Laura Lee

12/9/2011

Advisor: Dr. David Carlini

University Honors in Biology

Fall 2011

A Comparative Genetic Analysis of Three Norwegian *Gammarus lacustris* Populations

Abstract

A comparative genetic analysis of three Norwegian populations of *Gammarus lacustris* was undertaken in order to determine whether or not these populations were representatives of the same species despite the morphological variations found in the cave dwelling population. Two of the three populations occurred in surface freshwater lakes – Lake Lille Lauarvatn (LL) and Lake Ulvenvannet (UL) – and the third occurred in the subterranean freshwater in Lower Sandågrotta Cave (GG) and showed classic troglomorphic traits. This was accomplished through amplification and sequencing of the cytochrome c oxidase 1 (CO1) region of the mitochondrial DNA. Based on phylogenetic trees constructed using the maximum likelihood and maximum parsimony methods, as well as two species screening threshold calculations (SST) from Jukes-Cantor distances it was possible to conclude that these three populations were all indeed representatives of *G. lacustris*. Additionally, it was possible to infer hydrological linkage between the UL and GG populations based on the low level of divergence between the two populations. This also demonstrated morphological variation independent of sequence divergence. As such, these populations of *G. lacustris* may successfully serve as a model for the study of adaptive evolution.

Introduction

A comparative genetic analysis of three Norwegian populations of *Gammarus lacustris* was performed in order to contribute to the knowledge of the genetic structure of this species. Among the three populations sampled, two were surface dwelling populations that were potentially hydrologically distinct from one another. These occurred in two lakes: Lake Lille Lauarvatn and Lake Ulvenvannet. The third population occurs in a subterranean body of water in The Lower Sandågrotta Cave. This third population showed significant morphological variation from the two previous populations, occurring in the form of reduced eyes, fewer ommatidia, larger body size, longer antennae, and reduced pigmentation. The two lake populations share morphology (Kjartan Østbye Personal Communication 2011). Initially it was uncertain whether this cave population was a subterranean population of *G. lacustris* or whether it was actually some other amphipod crustacean species. Determining whether or not this third population was indeed *G. lacustris* based on genetic analysis was necessary to determine if these populations could serve as a system to study speciation. This was in part intended to help identify a model that can be used to study the nature of intraspecific variation, speciation, and adaptive evolution as it occurs.

G. lacustris has been used in previous studies of intraspecific variation in part because of its wide range and patchy distribution (Meyran 1998). Populations of *G. lacustris* occur in freshwater mountain lakes in France, central Europe, Italy, and Scandinavia. Often lakes containing *G. lacustris* contain genetically isolated populations, and each of these populations is subjected to widely differing environmental conditions and subsequent selection pressures (Meyran 1998). *G. lacustris* has been studied before because of this property through amplification and sequencing of the cytochrome oxidase subunit I (COI) region of the mitochondrial DNA.

Gammarus minus, a related species, is a eutroglophile. This means it has been shown to form subterranean populations with classic troglomorphic morphological variations without undergoing reproductive isolation (Carlini et al. 2009). These troglomorphic adaptations include reduced eyes, fewer ommatidia, larger body size, longer antennae, and reduced pigmentation. It has a similarly wide range and patchy distribution in North America. Based on COI sequence data from individuals sampled from five cave and ten surface populations of *G. minus*, it was found that hydrologically linked populations were more genetically similar than morphologically similar populations, which inhabited distinct water basins (Carlini et al. 2009). If the Lower Sandågrotta Cave population is identified as a *G. lacustris* population through COI sequence data, this can serve as a similar system with which to study adaptive evolution. Models such as this, in which morphological variation does not coincide with distinct genetic variation are rare and therefore present a valuable research opportunity.

