Using Nuclear DNA to Investigate the Phylogenetics of Californian Threespine Stickleback Populations

Thomas Frandsen
University of Redlands

Follow this and additional works at: https://inspire.redlands.edu/cas_honors

Part of the Molecular Biology Commons

Recommended Citation

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 License
This material may be protected by copyright law (Title 17 U.S. Code).
This Open Access is brought to you for free and open access by the Theses, Dissertations, and Honors Projects at InSPIRe @ Redlands. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of InSPIRe @ Redlands. For more information, please contact inspire@redlands.edu.
Using Nuclear DNA to Investigate the Phylogenetics of Californian Threespine Stickleback Populations
By Tom Frandsen

Senior Honors Thesis
Advisor: Dr. Linda Silveira
University of Redlands – May 2006
Abstract

The threespine stickleback, *Gasterosteus aculeatus*, is a diverse superspecies that exists mainly in coastal temperate zones of the Northern hemisphere with a variety of life history modes. Past phylogenetic analyses of the populations of threespine stickleback have been completed using mitochondrial DNA (mtDNA) data. However, mitochondrial introgression has been observed in many fish species, including threespine stickleback, which causes phylogenetic studies using mtDNA to be false. Here, mitochondrial introgression refers to the movement of mitochondrial genes from one population to another population.

Nuclear DNA (nDNA) was analyzed in this study to assess the validity of the mtDNA phylogenetic studies of Californian populations. Up to three nuclear loci containing a total of 41 single nucleotide polymorphisms (SNPs) were analyzed from five Californian populations of threespine stickleback and compared to stickleback from Pacific North America. Neighbor-joining trees were constructed and showed different relationships than found in the mtDNA-based study. Californian populations were found to be distinct from a British Columbian population and two different clades of Californian populations were found, including a Lost Lake/Shay Creek/Fillmore clade and a Holcomb Creek/*williamsoni* clade.

These results, however, are inconclusive due to a large amount of intrapopulation variation at the sites assayed. Lost Lake seems to contain an unusually large amount of variation that makes it unique. Further study, though, must be completed to determine the validity of the these results and those from the previous mtDNA-based studies. Specifically, Shay Creek Stickleback may be confirmed as being closely related to *Gasterosteus aculeatus williamsoni*, which is already on the endangered species list or Lost Lake may be found to be another unique population. Conservation efforts may then be made to protect Shay Creek and/or Lost Lake Stickleback.
Introduction

The threespine stickleback, *Gasterosteus aculeatus*, is a species complex containing many populations from different geographical habitats (Bell & Foster, 1994). This includes thousands of phenotypically varying populations found in the Northern Hemisphere in coastal marine waters as well as a variety of freshwater locations (Bell & Foster, 1994). Their range stretches around the Pacific Rim, from Baja California around to the Japanese Sea, and they are widespread along the Atlantic coast, from the Chesapeake Bay to Greenland and around to Europe and the Mediterranean Sea (Bell & Foster, 1994).

Morphologically, the threespine stickleback is a small, streamlined fish that are typically five to eleven centimeters in length (Bell & Foster, 1994). It has extensive protective armor on its body that distinguishes this fish from others. One form of armor is its namesake spines along the dorsal portion of the fish. There are two large, serrated spines and one smaller spine located in front of the dorsal fin (Bell & Foster, 1994). Most populations of threespine stickleback also have bony armor plates, which are also called lateral plates. These structures lie directly under the skin and look like bony scales that run along the side of the fish (Colosimo *et al.*, 2005). The number of these armor plates varies among the populations of stickleback. There are complete morphs that typically have a continuous row of armor plates from head to tail (32 – 36 plates), partial morphs that have a gap in the middle of the row of plates, and low morphs that have few (4 – 9 plates) or no plates present (Hagen & Gilbertson, 1972; Bell & Foster, 1994). The threespine stickleback species complex also includes a wide range of colors in the different populations. The dorsal portion of the body ranges from brown to green and the
ventral side is normally pale. During reproductive conditioning, males obtain a red coloration on ventrolateral parts of the body, which vary in intensity and extent of coloration between different populations (Bell & Foster, 2004). Thus, there is much phenotypic variation among threespine stickleback populations.

Populations of threespine sticklebacks also vary in life history modes. They have one of three life histories: fully marine, anadromous (they live in seawater and migrate into freshwater areas to breed), and resident freshwater (Orti et al., 1994; Bell & Foster, 1994). The freshwater populations are thought to have originated from marine and anadromous sticklebacks that migrated inland (Bell & Foster, 1994).

The divergence of the three life histories from ancient forms of *G. aculeatus* differs greatly. The fossil record shows that the species began to diversify at least 10 million years ago (Bell, 1994). The marine and anadromous forms have changed little during this time (Bell, 2001). However, the freshwater populations have been rapidly evolving. These populations come in two forms. One type of freshwater stickleback colonized new habitats after the receding of the glaciers in the last ice age (late Pleistocene), which occurred approximately 12,000 to 13,000 years ago (Bell & Foster, 1994; Johnson & Taylor, 2004). The other population type is thought to have colonized new habitats before this glaciation and survived in glacial refugia (O’Reilly et al., 1993; Orti et al., 1994; Thompson et al., 1997). The refugia were areas free of ice during the last glacial advance. Several different populations have been reported to have evolved in the refugia on and around the Queen Charlotte Islands off the coast of British Columbia (O’Reilly et al., 1993). The fastest evolving of the two types of freshwater populations is the post-glaciation colonizing type (Bell, 2001). However, no populations from any of
the life-histories have diverged so much in the past 10 million years that they have created a new, separate species. This complex relationship of evolving rapidly in molecular terms, yet staying fairly phenotypically similar has been called a "paradox" (Bell, 2001).

Similar environmental pressures on different populations of sticklebacks have caused parallel evolution of the populations to occur (Thompson et al., 1997). This means that independent lineages of sticklebacks have colonized different areas, yet evolved similar traits. The result of this type of evolution is that similar morphology results in different populations, however, the populations may only be distantly related. This makes it very difficult to determine which populations are more closely related based on the morphology of threespine sticklebacks. Research reported by Thompson et al. (1997) provides an excellent example of this idea. In this study, three parapatric pairs of sticklebacks from the northeastern Pacific were analyzed. In this case, the term parapatric refers to a lake form of stickleback living adjacent to a stream form of stickleback (Purves et al., 2002; Thompson et al., 1997). These three pairs of sticklebacks all lived in different lakes with an adjacent stream and the three lake/stream locations are within 400 km of each other. In comparing the morphology of the three pairs, the lake forms all looked alike and the stream forms all looked alike, but were distinct from the lake forms. However, when molecular methods were used to analyze the phylogeny of these fish, it was found that the streams forms were only distantly related to each other. In addition, the lake forms were also found to be only distantly related to each other. Thus, parallel evolution occurred, in which the same phenotypes were produced, but the populations were actually phylogenetically divergent.
The mechanism of parallel evolution in threespine stickleback shows that using morphological characteristics to determine phylogeny can be problematic. The rapid divergence of many populations also confounds the problem because it is unknown which morphological traits are shared with a common ancestor or derived separately (Haglund et al., 1992). Due to these difficulties, molecular methods of determining phylogeny have been used. Data based on protein and DNA sequences is particularly advantageous for many reasons. One is that these are unambiguously inherited entities. The character state of this data is also unambiguous and much more abundant than morphological data (Graur & Li, 2000). Many different molecular methods have been used to study threespine stickleback phylogenetics.

Allozyme analysis was the first type of molecular method used to determine stickleback phylogeny (Haglund et al., 1992). Allozymes are different forms of the same enzyme that are made due to different alleles of a gene. These different forms of the enzyme may be present in different populations due to evolution of the protein. In order to determine the particular allozymes an organism may have, a native gel must be made, which separates based on both charge and size (Haglund et al., 1992). Electrophoretic data from the studies can be used to determine relationships of the different stickleback populations and the information can be presented in a phylogenetic tree (Haglund et al., 1992; Avise, 1989). Although, this type of study can be very useful, it can be misleading because the same electric charge of a protein can be produced by many different combinations of amino acid sequences (Avise, 1989). Since the amino acid sequence may significantly differ, there may actually be only a distant relationship between two populations of sticklebacks, when allozyme analysis would imply that the two
populations are closely related (Orti et al., 1994). This type of analysis is also insensitive to synonymous substitutions, in which a DNA mutation does not change the amino acid. Thus, evolutionary relationships found using electrophoretic mobility can not be safely inferred (Avise, 1989).

The second type of analysis used to study phylogenetics in threespine sticklebacks was to directly examine the DNA sequence from different populations. DNA is favored in phylogenetic analysis of sticklebacks for many reasons. One is that it is specific to an individual, while allozymes and morphology can be shared by distantly related organisms (Avise, 1989). Thus, the DNA sequence can show the true underlying characteristics of a gene as opposed to the gene product or how that gene is expressed into a phenotype.

Examining DNA sequence is also advantageous because of the usefulness of the molecular clock. The molecular clock concept proposes that the rate of molecular evolution is approximately constant over time for a given gene (Graur & Li, 2000). This allows the time of divergence between two species to be calculated. Furthermore, different genes will evolve at different rates (Graur & Li, 2000). This is useful because different genes could be used to answer a wide range of phylogenetic questions. For example, genes that mutate slowly can be used to determine phylogeny of distantly related species. Genes that mutate at a high rate can not answer this question because the sequences from those species would be so different that relationships could not be inferred.

The preferred DNA type to use in phylogenetics studies has been mitochondrial DNA (mtDNA) (Patterson, 2002; Avise, 1989). Mitochondrial DNA is advantageous over nuclear DNA (nDNA) because it does not recombine (Avise, 1989). Mitochondrial
DNA also has a mutation rate that is about ten times greater than the rate found in nDNA (Graur & Li, 2000; Li, 1997; Avise, 1989). This is helpful in looking at population level relationships because this relatively faster rate of evolution is expected to provide better resolution of phylogenetic relationships among recently diverged taxa. However, the evolution of the mtDNA between populations can not be so different that the sequences can not be aligned and compared (Hillis et al., 1993). Thus, the sites being compared have to be similar enough among the taxa to ensure that they are homologous, but must be different enough so phylogenetic relationships can be inferred. With these concepts in mind, many evolutionary studies have used the mitochondrial cytochrome b gene to assay evolutionary relatedness among populations (e.g. Chow et al., 1995; Orti et al., 1994; Sullivan et al., 2004). This gene is very useful because it is conserved in all organisms and it mutates at a rate that works well for studying the evolution of different populations of a species (Bell & Foster, 1994).

Once a comparison of a particular sequence has been made between the taxa in a study, this data must be presented in a form that is clear and concise. Phylogenetic trees provide a means to present this data. The data from the sequence comparison is converted into a meaningful, evolutionary map that scientists can analyze and see phylogenetic relationships among the taxa of interest.

