterior distribution of $D^\psi$. If the 95% HDCI does not include the value of zero the null model can be rejected, suggesting that non-neutral dynamics are shaping the GFS.

**The Biological Significance of Deviations from the Neutral Model**

Under this model the expected value of $D^\psi$ is 0, significant deviations from this expected value in both the positive and negative direction can be observed in data. These deviations suggest different forces that may be acting on the observed system. Deviations in the negative direction occur when there is an excess of low frequency individuals, suggesting that there has either been a recent increase of sexual reproduction, pervasive advantage of asexually derived individuals or a population expansion. Alternatively, deviations in the positive direction occur when there is an excess of middle frequency individuals, suggesting that there has been a recent increase of asexual reproduction, pervasive advantage of sexually derived individuals or a population decline. Biological forces such as demographic change should manifest in both the GFS and the allele frequency spectrum, where as pervasive forces that are acting differentially on sexually or asexually derived individuals should only manifest as deviations in $D^\psi$. The above outlined hypotheses for explaining deviations in $D^\psi$ can therefore be partially resolved by contrasting patterns observed at the genotype level with patterns observed at the allelic level. Then by noting the presence or absence of genotype-allelic concordance different biological explanations for deviations in $D^\psi$ might be favored.
The Clonalescent

We created the program The Clonalescent to implement the analyses described above on genetic data. This program simulates data under the infinite allele model and provides draws from the posterior distributions of $\psi$, $D^\psi$, $F$, $S_e$ and $E[N_g]$. The user specifies a number of draws from the posterior distribution for the program to make for each statistic and then it accepts one or more GFS and, optionally, a $N_e$ value for each GFS as input. The program combines a defuse prior, $G_o=1$, $n=2$ following the rule of succession (Laplace 1814), with the observed data and then uses equation [4-6] to evaluate values of $\psi$ proposed from a Gaussian distribution. This process is continued in a Markov Chain Monte Carlo approach, accepting and rejecting proposed values of $\psi$, to yield an approximation of the posterior distribution of $\psi$. The program then uses values from the posterior distribution of $\psi$ to simulate a new $[\text{GFS}|n]$ for each requested draw from the posterior. These simulated GFSs are therefore variable in their GCs and $G_o$ values capturing the expected variance in GC and $G_o$ based on the data, these simulated GFSs can then be used to approximate the posterior distribution of other summary statistics that are generated from GFSs such as $D^\psi$, $F$, $S_e$ and $E[N_g]$. $F_o$, for example, is calculated for the observed GFS using equation [4-8] and $E[F|\psi, n]$ is estimated by calculating $F_o$ on each of the simulated GFSs yielding a posterior distribution for $E[F|\psi, n]$. This posterior distribution is then used in equation [4-7] to produce a posterior distribution for $D^\psi$. When a value of $N_e$ is provided, equation [5] is evaluated using the posterior distribution of $\psi$ and the provided $N_e$ value to produce a posterior distribution of $E[N_g]$. The posterior distribution of $S_e$ is produced by evaluating equation [4-4] with the
provided \( N_e \) value and the posterior distribution of \( E[F|\psi, n] \). The resulting draws from
the posterior distribution can be exported for evaluation in other statistical software.

The program is implemented in Python which allows for parallel processing of
simulations. This program is available for download at
thecoalescent.com/theclonalescent. Additionally, individual functions for running these
analyses in R are available at thecoalescent.com/theclonalescent.

**Background on Simulations**

We simulated genotype frequency spectrums from a population of 3,000
individuals. In order to compare the performance of the various summary statistics we
simulated data across a broad range of sample sizes and underlying sex rates. We
conducted simulations for \( N_g \) values of 50, 100, 200, 300, 600, 1000, 1500, 2000, 2500;
for each of these values of \( N_g \) we produced GFS based on samples sizes of 20, 30, 50,
100, 300, 1000, 1500, 2000, 2500 individuals. 100,000 GFSs were then made for each
combination of \( N_g \) and sample size totaling 8,100,000 simulations. We calculated \( \psi \), \( D^\psi \),
\( F \), \( S_e \), \( E[N_g] \), \( G_o/n \) and \( \beta \) on the resulting GFS. These GFSs and all summary statistics
calculated on them can be found in the supplementary data file SimulatedGFS3000.csv
Dryad XXXXX.

**Background on *Pocillopora damicornis* data from Gorospe and Karl (2013)**

Gorospe and Karl (2013) exhaustively sampled (2,352 colonies) *Pocillopora
damicornis* in a single patch reef in Kane’ohe Bay, Hawai’i, USA. These colonies were
genotyped at 6 polymorphic microsatellite loci, which resolved 78 unique genotypes
(Gorospe et al. 2015). This data was available at DRYAD (doi:10.5061/dryad.6d8g5).
The authors found that just seven genotypes comprised over 70% of the genotypes on the
reef. In their analysis of spatial relatedness, Gorospe et al. (2015) found that both the high frequency and skewed distribution of clone frequencies contributed to the inaccuracy of G/n.

**Background on *Populus tremuloides* data from Mock *et al.* (2008)**

Mock *et al.* 2008 sampled 2 study areas, Swan Flats (n=812) and Fish Lake (n=561). These sites were sampled by projecting evenly spaced (50-m) grid coordinates over tree stands in each area; if aspens were present, a sample was collected within a 5-m radius of the grid coordinate. At Swan Flats a single sample was taken at each coordinate, however, at Fish Lake both an overstory tree (n=298) and an understory tree (n=263) were sampled at each point when present. Samples were genotyped at nine microsatellite loci which returned both diploid (SF= 136, FL=51) and triploid (SF= 53, FL=9) unique MLGs. Data was obtained from Karen Mock upon the request of the present authors.

**Background on *Plasmodium falciparum* data from Phyo *et al.* (2012)**

Phyo *et al.* 2012 sampled from patients infected with *Plasmodium falciparum* (n=1731) at 4 clinics along the Thai-Burma boarder over a 10 year period. During this time there was a ~12 fold decline in transmission. Nkhoma *et al.* (2013) later analyzed population genetic patterns that correlated with declining transmission, focusing on the Mawker-Thai and MaelaCamp clinics which had sufficient sampling both before (Mawker-Thai = 268; MaelaCamp = 252) and after (Mawker-Thai = 271; MaelaCamp = 121) significant declines in transmission occurred (2001-2004 and 2007-2010). These samples were obtained from filter paper blood spots from malaria patients attending these clinics. DNA was extracted from the filter paper and genotyped at 96 single nucleotide polymorphisms using the Illumina GoldenGate platform. Data was obtained from Tim J.
C. Anderson upon request of the present authors. These data contained samples that arose from single infections as well as infections from multiple *P. falciparum* genotypes. In the present analysis we only used samples that arose from single multi-locus genotypes of *P. falciparum*.

**Estimating the Expected number of unique individuals and the Sex rate**

The values of $\psi$ and $D^\psi$ can be estimated from the GFS. Optionally, the number of unique individuals expected in a sample ($E[N_g]$) and the per individual sex rate ($S_e$) can then be estimated given a new sample size ($N$) using equations [4-4] and [4-5]. We envision multiple ways in which these equations might be evaluated to achieve meaningful inference and depending on the motivation for the chosen value of $N$ the resulting output can have a variety of interpretations. A few examples: An estimate of the census size of a population ($N_c$) may be used, in which case inputting this value will return an $E[N_g]$ that reflects the expected number of unique individuals in the full population. Alternatively, the user may have an estimate of the effective population size ($N_e$), perhaps even from the same data that was used to generate the GFS, inputting this value will return an $E[N_g]$ that reflects the effective number of unique individuals in the population ($N_{ue}$) under the model used to estimate $N_c$. Finally, the user may simply wish to explore how many unique individuals they might expect to find if they altered their sample size, in which case inputting the new sample size will return the $E[N_g]$ value for that new sample size.

Several approaches can be used to estimate $N_c$ and the choice of approach will generally depend on the type of data available. In our analysis we use two distinct approaches, the first relying on allele frequency data and the second relying on variance
effective size. In our analysis of *Pocillopora damicornis* the allele frequencies of the 6 microsatellite loci were used to estimate \( N_e \) from the data using the program Theestimate (Haasle and Payseur 2010). This program uses 5 different models to estimate theta from the data: 1) the number of alleles (theta-NA), 2) the exponential approximation to the microsatellite allele frequency spectrum (theta-EFS), 3) mean allele frequency (theta-xbar), 4) expected homozygosity (theta-H), 5) Xu and Fu’s estimator (theta-F) (Xu and Fu 2004). We used all five methods to estimate \( \theta \) and then calculated \( N_e \) based on an assumed mutation rate of \( 4.5 \times 10^4 \) (Whittaker *et al.* 2003). In our analysis of *Plasmodium falciparum* the original authors supplied values of \( N_e \) for the two time periods (Nkhoma *et al.* 2012). These values were based on the fluctuation in allele frequencies between adjacent years and were calculated using the program MLNE v.1547 (Wang 2001). Nkhoma *et al.* (2012) did not observe significant differences in \( N_e \) between time periods so we elected to use the mean \( N_e \) calculated by the original authors.

When making comparisons between populations we suggest comparing values of \( \psi \) and \( D\psi \) which do not require estimates of effective population size or mutation rate. For this reason, given no other information they are the most appropriate statistics to compare. Alternatively, using the same values of \( N_e \) for the two populations under investigation functions as a transformation of \( \psi \) to the scale of \( E[N_e] \) and/or \( S_e \). The resulting analysis is equivalent to saying “if the observed differences in the GFSs are entirely due to differences in \( S_e \) that difference would be…”. We find this approach valid when the results are understood in the light that any underlying difference in \( N_e \) will be partitioned into a difference in \( E[N_e] \) and/or \( S_e \), potentially erroneously. Additionally, the absolute values of \( E[N_e] \) and/or \( S_e \) will also be effected by the accuracy of the \( N_e \) value
chosen and so those values must be evaluated with the same caveats that surround the estimates of $N_e$.