Identification of representatives of the genus *Gammarus* has proven difficult in the past due to several factors. There is often a high degree of morphological variation between different species within the genus and different individuals within a species. This is exacerbated by the high degree of ecological diversity between *Gammarus* populations (Hou et al. 2009). The difficulty of morphological identification in a cryptic genus such as this necessitates the use of alternative methods of identification such as COI Barcoding. COI is an optimum region of DNA for a study of this kind for several reasons. The COI region is widely accepted as the region of DNA that is most useful in identification of species. It is well understood and the flanking regions of the COI gene are highly conserved among species (Hebert 2003; Kress 2008). This allows for the same or similar primers to be used to amplify this region in many species. Additionally, the high rate of 3rd position base substitution allows for a greater rate of molecular

evolution than other easily amplified regions of DNA such as the 12s or 16s rDNA regions. The rapid rate of molecular evolution in this region makes it ideal for determining distinction between species as well in addition to identifying conspecifics. COI is also a short enough region of DNA that it can be easily amplified and analyzed. Choosing this widely used region may also allow for comparison with other studies of *G. lacustris* in which COI was used more easily than if another region were chosen (Hebert 2003; Kress 2008). In studies such as this, in which only a few samples are available over a small geographic distance it is most useful to set the species screening threshold (SST) as ten times the average intrapopulation COI sequence divergence (Witt et al. 2006).

Methods

The live samples were obtained during July and August 2010. Cave samples were obtained using a torch lamp and a 15 cm by 10 cm dip net with small mesh size. To collect samples from the lake-dwelling populations a larger 30 cm by 30 cm net was used with small mesh size. The individuals to be sampled were excited out of hiding by the researchers, who would kick along the lake bottom as they collected samples. Roughly thirty individuals were sampled from each population. Lake Lille Lauarvatn has coordinates +59° 32' 57.76", +9° 38' 49.09", while Lower Sandågrotta Cave has coordinates +59° 32' 34.62", +9° 38' 59.27". These are the two hydrologically linked populations. The third population, Lake Ulvenvannet, has coordinates +59° 48' 55.11", +10° 20' 58.52". The map with the locations of these populations, created using Google Earth (Fig. 1). The specimens were examined for morphological variation and then preserved in 100% ethanol.

Figure 1. This map of the populations' locations was generated in Google Earth. The Distance between Lower Sandågrotta (pink) Cave and Lake Lille Lauarvatan (yellow) is roughly 0.734km, also calculated in Google Earth. The distance between these to populations and the Lake Ulvenvannet (blue) population is approximately 49.5km. Distances were also calculated in Google Earth.

For molecular analysis, The DNA was extracted from the samples with a DNeasy Blood and Tissue Kit from QIAGEN. Once extracted, genomic DNA was stored frozen. The PCR protocol was optimized for the amplification of these samples, based on early attempts in which the PCR product failed to run out on the gel (Fig. 2) indicating the mitochondrial sequence of interest had failed to properly amplify.