There are two types of data that are used to construct phylogenetic trees. One type of data is character data (Hall, 2001). A character is a defined taxonomic feature that can have one out of two or more unambiguous states (Graur & Li, 2000). Examples of this type of data include DNA nucleotides, amino acids, or the presence or absence of a trait. The other type of data is called distance data. To use distance data, two
sequences are compared and the amount of dissimilarity between the sequences is used to find a distance (Baxevanis & Ouellette, 2005; Graur & Li, 2000). In this way, character data is converted to distance data. For example, if a 500 base pair sequence from the cytochrome b gene from one population of threespine stickleback is compared to the same sequence in another population, there may be eight base pairs in which the two sequences differ. This distance of “eight” can be divided by the total number of sites (500) to give a percent dissimilarity (8/500 multiplied by 100%) (Graur & Li, 2000). This procedure is then repeated for all of the populations in the study, so that every pairwise comparison possible has been made. This data can then be put into the form of a matrix, in which a table is made with all of the populations on both the x and y-axes and the distance values are filled in the middle of the table. The matrix is then used to create phylogenetic trees.

In converting character data to distance data, a lot of information is lost because each pairwise sequence comparison has been reduced to a single number. However, distance data can accurately reflect phylogenies. The true tree could be made if all of the evolutionary substitutions were present in the sequences (Baxevanis & Ouellette, 2005). However, multiple substitutions can occur at the same site or a back mutation could take place, in which a substitution occurs at a site and then later reverts back to the original state (Hillis et al., 1993; Baxevanis & Ouellette, 2005). If many multiple substitutions at a site or back mutations occur, then the true evolutionary history can not be reflected by the distance data (Baxevanis & Ouellette, 2005). There are ways to estimate the number of these “unseen” substitution events and correct for this problem (Graur & Li, 2000; Baxevanis & Ouellette, 2005). Distance data, though, reflect the true evolutionary
history most accurately if there is a small evolutionary distance between two taxa (meaning a small number of sequence differences), thus, lowering the probability of unseen events (Hillis et al., 1993). Distance data does have advantages over character data, including the fact that it can produce a phylogenetic tree much faster than character data (Graur & Li, 2000). Therefore, both types of data are useful and should be analyzed when inferring phylogenetic relationships.

There are many different methods of building trees using distance data. The simplest method is called Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) (Graur & Li, 2000; Hall, 2001). The tree is made using the clustering method, in which the most similar taxa are sequentially grouped together (Hall, 2001). To begin this process, a distance matrix is generated with the distance data as explained previously. Then, the pair of taxa with the smallest distance between them is clustered together to form a branch on the tree. A new matrix is then created in which there is one less entry. One less entry results because the newly clustered taxa are placed into a single entry in the matrix. An arithmetic mean is used to calculate the distance between this cluster and the remaining taxa on the matrix. The procedure is then repeated until all of the taxa are placed on the tree (Hall, 2001). This method, thus, builds the tree one branch at a time. When grouping two taxa together, this method also assumes that the distance between two taxa is equal, which has been dismissed as very unlikely by phylogeneticists. For this reason, this type of tree building method is rarely used today (Hall, 2001). However, it provides a good example of how a phylogenetic tree is made with distance data.

Neighbor-joining is one of the most popular tree building methods and probably the most popular distance-based method. It is similar to UPGMA in that it uses a
distance matrix. However, neighbor-joining uses a star tree, which is an undifferentiated tree that gets its name because it is shaped like a star. Using this tree, taxa that have the smallest distance are joined together on a branch (Baxevanis & Ouellette, 2005). The program then calculates the distance from each taxon to the node that joins them (Hall, 2001). Unlike UPGMA, neighbor-joining allows the distances between each taxa and the node to be different lengths. A new matrix is then created and the process is repeated until all branches have been added to the tree. The best tree in this method will be the tree with the shortest branch lengths (Li, 1997). Overall, distance methods will yield a single, best tree at the end of the search (Hall, 2001).

There are two character-based methods that are very popular. These include maximum parsimony (also called parsimony) and maximum likelihood (ML) (Hall, 2001). Parsimony uses the underlying assumption that the best tree is the one that describes the data with the least number of steps (Graur & Li, 2000). However, this method differs from distance methods in that equally parsimonious trees can be produced (Graur & Li, 2000). This means that there are at least two trees that have an equally small number of steps to explain the data (Baxevanis & Ouellette, 2005). In this method, all of the sites in a sequence are first classified (Graur & Li, 2000). A site in the sequence could be invariant among all the taxa or it could be variant. The invariant sites are immediately thrown out of the analysis because without any differences at the site, separation of the taxa into evolutionary groups can not be made. Thus, phylogenetic relationships could not be determined from this site. Among the variant sites, there are informative and uninformative sites (Graur & Li, 2000). The informative sites favor one tree over another, while in the uninformative sites, that distinction is not present (example
given in Appendix A). Therefore, maximum parsimony only uses informative sites in the analysis, which is only a subset of the original data set. This method then creates a phylogenetic tree for each informative site. In the exhaustive searching method of parsimony, all possible trees are created and the best tree is the one with the smallest number of substitutions to explain that tree (Graur & Li, 2000). This process is repeated for all of the informative sites and the program keeps track of the number of substitutions for each tree. The most parsimonious tree will be the one with the smallest number of substitutions.

In addition to the exhaustive search method, there are other tree-searching methods that allow maximum parsimony to find the best tree faster. Branch-and-bound searching is an option that can be used, in which the best tree is always found and in a shorter amount of time (Graur & Li, 2000). In this search method, an upper bound length is set in the program, so that once a tree that is being built goes over this set length the programs stops building that tree and moves onto building a new tree. If a new tree has a shorter length than the upper bound, this new tree length becomes the upper bound. Another type of tree search that can be used is the heuristic search (Graur & Li, 2000). However, not all trees are searched in this method and the best tree is not guaranteed to be found. This search starts with a random input tree and takes branches off of the tree and roots them in other places in process called branch swapping (Graur & Li, 2000). The new tree is then assessed and the process is repeated until the best tree is found.

Maximum likelihood is also based on character data. However, the underlying assumption in this method is that the best tree is the one with the highest probability of observing the data. Therefore, ML uses probability to determine the best tree. Unlike
maximum parsimony, ML uses every site in the phylogenetic analysis (Baxevanis & Ouellette, 2005). At every site there is a probability associated with a possible mutation or lack of mutation. All possible trees are then created and a calculation is made for every site that determines the probability of a particular tree occurring. For a given tree, the probabilities of it occurring at every site are multiplied to give an overall probability, L (Graur & Li, 2000). The best tree is the one with the highest probability. This method is considered to be the most accurate method for determining phylogenies (Baxevanis & Ouellette, 2005). The major disadvantage of ML is that it is the most computationally intensive (Graur & Li, 2000). However, the branch-and-bound and heuristic searching methods can also be applied to ML to complete the analysis faster (Hillis et al., 1993).

Regardless of the method used to create a phylogenetic tree, the statistical confidence of the tree must be determined. A process called bootstrapping is used to do this. Bootstrapping tells how reliable a tree is as well as which specific branches of the tree are reliable (Hall, 2001). It falls into a category of resampling with replacement methods (Graur & Li, 2000). It works by creating a new alignment of sequences and from this alignment a new tree is made (an alignment refers to the sequences of the taxa lined on top of each other so that homologous sites match up (Baxevanis & Ouellette, 2005)). From the original alignment, a site is randomly picked and put into a new alignment. This step is repeated, in which another site is randomly picked from the original alignment to be in the new alignment. This second site could be the same as the first site. This process is repeated until the new alignment contains the same number of sites as the original alignment. A phylogenetic tree is built by the same previously used method (e.g. maximum parsimony). The new tree is assessed to see if the original clades
or groupings are still present (Hall, 2001; Felsenstein, 1985). This whole process is repeated a large number of times (normally 100 for ML and 1000 for other tree-building methods) and a percentage is determined for how often a particular clade appears in all of the bootstrapping replicates (Hall, 2001). There is no specific cut-off in bootstrap values that make a clade significant. However, it is generally accepted that with a 90% or greater bootstrap value, there is confidence that a branch on a tree is correct (Hall, 2001).

Through the use of molecular data to create phylogenetic trees, there have been many studies that have analyzed the phylogeny of the threespine stickleback. One of these studies was a global survey of the phylogenetic relationships of threespine sticklebacks reported by Orti et al. (1994). These researchers amplified a portion of the cytochrome b gene using PCR, sequenced the gene, compared the sequences from the different populations, and then created a phylogenetic tree based on the similarities and differences between the sequences. The overall result was that two major clades of sticklebacks were seen in the tree. There was a Japanese clade that contained all of the Japanese populations sampled and three populations from North America. The other clade was a Euro-American clade that contained populations from only Europe and North America. In this study, several British Columbian populations were assessed as well as a population living in the Ventura River, located near Los Angeles, California (Orti et al., 1994).

Recent studies at the University of Redlands expanded upon the results obtained by Orti et al. (1994) to include other Californian populations besides the Ventura River population. These populations include Fillmore Stickleback, Shay Creek Stickleback, Holcomb Creek Stickleback, \textit{G. aculeatus williamsoni}, and Lost Lake Stickleback.
(Patterson, 2002; Gunther, 1998; Perez, 1999; Tara Walker, University of Redlands, personal communication). These populations were placed into the phylogenetic tree made by Orti et al. (1994) so that evolutionary relationships could be determined. It was found that the Ventura River Sticklebacks, Holcomb Creek Sticklebacks, and Fillmore Sticklebacks all grouped together to form a clade in the tree. The williamsoni population and Shay Creek Sticklebacks formed a separate clade on the tree. The williamsoni and Shay Creek populations were also found to be more closely related to a group of sticklebacks in British Columbia and Alaska than to other California-residing sticklebacks (Patterson, 2002). Lost Lake Sticklebacks were later determined to be closely related to williamsoni and Shay Creek populations and, thus, grouped with those populations (Walker, personal communication).

The relationship of Southern Californian populations to sticklebacks in British Columbia introduces an interesting predicament of how the fish separated such a great geographical distance. One theory of how this could have occurred is based on glaciers receding or shifting during the last ice age. It is proposed that a population of sticklebacks migrated from British Columbia to Southern California via streams that were created by the movement or melting of glaciers (Patterson, 2002). It is interesting to reassess the stickleback relationships because additional evidence could support various theories of population origin like this intriguing idea proposed by Patterson.

These relationships among the Southern Californian populations have important consequences for the conservation of the threespine stickleback. Currently, G. a. williamsoni is listed by both state and federal agencies as an endangered species (Patterson, 2002). Since past studies at the University of Redlands found that both Shay
Creek and Lost Lake Sticklebacks are very closely related to the *williamsoni* population, this provides evidence that Shay Creek and Lost Lake populations should both be included as endangered species because of their similarity to *G. a. williamsoni* (Patterson, 2002).