**Results**

**Simulated Data**

Figure 4-1 shows the accuracy of $\psi$, $D^\psi$, $G/n$ and $\beta$ at various sample sizes and clonal frequency with simulated data. The population size was 5000 for all simulations, each line represents the same set of simulated genotypes sampled 20, 30, 40, 50, 100, 200, 500, 1000, 2500 and 4500 times. The value of $G/n$ was highly biased by sample size at all clonal frequencies; as discussed by Gorospe et al. 2015, $G/n$ requires nearly population level sampling to approach the true value (Fig. 4-1a). The slope of the Perato distribution ($\beta$) was not as biased by sample size as $G/n$ but still showed a substantial effect of sample size on the inferred value (Fig. 4-1b), this was also noted by Nkhoma et al. (2012) in their analysis of *Plasmodium*. Note that the plots of ($\beta$) do not extend to all sample sizes for all clonal frequencies. This is because sample sets at these sample sizes were composed of 100% unique individuals which produces no slope, for this reason a value could not be plotted. Similarly, the datasets with 100% unique individuals highly skew the estimates of $\psi$, however, once a population is sampled sufficiently such that multiple clones appear in the dataset the estimate of $\psi$ is unbiased by sample size (Fig. 4-1c). Estimates of $D^\psi$ are also generally unbiased by sample size, however the statistic performs poorly when the underlying number of unique individuals is high and the sample size is low, again because the sample set is composed of 100% unique individuals.
Figure 4-1) Performance of Psi, G/N and the slope of the Peto distribution corresponding to elements A, B and C respectively. X-axis is the number of samples, y-axis is the estimated value of the summary statistic. Different colors represent different numbers of underlying unique individuals in a population of 5,000 individuals. Values calculated with a sample size of 4,500 can be assumed to be near the true value, if there was no bias we would expect to see a line with a slope of zero for all plots.
**Pocillopora damicornis data from Gorospe and Karl (2013)**

We used this dataset to examine the accuracy of $\psi$, $\beta$, and $G/n$ at various sample sizes. We calculated $\psi$, $\beta$, and $G/n$ for the full dataset then subsampled from the larger dataset and recalculated $\psi$ and $G/n$ for the subsampled data. Figure 4-3 shows the values of $\psi$, $\beta$, and $G/n$ calculated for subsampled data compared to the value calculated from the full dataset (Fig. 4-3a and b) and the absolute percent error based on that value (Fig. 4-3c and d). Subsample estimates of $\psi$ tended to slightly underestimate the true value of $\psi$ ((Fig. 4-3a), however sample sizes between 15 and 50 (sample sizes commonly employed in population genetic studies) showed a range of 200%-1% error from the true value of $\psi$ with an average of 49% error (Fig. 4-3c). In contrast, subsampled estimates of $G/n$ showed a strong bias toward overestimating the true value of $G/n$ at sample sizes lower than 2300 (Fig. 4-3b). At sample sizes of 15 and 50 the estimated value of $G/n$ ranged from 25,000% to 800% greater than the true value (Fig. 4-3d).

We then used the full dataset $n=2352$ to calculate the values of $\psi$, $D^\psi$, $E[N_g]$, and $S_c$ based on all 5 theta models in the program Thesimate (Haasle and Payseur 2010) (Table 4-1). We then randomly sampled the dataset to produce 5 different samples of 30, 50 and 100 individuals. This produced a total of 15 simulated samplings of *Pocillopora damicornis* at Kane’ohe Bay with sample numbers similar to those gathered in typical population genetic studies. We then calculated $E[N_g]$ for each of these samples by estimating $\psi$ from the GFS of the sample and $N_c$ calculated from the allele frequencies of the sample using the theta-NA model in thestimate (Figure 4-4).
**Figure 4-2)** Performance of $\Psi$ and $G/n$ on the *Pocillopora damicornis* dataset. Dotted line shows the value calculated on the full population set, the blue line is the fitted mean of the values for each sample size along the x-axis. The Y axis is the value calculated for the summary statistics in a) and b). The y-axis is the log error in c) and d).
Table 4-1) Psi and D Psi values calculated for the full dataset of Gorospe and Karl 2015 and E[Ng] and Se values calculated with all 5 Ne models

<table>
<thead>
<tr>
<th></th>
<th>MLE</th>
<th>95% HDCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psi</td>
<td>15.6709</td>
<td>(11.858-19.526)</td>
</tr>
<tr>
<td>D Psi</td>
<td>0.266</td>
<td>(-0.424-0.750)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>theta-NA</th>
<th>theta-NA</th>
<th>theta-NA</th>
<th>theta-NA</th>
<th>theta-NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% HDCI</td>
<td>(65.919-95.244)</td>
<td>(65.919-95.244)</td>
<td>(65.919-95.244)</td>
<td>(65.919-95.244)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>theta-NA</th>
<th>theta-NA</th>
<th>theta-NA</th>
<th>theta-NA</th>
<th>theta-NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>73.024</td>
<td>73.024</td>
<td>73.024</td>
<td>73.024</td>
</tr>
<tr>
<td>95% HDCI</td>
<td>(59.195-96.703)</td>
<td>(59.195-96.703)</td>
<td>(59.195-96.703)</td>
<td>(59.195-96.703)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E[Ng]</th>
<th>E[Ng]</th>
<th>E[Ng]</th>
<th>E[Ng]</th>
<th>E[Ng]</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% HDCI</td>
<td>(64.310-95.576)</td>
<td>(64.310-95.576)</td>
<td>(64.310-95.576)</td>
<td>(64.310-95.576)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Se</th>
<th>Se</th>
<th>Se</th>
<th>Se</th>
<th>Se</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.003080</td>
<td>0.003080</td>
<td>0.003080</td>
<td>0.003080</td>
</tr>
<tr>
<td>95% HDCI</td>
<td>(0.00133-0.00493)</td>
<td>(0.00133-0.00493)</td>
<td>(0.00133-0.00493)</td>
<td>(0.00133-0.00493)</td>
</tr>
</tbody>
</table>
Figure 4-3 Distributions of $E[Ng]$ values estimated from sample sizes of 30, 50, and 100 samples from the *Pocillopora damicornis* data from Gorospe and Karl (2013). a) shows the $E[Ng]$ values estimated using $N_e$ from the 1st model from the Haasl and Payseur (2010); each distribution represents an independent subsample from the full dataset. b) shows the $E[Ng]$ values estimated using the census size ($N_c=2352$) each distribution represents an independent subsample from the full dataset.
**Populus tremuloides** data from Mock *et al.* (2008)

We estimated $\psi$ for diploid and haploid individuals at both Swan Flats and Fish Lake (Fig. 4-4). Estimates of $\psi$ were very similar between Swan Flats and Fish Lake (Fig. 4-4a). However, in both locations $\psi$ was significantly higher in diploid individuals than haploid individuals (Fig. 4-4b). We also estimated $\psi$ and $D^\psi$ for both understory and overstory individuals at Fish Lake (Fig. 4-4c and d). There was no statistically significant difference in $\psi$ between understory and overstory individuals (Fig. 4-4c). Estimates of $D^\psi$ did not significantly depart from the expected value of zero for any group. Pairwise comparisons between groups showed no significant difference in $D^\psi$, however the maximum likelihood estimate of $D^\psi$ was higher for both overstory ploidy groups when compared to their respective understory group (Fig. 4-4d).

**Plasmodium falciparum** data from Phyo *et al.* (2012)

We estimated $\psi$ at both Mawker-Thai and MaelaCamp for the time periods before and after malaria transmission rates declined (Fig. 4-5). We found a significant decrease in $\psi$ at both Mawker-Thai and MaelaCamp after malaria transmission rates declined. Using the effective population sizes for these populations published by Nkhoma *et al.* (2013) we inferred the rate of sexual recombination for these collections (Fig. 4-5b). We inferred the rate of sexual reproduction in the period prior to the reduction in transmission rate at Mawker-Thai and MaelaCamp to be 0.1947 (95%HDCI 0.1307-0.2569) and 0.2805 (95%HDCI 0.3972-0.1806) respectively. In the period after transmission rates were reduced the rate of sexual reproduction was 0.0349 (95%HDCI 0.0229-0.0478) at Mawker-Thai and 0.0390 (95%HDCI 0.0173-0.0608) at MaelaCamp. Therefore both
Figure 4-4) Psi and DPsi values for the Aspen data, Samples from Fish Lake are divided into understory (S) and overstory (B) collections and Swan Flats (USF) is below. Diploid and Triploid individuals in the same collection are plotted separately on each line and are labelled D and T in the legend.
Figure 4-5) a) Psi, b) $E[Ng]$ and c) DPsi values for the *Palasmodium* dataset. Locations marked on the y-axis and values marked on the x-axis. Sample periods plotted on the same line and marked as T1 (before transmission decline) and T2 (after transmission decline). Jagged shape of DPsi plot of MLA T1 is due to the low number of Genotype Configurations that are likely under the Psi value observed which produces some rather discreet values to populate the posterior.
locations showed a significant decrease in sexual recombination rate while there was no significant difference in rate between locations during the same time period.

In our analysis of $D^\psi$ we did not find a significant difference between the periods before and after the reduction in transmission rate. However, the MLE of $D^\psi$ in the period after the reduction in transmission rate was significantly positive, indicating a higher number of low frequency individuals than expected under the neutral model.