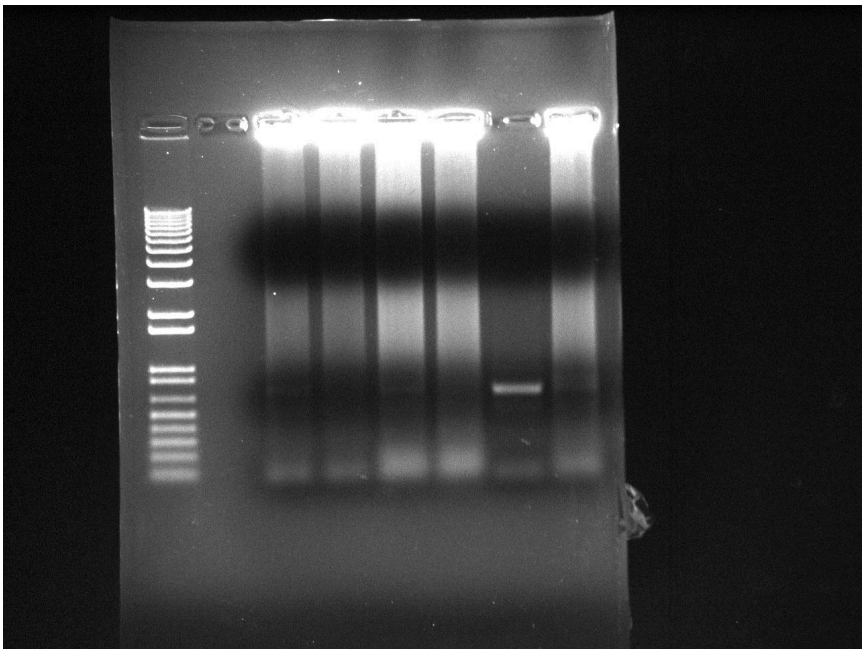

Figure 2. As shown here illuminated by ultraviolet light a large proportion of PCR product remained in the wells after application of the current and the PCR product that did travel from the wells failed to separate into distinct bands.

The process was optimized by simultaneously varying the concentration of genomic DNA used as the template for PCR and the annealing temperature. This was done for two samples: LL1 and GG2. Dilutions were made of each sample at 0.1x, 0.2x, 0.5x, and 1x. Annealing samples were tested at 45.3ºC, 47.1ºC, 50ºC, 53.3ºC, and 56.1C. The PCR products of each sample were run out in a 0.8% agarose gel in volumes of 5 µl. Varying both genomic DNA concentration and annealing temperature in this way produced 40 PCR products compared within this gel. It was found that a concentration of 0.1x and an annealing temperature of 45ºC was the optimum condition of PCR of these samples (Fig. 3). This optimized PCR protocol was employed to amplify all sequences with improved (Fig. 4), but still variable success.

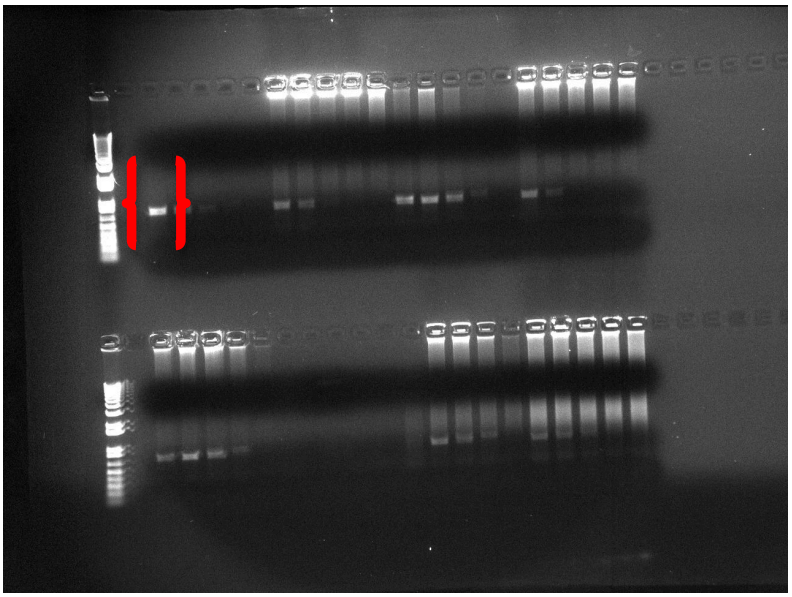

Figure 3. The two-combed gel includes all possible dilution and temperature combinations for two samples. The bracketed band is the most clearly separated with the least PCR product remaining in the well and was chosen as representative of the optimum conditions for amplification.

In order to amplify the COI region of the mitochondrial DNA, 1 µl of genomic DNA was diluted with 9 µl of PCR water to produce a 0.1x dilution. 1 µl of this diluted genomic DNA was then used as the template for a 25 µl PCR reaction with an annealing temperature of 45ºC. The amplified COI region was then isolated by running the PCR product out in an 0.8% agarose gel.