As definitive as the cytochrome b phylogenetics results may seem, there is a fault with the mtDNA studies. It has been documented in many fish species that mitochondrial introgression has occurred (For example, Bagley & Gall, 1998; Bernatchez *et al.*, 1995; Chow & Kishino, 1995; Sullivan *et al.*, 2004; Takahashi *et al.*, 2003), including in threespine stickleback (Yamada *et al.*, 2001). Mitochondrial introgression is the introduction of new mitochondrial genes into a population by backcrossing hybrids between two populations (Rissler, 1993). This means that two different species or populations may come in contact with each other at some time and mate. It is then possible that one of the species’ genome will be changed so that it now contains mtDNA from the species with which it just mated. This is possible because mtDNA is inherited maternally, so that all progeny get their mtDNA from their maternal parent. To illustrate this concept, an extreme case of mtDNA introgression was reported by Bernatchez *et al.* (1995). It was found that the mtDNA from Arctic char and a certain population of brook trout was the same. This could be possible if a few female Arctic char mated with male brook trout. The progeny would have Arctic char mtDNA and a cross of char and brook trout nDNA. If these female progeny mated with the parental brook trout (backcross), it is possible to retain the mtDNA from the char, while reintegrating the brook trout nDNA in a species’ nuclear genome (Bernatchez *et al.*, 1995).
This idea of mitochondrial introgression can thus confuse the data obtained using mtDNA. As seen in scientific literature, if phylogenetic analysis is completed with nDNA and compared to analyses that used mtDNA, the results can be different (Bagley & Gall, 1998; Bernatchez et al., 1995; Chow & Kishino, 1995; Sullivan et al., 2004; Yamada et al., 2001). Therefore, it is possible that mitochondrial introgression occurred between the populations studied by Orti et al. (1994), Patterson (2002), and Walker (personal communication). This possibility is amplified by the fact that mitochondrial introgression is believed to have occurred between different Japanese populations of sticklebacks (Yamada et al., 2001). The opposing idea to mitochondrial introgression is that the past phylogenetic relationships are correct. This means that new populations were created by groups of stickleback migrating to previously unfounded areas.

In 2005, Colosimo et al. reported a study that used nDNA to determine phylogeny of threespine sticklebacks. In this study, researchers made a phylogenetic tree using 25 nuclear loci, which contained 193 single nucleotide polymorphisms (SNPs). A SNP is another name for a single-base pair substitution (Baxevanis & Ouellette, 2005). In this study, maximum likelihood, maximum parsimony, and another tree-building method called Baysian analysis were used to create phylogenetic trees of populations from around the world. However, the only Southern Californian population included in the study was williamsoni. This paper reports what seems to be the first study that used sequenced nuclear DNA to determine phylogeny of sticklebacks. My research project was able to develop from this framework.

The goal of my research was to use this paper as a template and reassess the phylogeny of Southern Californian populations of threespine stickleback. In this way, the
validity of the phylogenetic relationships found using mtDNA could be determined by analyzing nDNA in the Southern Californian populations of sticklebacks and comparing those results to the previously completed studies. If the same relationships were seen in both studies, then it could be determined that the previously found phylogeny was correct. If different relationships were seen between the studies then it is possible that mitochondrial introgression occurred. Colosimo et al. (2005) was also useful in this study because it analyzed populations from California, Washington, British Columbia, and Alaska. Furthermore, two of these populations, *G. a. williamsoni* and Friant, California, are actually two of the same groups studied in the Californian stickleback phylogenetic analysis by Patterson (2002) and Walker (personal communication). Therefore, the Southern Californian populations could be compared to these northern populations in order to determine if mitochondrial introgression took place and to determine the true relatedness of these populations.

Due to the time restrictions and the cost required to repeat Colosimo et al.’s analysis with Southern Californian populations, a subset of SNPs was picked from their study to analyze. Three of the 25 loci were chosen, which contained a total of 41 SNPs. A phylogenetic tree was made by comparing the sequences of the different populations. This tree contained the Southern Californian populations and the relevant populations from Colosimo et al. (2005). This includes the population from British Columbia to which three of the Southern Californian populations were closely related (Patterson, 2002). Thus, this project could either confirm or challenge the relationships found based on mtDNA.
This overall project of determining phylogenetics of the Southern Californian populations is important because it can confirm the existence of populations that are evolutionarily similar to *G. a. williamsoni*, which is on the endangered species list. Shay Creek Stickleback and Lost Lake Stickleback were found to be closely related to *G. a. williamsoni* in the mtDNA-based study. A study analyzing nDNA was necessary to determine the validity of this conclusion. If the results from both studies coincide, then a strong argument can be made to also include these other two populations of stickleback as endangered. The enlistment of these populations as endangered could then aid in preserving biological diversity of the world, which is the major objective of conservation (Purves et al., 2002).

**Materials and Methods**

A research protocol was found that could be used as a framework for using nDNA to study the phylogeny of threespine stickleback populations. This study would also preferably include populations from British Columbia and Alaska. It could then be determined if the Shay Creek Stickleback and *williamsoni* are actually closely related to the British Columbian and Alaskan populations. The research protocol used by Colosimo *et al.* (2005) fit these criteria. This study used nuclear data obtained from 193 SNPs from 25 nuclear loci in the stickleback genome. This data was obtained from Jeremy Schmutz, a co-author of Colosimo *et al.* (2005).

Due to labor and financial constraints of sequencing so many loci from the specimens, three loci were chosen to be analyzed in this study. The loci chosen contained many SNPs within each locus and many of these SNPs were informative,
which means that differences in sequence could be seen between the Pacific North American populations. The loci chosen from the paper include loci 95451, 95452, and 95470. These names were used in the sequencing project by Colosimo et al. (2005). Using the primer names used to amplify these loci, the loci names of 95451, 95452, and 95470 were converted in this study to Locus A, Locus B, and Locus H, respectively.

Table 1 outlines the changes in nomenclature between this study and the Colosimo et al. (2005) study.

Table 1: Differences in Naming used in this Study and Colosimo et al. (2005)

<table>
<thead>
<tr>
<th>Naming Used in Colosimo et al. (2005)</th>
<th>Corresponding Naming used in this Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer Names</td>
<td>Locus Name</td>
</tr>
<tr>
<td>P6-A10-F.f &amp; P6-A10-F.r</td>
<td>95451</td>
</tr>
<tr>
<td>P6-B03-F.f &amp; P6-B03-F.r</td>
<td>95452</td>
</tr>
<tr>
<td>P7-H05-F.f &amp; P7-H05-F.r</td>
<td>95470</td>
</tr>
</tbody>
</table>

Legend: The use of one letter for each locus and its corresponding primer set is derived from the third letter in the primer names used by Colosimo et al. (2005). For example, the name “Locus A” was derived from the “A” in P6-A10-F.

To aid in picking the loci with the greatest number of informative sites, the SNP data from all of the loci and all of the populations analyzed in Colosimo et al. (2005) were organized in an Excel spreadsheet (Microsoft Corporation, 2002). A matrix was made showing the genotypes of the sticklebacks at each of the 193 SNPs. This organization allowed for easy visual scanning of the data to look for intraspecific differences and interspecific differences in genotype at a particular SNP. The loci that were chosen showed differences when comparing Pacific North American populations to each other. These loci also contained sites that seemed to have a low amount of
intrapopulation SNP variation. Loci A, B, and H were determined to contain the most useful sets of SNPs, in which 41 total SNPs were observed. After the scheme of the experiment was determined, it could be carried out on the stickleback specimens.

The stickleback specimens used in this study came from four Southern Californian populations and one Central Californian population. The four Southern Californian populations came from Shay Creek, Holcomb Creek, the Fillmore fish hatchery, and *Gasterosteus aculeatus williamsoni* came from the headwaters of the Santa Clara River (Patterson, 2002). The Central Californian population came from Lost Lake. Figure 1 and Figure 2 show the locations of these populations (Donley, 1979). Patterson, 2002).

The Shay Creek, Holcomb Creek, Fillmore, and Lost Lake samples were collected by Dr. James Malcolm of the University of Redlands. These fish were killed on site with MS222 or killed in the laboratory. The *G. a. williamsoni* specimens were given to this study by Shauna Bautista, who is a fisheries biologist working in the Angeles National Forest. Unlike the other samples, the *williamsoni* specimens were found desiccated floating in the river or located on the bank (Patterson, 2002). All samples were frozen at -70°C until DNA isolation took place. Some of the fish sequenced in this analysis were used in the previously completed cytochrome b study (Gunther, 1998; Perez, 1999; Patterson, 2002). These include Fill 4, Fill 5, Hol 2, Hol 6, Shay 8, Shay 9, Will 3, and Will 4. The Lost Lake specimens were different from those used in the analysis by Walker (personal communication). These new specimens were labeled LL22 and LL23.
Figure 1: Location of Southern Californian Populations Analyzed in this Study

Legend: The Fillmore population and *G. a. williamsoni* are both found in the Santa Clara River. *G. a. williamsoni* are found in the upper portion of the river drainage, while Fillmore is found further downstream. Holcomb Creek Stickleback are located northeast of Big Bear Lake, while Shay Creek Stickleback are found northwest of Big Bear Lake. The map was adapted from Donley, Allan, Caro, & Patton (1979).
Figure 2: Location of Lost Lake Stickleback

Legend: Lost Lake Stickleback are located about 19 miles north of Fresno, near Friant, California. This lake itself is adjacent to the San Joaquin River. Map was adapted from Eureka (2006).
Genomic DNA was isolated from the specimens. This was completed using the QIAamp Tissue Kit (QIAGEN Inc., Valencia, California). White muscle tissue was utilized to obtain DNA in the Fillmore, Holcomb Creek, Shay Creek, and Lost Lake samples. In the degraded *williamsoni* samples, organs and vertebrae were used to obtain DNA. Tissue was excised from the samples using a razor blade and placed into a microcentrifuge tube. Tissue sample sizes ranged from 256 mg to 347 mg. The protocol from the kit was then followed, including the optional step of applying RNase to the samples (QIAGEN Inc., Valencia, California). DNA was eluted in a final volume of 100 μL.

Verification of DNA isolation was completed using gel electrophoresis. Samples were run on a 0.8% agarose gel in 1x TBE electrophoresis buffer or 1x TAE electrophoresis buffer with 1% ethidium bromide (Sambrook & Russell, 2001). A one kilobase DNA ladder was also ran as a positive control and to approximate concentration of the isolated genomic DNA. Gels were digitally photographed using the GelDoc-It Imaging System with a Firstlight UV transilluminator and visualized using LabWorks Image Acquisition and Analysis Software (UVP BioImaging Systems Inc., Upland, California). Concentrations of the isolated DNA varied from approximately 0.6 ng/μL to 12 ng/μL as estimated from comparison to the DNA ladder standards.

