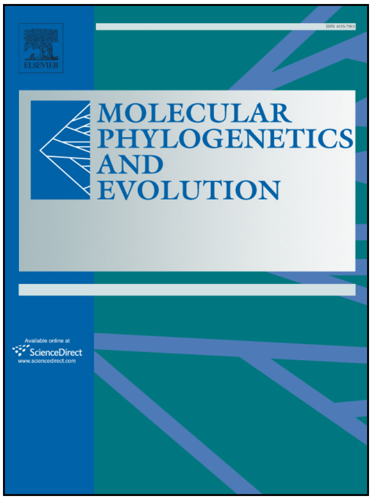


The *Hyalella* (Crustacea: Amphipoda) species cloud of the ancient Lake Titicaca originated from multiple colonizations

Sarah J. Adamowicz, María Cristina Marinone, Silvina Menu Marque, Jeffery W. Martin, Daniel C. Allen, Michelle N. Pyle, Patricio R. De los Ríos-Escalante, Crystal N. Sobel, Carla Ibañez, Julio Pinto, Jonathan D.S. Witt

PII: S1055-7903(17)30154-9
DOI: https://doi.org/10.1016/j.ympev.2018.03.004
Reference: YMPEV 6076

To appear in: *Molecular Phylogenetics and Evolution*

Received Date: 18 February 2017
Revised Date: 13 February 2018
Accepted Date: 5 March 2018

Please cite this article as: Adamowicz, S.J., Cristina Marinone, M., Menu Marque, S., Martin, J.W., Allen, D.C., Pyle, M.N., De los Ríos-Escalante, P.R., Sobel, C.N., Ibañez, C., Pinto, J., Witt, J.D.S., The *Hyalella* (Crustacea: Amphipoda) species cloud of the ancient Lake Titicaca originated from multiple colonizations, *Molecular Phylogenetics and Evolution* (2018), doi: https://doi.org/10.1016/j.ympev.2018.03.004

The *Hyalella* (Crustacea: Amphipoda) species cloud of the ancient Lake Titicaca originated from multiple colonizations

Authors:

Sarah J. Adamowicz[12*], María Cristina Marinone[3], Silvina Menu Marque[3], Jeffery W. Martin[14], Daniel C. Allen[1], Michelle N. Pyle[2], Patricio R. De los Ríos-Escalante[56], Crystal N. Sobel[2], Carla Ibañez[7], Julio Pinto[7], Jonathan D.S. Witt[1]

[1] Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada.

[2] Present address: Biodiversity Institute of Ontario & Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada.

[3] Departamento de Biodiversidad y Biología Experimental, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón II, 4to Piso, C 1428 EHA, Buenos Aires, Argentina.

[4] Present address: Kingston General Hospital, Department of Family Medicine, 76 Stuart Street, Kingston, Ontario, K7L 2V7, Canada.

[5] Universidad Católica de Temuco, Facultad de Recursos Naturales, Escuela de Ciencias Ambientales, Laboratorio de Ecología Aplicada y Biodiversidad, Casilla 15-D, Temuco, Chile.

[6] Nucleo de Estudios Ambientales, Universidad Católica de Temuco, Temuco, Chile.

[7] Instituto de Ecología, Universidad Mayor de San Andrés, Casilla 10077, La Paz, Bolivia.

Email addresses of all authors: sadamowi@uoguelph.ca, marinone@bg.fcen.uba.ar, silvina@bg.fcen.uba.ar, 7jm53@queensu.ca, danallen26@hotmail.com, mpyle@mail.uoguelph.ca, prios@uct.cl, csobel@uoguelph.ca, ibanezluna@yahoo.com, julio.julpin@gmail.com, jwitt@uwaterloo.ca

* Corresponding author email: sadamowi@uoguelph.ca.

Abstract

Ancient lakes are renowned for their exceptional diversity of endemic species. As model systems for the study of sympatric speciation, it is necessary to understand whether a given hypothesized species flock is of monophyletic or polyphyletic origin. Here, we present the first molecular characterization of the *Hyalella* (Crustacea: Amphipoda) species complex of Lake Titicaca, using COI and 28S DNA sequences, including samples from the connected Small and Large Lakes that comprise Lake Titicaca, as well as from a broader survey of southern South American sites. At least five evolutionarily distant lineages are present within Lake Titicaca, which were estimated to have diverged from one another 12-20 MYA. These major lineages are dispersed throughout the broader South American *Hyalella* phylogeny, with each lineage representing at least one independent colonization of the lake. Moreover, complex genetic relationships are revealed between Lake Titicaca individuals and those from surrounding water bodies, which may be explained by repeated dispersal into and out of the lake, combined with parallel intralacustrine diversification within two separate clades. Although further work in deeper waters will be required to determine the number of species present and modes of diversification, our results strongly indicate that this amphipod species cloud is polyphyletic with a complex geographic history.