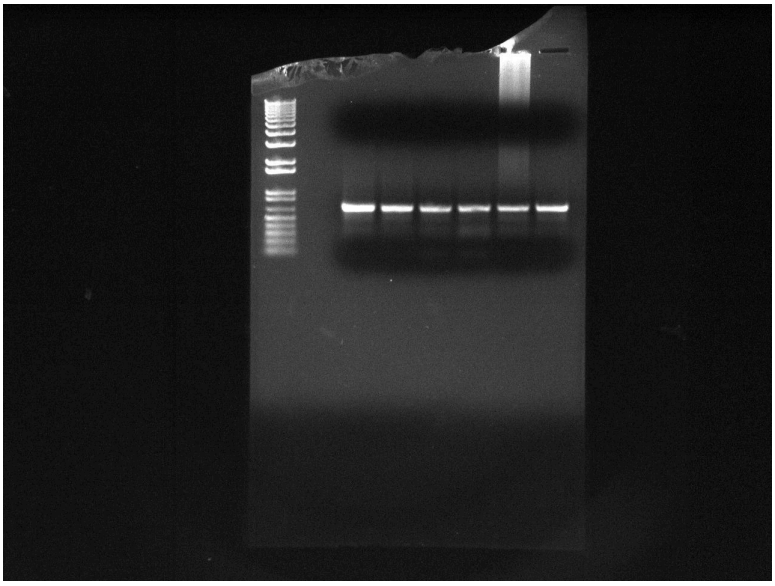

Figure 4. A gel run with PCR product amplified using the optimized PCR protocol is shown. Lanes 3-6 and 8 show distinct bands. Though the wells have been broken off for lanes 1-6, lane 8 shows the minimal PCR product remaining after exposure to the electrical field.

The bands produced during gel electrophoresis containing the COI region were viewed under ultraviolet light and excised from the gel. DNA was purified from the excised bands using a MinElute Gel Extraction Kit from QIAGEN. Once extracted, the concentration of DNA was determined using 2 µl of this final product with a Nanovue microspectrophotometer. The concentration was obtained with the Nanovue three times, with the average of the three taken as the final accepted concentration. The concentration of DNA was visually estimated by running the same 2 µl of the final product in an agarose gel. Based on a visual estimate of the gel, the concentrations of DNA in the purified PCR product were likely to be lower but still sufficient for sequencing. Once enough samples were amplified they were sent to Genewiz, Inc., a commercial DNA sequencing service.

After amplification and gel purification of all samples 22 sequences were successfully obtained. Sequences were obtained in two directions for each individual sample using a sequencing primer from each end. Final sequences were obtained by aligning both the forward

and reverse sequences to produce consensus sequences. Of these sequences, seven were from the GG population, ten were from the LL population, and the remaining five were from the UL population. No gaps were found in the sequences and allowing them to be aligned manually to produce a multiple sequence alignment. The final COI consensus sequences were 658 base pairs long. A multiple sequence alignment was generated using a *G. lacustris* sequence obtained from BLAST (Altschul et al. 1990) from Yellowstone National Park, referred to here as Yellowstone, as the outgroup. The likelihood settings, based on the Hasegawa, Kishino and Yano model (Hasegawa et al. 1985), were determined in Modeltest (Posada & Crandall, 1998). Each tree was generated in PAUP (Swofford 2003), using the 22 sequences obtained from the three populations of interest. Each nodal value represents a bootstrap value generated using 100 replicates (Felsenstein 1985). The trees were then formatted in Mesquite (Maddison & Maddison 2007).