In order to amplify the nuclear loci analyzed in this study, the Polymerase Chain Reaction (PCR) was used. Primers used to amplify Loci A, B, and H were the same as those used in Colosimo *et al.* (2005). The primers used to amplify Locus A were P6-A10-F.f (5’ GGAGGAAGTGTGAAGAGTGCAA 3’) and P6-A10-F.r (5’ GTTCAAGCTGTTGCACATGGAC 3’). The primers used to amplify Locus B were P6-
B03-F.f (5' TGAACACGCAATCCCTTCATTA 3') and P6-B03-F.r (5' TCGGAGGTACGACCTTTTC 3'). The primers used to amplify Locus H were P7-H05-F.f (5' CCAGTGTGAGCATCTCTATCCG 3') and P7-H05-F.r (5' CTCTGCCTCAGGGAGCAGATT 3'). Amplification of the genomic stickleback DNA was completed using the QIAGEN Taq PCR Master Mix Kit (QIAGEN Inc., Valencia, California). A total reaction size of 100 μL was used for the Lost Lake samples. These samples contained 10 μL of isolated DNA ranging in concentration from approximately 4 ng/μL to 12 ng/μL. For all of the other stickleback samples, a total reaction size of 50 μL was used. In these samples, 5 μL of isolated DNA ranging from approximately 0.6 ng/μL to 8 ng/μL in concentration was used. Both sizes of reactions had the same final concentrations of other reagents, including 2.5 units of Taq DNA polymerase, 1x Qiagen PCR Buffer (containing a final concentration of 1.5 mM MgCl₂), 200 μM of each dNTP, and 0.25 μM of each primer. The reactions were placed in the PCR machine with the "pre-heat lid" option selected and were run with the following program: 95° C for 10 minutes, 94° C for 30 seconds, 63° C for 30 seconds, 72° C for 23 seconds, then repeat the last three steps 35 times, and finish with 72° C for 3 minutes and 30 seconds. PCR was found to work only when the lid was pre-heated. It is thought that without pre-heating the lid, the sample would be in the gas phase for most of the sequencing reaction.

To determine if amplification of the loci had taken place, some of each PCR product was loaded onto a 2% agarose gel in 1x TBE electrophoresis buffer or 1x TAE electrophoresis buffer and ran with 1% ethidium bromide in the gel and in the running buffer. A preparatory gel was then made to isolate the PCR product. For the 50 μL
reactions, all remaining PCR product was run on the preparatory gel. For the 100 μL reactions, 50 μL of the PCR product was used. The preparatory gel was viewed using a UV transilluminator and the PCR product bands were cut out of the gel using a razor blade. Each PCR product was extracted and isolated from the gel using the QIAquick Gel Extraction Kit (QIAGEN Inc., Valencia, California). The isolated DNA was eluted into a final volume of 30 μL. The isolated DNA was run on another 2% agarose gel (as previously described) to verify that gel extraction was successful. A one hundred basepair DNA ladder was run with the samples in order to determine the concentration of each PCR product. The concentration is needed because it allows how much DNA is necessary for the DNA sequencing reactions to be known. Almost all of the PCR products ranged between 15 ng/μL and 90 ng/μL in concentration, but some samples were as low as 4 ng/μL.

The isolated template DNA was sent to be sequenced at the California State University Northridge sequencing facility. The primers used to sequence the loci were the same as those described above (P6-A10-F.f and P6-A10-F.r for Locus A, P6-B03-F.f and P6-B03-F.r for Locus B, and P7-H05-F.f and P7-H05-F.r for Locus H). Primers were sent in concentrations of 1 pmol/μL and approximately 3 μL of each primer was necessary for each reaction. The sequencing reactions required 10 ng of DNA per 100 bases being sequenced, so the amount of template sent for each sample varied because the concentrations of the templates differed. The reported sizes of the sequenced products are as follows: 809 bp for Locus A, 465 bp for Locus B, and 584 bp for Locus H.
The sequence results were returned from California State University Northridge in electropherogram format. These were viewed using the EditView Program (Applied Biosystems, Foster City, California). The reliable portion of each sequence was copied from the electropherogram into the program MacVector version 7.2 (Accelrys, 2001). This format of the sequence could then be imported into AssemblyLIGN Version 1.0 (Oxford Molecular Group, Inc., 1998). In the AssemblyLIGN files, the forward and reverse sequences for each sequencing reaction were imported and aligned along with the consensus sequence of the corresponding locus (provided by Jeremy Schmutz of the Colosimo et al. (2005) study). The SNP positions in the loci were also provided by Jeremy Schmutz and could be easily observed once the Colosimo et al. (2005) consensus sequence was aligned with the sample sequences. By observing the position where the Colosimo et al. (2005) consensus sequence started, it was known where nucleotide #1 was in the stickleback sample sequences. Thus, the other SNP positions could be determined by counting along the consensus sequence. The genotype at a SNP was recorded and checked in the EditView electropherogram file to verify accuracy and to look for possible heterozygosity at the site. If a nucleotide was not properly read by the sequencing machine, the base was changed in the AssemblyLIGN program (Oxford Molecular Group, Inc., 1998).

Before phylogenetic trees containing the Californian populations were made, phylogeny of the populations from the Colosimo et al (2005) research was determined using only the SNPs analyzed in this study. Therefore, only the 41 SNPs from Loci A, B, and H were used to make phylogenetic trees instead of using the 193 SNPs as previously reported. This was completed to compare the tree used in this study to the tree obtained
in Colosimo et al. (2005), which would help determine if the SNP analysis completed in this study is valid. For each population, a consensus sequence of the SNPs was produced and put into FASTA format. For sites in which a SNP base varied in a population, standard DNA ambiguity codes were used to account for this variation. These sequences were then aligned using ClustalW at the European Bioinformatics Institute website (http://www.ebi.ac.uk/clustalw/) and saved using the PHYLIP output format, which could be imported into the tree-building program.

PAUP v. 4.0b10 (Swofford, 1991) was used to create a maximum likelihood tree of this data. The following model of evolution was used: two parameter substitution model and a gamma distribution of evolutionary rates across sites, with 6 rate parameters in the discrete approximation to the gamma distribution. JASE was selected as the outgroup, as seen in the tree from Colosimo et al. (2005). Four equally likely trees were found and a consensus tree of those four trees was made.

The SNP data from the Californian populations was then investigated. This analysis included the samples from Colosimo et al. (2005) along with two Fillmore Sticklebacks, two Holcomb Creek Sticklebacks, two Lost Lake Sticklebacks, two Shay Creek Sticklebacks, and one G. a. williamsoni sequenced in this study. In this analysis, individual fish sequences were used to create trees instead of creating a consensus sequence for each population. This was completed in order to obtain trees that were more statistically significant and to observe intrapopulation variation. However, an intrapopulation consensus sequence was made if two fish had the same SNP nucleotides. Standard DNA ambiguity codes were used where necessary to account for heterozygosity.
at a SNP. After all possible intrapopulation consensus sequences were made, the sequences were aligned using ClustalW and saved using the PHYLIP output format.

PAUP v. 4.0b10 (Swofford, 1991) was used to create neighbor-joining trees of this data. These were created instead of maximum likelihood because of time constraints (neighbor-joining trees take the computer little time to make and maximum likelihood can take days to months). Using this method, many trees with different combinations of populations were produced to analyze certain phylogenetic relationships. For example, a tree using all of the Pacific North American Populations was made along with a tree that only included Southern Californian populations. A neighbor-joining/UPGMA search method was used for all trees and the maximum likelihood model of evolution was used for almost all of the trees. However, the Jukes-Cantor model of evolution was used in one tree to see if any differences in the topology arose and no significant difference in branching patterns were seen. The clade containing WALL, COND, and PAXB, as seen in Colosimo et al. (2005), was selected as the outgroup. Bootstrapping was completed on all neighbor-joining trees using 1000 replicates (Felsenstein, 1985).

In analyzing the sequence data, an open reading frame of Locus H was found in order to determine the codon position of Locus H-site 507. This was completed using the open reading frame finder at the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).
Results

DNA was isolated, amplified, and sequenced from a total of ten sticklebacks representing five populations. Most of the specimens sequenced in this research were the same specimens used in the cytochrome b phylogeny project. The Lost Lake samples were the only sticklebacks that were not the same fish specimens used in the cytochrome b study (Table 2). All three loci were able to be amplified from only six fish representing three populations (Holcomb Creek, Fillmore, and Lost Lake Stickleback). However, the sequence obtained from one of the Fillmore specimens (Fill 5) at Locus A was of such low quality that it could not be used. This situation also occurred with Will 4 resulting in a lack of sequence data from this fish specimen.

Table 2: Loci Sequenced from Stickleback Specimens and Used to Create Trees

<table>
<thead>
<tr>
<th>Population</th>
<th>Specimen Name</th>
<th>Loci Sequenced from the Specimen</th>
<th>Loci that Could be Read and Reported with Accuracy</th>
<th>Was this Specimen used in Cytochrome b Study?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fillmore</td>
<td>Fill 4</td>
<td>A, B, &amp; H</td>
<td>A, B, &amp; H</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Fill 5</td>
<td>A, B, &amp; H</td>
<td>A &amp; B</td>
<td>Yes</td>
</tr>
<tr>
<td>Holcomb Creek</td>
<td>Hol 2</td>
<td>A, B, &amp; H</td>
<td>A, B, &amp; H</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Hol 6</td>
<td>A, B, &amp; H</td>
<td>A, B, &amp; H</td>
<td>Yes</td>
</tr>
<tr>
<td>Lost Lake</td>
<td>LL22</td>
<td>A, B, &amp; H</td>
<td>A, B, &amp; H</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>LL23</td>
<td>A, B, &amp; H</td>
<td>A, B, &amp; H</td>
<td>No</td>
</tr>
<tr>
<td><em>williamsoni</em></td>
<td>Will 3</td>
<td>B &amp; H</td>
<td>B &amp; H</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Will 4</td>
<td>B &amp; H</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>Shay Creek</td>
<td>Shay 8</td>
<td>H</td>
<td>H</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Shay 9</td>
<td>H</td>
<td>H</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Legend: Fillmore, Holcomb Creek, and Lost Lake samples were sequenced at all three loci. However, only 5 out of 6 specimens yielded usable sequence data for all loci. *G. a. williamsoni* was only amplified at Locus B and Locus H, while Shay Creek samples were only sequenced at Locus H. Sequence data from Will 4 was of such low quality that it was unusable. Samples LL22 and LL23 were not used in the previous study based on mtDNA.

Three samples (Shay 8, Shay 9, and Will 3) could be amplified at only one or two of the loci, which allowed only a partial set of SNP data to be obtained for these fish...
(Table 2). It is possible that if PCR and band extraction were repeated for these samples, all loci could be isolated. However, due to time constraints it was not possible to obtain the full SNP dataset from these samples.

The sequences of the loci were received in an electropherogram format. The forward and reverse sequences were aligned and the SNP positions were identified, recorded, and confirmed in the electropherograms. Table 3 shows the genotypes recorded at each SNP for all of stickleback samples sequenced in this project.