Discussion

Clone Diversity: Richness, Evenness and Heterogeneity

In a comprehensive review of methods and statistics used to describe clonality Arnaud-Haond et al. (2007) showed that the information contained in many commonly used statistics was highly correlated. The authors recommended that future studies report just three indices to summarize the richness, the evenness and heterogeneity of clonal diversity. Specifically, they suggest $R$ (Dorken and Eckert 2001) for summarizing clonal richness, Simpson evenness as $V$ (Hurlbert 1971) and the slope of the Pareto distribution ($\beta$) to depict heterogeneity. Here we propose the alternate summary statistics $\psi$ and $D^\psi$ for summarizing clonal diversity. We show in Figure 1 that $\psi$ and the Pareto distribution perform very similarly at modeling clonal data. Indeed we suspect that the summary statistics proposed by Arnaud-Haond et al. (2007) and those proposed here contain very similar information. The benefit of using $\psi$ and $D^\psi$ for describing clonal data is not that they contain more information than previously used summary statistics; rather, their major benefit is that they facilitate biologically meaningful inference. Deriving these statistics from neutral population genetic models allows values of $\psi$ to be used to calculate a population sex rate or asexual rate or be used to make predictions about the
number of unique genotypes in the population. Further, we ground the expected shape of the clone frequency spectrum in a neutral model allowing the value of $D^w$ to be calculated for a single population, facilitating meaningful inference about biological phenomena that skew this distribution from its expected value of zero under neutral dynamics.

Our analysis of bias of estimates of $g/N$, $\beta$ and $\psi$ show that in general clonal analyses perform poorly when there are no clones in the data (Fig 4-1). While seemingly trivial it does mean that researchers can be confident in a non-biased estimate from their data so long as several clones are present in the data. Further, this bias was significantly less for $\psi$. When the number of clones observed is very low or when clones are absent, the estimation of $\psi$ must integrating over all possible values of $\psi$ which can not be ruled out by the sample size. There are, however, an infinite number of large values for $\psi$ because $\psi$ can take on any value from 0 to $\infty$. Thus, we suggest and use, a defuse prior ($G_0=1$, $n=2$ in equation [4-6]) on the Ewens distribution when evaluating observed data to approximate the posterior of $\psi$. This insures that when no clones are observed the posterior is still wide over large values of $\psi$ but the max likelihood estimate of $\psi$ is not overly inflated. The resulting posterior distribution gives reliable densities in its lower density interval, which allows for accurate comparisons between posteriors when hypothesis testing two collections in which one contains clones and one does not. In this scenario it is possible to reject the null hypothesis of equal $\psi$ values if the sample sizes are sufficiently large. Such a result could not be obtained by comparing the slope of the Perato distribution.

The max likelihood estimate of $\psi$ when the number of clones is very low or when clones are absent is still biased upward in this case. In our simulated data sample sizes
were generally outside of the region of bias when at least one clone was observed 3 times and another 2 times. Additionally, even though there is no biased of sample size, the variance in these statistics does decrease with sample size so the confidence interval of estimates does diminish with more samples, as it does with most statistics. For this reason researchers should perform a power analysis to decide the appropriate sample size for their datasets when the goal of the analysis is to make a comparison rather than aiming to just make an unbiased estimate

**Clonality in *Pocillopora damicornis***

A collection of 2352 samples of *Pocillopora damicornis* from a nearly exhaustively sampled, single reef in Kane’ohe Bay, Hawai’i, offered a rare opportunity to examine the accuracy of population summary statistics. Population summary statistics describe the parameters of an idealized model population that are concordant with the observed data from a natural population. While these summary statistics rarely align directly with ground truthed data, they provide a useful means for making comparisons across populations. They also allow biologically meaningful processes to be highlighted and further examined when population summary statistics diverge strongly from observed data. For instance, effective population size commonly diverges from census size in a diverse range of taxa. This observation has been used to argue for pervasive directional selection across the genome (Corbett-Detig et al. 2015); an effect that may alternatively or additionally be explained by skewed offspring distributions (Eldon and Wakely 2006). For this reason, even though it is a desirable feature, testing the accuracy of population summary statistics against ground truthed data is not the best metric of its utility. Rather, effective population summary statistics should allow comparisons to be made between
collections by being unbiased and facilitate inference by directly relating to the processes that explain their value.

In our analysis we found that $\psi$ was very accurate even using relatively small sample sizes at estimating the true value of $\psi$. For example, the 5 subsample replicates with $n=30$ had a range of MLEs from 7.357 to 16.925, and 4 of the 5 95% HDICs contained the MLE for the full dataset (15.671). The estimates of $\psi$ showed a slight bias toward under estimating the true value of $\psi$. As with any statistic, the accuracy of the estimate of $\psi$ was improved with more samples, however the largest change in accuracy occurred in very low sample sizes so sample sizes as low as 50 were effective at estimating $\psi$. Further, there was no bias in estimation due to sample size, a pattern that we observed in $G/n$ and $B$.

We then used equation [4-5] to test the accuracy of our model for estimating the true number of unique genotypes of *Pocillopora damicornis* at Kane’ohe Bay, Hawai’i. We estimated $N_e$ by using the program THESTIMATE (Haasl and Payseur 2010), and based on this $N_e$ calculated a MLE of 44.75 given a sample of 30 with a 95% HDIC of 17.38-89.57 individuals, while the population level sampling gave a value of 78 unique individuals. We caution that the effective number of genotypes may not always be equal to the actual number of unique genotypes in a population. We expect that the value might vary depending on the ability of researchers to estimate $N_e$ in their system and the value of $D^\psi$ of a sample, where equation [4-5] is biased when $D^\psi$ does not equal 0 and will tend to over-estimate $N_g$ when $D^\psi$ is positive and under estimate $N_g$ when $D^\psi$ is negative. Our estimate here was reasonably accurate because the $N_e$ values returned by the program THESTIMATE (Haasl and Payseur 2010) for the five samples of 30 individuals were
1,333, 889, 889, 1,074 and 889 based on our assumed mutation rate of $4.5 \times 10^4$. This value was less than the observed census size of 2,352 but near the same order of magnitude resulting in a slight downward bias to the estimate of $N_g$. It is therefore worth noting that evaluating equation [4-5] with the census size of 2,352 results in an $E[N_g]$ of 57.41 given a sample of 30 with a 95% HDIC of 22.71-113.06 individuals which is closer to the observed number of unique individuals (78) than the $N_c$ based approach. It is worth noting, however, that the difference of ~1,000 individuals between our estimated $N_c$ and $N_c$ only resulted in a difference of 12.6 unique genotypes in the MLE of $E[N_g]$ demonstrating that there is also a measure of robustness when using equation [4-5] to estimate $E[N_g]$ in populations (i.e. large values of N), particularly when $S_e$ is low.

Additionally, when we re-analyze the same GFS with a new or alternative values of N (i.e. going from $N_c$ Fig. 4-4a to $N_c$ in Figure 4-4b) it is primarily the scale of the posterior distribution that is changed. This has the effect of changing the MLE of $E[N_g]$ but the significance of any comparisons made between separate populations would have returned nearly identical results.

This $N_c$ based estimate of $E[N_g]$ ($E[N_g]=57.41$) was still lower than the $N_g$ observed when the total population was sampled ($N_g=78$). We attribute this to a slightly positive value of $D^\psi$, which we calculated from the full population dataset (Fig. 4-6).

Under the present scenario of analysis ($n=30$) $D^\psi$ cannot be shown to be significantly positive and so this aspect of the population cannot be taken into account, indeed $D^\psi$ could not be shown to be significantly positive even for the full dataset (Table 4-1). For this reason we analyze $E[N_g]$ with the hyperparameter of $D^\psi=0$, however it may be possible in the future to integrate our uncertainty about $D^\psi$ into the analysis by treating
Dψ as a parameter. Such an approach would likely increase the MLE of E[Ng] closer to the observed Ng because the MLE of Dψ was slightly positive even though it was not significant, this, however, would come at the cost of wider credibility intervals for E[Ng].

Finally, researchers should not over value the correspondence between our estimate of Ng and ground truth. Rather, we wish to highlight how using ψ and Dψ facilitated the conversion of population parameters to values expressed in terms of individuals and these values could be checked against null expectation values to make inferences about the biology of an organism. We believe these inferences are made richer by this method of analysis than if we summarized these data with R (Dorken and Eckert 2001) V (Hurlbert 1971) and the slope of the Pareto distribution (β) (Arnaud-Haond 2007).

**Clonal and sexual reproduction in Populus tremuloides**

The establishment of new stands of quaking aspen (Populus tremuloides) generally occurs following a disturbance event such as a fire or a change in localized climate. These events are highly episodic, generally discreet in nature and affect a large area. Following the initial establishment of a stand it has been traditionally thought that the established trees exclude new seedlings and competition among individuals eventually results in the dominance of single clonal individuals(Mock et al. 2008). For this reason, stands of aspen are traditionally thought of as even-aged and mono-clonal. Indeed, the Fish Lake location examined in this study is the home of the largest living organism documented; the “Pando clone” (Mitton and Grant 1996, Kemperman and Barnes 1976). These traditional views of aspen stand dynamics are being challenged, however, by a growing number of studies. Mock et al. (2008) showed that stands of aspen
are comprised of a number of unique individuals of different sizes and concluded that recent sexual reproduction must contribute substantially to stand variation. Additionally, the authors noted that the largest genets were triploid rather than diploid. In a recent study DeRose et al. (2015) demonstrated that stems of triploid aspens grow more rapidly than stems of diploids. In our analysis we found a significantly higher rate of sex in diploids than triploids. It follows then that clonal reproduction occurs at a higher rate in triploids than diploids (Fig 4-5a).