Results and Analysis

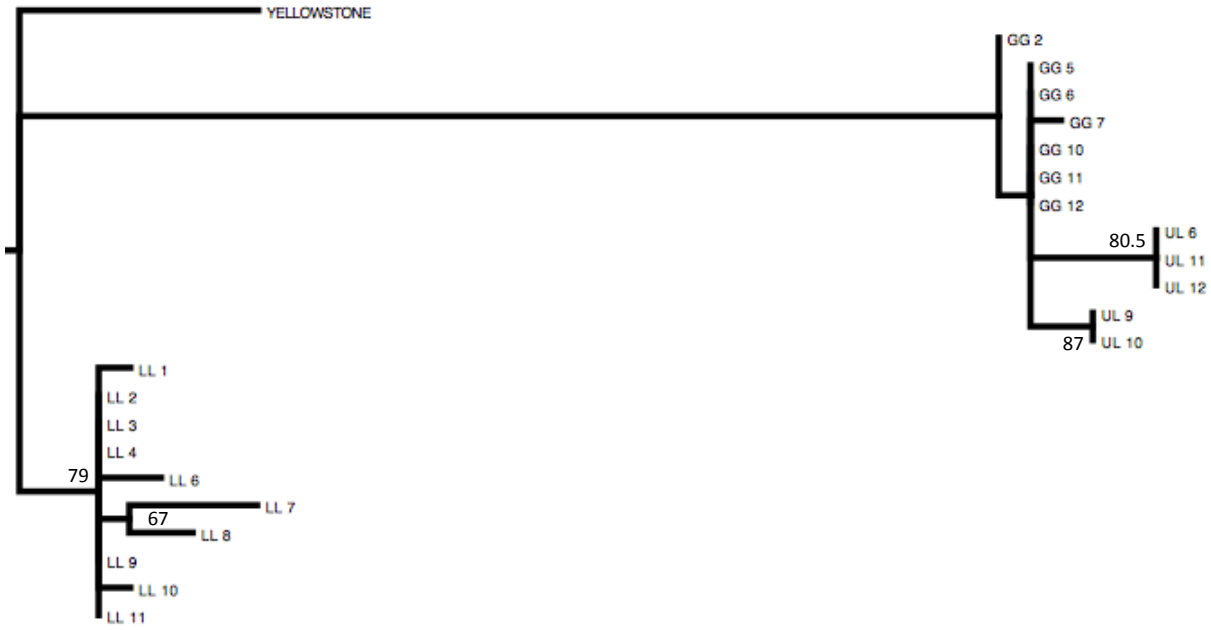

Figure 5. Maximum Likelihood Tree with bootstrap values generated in PAUP. This tree was generated using 22 CO1 sequences from the populations of interest with a sequence retrieved from BLAST (a *G. lacustris* sequence from Yellowstone National Park). Nodal values are bootstrap support values generated from 100 replicates. Branch lengths are shown proportional to distance.

Of 658 positions, 46 were variable across all sequences and 37 of those sites were parsimony informative. The average base frequencies were A=0.22710, C=0.22110, G=0.21150, and T=0.34030. The transition to transversion ratio was 3.2271 with a κ value equal to 6.4459982. Both the Maximum Likelihood tree and the Maximum Parsimony Tree share identical topology. The LL population is clustered as a monophyletic clade in both models with reasonably well supported bootstrap values: 74% in the likelihood tree (Fig. 5) and 96% in the parsimony tree (Fig. 6).

Both trees show the UL and GG populations most closely associated with each other. Other than this difference each tree clusters the LL and GG populations monophyletically with high bootstrap support. Each tree also clusters the GG and UL populations, while the LL population exists as a separate clade. Based on the geographic locations of the three populations (Fig. 1) it was expected that the LL and GG populations would show the least interpopulation divergence. Alternatively, the possibility that LL and UL populations would show the least interpopulation divergence was considered. This is due to the morphological similarity that the GG cave population does not share with the other two. Given the results of the phylogenetic analysis it may be possible that the GG and UL populations experience more interpopulation interaction