This SNP data was then converted into a table that could be used to possibly determine similarities and differences in sequence between populations (Table 4). This table was made by creating a consensus sequence for all samples used in a given population and shows only variant sites among the Californian populations. Friant, California and *G. a. williamsoni* population data reported in Colosimo et al. (2005) are also shown in the table for comparison because these two populations were examined in the study reported here (the Friant population and the Lost Lake population are actually from the same location, while *G. a. williamsoni* from both studies are from the same river).
### Table 3a: SNP Composition of Californian Stickleback Populations at Locus A and Locus B

<table>
<thead>
<tr>
<th>Stickleback Sample</th>
<th>SNP Sites for Locus A and their Genotype</th>
<th>SNP Sites for Locus B and their Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fill 4</td>
<td>CCGCCTGGCCTT</td>
<td>ATGCGCCGGGAGG</td>
</tr>
<tr>
<td>Fill 5</td>
<td>Unreadable Sequence</td>
<td>TATGCGCCGGGAGG</td>
</tr>
<tr>
<td>Hol 2</td>
<td>CCGCGCGCGCT</td>
<td>TATGCGCCGGGAGG</td>
</tr>
<tr>
<td>Hol 6</td>
<td>CCGCGCGCGCT</td>
<td>TATGCGCCGGGAGG</td>
</tr>
<tr>
<td>LL22</td>
<td>CCGCGCGCT</td>
<td>TATGCGCCGGGAGG</td>
</tr>
<tr>
<td>LL23</td>
<td>CCGCGCGCT</td>
<td>TATGCGCCGGGAGG</td>
</tr>
<tr>
<td>Will 3</td>
<td>Not Sequenced</td>
<td>TATGCGCCGGGAGG</td>
</tr>
<tr>
<td>Will 4</td>
<td>Not Sequenced</td>
<td>Unreadable Sequence</td>
</tr>
</tbody>
</table>

### Table 3b: SNP Composition of Californian Stickleback Populations at Locus H

<table>
<thead>
<tr>
<th>Stickleback Sample</th>
<th>SNP Sites for Locus H and their Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fill 4</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>Fill 5</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>Hol 2</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>Hol 6</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>LL22</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>LL23</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>Will 3</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>Will 4</td>
<td>Unreadable Sequence</td>
</tr>
<tr>
<td>Shay 8</td>
<td>GCACGCTCAAGGGC</td>
</tr>
<tr>
<td>Shay 9</td>
<td>GCACGCTCAAGGGC</td>
</tr>
</tbody>
</table>

**Legend:** The stickleback sample names in the table refer to the following populations: Fill = Fillmore, Hol = Holcomb Creek, LL = Lost Lake, Will = G. a. williamsoni, and Shay = Shay Creek. The SNP sites numbers are read vertically (ex. the first SNP site in the table for Locus A is site 142). Single letters at a site indicate that the sample was homozygous for the described nucleotide. Heterozygous sites have the two nucleotides separated by a slash. G. a. williamsoni samples were not sequenced for Locus A and Shay Creek samples were not sequenced for Locus A or Locus B. Sequences for Locus A in Fill 5 and Loci B & H in Will 4 were unreadable.
Table 4: Nucleotide Composition at Variant Sites in Californian Populations

<table>
<thead>
<tr>
<th>Population or Specimen Name</th>
<th>Sites from Locus A</th>
<th>Sites from Locus B</th>
<th>Site from Locus H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 4</td>
<td>2 3</td>
<td>1 2 5</td>
</tr>
<tr>
<td></td>
<td>7 7</td>
<td>8 3</td>
<td>9 2 0</td>
</tr>
<tr>
<td></td>
<td>3 9</td>
<td>3 9</td>
<td>1 8 7</td>
</tr>
<tr>
<td>Fillmore</td>
<td>T C</td>
<td>G G</td>
<td>S G M</td>
</tr>
<tr>
<td>Holcomb</td>
<td>C Y</td>
<td>. .</td>
<td>G . C</td>
</tr>
<tr>
<td>Shay Creek</td>
<td>------------------</td>
<td>C</td>
<td>. .</td>
</tr>
<tr>
<td>Lost Lake</td>
<td>Y T</td>
<td>R .</td>
<td>G . H</td>
</tr>
<tr>
<td>Friant</td>
<td>Y Y</td>
<td>R A</td>
<td>C . C</td>
</tr>
<tr>
<td>Will 3</td>
<td>----</td>
<td>. .</td>
<td>G . A</td>
</tr>
<tr>
<td>WSMO</td>
<td>C .</td>
<td>. A</td>
<td>. A</td>
</tr>
</tbody>
</table>

Legend: The population abbreviations from the table correspond to populations from the following locations: Fillmore = Fillmore Stickleback; Holcomb = Holcomb Creek Stickleback; Shay Creek = Shay Creek Stickleback; Lost Lake = Lost Lake Stickleback; Friant = Friant, California, reported in Colosimo et al. (2005); Will 3 = G. a. williamsoni fish specimen #3 sequenced in this study; WSMO = G. a. williamsoni from the Santa Clara River, California reported in Colosimo et al. (2005). The SNP sites numbers are read vertically (ex. the first SNP site in the table for Locus A is site 373. Typical nucleotide symbols (A, T, C, or G) at a site indicate that the population was homozygous for the described nucleotide. Heterozygosity within a population is denoted using the following standard ambiguity codes: M = A/C, R = A/G, S = C/G, Y = C/T, H = A/C/T. The dot at each site represents the same nucleotide as reported in the sequence for Fillmore. Dashes in the Shay Creek and Will 3 sequences indicate that these SNP genotypes were not determined.

As seen in Table 4, there is a large amount of variation within each population as indicated by the heterozygosity codes. In comparing different populations, there is also much variation. In fact, few conclusions can be drawn from these variant sites by simply observing the table due to excessive variation between and within populations. However, one interesting point from Table 4 is that the G. a. williamsoni sample sequenced in this study differed at four out of five variant sites when compared to the G. a. williamsoni samples from Colosimo et al. (2005). The comparison of the Lost Lake sequence with the Friant sequence yields a similar result in that four out of seven possible SNPs differ between the populations. Thus, both of these comparisons show large amounts of differences in the sequences obtained from the same population.
Intrapopulation variation is seen in five of the 41 total SNP sites, including Locus A-site 373, Locus A-site 479, Locus B-site 283, Locus H-site 191, and Locus H-site 507 (Table 4). At four of these sites, one or two Californian populations are heterozygous, but at least one other North American population is also heterozygous at this site. Thus, it seems valid that this variation be present in some Californian populations. Site 507 from Locus H is different from the other sites because there appears to be considerable variation, including a trimorphic genotype seen in the Lost Lake Stickleback. However, it was difficult to determine the correct genotype at this site for the Lost Lake specimens when the electropherogram was examined. Therefore, additional sequencing reactions and further sampling would help confirm this unique situation.

In an attempt to explain the quickly evolving Locus H-site 507, open reading frames (ORFs) were found for Locus H. It was hypothesized that this site may mutate quickly if it is located at the third position in a codon. One large open reading frame of 180 nucleotides long was determined, in which site 507 was in the third codon position. A BLAST search of this ORF yielded no homologous sequences.

The SNP data was then used to build phylogenetic trees. Before the data obtained in this study was inserted into the Colosimo et al. (2005) dataset, a maximum likelihood tree containing only the taxa from the Colosimo et al. (2005) tree was made. In contrast to the published tree, this tree was made using only the data from the 41 SNPs analyzed in this study. The difference in the branching patterns between the two trees could help estimate the validity of using only a subset of SNP data to determine threespine stickleback phylogenetics. If the branching pattern was approximately the same, then it could be inferred that this analysis may be a good indicator of the true stickleback
phylogeny. This tree was produced using the method outlined in Colosimo et al. (2005) by having a single consensus sequence for each population. Figure 3 shows the maximum likelihood tree with the Colosimo et al. (2005) taxa, which can be compared to the tree reported by Colosimo et al. (2005) shown in Figure 4.

In comparing the two trees, a major difference is seen. Figure 3 shows a tree in which many branches split off from a single node or common ancestor of the tree, while Figure 4 does not show this branching topology. The type of pattern in Figure 3 is called an unresolved polytomy or polytomy, which means that the evolutionary relationships of the populations contained in that clade can not be deduced with statistical confidence. Unresolved polytomy resulted in this case because four equally likely trees were made using this method (Appendix B contains these trees- Figures A-4, A-5, A-6, and A-7). A consensus tree is made by combining these four trees and if there is a disagreement in branching pattern between two of the equally trees, a polytomy is produced in the program. The tree shown in Figure 3 is, thus, the consensus tree. The major similarity between the two trees is that the PAXB, WALL, and COND populations all branch off separately from the other taxa at the first branching point in the tree. The polytomy produced in this tree, though, proved to be a common characteristic of subsequently built trees.
Figure 3: Maximum Likelihood Consensus Tree of World Populations of Three-spine Stickleback Using Data from 41 SNP Sites

Legend: The abbreviations correspond to the following populations: AKMA = Alaska Marine; LITC = Little Campbell River in Vancouver, Canada; WSMO = *G. a. williamsoni* from Santa Clara River, California; NOST = Norway Stream; BLAU = Blautaver, Iceland; FADA = Loch Fada, Scotland; FRIL = Friant, California; SCX = Schwale, Germany; JAMA = Japan Marine; NAKA = Nakagawa Creek, Japan; OMPL = Olmstead Park, Jamaica Plain, Massachusetts; LLOY = Demarest Lloyd State Park, Massachusetts; NHR = New Harbor River, Nova Scotia, Canada; PAXL = Paxton Lake Limnetic from Texada, Island, Canada; GJOG = Gjogur, Iceland; COND = Conner Creek, Washington; PAXB = Paxton Lake Benthic from Texada, Island, Canada; WALL = Wallace Lake, Alaska; JASE = Japan Sea. Bootstrapping was not completed on this tree due to time restraints.
Legend: See Figure 3 for abbreviation descriptions. The Atlantic and Pacific clades of the threespine stickleback are outlined on the tree. Posterior probabilities greater than 95% are shown, with corresponding branches in bold. Posterior probabilities are similar to bootstrap values and were used on this tree due to the tree-building method (Bayesian analysis). Thus, the high posterior probabilities on certain branches indicate the branching pattern is highly statistically supported.
After the tree shown in Figure 3 was made, neighbor-joining trees were made using the data obtained in this study, including the Pacific clade stickleback data obtained from Colosimo et al. (2005). Table 5 shows the consensus sequences for each population that was used to create these trees. This set of populations was found to be variant at 27 of the 41 SNPs. Individual fish data was used to create the tree, which could allow higher bootstrapping values to be obtained and for intrapopulation variation to be seen. If two or more fish within a population had the same sequence, then the fish data was combined into a single sequence.