In order to test if there is a significant difference in genotypic diversity between new growth and older growth, Mock et al. (2008) sampled from both the understory and the overstory at each sampling point at Fish Lake. Here we show that there is no significant difference in the contribution of sexually derived and asexually derived individuals in populations of the same ploidy taken from the understory and overstory. This suggests that individuals of sexual and asexual origin establish themselves with equal success, at least between the time period that separates the two size classes sampled. Our analysis of $D^\psi$ supports this conclusion, suggesting that the equivalence of sexual and asexual individuals is not just a recent pattern but a long term attribute of this system. That is, by observing an equal sex rate between the overstory and understory we can infer that there is no advantage to either group in the time it takes for understory individuals to become overstory individuals. In our analysis we found that diploids and triploids from both sites, and from the overstory and understory all had values of $D^\psi$ that were not significantly different from zero which suggests that the equivalence is long standing.
It would be logical to presume that the existence of expansive, dominant clones such as the Pando clone must be the result of an advantage that these individuals have over other individuals in the population. If the established clonally reproducing individuals were more likely to have advantageous allelic combinations than the newly recombined genotypes of the sexually derived individuals, the advantage of asexually derived individuals would be pervasive. That is, the effect would influence the individuals survivorship and reproductive success even after establishment. If this were the case, the number of sexually derived individuals in the population would be proportionally greater than their contribution to the genealogy of the population. This would appear in the GFS as an excess of low frequency individuals causing $D^w$ to be negative.

It might also be logical to presume that the contribution of sexually derived individuals would decrease over time as the dominant genotypes exclude the establishment of new genotypes. If the contribution of sexually derived individuals was diminishing over time we would expect a higher number of sex events at deep points in the genealogy relative to the number observed at the more recent points. This would result in a higher proportion of middle frequency individuals, causing $D^w$ to be positive. The shape of the genotype frequency spectrums, however, suggest that the relative contribution of sexually and asexually derived individuals has remained constant over a long period of time; the relative frequency of high, middle and low frequency genotypes is consistent with this conclusion.

It is important to note that this does not mean that there is no advantage of asexual reproduction over sexual reproduction in this system. Clearly, the reason that the sex rate
is so low, and the clonal rate is so high, is that the number of successful individuals created from the shoots of already established individuals is higher than the number of successful individuals created from seed each generation. What should be inferred from a $D^\psi$ value of zero is that once established, individuals that originate from sexual or asexual reproductive events are equivalent in their survivorship and reproductive success; there is no intrinsic advantage to being asexually or sexually derived once established. Here “establish” means reaching the stage at which an individual might be collected as an understory individual. Indeed, a reasonable explanation of the data might be that the advantage asexual propagules have over new sexual propagules has already occurred in the life period before individuals were large enough to be sampled by Mock et al. (2008) as understory individuals. If so, we would hypothesize that a sampling effort that collected the smallest size class of individuals including new sprouts and newly emergent asexual shoots would return both a higher estimate of $\psi$ and a significantly negative value of $D^\psi$. Alternatively, if there was no difference in $\psi$ or $D^\psi$ in such a sample set it could be inferred that the clonal dynamics are primarily driven by the relative frequencies of new sprouts and asexual shoots produced rather than any advantage of one class over the other.

**Declining Transmission and Sex rates in Malaria**

The malaria causing parasite, *Plasmodium falciparum*, has an obligate sexual life cycle. Haploid parasites replicate mitotically in human hosts. These cells then produce male and female gametes which undergo syngamy in the midgut of the mosquito to form a diploid zygote (Nkhoma et al. 2012). This zygote then undergoes meiosis to form new infective haploid cells. While it would seem like this obligate sexual life cycle would lead
to high clonal diversity, the creation of new recombinant individuals requires both male and female gametocytes from different clones to be present in the midgut of the mosquito to undergo syngamy. If the male and female gametes in the midgut of the mosquito are from the same *P. falciparum* genotype these gametes fuse and then undergo meiosis to produce cells with identical genotypes to their parent, unaffected by recombination.

The frequency of sexual reproduction by the outcrossing of genotypes is therefore related to the frequency of mosquitoes that are infected with two different *P. falciparum* genotypes, either sequentially or from hosts that are infected with multiple *Plasmodium* genotypes. Differences in clonal frequency coincide with transmission rate, with areas of low transmission showing a higher frequency of clonal individuals in the population (Anderson et al. 2000). It is for this reason that Nkhoma et al. (2012) hypothesized that population genetic parameters might show a measurable response to lower transmission rates of *P. falciparum* in a region that has undergone significant decline in transmission rate. Concentrating on two locations with substantial sampling before and after transmission decline (Mawker-Thai and MaelaCamp) the authors showed a significant decline in genotypic richness and the slope of the Pareto distribution (β). The authors did not, however, observe a difference in effective population size (Nkhoma *et al.* 2012).

In our analysis we similarly show a decline in ψ following the reduce rate of transmission using the effective population sizes calculated by Nkhoma *et al.* (2012) for these populations. Converting the sex rate into a ratio of sexual outcrossing to sexual inbreeding we infer that the ratio fell from 5.29:1 (3.60-7.07 : 1 95% HDCI) for MKT and 3.72 : 1 (2.34-5.21 : 1 95% HDCI) for MAL before the reduction to 29.68 : 1 (18.91-
40.48 : 1 95% HDCI) and 29.18: 1 (13.71-45.22 : 1 95% HDCI) respectively after the reduction in transmission rate, a 6.5 fold reduction in sexual reproduction over-all. The frequency of multiple clone infections decreased from 50% to 20% during the same time period. This suggests an exponential relationship between the frequency of sexual reproduction in *Plasmodium falciparum* and the infection rate. Such a relationship would make intuitive sense because the frequency of mosquitos that encounter two infected hosts or hosts with multiple genotype infections is related to the square of the number of infected hosts if infection was random. For a number of reasons we could not expect this relationship to be perfect since infected hosts are not randomly distributed and mosquitos with multiple genotype infections might contribute disproportionately to transmission (Pollitt et al 2015) potentially increasing the population sex rate.

In many ways our analyses mirror the analyses Nkhoma *et al.* (2012) performed with previously published clonal analytical methodologies such as the genotypic richness index and the slope of the Pareto distribution. Nkhoma *et al.* (2012) suggested that since these population genetic parameters are responsive to changes in transmission rate they might provide useful information for designing treatment strategies, as well as useful metrics for judging the success of those efforts over time. Nkhoma *et al.* (2012), however, also showed a significant sample size bias of both the genotypic richness index and the slope of the Pareto distribution. Indeed, the authors showed that it might be necessary to collect more than 800 samples in order to avoid the biases of these statistics. This poses a major hurdle to monitoring these values in the real world; in this sample set the locations with the best sampling had 50-150 samples per year. Using the analytical methods we propose, samples of this size or even smaller could provide useful estimates
that are responsive to changes in the sexual recombination rate. This is an important result since it suggests that the number of unique malaria strains and the population sex rate could be monitored over time with modest sample sizes at relatively low costs using the presently available genetic technology.

**Conclusions**

In organisms with mixed modes of reproduction the relative frequency of sexual reproduction to clonal reproduction is a product of intrinsic biological factors and external influences. For this reason, estimates of the rate of sexual recombination can provide a number of interesting insights about the biology of organisms with mixed modes of reproduction. Calculating $\psi$ at various spatial scales in partially clonal organisms can provide insights into the scale of movement or migration of asexual individuals. Comparing $\psi$ between age-classed individuals can provide novel insight into how sexual and asexual individuals contribute to the total population both at short and long time scales. Indeed, the effect of any process on the relative rate of sexual recombination can be made given appropriately collected samples.

The utility of this analytical approach is not limited to making comparisons between two sample sets. Inferences can also be made about single sample sets by comparing the frequencies of genotypes observed to the frequencies expected under a neutral model. Since this neutral model is built from the processes of birth, death, reproduction and recombination of genotypes we can make meaningful inferences about populations that significantly deviate from these expectations even when there is no other population to compare it to.
While these methods do share some of the utility of previously described methods they offer some key advantages. Perhaps the most important being that the parameters of this model can be readily expressed in biological units. The population sex rate can be expressed in terms of a percentage of replication events that are sexual or converted to a ratio of sexual to asexual events in the genealogy of the population. At the same time our approach also provides an estimate of the expected number of unique individuals in the population. Importantly, the approach accomplishes this with better accuracy and less bias than previous methods; allowing meaningful estimates to be made with relatively modest sample sizes.
References


Chapter 5

MICROSATELLITE LOCI FOR THE HOST-GENERALIST
“ZOOXANTHELLA” SYMBIODINIUM TRENCHI AND OTHER CLADE D SYMBIODINIUM.