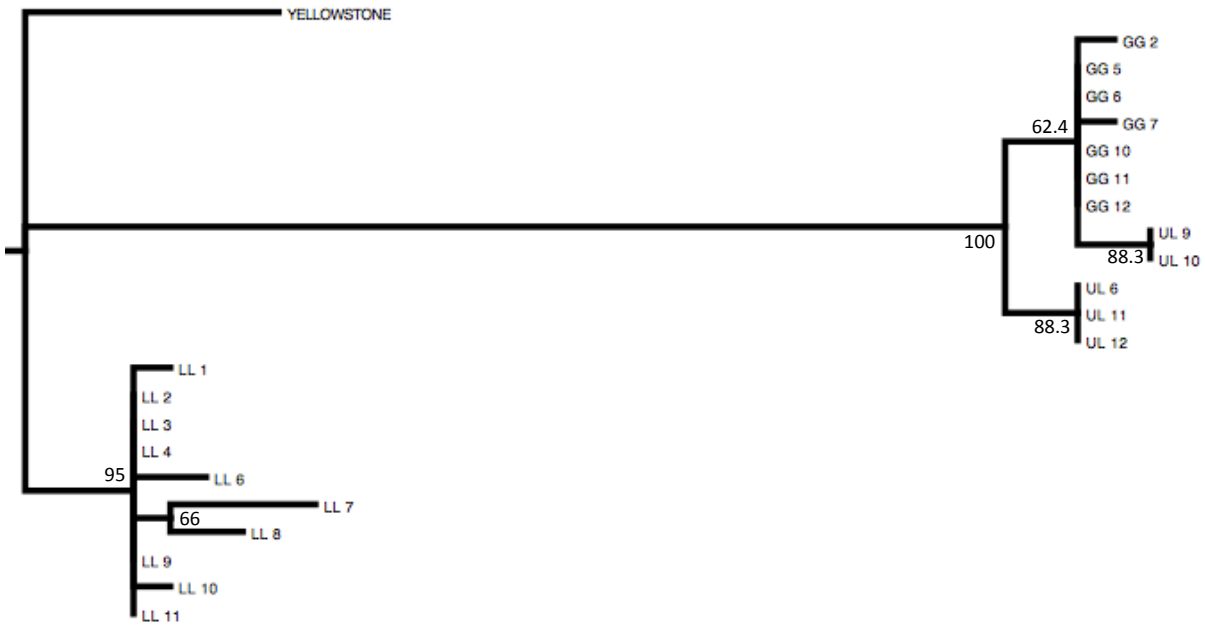

Figure 6. Maximum parsimony tree with bootstrap values generated in PAUP. This tree was generated using 22 CO1 sequences from the populations of interest with a sequence retrieved from BLAST (a *G. lacustris* sequence from Yellowstone National Park). Nodal values are bootstrap support values generated from 100 replicates. Branch lengths are shown proportional to distance.

than the LL and GG populations despite their relative geographic proximities. The intrapopulation distance for GG is much lower than that of the other two populations, 0.09±0.02% for GG versus 0.36±0.05% for LL or 0.5±0.1% for UL (Table 1). Such low intrapopulation genetic variation may be indicative of a founder effect. Perhaps the GG population was founded by a representative of the UL population at some point previously, given the low level of interpopulation distance between the GG and UL population as opposed to the distance between the GG and LL populations, 0.53±0.03% versus 5.03±0.03%. This also suggests that Lake Ulvenvannet and The Lower Sandågrotta Cave are part of the same water basin.

Mean Distance Between Populations				Distance Within Populations			Average Intrapopulation Distance	Species Screening Threshold
Population 1	Population 2	Dist	SE	Populations	Dist	SE		
GG	LL	5.03	0.03	GG	0.09	0.02	0.31±0.04	3.1±0.4
GG	UL	0.53	0.03	LL	0.36	0.05		
LL	UL	5.13	0.04	UL	0.5	0.1		

Table 1. The mean distance between groups and within groups are shown along with standard error. Analyses were conducted using the Jukes-Cantor model. Values are given as percentages. The SST was calculated by obtaining the average intrapopulation distance and multiplying by ten.