Table 5: Nucleotide Composition at Variant Sites in Californian Stickleback Populations and Populations from Colosimo et al. (2005)

<table>
<thead>
<tr>
<th>Population Name from Colosimo et al. (2005) or Lost Lake Specimen Name</th>
<th>Sites from Locus A</th>
<th>Sites from Locus B</th>
<th>Site from Locus H</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRIL</td>
<td>C C R C Y C G Y T T</td>
<td>T C G C G R A T G</td>
<td>C A G C A G G C</td>
</tr>
<tr>
<td>Lost Lake</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>Shay</td>
<td>G . . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>WMSO</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>Will 3</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>Holcomb</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>Fillmore</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>LITC</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>AKMA</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>JAMA</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>NAKA</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>PAXL</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>PAXB</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>WALL</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
<tr>
<td>COND</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
<td>. . . . . . . . . . .</td>
</tr>
</tbody>
</table>

Legend: The population abbreviations from the table correspond to populations from the following locations: FRIL = Friant, California; Lost Lake = Lost Lake, California; WSMO = G. a. williamsoni from Colosimo et al. (2005) (Santa Clara River, California); Will 3 = G. a. williamsoni sequence from this study (Santa Clara River, California); Holcomb = Holcomb Creek, California; Fillmore = Fillmore Hatchery population, Fillmore, California; LITC = Little Campbell River in Vancouver, Canada; AKMA = Alaska Marine; JAMA = Japan Marine; NAKA = Nakagawa Creek, Japan; PAXL = Paxton Lake Limnetic from Texada, Island, Canada; PAXB = Paxton Lake Benthic from Texada, Island, Canada; WALL = Wallace Lake, Alaska; COND = Conner Creek, Washington. The SNP sites numbers are read vertically (ex. the first SNP site in
the table for Locus A is site 142). Typical nucleotide symbols (A, T, C, or G) at a site mean that the sample was homozygous for the described nucleotide. Polymorphic sites are denoted by the following standard ambiguity codes: K = G/T, M = A/C, R = A/G, S = C/G, Y = C/T, H = A/C/T, V = A/C/G. The dot at each site represents the same nucleotide as reported in the sequence for FRIL. Dashes in Will 3 and Shay Creek rows indicate that these loci were not sequenced.

Figure 5a shows a tree containing all of the Pacific clade from Colosimo et al. (2005) along with Fillmore, Holcomb Creek Stickleback, and Lost Lake Stickleback. Only samples that were sequenced for all three loci were included in this tree. Thus, Shay Creek and Will 3 specimens were not included. The branch lengths of this tree depict the divergence in sequence between the taxa and can be useful in inferring phylogenetic relationships because branch length correlates to evolutionary relatedness. The outgroup of this tree was set to include PAXB, COND, and WALL as seen in Figure 4 (the tree obtained in Colosimo et al. (2005)).

There are interesting relationships that can be seen from this tree. One is that williamsoni appears to be quite distant from Fillmore, Holcomb, and Lost Lake. This contrasts past results, in which Lost Lake was closely related to williamsoni. Furthermore, Lost Lake and Fillmore actually seem to be closely related on this tree as shown by the LL 22, Fill 4, Fill 5 clade, with LL 23 being a close distance to this clade. Finally, a very intriguing result is that the Lost Lake fish sequenced in this study appear to be distinct from the Friant population surveyed in Colosimo et al. (2005) and these two taxa actually come from the same area. These relationships, however, are not well supported when the tree is bootstrapped (Figure 5b). All of the Californian populations are seen in the large polytomy of the tree, which means that the relationships shown in Figure 5a are not statistically supported.
Figure 5a: Neighbor-Joining Tree of Pacific Clade and Californian Populations of Threespine Stickleback

Legend: The abbreviations correspond to the following populations: FILL = Fillmore, California; HOL = Holcomb Creek, California; LL = Lost Lake, California; FRIL = Friant, California from Colosimo et al. (2005); WSMO = G. a. williamsoni from Colosimo et al. (2005) (Santa Clara River, California); LITC = Little Campbell River in Vancouver, Canada; AKMA = Alaska Marine; JAMA = Japan Marine; NAKA = Nakagawa Creek, Japan; PAXL = Paxton Lake Limnetic from Texada, Island, Canada; PAXB = Paxton Lake Benthic from Texada, Island, Canada; WALL = Wallace Lake, Alaska; COND = Conner Creek, Washington. Numbers following the abbreviations of each population refer to the specimen number(s) that have this haplotype (e.g. WMSO1 = specimen #1 from WMSO; WMSO2/3 = Specimens 2 and 3 from WMSO, etc.). COND is unique in that this population contains only one haplotype. Dotted lines on the tree are present to clarify the position of taxa on the tree. These lines do not refer to actual distance as the solid lines indicate.
Figure 5b: Bootstrapped Neighbor-Joining Tree of Pacific Clade and Californian Populations of Threespine Stickleback

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen.
Due to the large amounts of polytomy occurring in the data, different trees with varying taxa were made in an attempt to ask specific evolutionary questions. For example, it could be seen if Fillmore and Holcomb Creek Stickleback are actually in a separate clade from Shay Creek, Lost Lake and _williamsoni_ as concluded in the cytochrome b analysis by Patterson (2005) and Walker (personal communication). Figures 6 – 8 were created to answer these types of questions.

Figure 6a shows the relationships between the Californian populations that were sequenced for all SNPs (Lost Lake, Holcomb, and Fillmore) and the British Columbian population, LITC (Little Campbell River, Vancouver, Canada). This is an important comparison because the LITC population was found in the cytochrome b analysis to be closely related to _G. a. williamsoni_ and Shay Creek Stickleback. If these relationships are true, then two clades would be seen in the ingroup of Figure 6a, including a LITC/williamsoni/Lost Lake clade and a Holcomb/Fillmore clade.

This situation, however, is not seen. Almost all LITC populations branch away from the Californian populations, which causes a separate clade to be formed. It is also observed in this tree that Lost Lake is more closely related to Fillmore than to _G. a. williamsoni_. These relationships inferred from Figure 6a, however, are not supported by the bootstrapped tree shown (Figure 6b). A large polytomy is seen and, thus, the relationships found in Figure 6a are not statistically supported.
Figure 6a: Neighbor-Joining Tree of British Columbian and Californian Populations of Threespine Stickleback

Legend: See Figure 5a for abbreviation descriptions. The COND, WALL, and PAXB sequences were set as the outgroup. LITC1 was not included because it was so different from other sequences. This tree, though, was made with LITC1 and is reported in Appendix B- Figure A-8.

Figure 6b: Bootstrapped Neighbor-Joining Tree of British Columbian and Californian Populations of Threespine Stickleback

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen.
The tree in Figure 7a was made to determine the relationship of Shay Creek Stickleback to the other populations. This tree is similar to the one presented in Figure 6a, but differs in that LITC was not included. The branching pattern shows that Shay Creek appears to be similar to the Lost Lake/Fillmore clade found in Figure 6a. It may seem that LL23 is not included in this grouping, but the length of that branch is very short. Therefore, the conclusion can be made that the Shay Creek Stickleback seem to look like Lost Lake and Fillmore. The bootstrapped tree shown in 7b does show an interesting branch, in which Lost Lake, Fillmore, and Shay Creek samples all group together. However, the bootstrap value is low, which means that the branching patterns seen in Figure 7a are not statistically confident. Figure 7a also shows the Will 3 specimen to be grouped with the williamsoni population sequenced in Colosimo et al. (2005).

The tree in Figure 8a was made from the interesting observation that the Lost Lake population and the Friant population from Colosimo et al. (2005) do not actually group together on the tree in Figure 5a. Figure 8a shows that the Lost Lake fish branch away from the Friant Stickleback, which is not expected because these fish come from the same area. Furthermore, the bootstrapped tree (Figure 8b) actually shows a higher bootstrap value of 85% for the Lost Lake samples branching away from the others. This indicates confidence in this branching pattern. More trees were made based on this tree, in which divergent taxa were pruned from the tree (Appendix B- Figures A-10, A-11, and A-12). This allowed even higher bootstrap values to be observed.
Figure 7a: Neighbor-Joining Tree of Californian Populations of Threespine Stickleback

Legend: See Figure 5a and Table 5 for abbreviation descriptions. The tree was made with Shay Creek sequences containing only 13 SNPs (Locus H SNPs), while all other taxa had the typical 41 SNPs. The COND, WALL, and PAXB sequences were set as the outgroup.

Figure 7b: Bootstrapped Neighbor-Joining Tree of Californian Populations of Threespine Stickleback

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen.
Legend: See Figure 5a for abbreviation descriptions. The COND, WALL, and PAXB sequences were set as the outgroup.

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen.
Discussion

Past research of *G. aculeatus* phylogenetics at the University of Redlands has been completed by analyzing the mitochondrial cytochrome b gene. Results of this study showed that *G. a. williamsoni* and Shay Creek Stickleback are very closely related, with Lost Lake also being closely related to these two populations (Patterson, 2002; Walker, personal communication). Holcomb Creek Stickleback and Fillmore Stickleback were found to be in a separate clade. In addition, *williamsoni*, Lost Lake and Shay Creek Stickleback were found to be more related to British Columbian and Alaskan populations than to the Fillmore and Holcomb Creek Sticklebacks. A possible explanation for these results is that mitochondrial introgression occurred, which would skew the phylogenetic results. The research completed here sought to confirm or dispute these results by analyzing nuclear DNA.

Three nuclear loci were amplified and sequenced in this study in order to obtain data from 41 SNPs. Amplification of all three loci was successful from only six fish representing three populations (Holcomb Creek, Fillmore, and Lost Lake Stickleback). However, the sequence obtained from Fill 5 at Locus A was of such low quality that it could not be used. The sequence data for Will 4 was also unreadable in the electropherogram, which caused no sequence data to be obtained from this fish specimen. The probable reason for the low quality sequence data in these samples was that the amount of DNA in these sequencing reactions was low compared to the samples that could be sequenced. This fact is reflected by the concentrations of the isolated PCR product used in the unreadable sequencing reaction (4 ng/μL – 8 ng/μL) versus the accurate, readable sequencing reactions (8 ng/μL – 90 ng/μL). The low concentration of
isolated DNA would then cause a small amount of DNA to be present in the sequencing reaction.

In analyzing the sequence data, intrapopulation variation was found to occur at five different sites, including Locus A-site 373, Locus A-site 479, Locus B-site 283, Locus H-site 191, and Locus H-site 507 (Tables 3a and 3b). This variation was seen among three different populations and significantly contributed to the quality of the phylogenetic trees later produced. Three of the sites caused variation within the Lost Lake Stickleback, while the variation was also found in Fillmore at two sites and Shay Creek at one site. Variation at Locus A-site 373, Locus A-site 479, and Locus H-site 191 was not surprising because a variety of genotypes were seen in the Colosimo et al. (2005) populations in all three of these sites. For example, heterozygosity at Locus A-site 373 was very possible because the nucleotides for that site, thymine and cytosine, are both found in the populations analyzed by Colosimo et al. (2005). Furthermore, there are many sticklebacks in Colosimo et al. (2005) that show heterozygosity at this position.