Abstract

Nine new polymorphic microsatellites were developed for Symbiodinium trenchi (sensu type D1a). These loci were tested on populations of S. trenchi from corals in Palau and 3 to 19 alleles were observed at each haploid locus with an average of 7 alleles. Many of the primer sets successfully amplified loci within other members of Symbiodinium clade D, demonstrating their utility across the group. Clade D Symbiodinium spp. are generally regarded as thermally tolerant and are common in reef habitats with warm, turbid, and/or variable environmental conditions. These population genetic markers are therefore useful for investigating how these symbionts respond to climate change through geographic and possibly host-range expansion. Additionally, the cross-lineage utility of these markers should help delineate the evolutionary relationships among members of this clade.
Symbiodinium clade D comprises genetically and ecologically distinct lineages found throughout the tropical Indo-Pacific. One lineage in particular, Symbiodinium trenchi, appears to be a host generalists, associating with various host taxa representing several anthozoan orders. Most others in the group, however, are apparently specialists, typically associating with a single host genus, or species (LaJeunesse et al. 2010a). Corals harboring members of this clade appear less susceptible to bleaching (Jones et al. 2008; LaJeunesse et al. 2010b). Consequently, Symbiodinium spp. in clade D are more common in reef habitats exposed to high thermal fluctuations and changes in turbidity (Fabricius et al. 2004; Abrego et al. 2009; LaJeunesse et al. 2010a). It is possible that climate warming may be facilitating the spread of these symbionts to new regions and to different coral taxa. This ecological response may represent a fundamental process by which corals and their symbionts adjust through time to major climate oscillations (Baker et al., 2004; LaJeunesse et al. 2009). Population level investigations of dispersal and connectivity are therefore critically important in understanding how coral symbioses will respond to climate change. In order to investigate dispersal and diversity, eight microsatellite loci were previously developed for this group (Pettay and LaJeunesse 2009). Only a subset of these eight microsatellites, however, amplify fragments within each of the various members of Clade D (Pettay and LaJeunesse 2009). For the purpose of having an appropriate number of polymorphic loci to investigate geneflow and diversity within and among the various lineages of clade D, we have developed nine additional microsatellite markers for the group.
We used data generated by 454 sequencing as a library to search for repetitive sequence motifs. Genomic libraries were prepared from 3,000 ng of double-stranded DNA from three different cultures of *S. trenchi* (type *D1a*, cultures A002, A003, MTB42) and one culture of *Symbiodinium* sp. ITS type D5 (culture PHMSTD4) using the Nextera DNA Sample Prep Kit (Epicentre Biotechnologies, Madison WI) and sequenced on a 454 GS-FLX sequencer (Roche Diagnostics Corporation, Indianapolis, IN) utilizing the Titanium Sequencing Kit (Roche). Sequences containing tri- or tetranucleotide repeat motifs with a minimum copy number of 7 and a minimum of 20 bp flanking sequences up and downstream of the repeat were identified using the Tandem Repeat Database (Gelfand *et al.* 2006). The software Primer3 (http://primer3.sourceforge.net/) was used to design 48 PCR primers sets from these putative microsatellite sequences.

These PCR primers were then screened using DNA from samples that had been previously confirmed as Clade D by ITS2 DGGE fingerprinting. These samples were selected for their diverse geographic origin (Palau, La Paz Mexico and Japan) as well as their divergent ITS sequence (*D1a, D1* and *D15* respectively). PCR reactions were preformed using 1 µl of 25-50 ng of template DNA in a reaction solution containing standard Taq buffer (New England Biolabs; 10mM Tris-HCl (pH 8.3), 50 mM KCl, 2.5 mM MgCl) and 0.075 U/µl of Taq (New England Biolabs) and bovine serum albumin (BSA, 0.1 mg/ml). Primer concentrations of 0.2 µM of the tailed forward primer (see Table 1), 1 µM of the reverse primer, and 1 µM of the dye labeled T-oligonuclotide were used for amplifying PCR products. All loci were amplified using the following
touchdown thermal cycle profile: 94°C 2 min (1 cycle); 94°C 30 sec, 5°C above the primer specific annealing temperature (see Table 1) for 30 sec -0.5°C per cycle , 72°C 30 sec (10 cycles) 94°C 30 sec, primer specific annealing temperature for 30 sec, 72°C 30 sec (20 cycles); 72°C 5min on a Mastercycler gradient thermal cycler (Eppendorf, Westbury, NY).

Nine polymorphic loci were identified in *Symbiodinium trenchi* samples from Palau (Table 1). All nine loci were polymorphic with effective number of alleles at each locus ranging from 1.343 to 11.1. For each sample a single multilocus genotype was recovered. A linkage disequilibrium test was preformed between all of the loci presented in this study as well as the loci developed by Pettay and LaJeunesse (2009). All but one pairwise test were insignificant at a Bonferroni corrected $\alpha = .0006$. One pair of loci, the previously reported D1Sym14 and D1Sym93 from the present study, did show significant linkage. Screening of these loci on *Symbiodinium glynii* (ie. *D1*) from La Paz revealed that three of these loci were polymorphic within the population, while two were fixed for an allele and four did not amplify a product despite using a range of reaction conditions (Table 2). Subsequent screening of other eastern Pacific populations, however, has shown polymorphisms in D1Sym88 and D1Sym92 within S. glynii (unpubl. data). A subset of these loci were also shown to be polymorphic in *Symbiodinium* ITS type *D15* found in the coral *Oulastera crispata* from Japan, three of the loci did not amplify under various reaction conditions and one was fixed within the samples screened. Most of the DNA used to develop the 454 sequence library was from *Symbiodinium trenchi* and may explain the failure of several loci to amplify and the lower allelic diversity found in the
other D types (Table 2). This trend suggests that there is some bias that occurs when these markers are used on clade D *Symbiodinium* other than *S. trenchi*. Further, screening of *S. trenchi* populations from Thailand and Zanzabar also amplified products at each locus (data not shown). This suggests that amplification failure in the other D types is not a geographic effect. Rather, we believe it is more likely due to mutations at the priming site, an effect that is common in cross species microsatellite applications (Primmer et al. 1996, Primmer and Merila 2002). The effects of this bias should be considered when using these markers in cross ‘species’ comparisons.

We found that for each sample of *S. trenchi*, as has been previously reported by Pettay and LaJeunesse (2009), two allele peaks usually amplified at each locus. For locus D1Sym77 a total of four peaks were typically recovered. Samples of *S. glynni* and *D15* both returned single peaks for all loci with the exception of Sym77, which returned two peaks. These results were consistent with analyses of 5 independently acquired isoclonal cultures indicating differences in peak number originated from differences in genome organization. As with most all dinoflagellates, *Symbiodinium* are haploid (Santos and Coffroth 2003). We therefore presume that the most recent common ancestor of Clade D underwent a gene duplication that affected the D1Sym77 locus, resulting in two loci that are consistently co-amplified. Apparently, the *S. trenchi* lineage then underwent a complete genome duplication. Therefore the two alleles amplified at each locus correspond to recently duplicated loci. The “ploidy” of *S. trenchi* must therefore be considered when scoring these loci and conducting population genetic analyses.

Additionally, we observed large differences in fragment length between Clade D
types (Table 2). Although the difference in fragment length could be explained by long standing geographic structure, taken together with the observations of gene and genome duplication, as well as the ITS2 sequence divergence, it strongly supports the hypothesis that the many subclade members of clade D are evolving along separate evolutionary trajectories. It has been suggested that each of these entities, characterized by their dominant ITS sequence, are reproductively isolated, ie ‘species’ (LaJeunesse 2001; Pettay and LaJeunesse 2007, 2009, LaJeunesse et al. 2010a). The addition of the nine new loci should provide sufficient resolution to evaluate the diversity within and the relationship among members of clade D. Resolving these relationships will inform the interpretation of past and future ecological data and aid in inferring how these symbionts are responding to climate change through range expansion. (Baker et al. 2004; LaJeunesse et al. 2009).

Acknowledgments

This research was funded in part by The Pennsylvania State University and the National Science Foundation (OCE-0928764) as well as the IOC-UNESCO-World Bank targeted working group on coral bleaching. We would like to thank Yi-T Lien and Allen Chen for providing samples of *Symbiodinium* from Japan and Dustin Kemp for helping collect samples in Palau.
References


LaJeunesse, TC, Pettay, DT, Sampayo, EM, Phongsuwan, N, Brown, B., Obura, D., Hoegh- Guldberg, O. & Fitt, WK (2010a) Long-standing environmental conditions, geographic isolation and host – symbiont specificity influence the relative ecological dominance and genetic diversification of coral endosymbionts in the genus *Symbiodinium*. *J. Biogeogr.* 37, 785–800


Santos SR, MA Coffroth (2003) Molecular Genetic Evidence that dinoflagellates belonging to the genus *Symbiodinium* Freudenthal are haploid. *Biol Bull* 204:10-20
<table>
<thead>
<tr>
<th>Locus</th>
<th>Motif</th>
<th>Primer Sequences (5'-3')</th>
<th>$T_a$ (˚C)</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1Sym_66</td>
<td>(TCT)$_n$</td>
<td>F: T1CTCTGGGATGAGCCCCCGAA</td>
<td>57</td>
<td>HQ894461</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: GACAGTGTTCATGGTCGCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_67</td>
<td>(ATC)$_n$</td>
<td>F: T1-GAATCCAGATGAGCTGCTG</td>
<td>57</td>
<td>HQ894456</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: CAAAGGTAGCCGATTTGTCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_71</td>
<td>(GTT)$_n$</td>
<td>F: T2-CCACTCAAACGGTGACCCA</td>
<td>55</td>
<td>HQ894462</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: CAGACACAATGTGTCACAGGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_77</td>
<td>((TTC)$_3$ CT (CTTCCCT)$_2$ C (TTC)$_4$</td>
<td>F: T1-CCACTGAGATGTAGTGAA</td>
<td>58</td>
<td>HQ894457</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: ACCGATGGGTTGTGACTCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_82</td>
<td>(ACT)$_{12}$</td>
<td>F: T3-AGACTTCGCAAGCACAC</td>
<td>56</td>
<td>HQ894463</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: CGCATCTACTGGTGTCACAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_87</td>
<td>(GAAG)$_2$</td>
<td>F: T2-CCTATGACTCAAGGTGACG</td>
<td>57</td>
<td>HQ894458</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: AGACATACCTCGTCGGTGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_88</td>
<td>(TTTC)$_7$</td>
<td>F: T3-TTTGCACTGATGTGCTGCA</td>
<td>55</td>
<td>HQ894459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: GTTTGCAAGCCATCCCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_92</td>
<td>(CCTA)$_6$ (CCTG)$_3$</td>
<td>F: T2-GGTTTTGAACTAGATCCCT</td>
<td>58</td>
<td>HQ894460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: TGGGATGCTCTGGGCGAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1Sym_93</td>
<td>(AAAC)$_6$</td>
<td>F: T3-GCTCAAGGGACTGGGTG</td>
<td>58</td>
<td>HQ894464</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: TGTCAGAAGGGAGCGCTG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- D1Sym primers were developed from 454 sequence data and amplified with tail sequences on the 5' ends of each forward primer (T1 = ggctaggaaaggttaggc, T2 = TCATACTGTCTCTCACGTAAAC, and T3 = ACCAACCTAGGAACACAG). Amplifications included these primers along with separate dye labeled tail oligonucleotides (T1 with PET, T2 with 6FAM, and T3 with NED).
Table 2. An analysis of 3 *Symbiodinium* Clade D types. We report the allelic size range for each *Symbiodinium* Clade D type, the number of alleles (N<sub>A</sub>) and effective number of alleles (A<sub>e</sub>), within each population.