Polytomies are unresolved within populations, reflecting the low level of intrapopulation variation. However more information concerning the relationships among and between populations can be gathered by examining the net and mean distances between populations and the distances within populations (Table 1) calculated using the Jukes-Cantor model (Jukes & Cantor 1969). The Jukes-Cantor model was employed because it corrects for multiple mutational hits and takes account for any potential back mutations to the ancestral sequence. This provides a more accurate determination of distance than counting the observed number of nucleotide differences (Templeton 2006). Here the average intrapopulation distance is equal to 0.31±0.04%. Therefore the SST is equal to 3.1±0.4% (Table 1). Worldwide *G. lacustris* intraspecific distance in the COI region averages at 1.6%±0.3% with a maximum distance of 3.5% (Hou et al. 2009). This would allow for a divergence threshold of 16±3.0%.

Discussion

The distance between populations is clearly supported by both the maximum likelihood and the maximum parsimony tree (Fig. 5, Fig. 6). Each tree shows the same topology, with a highly supported UL and GG monophyletic clade as well as an LL monophyletic clade. Additionally, within the UL population each tree shows a highly supported node grouping UL 9 and UL 10 as

well as a highly supported node grouping UL 6, 11, and 12. Both trees also resolved a separate node within the monophyletic LL clade grouping LL 6 and LL 7. That these topologies are reproduced in both trees suggests they reflect the true topological relationship between these populations. However the clustering of the UL and GG populations was unexpected based on the geographic distances between populations (Fig. 1). It would have been reasonable to expect a clustering of the LL and GG populations given the distance between the two populations was less than one kilometer, while the distance between the UL and GG populations was nearly fifty kilometers. Despite this unexpected result, each interpopulation distance does not exceed the distance expected for conspecifics. Additionally, the straight line distance between each population may not actually reflect the intricacies of the underground hydrological network. The existence of species in which great morphological deviation occurs without concordant genetic deviation provides a unique opportunity for the study of adaptive evolution. In the absence of distinct morphological standards used to define the species, COI barcoding is a well-supported method for determining species membership. It appears, based on COI sequence data, that these three populations are all representatives of *Gammarus lacustris*. Different conclusions concerning the identification of these three populations as representatives of *Gammarus lacustris* may be reached based on the method used for determination of a divergence threshold. Because these sequences represent a small geographic sample it is reasonable to calculate the SST value as equal to ten times the mean intrapopulation divergence as opposed to the mean intraspecific divergence, which is the figure generally employed as the divergence threshold (Witt et al. 2006). Based on the SST calculated using only local *G. lacustris* sequence divergence it appears that these three populations should not be members of the same species. The average interpopulation distance between GG and LL as well as that between LL and UL exceed the

SST. However the average interpopulation distance between GG and UL is under the SST (Table 1). This suggests GG and UL are representative populations of the same species, whereas LL belongs to another. Based on morphological observations the LL and UL populations are both *G. lacustris*. Therefore all three of these populations should be representatives of *G. lacustris*.

To better delineate between intra- and interspecific diversity it is necessary to calculate SST with large sample sizes. Thus, a more conservative method of determining a SST would be to use intraspecific rather than intrapopulation diversity. According to a previous estimation of worldwide *G. lacustris* variation estimated the average intraspecific variation to be 1.6±0.3% (Hou et al. 2009). This method provides an SST of 16±3.0% and greatly exceeds the interpopulation divergence between each of the current populations of interest. Measured by this standard each population is comprised of the same species, *Gammarus lacustris*. Even with much greater samples sizes, the determination of SST in this way is still highly contentious. This method of calculating SST relies on the assumption that a gap exists between levels of intra- and interspecific diversity known as the barcoding gap. The existence of such a gap cannot be assumed, as was demonstrated within the well-defined cowry group and may also cause false identification as well as masking of distinct species (Meyer & Paulay 2005). Additionally, potential limitations to the utility of mtDNA include retention of ancestral polymorphism, male-biased gene flow, selection on any mtDNA nucleotide (as the whole genome is one linkage group), introgression following hybridization, and paralogy resulting from transfer of mtDNA gene copies to the nucleus as noted by Mortiz and Cicero (2004). Therefore the employment of SST without the inclusion of alternative taxanomic analysis is arguably unrealiable (Mortiz & Cicero 2004). Furthermore, mtDNA within arthropods may be subject to greater bias due to maternally inherited symbionts. These microbial symbionts may alter patterns of mitochondrial