The variation at the other two sites, Locus B-site 283 and Locus H-site 507, was less expected. The Lost Lake samples accounted for the variation at Locus B-site 283 by having unique genotypes. Out of the 91 stickleback specimens sequenced in Colosimo et al. (2005), 89 were homozygous for guanine at this position. Two were homozygous for adenine. The Lost Lake samples are unique in that LL22 is homozygous for adenine and LL23 is heterozygous with adenine and guanine present. Furthermore, this substitution was not present in any other Californian populations. Thus, it seems there has been an evolutionary substitution at this position. Sampling more Lost Lake Sticklebacks will confirm this unique result.
Locus H-site 507 produced very interesting results. The site was found to be trimorphic within a single population, which is uncommon for single nucleotide polymorphisms (SNPs). LL22 was heterozygous AT at the site, while LL23 was homozygous CC at the site. Other populations in Colosimo et al. (2005) had both thymine and cytosine present at this site, with genotypes of AA, AC, or CC. Thus, the presence of a thymine at this site is peculiar. However, the Japan Marine population studied in Colosimo et al. (2005) was also trimorphic for this site, with genotypes of CC, AC, and GG. This means all possible nucleotides have been observed at this SNP in comparing a variety of stickleback populations. Therefore, this SNP site seems to mutate at a particularly high rate.

This finding at Locus H-site 507 is reasonable, though, because different parts of a gene and different codon positions mutate at different rates (Graur & Li, 2000). Since all of the nuclear loci in this analysis are expressed in threespine sticklebacks, this site may be one that mutates at a high rate in a gene. It is possible that this site is the third position in a codon and these positions have a greater substitution rate than others in the codon (Graur & Li, 2000). These types of substitution are also often allowed because most cause synonymous substitutions. In an attempt to explore this possibility further, a possible open reading frame was found in this locus, in which site 507 corresponded to the third position in the codon. Thus, this hypothesis is highly possible. As the threespine stickleback genome becomes more annotated in the future, the definite position of this site should be determined to help explain this hypervariation.

Intrapopulation variation can also be seen in Table 4. In comparing the G. a. williamsoni sample sequenced in this study (Will 3) to the G. a. williamsoni samples
from Colosimo et al. (2005), it is observed that these sequences differ at four out of five possible variant sites. This is strange that there would be so much variation within one population. This result, however, cannot be considered very significant because there was only one fish used from *williamsoni* in this study and the phylogenetic trees produced did not show significant differences in these specimens. It is possible, though, that the two types of *williamsoni* were sampled from different places in the Santa Clara River and each location has a specific genotype. Further sampling could help determine the correct hypothesis. A similar result is seen in comparing the FRIL (Friant, California) sequence to the Lost Lake sequence obtained in this study. These two samples came from the same area, yet they differ at four out of seven possible SNPs.

Significant amounts of variation between populations were also found in examining the sequence data (Table 4 and Table 5). By comparing the Californian populations that were sequenced for all SNPs (Lost Lake, Fillmore, Holcomb Creek, and *G. a. williamsoni* reported in Colosimo et al. (2005)), it is seen that each pair of populations differs in sequence in at least three out of the seven variant sites. Lost Lake and *williamsoni* differ at all seven variant sites. These differences made it unfeasible to determine obvious similarities between the populations by simply observing the tables.

The sequence data was then used to produce phylogenetic trees (Figures 3–8). The first phylogenetic tree (Figure 3) was created using only the data from Colosimo et al. (2005). However, the subset of 41 SNPs sequenced in this study was used to create the tree, instead of all 193 SNPs. This tree could then be compared to the tree reported in Colosimo et al. (2005) to help determine the validity of using only a subset of SNPs to infer phylogeny (Figure 4). A comparison of Figure 3 and Figure 4 shows that the tree
made from the subset of SNPs has a large polytomy. This means that relationships of taxa within the polytomy can not be deduced with statistical confidence. This result, however, can be explained by the method in which the SNPs were chosen. The SNPs were chosen in order to tell differences between the Pacific North American populations. Thus, they were not picked to tell differences among the populations worldwide and the analysis should not be concluded to be invalid from the topology shown in this tree.

Different phylogenetic trees containing the Californian populations and the Pacific North American and Japanese populations were then made. In these trees, individual fish sequences were used instead of the population consensus sequences as completed in the tree shown in Figure 3. It was thought that having the individual fish on the tree would cause a more clearly defined branching pattern because less ambiguity codes would be used in the sequences. Ambiguity codes would only be used for heterozygosity in a single fish instead of any variation within a population. Individual fish sequences were also used to help determine the intrapopulation variation.

Another difference between Figure 3 and the new trees in Figures 5 – 8 was the type of tree-building method used. The tree shown in Figure 3 was made using maximum likelihood analysis. Bootstrapping was not completed on this tree because it could take weeks to months for this data to be collected. Figures 5 – 8 were made using the neighbor-joining method, which can build and bootstrap a tree in a matter of seconds. It would be preferable to use maximum likelihood or Bayesian analysis for these trees because these are both highly regarded methods, but time restraints again limited this possibility. Parsimony was not used to create phylogenetic trees because too many
equally parsimonious trees could be produced, which would cause more polytomies to be seen.

The tree shown in Figure 5a was first made and contains the Californian and Pacific clade populations from Colosimo et al. (2005). The bootstrapped tree, however, shows polytomies that cause the branching patterns seen in Figure 5a to not be statistically significant. Due to the reoccurring problem of the polytomies seen in the phylogenetic analyses, trees with fewer taxa were produced in order to determine specific phylogenetic questions. It was thought that these trees might be able to produce higher bootstrap values and, thus, cause less polytomies to be seen. These phylogenetic trees are shown in Figures 6a, 7a, and 8a.

Figure 6a is useful because it tests the conclusion found by Patterson (2002) and Walker (personal communication) that williamsoni, Shay Creek, and Lost Lake are more closely related to British Columbian stickleback than to Fillmore or Holcomb. This is possible because LITC seen in Figure 6 corresponds to one of same British Columbian populations analyzed in the cytochrome b study (Patterson, 2002; Walker, personal communication). By examining the tree in Figure 6, it is seen that the results obtained in this study are different from those found in the cytochrome b study in that LITC actually branches away from the Californian populations. Thus, the Californian populations are shown to be derived separately from the British Columbian population.

Figures 6a and 7a were all useful in testing the relationships of Californian populations to each other. It was found in the cytochrome b study that Fillmore and Holcomb grouped together while williamsoni, Shay Creek and Lost Lake grouped together (Patterson, 2002; Walker, personal communication). The results from the study
reported here, though, show that the populations are grouped into different clades than previously found. Figure 6a shows that two clades were still found, however one contains Lost Lake and Fillmore and the other clade contains Holcomb and *williamsoni*. These relationships are also supported in Figure 7a. Figure 7a also shows that Shay Creek Stickleback group with the Lost Lake and Fillmore clade. Thus, the results obtained from this study using nDNA are different from the results obtained in the studies using mtDNA to analyze stickleback phylogenetics.

The branching patterns of the trees in Figures 6a and 7a can not be thought of as highly supported. In these trees, the bootstrapped tree produced a large polytomy. Some branching patterns may be seen in the bootstrapped trees, but the bootstrap values are so low that the branches are still not highly supported. In fact, the only truly supported bootstrap values seen in these trees are the values distinguishing the ingroup taxa from the outgroup taxa. For example, in Figure 6b, there is a 96% bootstrap value distinguishing WALL, PAXB, and COND samples as in a distinct clade, which is about the cut-off of 90% for an “A” grade.

Figures 6a and 8a both show how the Lost Lake fish obtained by James Malcolm differ from the Friant population reported in Colosimo *et al.* (2005). These fish were both sampled in the vicinity of Lost Lake Recreation Area near Friant, California. Furthermore, from descriptions of the sampling locations, these samples could not have been taken more than two miles apart from each other (Colosimo *et al.*, 2005; James Malcolm, personal communication). Therefore, it is strange that these populations would be so distant from each other on the tree. Figure 8b even shows some statistical confidence with a high bootstrap value on the branch separating the Lost Lake fish from
the others. Taking the most divergent taxa off of the tree in Figure 8a produces even higher bootstrap values (Appendix B- Figures A-10, A-11and A-12). Thus, it seems that there is considerable variation within the Lost Lake/Friant population. This result suggests that Lost Lake/Friant may be a very unique population that is different from other stickleback populations because the variation is so great in this area. Future work will help determine the true variation of the Lost Lake population, which could lead to extending conservation efforts to this possibly unique stickleback population.

Similar to the Lost Lake/Friant results, the populations of *G. a. williamsoni* assayed in both studies appeared to have considerable variation from each other as shown by the results in Table 4. However, the results in the phylogenetic tree in Figure 7a showed that Will 3, sequenced in this study, grouped with the *williamsoni* samples from Colosimo *et al.* (2005). Considerable variation within this population, though, can not be ruled out because of the small sample size sequenced in this study. Further sampling of the *williamsoni* population will help resolve its true amount of variation.

Although the statistical significance of most data obtained in this study may be low, it is still interesting to compare the major findings of this study to the conclusions made in the cytochrome b analysis. Patterson (2002) found that Shay Creek Stickleback and *williamsoni* were more closely related to British Columbian and Alaskan populations than to the Fillmore and Holcomb Creek Stickleback. Walker (2004) found that Lost Lake Stickleback were also closely related to *williamsoni* and Shay Creek Stickleback. The results from the nDNA study, however, indicate that Lost Lake, Shay Creek, and Fillmore group together on the tree while *williamsoni* and Holcomb are in a separate clade. In addition, the study by Colosimo *et al.* (2005) sampled a British Columbian
population (LITC) that Patterson (2002) found to be closely related to williamsoni and Shay Creek. The results showing Californian population phylogeny based on nDNA and Colosimo et al. (2005), however, indicate that williamsoni, Shay Creek, and this British Columbian population are not as closely related as previously found. Thus, the results from the studies based on mitochondrial DNA differ from those based on nuclear DNA. All of these results may be explained by considering mitochondrial introgression.

To consider the possibility of mitochondrial introgression occurring, it must first be assumed that the results from both types of study (mtDNA vs. nDNA) are correct for each part of the genome. To explain how the mtDNA results could occur, it is possible that females from British Columbian populations migrated south to areas like Lost Lake. Patterson (2002) proposed that populations might migrate from British Columbia to California via streams created from glacial melting. Lost Lake could have already been inhabited by a separately derived group of stickleback. Therefore, these two populations would have distinct genomic differences. Mitochondrial introgression could occur in which the British Columbian females mated with the Lost Lake males to create progeny with a hybrid nuclear genome and a British Columbian mitochondrial genome (since mtDNA is inherited maternally). If subsequent backcrossing of the female progeny to the parental Lost Lake males occurred for several generations, then it is possible that the Lost Lake Stickleback would retain their nuclear genome, but now have a British Columbian mitochondrial genome. This also assumes that an evolutionary bottleneck occurs in which the British Columbian mitochondrial genome becomes fixed into the populations. This situation may seem unlikely, but mitochondrial introgression in fish has been reported by many studies (For example, Bagley & Gall, 1998; Bernatchez et al., 1995;
Chow & Kishino, 1995; Sullivan et al., 2004; Takahashi et al., 2003), including in threespine stickleback (Yamada et al., 2001).

The same introgression process could have also occurred with British Columbian females and williamsoni males. This would create a Southern Californian population containing British Columbian mtDNA. This mitochondrial genome would be the same as the one fixed into the Lost Lake population. However, the nuclear genomes would be distinct, which would account for the results found in this study. In addition to this introgression event, others would have also had to occur to explain the relationships found in the two studies. Thus, the limitation in using this hypothesis to explain the data is that multiple introgression events must have occurred, which may be unlikely.