* The number of individual samples for *Symbiodinium trenchi* was 33 however each locus showed 2 peaks so the allelic sample size was n=66. ** Analysis of D1Sym_77 consistently returned 2 peaks in *Symbiodinium glynni* and *Symbiodinium* type D15 and 4 peaks in *Symbiodinium trenchi*, the allelic sample size was therefore n=66 in *S. glynni* and S. D15 and n=132 in *S. trenchi*. *** Overall mean and standard deviations were calculated leaving out invariable loci.

<table>
<thead>
<tr>
<th>Symbiont</th>
<th>Population</th>
<th>Sample size</th>
<th>D1Sym_66</th>
<th>D1Sym_67</th>
<th>D1Sym_71</th>
<th>D1Sym_77</th>
<th>D1Sym_82</th>
<th>D1Sym_87</th>
<th>D1Sym_88</th>
<th>D1Sym_92</th>
<th>D1Sym_93</th>
<th>Over All Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Symbiodinium trenchi</em> ITS type D1a</td>
<td>Palau</td>
<td><em>n=33</em></td>
<td>288-300</td>
<td>125-140</td>
<td>163-191</td>
<td>175-209</td>
<td>231-301</td>
<td>240-246</td>
<td>235-250</td>
<td>124-132</td>
<td>140-157</td>
<td>2.503 (5.123)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N&lt;sub&gt;A&lt;/sub&gt;</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>19</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A&lt;sub&gt;e&lt;/sub&gt;</td>
<td>1.244</td>
<td>2.135</td>
<td>2.135</td>
<td>1.163</td>
<td>1.163</td>
<td>1</td>
<td>1</td>
<td>3.40 (1.140)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Over All Mean (SD)</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
<td>3.333 (0.577)***</td>
</tr>
<tr>
<td><em>Symbiodinium glynni</em> ITS type D1</td>
<td>La Paz</td>
<td><em>n=30</em></td>
<td>NA</td>
<td>131-137</td>
<td>181-193</td>
<td>252-268</td>
<td>235</td>
<td>128</td>
<td>190-202</td>
<td>190</td>
<td>190</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N&lt;sub&gt;A&lt;/sub&gt;</td>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>4</td>
<td>NA</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td>3.40 (1.140)***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A&lt;sub&gt;e&lt;/sub&gt;</td>
<td>2.174</td>
<td>2.632</td>
<td>2.632</td>
<td>1</td>
<td>1</td>
<td>3.125</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Over All Mean (SD)</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
<td>3.40 (1.140)***</td>
</tr>
<tr>
<td><em>Symbiodinium</em> type D15</td>
<td>Japan</td>
<td><em>n=12</em></td>
<td>291-297</td>
<td>144-147</td>
<td>251-262</td>
<td>262-280</td>
<td>190-202</td>
<td>190</td>
<td>190</td>
<td>190</td>
<td>190</td>
<td>190</td>
</tr>
</tbody>
</table>
|          |            | N<sub>A</sub> | 3 | 2 | NA | NA | NA | 3 | 5 | 4 | 1 | 1.
|          |            | A<sub>e</sub> | 2.174 | 2.632 | 2.632 | 1 | 1 | 3.125 | 1 | 4 | 1 | 1 |
|          |            | Over All Mean (SD) | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** | 3.40 (1.140)*** |
Chapter 6

EIGHT POLYMORPHIC MICROSATILITE LOCI FOR THE INDO-PACIFIC WIDE ZOANTHID, ZOANTHUS SANSIBARICUS

Abstract

Next generation sequencing allows rapid development of genetic markers for investigating the ecology and evolution of non-model organisms. In the present study we employed 454 sequencing and high through-put screening to ultimately generate 8 highly polymorphic microsatellites for Zoanthus sansibaricus, a common zoanthid that is widely distributed throughout the Indo-Pacific. Each locus was screened against cultured Symbiodinium as well as on individuals harboring distantly related symbiont species to discount contamination by dinoflagellate microsatellites. The range of Z. sansibaricus extends across the tropical and subtropical Indian and Pacific oceans. These population genetic markers will allow examination of dispersal and gene flow in this species across large tropical and subtropical oceanic ranges.
Introduction

Members of the genus *Zoanthus* are common in shallow water tropical and subtropical environments world wide, often occurring on reef crests and shallow reef fronts where they may dominate the benthic landscape (Irei et al. 2011). Like most tropical zoanthids, they maintain obligate symbioses with various species of photosynthetic dinoflagellate (*Symbiodinium spp.*), which may allow them to occupy different reef photic zones (Reimer et al. 2006, Kamezaki et al 2012). Within this genus, *Zoanthus sansibaricus* Carlgren 1900 is one of the most common species in the tropical Pacific and Indian Oceans (type locality = Zanzibar; reports from Japan: Reimer et al. 2004; Galapagos: Reimer & Hickman 2009, New Caledonia: Sinniger 2008, Singapore: Reimer & Todd, 2009 etc.). Phylogenetic evidence from COI (Reimer et al. 2004) and ITS2-rDNA (Reimer et al. 2007) indicate that various morphotypes once considered as separate species probably comprise a single widespread species. The application of population genetic markers would allow further testing of the species boundaries in this group. Here, we use a next-gen sequencing approach to rapidly generate microsatellite markers. These single copy, independently assorting nuclear markers provide the ability to examine clonal and population structure as well as identify the existence of reproductive boundaries in this group (i.e. species delimitation).

Materials and Methods

Development of microsatellite library

We used data generated by 454 sequencing at the Penn State University Genomics Core Facility (University Park, PA) to search for repetitive sequence motifs.
A genomic library was prepared from 3,000 ng of double-stranded DNA. Three samples of *Zoanthus sansibaricus* were pooled and prepared using the Nextera DNA Sample Prep Kit (Epicentre Biotechnologies, Madison WI) and sequenced on a 454 GS-FLX sequencer (Roche Diagnostics Corporation, Indianapolis, IN) utilizing the Titanium Sequencing Kit (Roche). Sequences containing di-, tri- or tetrancleotide repeat motifs with a minimum copy number of 7 and a minimum of 20 flanking nucleotides up and downstream of the repeat were identified using the Tandem Repeat Database (Gelfand *et al.* 2006). The software Primer3 (http://primer3.sourceforge.net/) was used to design 96 PCR primers sets from these putative microsatellite sequences of *Zoanthus sansibaricus*.

**Screening for variable loci**

Ninety-six primer sets (i.e. loci) were then screened using DNA from 8 *Zoanthus sansibaricus* samples from Otsuki (Kochi), Kushimoto (Wakayama), Yoron (Kagoshima) and Manza (Okinawa). PCR reactions were performed using 1 µl of 25-50 ng of template DNA in a reaction solution containing standard Taq buffer (New England Biolabs; 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 2.5 mM MgCl₂) and 0.075 U/µl of Taq (New England Biolabs) and bovine serum albumin (BSA, 0.1 mg/ml). Primer concentrations of 0.2 µM of the tailed forward primer (Tables 1), 1 µM of the reverse primer, and 1 µM of the dye labeled T-oligonucleotide were used for amplifying PCR products. All loci were amplified using the following touchdown thermal cycle profile: 94°C 2 min (1 cycle); 94°C 30 sec, 5°C above the primer specific annealing temperature (Table 1) for 30 sec -0.5°C per cycle, 72°C 30 sec (10 cycles) 94°C 30 sec, primer specific annealing temperature for 30 sec, 72°C 30 sec (20 cycles); 72°C 5 min on a Mastercycler gradient thermal cycler.
(Eppendorf, Westbury, NY). PCR product was then screened by capillary gel electrophoresis (Penn State University Genomics Core Facility, University Park, PA). Primer sets that returned repeatable, polymorphic peaks that were > 1,000 relative fluorescent units in intensity and within the size range expected were retained for further analysis.

**Testing for “contaminant” loci that amplify Symbiodinium**

Screening yielded a total of 10 polymorphic microsatellite loci from 8 individuals initially examined. Tests were then performed on *Symbiodinium goreau* cultures (rt113 and rt152), representative of Clade C sensu LaJeunesse 2001, and *Zoanthus sansibaricus* samples collected from Hachijo-jima, Tokyo, Japan that harbored a Clade D symbiont. Two loci produced strong amplification products from the DNA of cultured *Symbiodinium* and also failed to amplify the *Z. sansibaricus* samples from Hachijo-jima. Therefore, 2 of 10 loci were markers that amplified *Symbiodinium* Clade C, which were symbionts in the animals used to generate these loci. The remaining 8 loci, which did not amplify cultured symbiont DNA, amplified in the sample obtained from Hachijo-jima confirming their host origin.

**Zoanthus sansibaricus collections for population genetic analysis**

*Zoanthus sansibaricus* samples were collected from Okinawa-Jima Island, Japan in June 2011. Two sites were chosen, one on the west side of the island facing the East China Sea (Manza; n = 31), and the other on the east side facing the Pacific Ocean (Teniya; n = 31). Samples were collected by SCUBA and snorkel from the intertidal zone
to a maximum depth of 3 meters. Small clusters of 1-30 polyps/colony were individually bagged and labeled at the collection site. Colonies were separated by 5 to 100 meter. Upon returning to the University of the Ryukyus, samples were transferred into 99% ethanol for preservation. DNA was then extracted using a Qiagen DNA isolation kit from a single polyp from each sample.