polymorphisms and skew analysis of mtDNA sequences by either increasing or decreasing sequence diversity between species (Hurst & Higgins 2005). Consequently further study should be undertaken in order to confirm the results determined here by use of a SST. It may be useful to confirm this result through analysis of other highly conserved regions of DNA in order to reproduce this trend. For example, 28S rDNA has been used for this purpose in the examination of cryptic species in the *Hylella* genus (Witt et al. 2006). This would account for any biases stemming from use of mtDNA. Additionally, this assessment is meant to supplement morphological data concerning these populations and should not be considered a definitive statement regarding the species identity of these three populations. Rather, this evaluation was meant to confirm whether or not further study would likely be productive.

Based on these results it can reasonably be concluded that further study may prove beneficial in pursuit of a greater understanding of adaptive evolution. Additionally these topologies, as well as the interpopulation distances show the least divergence between the subterranean GG population and the surface-dwelling UL population (Fig. 5, Fig. 6. Table 1). This suggests there is a hydrological connection between the two populations. The low level of divergence between the two populations and the low level of divergence within the GG population itself may indicate the GG population was founded by a representative of the UL population and experienced a bottleneck effect. This suggests that these populations of *Gammarus lacustris* could successfully serve as a model for the study of divergent evolution.

Sources

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). "Basic local alignment search tool". *J Mol Biol* **215** (3): 403–410.

Carlini, D. B., Manning, J., Sullivan, P. G., & Fong, D. W. (2009). Molecular genetic variation and population structure in morphologically differentiated cave and surface populations of the freshwater amphipod Gammarus minus. *Molecular Ecology, 18, 1932-1945.*

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap Joseph Felsenstein. *Evolution*, *39*(4), 783-791.

Hasegawa M, Kishino H, Yano T (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". *J. Mol. Evol.* 22 (2): 160–74.

Hebert, P. D., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal British Society, 270, S96-S99.*

Hou, Z., Li, Z., & Li, S. (2009). Identifying Chinese Species of Gammarus (Crustacea:Amphipoda) Using DNA Barcoding. *Current Zoology*, *55*(2), 158-164.

Hurst, G. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proc. R. Soc. B*, *272*, 1525-1534.

Jukes, T. H. and C. R. Cantor (1969) Evolution of protein molecules. In H. N. Munro, ed., *Mammalian Protein Metabolism*, pp. 21-132, Academic Press, New York.

Kress, W. J., & Erickson, D. L. (2008). DNA barcodes: Genes, genomics, and bioinformatics. *PNAS, 105(8), 2761-2762.*

Maddison, W. P. & Maddison, D. R. (2007). Mesquite: a modular system for evolutionary analysis. Version 2.01 http://mesquiteproject.org

Meyer, C. P., & Paulay, G. (2005). DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLoS Biology*, *3*(12), 2229-2238.

Meyran, J. C., & Taberlet, P. (1998). Mitochondrial DNA polymorphism among alpine populations of Gammarus lacustris (Crustacea, Amphipoda). *Freshwater Biology, 39, 259-265*.

Mortiz, C., & Cicero, C. (2004). DNA Barcoding: Promise and Pitfalls. *PLoS Biology*, *2*(10), 1529-1531.

Posada, D., & Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics*, 14 (9): 817-818.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Templeton, A. R. (2006). *Population genetics and microevolutionary theory*. Hoboken, N.J.: Wiley-Liss.

Witt, J., Therloff, D., & Hebert, P. (2006). DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology*, *15*, 3073-3082.