Extensive introgression could be a possibility, though, as seen in many species of fish (For example, Bernatchez et al., 1995; Chow & Kishino, 1995).

The confidence of the data obtained must be taken into serious consideration, though, before major conclusions can be made about threespine stickleback phylogenetics. The low bootstrap values seen in the figures are fundamentally caused by the quality of the data set used to build the trees. Bootstrap values lower than 50% for a branch were shown as a polytomy, which are phylogenetically useless for determining relationships. This quality of the data set returns to the observation of the high amounts of intrapopulation variation and interpopulation variation as seen in Table 4 and Table 5, respectively. In completing a phylogenetic analysis, it is necessary for the intrapopulation variation to be less than the interpopulation variation. The reason is that if a population varies too much within itself and is compared to another population, than any differences in the sequences could be attributed to possible interpopulation
variation or to the wide range of variation within a single population. Thus, for a quality phylogenetic analysis to be completed, the intrapopulation variation must be small and the interpopulation variation be greater so that phylogenetic relationships could be deduced.

This problem of excess intrapopulation variation is also seen in the tree from Colosimo et al. (2005). It is not obvious from first observation of the tree, but the overall reliability of tree in this publication is actually ambiguous because the posterior probabilities were not reported for almost all of the branches. Posterior probabilities are reported on trees made using Bayesian analysis and can be thought of like bootstrap values, with high values (0.9 to 1.0) being very significant. Therefore, the major problem that was realized with this tree is that it shows a bifurcating (splits into two branches at a node), ideal branching pattern, but the statistical confidence of this branching pattern is not known. The authors only printed the posterior probabilities that were greater than 0.95, which is a high standard, but the other values should be present to show the reliability of certain branches.

Examination of the Colosimo et al. (2005) tree shown in Figure 4 could then help explain the poor bootstrap values seen Figures 5 – 8. In this study of Californian threespine stickleback populations, the Colosimo et al. (2005) tree is only important for the portion of the tree relating the williamsoni population with the LITC population (from British Coloumbia) and the Friant, California population (Figure 4). The reason is that all three of these populations were surveyed in the cytochrome b analysis and so comparison of this section of the tree to past results could help explain the true relationship in threespine stickleback (Patterson, 2002; Orti, 1994). Figure 4 shows LITC and the Friant
population (FRIL) to be more closely related to each other than either is to *williamsoni*. This is an interesting relationship to note, but the statistical confidence of this branching pattern is not shown on the tree which means the reliability of this branch is unknown. Thus, the authors of this paper could not determine the relationships of some of the key populations of interest using the complete SNP dataset.

This lack of clarity in the Colosimo *et al.* (2005) tree has important consequences for future work. There is a possibility that using the SNP data to create phylogenetic trees of Californian populations may never show statistically significant results because the confidence of the LITC, Friant, *williamsoni* branching pattern on this tree. The posterior probabilities may actually be fairly high, which provides hope that the Californian population phylogeny could be found using this set of SNPs. However, if the values are low then it is possible that this protocol may not actually yield the phylogenetic relationships of the Californian populations with statistical confidence. Thus, future work should be done to remake the tree using the same methods as reported in Colosimo *et al.* (2005) in order to determine the true reliability of the tree.

Due to the low confidence of the data obtained, much future work must be completed to determine the true phylogeny of the threespine stickleback. Work may be continued in creating phylogenetic trees with the data obtained using preferred methods such as maximum likelihood and Bayesian analysis. More fish could also be sampled and more SNPs could be sequenced from those reported in Colosimo *et al.* (2005). However, the potential usefulness of using the Colosimo *et al.* (2005) protocol to determine Southern Californian population phylogenetics may be limited as explained above.
More analysis can also be completed on the data obtained from the cytochrome b project. Yamada et al. (2001) completed different types of mathematical analyses on data obtained from sequencing a mitochondrial gene in three-spine stickleback to help determine that mitochondrial introgression occurred. This included finding nucleotide diversity and haplotype diversity values for populations and the number of effective female migrants between populations. A similar analysis could be completed on the data obtained in the cytochrome b project. However, smaller sample sizes in this study may limit the usefulness of the conclusions found from this type of analysis.

Future work in creating a phylogeny based on nuclear DNA could also be shifted from the current protocol adapted from Colosimo et al. (2005). Specifically, microsatellites have been used in three-spine stickleback phylogenetics and may prove useful in finding the relationships of the Southern Californian populations (Takamura & Mori, 2005). Furthermore, much research has been completed to find many microsatellite sites in the stickleback genome that could be useful in such analyses (Largiader et al., 1999; Rico et al., 1993; Taylor, 1998)

Although the purpose of this study was to elucidate the true phylogeny of the Southern Californian G. aculeatus populations, further study must be completed to verify the results. Currently, the data obtained from nuclear DNA point to different relationships than those found using mitochondrial DNA. In addition, the results from this study indicate that mitochondrial introgression may have occurred, which would alter the conclusions made in the cytochrome b study. Finally, the comparison of Lost Lake and Friant samples shows considerable variation at this location. Further study can help resolve the true uniqueness of Lost Lake Stickleback. Shay Creek Stickleback and/or
Lost Lake Stickleback could then be extended the same conservational protection that *G. a. williamsoni* receives on the endangered species list.
References


Acknowledgements

Dr. Linda Silveira

Dr. James Malcolm

Dr. Angela Burk-Herrick

Dr. Lisa Olson

Dr. Caryl Forristall

Dr. Dan Wacks

University of Redlands Biology Department

Jeremy Schmutz and David Kingsley of Stanford University

Jocelyn Buzzas
Appendix A: Creating a Maximum Parsimony Tree

In this example, the four taxa being compared are human, dog, cat, and threespine stickleback. The same DNA locus is sequenced from each species and the results are compared. Table A-1 below shows hypothetical results:

<table>
<thead>
<tr>
<th>Species</th>
<th>Nucleotide #1</th>
<th>Nucleotide #2</th>
<th>Nucleotide #3</th>
<th>Nucleotide #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Dog</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>Cat</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>Stickleback</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
</tbody>
</table>

The basic idea in parsimony is that the best tree is the one that explains the data in the fewest number of steps or mutations. In this type of tree-building method, sites that are variant and informative are used. Invariant sites do not show differences between the species and are, thus, unusable in parsimony (for example, nucleotide #1). Nucleotide 2 and 3 are both variant, but uninformative so they are not included in the analysis. To illustrate how this is possible, Figure A-1 and A-2 show the three possible unrooted trees for each site and how each requires the same number of mutations to explain the relationship.

Figure A-1: Possible Unrooted Trees to Explain Nucleotide #2

Legend: Possible unrooted trees are shown in black. The blue letters next to species names are the DNA bases for each species at this site (found in Table A-1). The black bases in parentheses are the nucleotides the common ancestor possessed. The blue lines shows where the substitutions have to occur in the tree for the branching patterns to be possible. The number of steps or mutations is also shown for each tree.
It is seen from Figure A-1 and Figure A-2 that sites #2 and #3 are uninformative because an equal number of substitutions is seen in all trees. Nucleotide site #4, however, is informative. In this case, one of the unrooted trees has a smaller number of mutations to explain the relationship (Figure A-3).

Once the informative sites are found in the tree, the number of steps required to build each tree at these sites are summed. The most parsimonious tree is the tree with the smallest number of steps to explain the data.
Figure A-3: Possible Unrooted Trees to Explain Nucleotide #4

Legend: See Figure A-1.
Appendix B: Additional Phylogenetic Trees

This section includes a variety of other phylogenetic trees that are interesting to note. Figures A-4, A-5, A-6, and A-7 show the maximum likelihood trees used to create the consensus tree seen in Figure 3. Figures A-8a and A-8b show the trees from Figure 6a and 6b with LITC1 included in the dataset. Figures A-9a and A-9b show a tree that contains the same ingroup taxa used in the phylogenetic analysis by Patterson (2002). Figures A-10, A-11, and A-12 all show trees derived from Figure 8. Each of these trees shows the statistical confidence of the Lost Lake branch getting larger as the most divergent taxa of the trees are pruned from the phylogenetic analysis.
Figure A-4: First Maximum Likelihood Tree of World Populations of Threespine Stickleback

Figure A-5: Second Maximum Likelihood Tree of World Populations of Threespine Stickleback

Legend: See Figure 3 for abbreviation descriptions.
Figure A-6: Third Maximum Likelihood Tree of World Populations of Threespine Stickleback

Figure A-7: Fourth Maximum Likelihood Tree of World Populations of Threespine Stickleback

Legend: See Figure 3 for abbreviation descriptions.
Figure A-8a: Neighbor-Joining Tree of British Columbian and Californian Populations of Threespine Stickleback with LITC1

Legend: See Figure 5a for abbreviation descriptions. The COND, WALL, and PAXB sequences were set as the outgroup.

Figure A-8b: Bootstrapped Neighbor-Joining Tree of British Columbian and Californian Populations of Threespine Stickleback with LITC1

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen.
Figure A-9a: Neighbor-Joining Tree of Southern Californian Populations

Legend: See Figure 7a.

Legend: 0.01 substitutions/site

Figure A-9b: Bootstrapped Neighbor-Joining Tree of Southern Californian Populations

Bootstrap

Legend: See Figure 5a for abbreviation descriptions. Shay refers to the Shay Creek Stickleback. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen.
Figure A-10a: Neighbor-Joining Tree Showing Lost Lake/Friant Population Variation without FRIL4

Legend: See Figure 5a for abbreviation descriptions. The COND, WALL, and PAXB sequences were set as the outgroup.

Figure A-10b: Bootstrapped Neighbor-Joining Tree Showing Lost Lake/Friant Population Variation without FRIL4

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen. Note the bootstrap value of 90% separating Lost Lake specimens from the other taxa.
Figure A-11a: Neighbor-Joining Tree Showing Lost Lake/Friant Population Variation without FRIL 3 and FRIL4

Legend: See Figure 5a for abbreviation descriptions. The COND, WALL, and PAXB sequences were set as the outgroup.

Figure A-11b: Bootstrapped Neighbor-Joining Tree Showing Lost Lake/Friant Population Variation without FRIL 3 and FRIL4

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen. Note the bootstrap value of 92% separating Lost Lake specimens from the other taxa.
Figure A-12a: Neighbor-Joining Tree Showing Lost Lake/Friant Population Variation without FRIL 3, FRIL4, and WMS04

Legend: See Figure 5a for abbreviation descriptions. The COND, WALL, and PAXB sequences were set as the outgroup.

Figure A-12b: Bootstrapped Neighbor-Joining Tree Showing Lost Lake/Friant Population Variation without FRIL 3, FRIL4, and WMS04

Legend: See Figure 5a for abbreviation descriptions. All bootstrap values over 50% are shown, while those less than 50% result in the polytomies seen. Note the bootstrap value of 94% separating Lost Lake specimens from the other taxa.