The allelic variation of each locus was evaluated from a total of 62 samples. Alleles were scored using GeneMarker v.1.91 (SoftGenetics). The number of alleles per collection site ranged from 6-25. Tests for departure from Hardy-Weinberg expectations (HWE) as well as linkage disequilibrium (LD) were preformed by GenPop on the Web http://genepop.curtin.edu.au/ (Raymond & Rousset 1995). Departure from HWE was observed in all loci at Manza (Table 2) and at most loci at Tenyia (Table 3) caused by a heterozygote deficiency. These highly variable loci spanned large size ranges. PCR is bias toward smaller fragments, the drop out of large alleles there for might explain the observed deficiency in heterozygotes. Tests of LD showed no significant (Bonferroni corrected alpha to 0.0018302) pattern of linkage among loci.

Conclusion
The markers generated in this study will be used in the investigation of connectivity, diversity, species identity and clonality of *Zoanthus sansibaricus*. Although not reef building, *Z. sansibaricus* is extremely widespread, with longitudinal distributions from the Red Sea and eastern coast of Africa to the Hawaiian Islands and the Galapagos, and extend into sub-tropical environments spanning latitudes from mainland Japan to the southern Great Barrier Reef (Reimer et al. 2006; Reimer & Hickman 2009; Reimer,
unpublished data). As such, investigations into patterns of clonal structure and connectivity between populations of *Z. sansibaricus* and its symbionts may be useful in examining patterns of relative connectivity and host-symbiont co-evolution across this entire region. The genotypic diversity of *Z. sansibaricus* observed in this small dataset was greater than the genotypic diversity reported for *Acropora palmata* and *Pocillopora spp.*, two scleractinians with wide geographic ranges and limited population genetic structure (Baums et al. 2005a, Baums et al. 2005b, Pinzon and LaJeunesse 2011). This remarkable degree of diversity would suggest a large effective population size perhaps from a high degree of connectivity across large geographic areas due to potentially long planktonic larval stages (Polak et al. 2011, Hirose et al. 2011). Investigation of gene flow across seemingly disjunct geographic locations will be an important area of future research.
References


Carlgren O (1900) Ostafrikanische Actinien,” Jahrbuch der Hamburgischen Wissenschaftlichen Anstalten, 17:21–144.


Table 1. Locus name, repeat motif, primer sequences, annealing temperature ($T_a$) and approximate size range of expected product for nine microsatellite loci for *Zoanthus sansibaricus*.

<table>
<thead>
<tr>
<th>Locus name</th>
<th>Fragment Length</th>
<th>Motif size</th>
<th>Primer Sequences (5’-3’)</th>
<th>$T_a$ (°C)</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zsan_2</td>
<td>186-259</td>
<td>3</td>
<td>F: M13-ATCGTGACGAGCCTAAATA</td>
<td>56</td>
<td>JX129533</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: AGTGGACAAGGCTGAACTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_14</td>
<td>72-114</td>
<td>4</td>
<td>F: M13-CTTGTACTGACGAGGG</td>
<td>54</td>
<td>JX129529</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: TGAGGTCACAAACGTAGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_30</td>
<td>119-167</td>
<td>3</td>
<td>F: M13-ATTATTAGGCTTGGCTGG</td>
<td>55</td>
<td>JX129527</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CGGAGGTTATACGGAGAAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_67</td>
<td>131-155</td>
<td>2</td>
<td>F: M13-ATGGGTAAATACGCACA</td>
<td>56</td>
<td>JX129526</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CACACACACACACAGTGAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_71</td>
<td>121-165</td>
<td>3</td>
<td>F: M13-ATTGTCTGTGGAAGCTTG</td>
<td>53</td>
<td>JX129528</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: AGCTTTATCTTTGTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_102</td>
<td>150-190</td>
<td>3</td>
<td>F: M13-AATTGTGAAGCATTTCGC</td>
<td>56</td>
<td>JX129522</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: TACAGGCCCTCAGTAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_104</td>
<td>150-215</td>
<td>3</td>
<td>F: M13-AAAACTTCTGCAACCG</td>
<td>55</td>
<td>JX129523</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GCAGACAGTGCTGCTTAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zsan_130</td>
<td>130-172</td>
<td>3</td>
<td>F: M13-CAAGCAGAAGTTGAATGGT</td>
<td>55</td>
<td>JX129531</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GCCGCATAATCCATTTGTA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Primers were developed from 454 sequence data and amplified with M13 tail sequences (TGTTAAAACGACGAGCCAGT) on the 5’ ends of each forward primer amplifications included these primers along with separate dye labeled M13 oligonucleotides.
**Table 2.** Number of alleles, Observed and expected heterozygosity for samples collected at Manza.

<table>
<thead>
<tr>
<th>Locus Name</th>
<th>Number of Alleles</th>
<th>Heterozygosity Observed</th>
<th>Heterozygosity expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zsan_2</td>
<td>24</td>
<td>0.621</td>
<td>0.952*</td>
</tr>
<tr>
<td>Zsan_14</td>
<td>8</td>
<td>0.300</td>
<td>0.700*</td>
</tr>
<tr>
<td>Zsan_30</td>
<td>25</td>
<td>0.677</td>
<td>0.952*</td>
</tr>
<tr>
<td>Zsan_67</td>
<td>9</td>
<td>0.538</td>
<td>0.704*</td>
</tr>
<tr>
<td>Zsan_71</td>
<td>16</td>
<td>0.839</td>
<td>0.924*</td>
</tr>
<tr>
<td>Zsan_102</td>
<td>21</td>
<td>0.679</td>
<td>0.927*</td>
</tr>
<tr>
<td>Zsan_104</td>
<td>21</td>
<td>0.556</td>
<td>0.944*</td>
</tr>
<tr>
<td>Zsan_130</td>
<td>17</td>
<td>0.615</td>
<td>0.879*</td>
</tr>
</tbody>
</table>

* Indicates significant departure from HWE with a Bonferroni corrected alpha of 0.0063912
Table 3. Number of alleles, Observed and expected heterozygosity for samples collected at Teniya

<table>
<thead>
<tr>
<th>Locus Name</th>
<th>Number of Alleles</th>
<th>Heterozygosity Observed</th>
<th>Heterozygosity expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zsan_2</td>
<td>21</td>
<td>0.786</td>
<td>0.937</td>
</tr>
<tr>
<td>Zsan_14</td>
<td>6</td>
<td>0.536</td>
<td>0.638*</td>
</tr>
<tr>
<td>Zsan_30</td>
<td>24</td>
<td>0.552</td>
<td>0.929</td>
</tr>
<tr>
<td>Zsan_67</td>
<td>6</td>
<td>0.500</td>
<td>0.752*</td>
</tr>
<tr>
<td>Zsan_71</td>
<td>13</td>
<td>0.704</td>
<td>0.857</td>
</tr>
<tr>
<td>Zsan_102</td>
<td>13</td>
<td>0.480</td>
<td>0.846*</td>
</tr>
<tr>
<td>Zsan_104</td>
<td>15</td>
<td>0.621</td>
<td>0.927*</td>
</tr>
<tr>
<td>Zsan_130</td>
<td>15</td>
<td>0.607</td>
<td>0.884*</td>
</tr>
</tbody>
</table>

* Indicates significant departure from HWE with a Bonferroni corrected alpha of 0.0063912
Chapter 7

MICROSATELLITE LOCI FOR SYMBIODINIUM GOREAUI AND OTHER CLADE C SYMBIODINIUM

Abstract

The genus *Symbiodinium* comprises a diverse group of dinoflagellates known for their obligate relationship with reef building corals. Members of the sub-genus ‘clade C’ are abundant, geographically wide-spread, and genetically and ecologically diverse. Reef corals associated with these *Symbiodinium* often experience severe physiological stress. The genotypic diversity, dispersal and genetic connectivity exhibited by this group are the subjects of an increasing number of population genetic studies utilizing microsatellites to better understand their ecology and evolution. In the present manuscript we describe 18 new microsatellite loci and test their utility across several genetically distinct, ecologically important members within clade C. Our utilization of these markers may provide further insight into the dispersal capability, processes of bleaching recovery as well as resolving species boundaries among phylogenetically distinct yet closely-related lineages.
The genetically diverse dinoflagellates in the genus *Symbiodinium* are often found in obligate symbiosis with reef building corals. This genus has been divided into a number of “Clades”, with the most ecologically dominant and diverse being Clade C (LaJeunesse et al. 2003; 2004ab; LaJeunesse et al. 2010a). While the group includes phylogenetic types that display narrow and broad thermal tolerances, members of clade C are often the most severely impacted during bleaching events (Abrego et al. 2008, Jones et al. 2008; LaJeunesse et al. 2010b). For this reason, their symbiosis biology, dispersal capabilities and genetic identities are of increasing importance to conservation efforts as climates change globally, further affecting reef coral communities.

Coral endosymbionts are increasingly being studied with multi-locus techniques because this approach offers the highest level of genetic resolution. High-resolution markers, however, are not yet available for most *Symbiodinium* spp. In the present study, we present 18 new loci for *Symbiodinium* Clade C; 14 loci developed on a monoclonal culture of *S. goreau*i (C1) and 4 loci developed from the symbiont specific to the zooanthid *Zoanthus sansibaricus* in the western tropical North Pacific referred to here by its diagnostic ITS2 sequences C3ee-ff, previously referred to as C1z (*sensu* Reimer 2008).

We used 454 sequencing at the Penn State Nucleic Acids Facility (University Park, PA) to generate two genomic libraries of >50,000 read lengths of 250-400 bp. The first sequence library was created from genomic DNA obtained from cultured *S. goreau*i (culture rt-113 from the Robert K. Trench collection). The second library was created from the pooled DNA from the symbionts of *Zoanthus sansibaricus*. Both libraries were prepared from 3,000 ng of double-stranded DNA, using the Nextera DNA Sample Prep
Kit (Epicentre Biotechnologies, Madison WI) and sequenced on a 454 GS-FLX sequencer (Roche Diagnostics Corporation, Indianapolis, IN) utilizing the Titanium Sequencing Kit (Roche). Sequences containing di-, tri- or tetranucleotide repeat motifs with a minimum copy number of 7, as well as a minimum of 20 nucleotides flanking the repeat, were identified using the Tandem Repeat Database (Gelfand et al. 2006). The software Primer3 (http://primer3.sourceforge.net/) was then used to design primer sets for candidate microsatellites.

These PCR primers were screened for their ability to consistently amplify DNA fragments of expected allelic size from the monoclonal strains of *S. goreaui* (rt-113 and rt-152) and from samples obtained from animal tissues containing Symbiodinium clade C. These samples were selected for their diverse geographic origin (Hawaii, USA, Great Barrier Reef, Australia and Okinawa, Japan) as well as their diagnostic ITS sequence (C1, C1c, C3 and C3ee-ff). PCR reactions with M13 labeled primers were preformed according to the methods described by Schuelke (2000). Loci were only retained if they were polymorphic within or between types of *Symbiodinium* types (Table 1). Dye labeled forward primers were then used to sample genetic diversity in populations of four sub-clade C types (Table 2). A linkage disequilibrium test was preformed between all loci with sufficient sampling within populations (see table 2) using the software GENEPOP (Raymond and Rousset 1995). All pairwise test were insignificant for linkage with the exceptions of SgrSpl_26 and Sgr_40 in C1c in Hawaii and Spl_1 and Spl_89 in C3ee-ff from Okinawa, Japan.

These loci can be utilized in a number of genetic applications. Many of the microsatellites amplified across more than one of the species we tested (Table 2). Most
loci were polymorphic across samples of a given phylogenetic type and will be useful for examining diversity gene-flow and genetic exchange. Several were polymorphic only between, and hence diagnostic of, certain phylogenetic types. (Table 2). These may be utilized in the future as tools for identifying species within clade C. Collectively, these microsatellites offer additional high-resolution genetic markers for researchers investigating *Symbiodinium* ecology and their use may provide new insights into species diversity, mechanisms for local adaptation, as well as, population dynamics during bleaching and recovery of reefs.


dominance and genetic diversification of coral endosymbionts in the genus

_Symbiodinium._ *J Biogeography* 37: 785-800.


Table 1. Locus name, size of product repeat motif, primer sequences, annealing temperature ($T_a$) and GenBank accession number for eighteen microsatellite loci for *Symbiodinium* Clade C.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Fragment Length</th>
<th>Repeat motif</th>
<th>Primer Sequences (5’-3’)</th>
<th>$T_a$ (°C)</th>
<th>GenBank accession #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sgr_9</td>
<td>232</td>
<td>GTT</td>
<td>F:TCTTTTGTGCTCTGGTTTG</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R:GGGAGACATTTGAAGACAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_19</td>
<td>152</td>
<td>TCA</td>
<td>F:AACACATGAAAACACGAC</td>
<td>50</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R:GATGGTAAAGAGTGAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_21</td>
<td>142</td>
<td>TCA</td>
<td>F: TGCTGATGTCCTGGTTATATC</td>
<td>53</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: TGATGCTACCTGGATATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_33</td>
<td>171</td>
<td>TA</td>
<td>F: ATGAAGCTGATGGGAATGAT</td>
<td>49</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GGGCTTCAATACCATAAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_34</td>
<td>158</td>
<td>TA</td>
<td>F: TGACCTTCAGCCAGTTTATATC</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CATGCTCAGCCAGGTGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_37</td>
<td>123</td>
<td>TA</td>
<td>F: TTTGGCATTTGGTGGCTCT</td>
<td>52</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: AAAGGACATACAGGAGCTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_40</td>
<td>198</td>
<td>ATAG</td>
<td>F: GCTGGTAAAGCTGATGTA</td>
<td>54</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GGGGAGAGATGTTGTATCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_13</td>
<td>188</td>
<td>TAG</td>
<td>F: GTCTAAAGGGAGATACATG</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CCATTACATGGTCTAGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_22</td>
<td>137</td>
<td>TAA/TCA</td>
<td>F: CGATCTATCTCATCCACAATA</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: TCAGGTGTGACACAGAAAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_24</td>
<td>129</td>
<td>GAT</td>
<td>F: GGAATGTTGGAGCATATG</td>
<td>50</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CCAATGGTGTACATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_25</td>
<td>122</td>
<td>AAG/TGA</td>
<td>F: CGTGGAGACAATGGAGTCA</td>
<td>49</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: TGCGGAAGAATGGGATC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_26</td>
<td>121</td>
<td>CAA</td>
<td>F: TTTCTTTTGTGACACACCAC</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CACTGAAAGACGAAATGCAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_30</td>
<td>214</td>
<td>TA</td>
<td>F: CCGACTACCTTTGCTAAG</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CAATGCTGACGACAGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_78</td>
<td>164</td>
<td>TA</td>
<td>F: TGAATTGCGCTGTATGTG</td>
<td>51</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CTCAGATTTTTGGAGACTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_1</td>
<td>137</td>
<td>TA</td>
<td>F: TGGATATGCTATGAAGAAA</td>
<td>50</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GGCCTTCCAGATGTTATT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_7</td>
<td>208</td>
<td>CAT</td>
<td>F: ACTTGAAGATCATGACAG</td>
<td>50</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GACGCTTTGAAAGTCTGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_16</td>
<td>170</td>
<td>AGG</td>
<td>F: CTGGAAGTGTGAGCATACAT</td>
<td>49</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: CATCATTTTTGCTATGTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_89</td>
<td>170</td>
<td>TA</td>
<td>F: CGGTTGGGTAAGATCATG</td>
<td>50</td>
<td>GQXXXXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: GGCATGCTTGGATTATT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 2

<table>
<thead>
<tr>
<th>Locus</th>
<th>C1 GBR Range</th>
<th>N</th>
<th>Na</th>
<th>C3 GBR Range</th>
<th>N</th>
<th>Na</th>
<th>C1c Hawaii Range</th>
<th>N</th>
<th>Na</th>
<th>CZ Okinawa Range</th>
<th>N</th>
<th>Na</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sgr_9*</td>
<td>254-271</td>
<td></td>
<td></td>
<td>169-178</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_19*</td>
<td>132-154</td>
<td>3</td>
<td>3</td>
<td>154-166</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_21*</td>
<td>124-136</td>
<td>5</td>
<td>7</td>
<td>182-215</td>
<td>52</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_33*</td>
<td>121-131</td>
<td>7</td>
<td>7</td>
<td>125-141</td>
<td>10</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_37*</td>
<td>142-157</td>
<td>4</td>
<td>4</td>
<td>128-137</td>
<td>42</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sgr_40</td>
<td>136-193</td>
<td>19</td>
<td>19</td>
<td>125-141</td>
<td>10</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_13</td>
<td>134-140</td>
<td>51</td>
<td>5</td>
<td>101-107</td>
<td>10</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_22</td>
<td>111-128</td>
<td>20</td>
<td>20</td>
<td>173-183</td>
<td>11</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_24</td>
<td>142-157</td>
<td>4</td>
<td>4</td>
<td>142-157</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_25</td>
<td>150-164</td>
<td>5</td>
<td>5</td>
<td>150-164</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_26</td>
<td>121-131</td>
<td>20</td>
<td>20</td>
<td>121-131</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SgrSpl_30</td>
<td>168-184</td>
<td>9</td>
<td>9</td>
<td>168-184</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_1*</td>
<td>125-136</td>
<td>7</td>
<td>7</td>
<td>125-136</td>
<td>7</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_16</td>
<td>142-157</td>
<td>19</td>
<td>19</td>
<td>142-157</td>
<td>19</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_33</td>
<td>134-140</td>
<td>51</td>
<td>5</td>
<td>134-140</td>
<td>51</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spl_89*</td>
<td>121-131</td>
<td>20</td>
<td>20</td>
<td>121-131</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Size ranges of 4 clade C types, N is the number of samples genotyped and Na is the number of alleles observed. Data in bold included in linkage-disequilibrium test.

* Allele sizes include 20bp M13 sequence used to fluorescently label PCR products with the method described in (Schuelke 2000).
VITA
Francis Condon Wham Jr. (Drew)
Fcwham@gmail.com

EDUCATION

PhD. in Biology
The Pennsylvania State University, University Park, PA
Dissertation Advisor: Todd C. LaJeunesse, Ph.D.

M.S in Marine Biology
College of Charleston, Grice Marine Laboratory, Charleston, SC
Thesis: Molecular Genetics of the Coral Holobiont
Thesis Advisor: Phillip Dustan, Ph.D.

B.S. in Biology
College of Charleston, Charleston, SC
Thesis: Symbiont Genotyping in Corals: Identifying Coral Associated Microbial Eukaryotes
Thesis Advisor: Phillip Dustan, Ph.D.

TEACHING EXPERIENCE

Pennsylvania State University, State College, PA
Instructor Invertebrate Zoology Laboratory
Undergraduate Project Mentor
College of Charleston, Charleston, SC
Biology of Coral Reefs: Teaching Assistant
Advanced Topics in Biology: Instructor
Biology 111 Lab: Teaching Assistant

AWARDS

Summer Institute of Statistical Genetics – University of Washington
National Science Foundation, East Asia and South Pacific Summer Institute
Braddock Award – Pennsylvania State University
Genomics Fellowship - College of Charleston
McLeod-Frampton Scholarship
College of Charleston Graduate Award
Summer Institute of Statistical Genetics – University of Washington
College of Charleston Graduate Poster Presentation Award