Keywords: ancient lakes; species flocks; sympatric speciation; evolutionary radiation; Lake Titicaca

1. Introduction

The world's ancient lakes have long been a source of fascination as a consequence of their extraordinary diversity and endemism (reviews by Brooks, 1950a, 1950b; Fryer, 1991; Martens, 1997; Schön and Martens, 2004; see Table S1). There are approximately 15 extant ancient lakes in the world, which have continuously held water for more than ca. 1 million years (e.g. Lakes Tanganyika, Malawi, Baikal, and Titicaca), while the majority of fresh water bodies are geologically much younger. Initially, evolutionary biologists typically focused upon the radiations of ancient lake endemic species flocks as examples of sympatric or intra-lacustrine (within-lake) speciation. However, a review of the evidence by Cristescu et al. (2010) indicated that allopatric vs. sympatric characterizations are too simplistic. Even one of the formerly best-established cases of sympatric speciation in lake environments, the cichlids of crater lakes in Cameroon, has recently been challenged by genomic evidence (Martin et al., 2015). Multiple colonizations, subsequent intralacustrine diversification (which may involve both sympatric and microallopatric components), and hybridization and introgression have been invoked as acting together, producing ancient lakes radiations (Herder et al., 2006; Cristescu et al., 2010; Martin et al., 2015).

Multiple spatial and genetic processes of diversification can also operate in tandem during ancient lake radiations. Despite their great age, ancient lakes do not provide a static backdrop for evolution. Dramatic water-level fluctuations as well as other physico-chemical changes have likely provoked extinction waves yet also provided colonization and diversification opportunities in multiple different ancient lake systems (Cristescu et al., 2010; Kroll et al., 2012). Moreover, water-level changes have resulted in the creation of isolated paleolakes, which have been connected to varying degrees over time, such as in the Lake Titicaca system (Mourguiart, 2000). This history may have promoted allopatric divergence followed by secondary contact or may have instigated a more complex series of events involving allopatric divergence coupled with extirpation/recolonization dynamics (Mourguiart, 2000). However, the evolutionary impacts of such historical limnological changes requires an in-depth exploration. These scenarios are further complicated by repeated dispersal to and from more distant water bodies, providing additional opportunities for allopatric divergence, gene flow, and reticulate evolution (e.g. Herder et al., 2006; Martin et al., 2015). Additionally, instantaneous reproductive isolation resulting from polyploidization and macro-chromosomal change can also play a role in the radiation of species flocks (Martens, 1997; Schön and Martens, 2004). All of these processes can increase total genetic variance and provide opportunities for phenotypic divergence. As a result of the complexities involved in the evolution of ancient lake species complexes, Cristescu et al. (2010) have advocated an approach employing multiple genetic and geographic scales, aimed at developing a broader understanding of radiations in ancient lakes.

Understanding the evolution of lacustrine species clouds foremost requires a determination of their phylogenetic status; do they represent monophyletic, paraphyletic, or polyphyletic aggregations? In turn, addressing this question requires a strong geographic representation of lineages from within the ancient lake, habitats adjacent to the lake, as well as more distant habitats. It is also advantageous to sample a variety of potential outgroup lineages, as complex phylogenetic patterns have been detected in prior studies of putatively monophyletic “species flocks” (e.g. Martin et al., 2015). Although a complex relationship between ancient lakes and their surrounding regions has been recognized for some time (see Martens, 1997), many investigations of lacustrine species flocks have been limited by an inadequate inclusion of samples derived from outside of the lakes. Moreover, because of the greater age and stability of ancient lakes, habitats outside of them are expected to experience greater rates of extinction, which can potentially mask historical patterns of lake colonization. Despite this confounding issue of extinction rates, studies that have included lineages derived from regions surrounding the lakes have often revealed strong signatures of multiple colonizations (reviewed and summarized in Table S1).

Ancient Lake Titicaca, situated on the South American Andean Altiplano in Bolivia and Peru, lies at an altitude of 3810 m and is the largest high-altitude navigable lake in the world (Wirrmann, 1992). While the lake dates back approximately 3 million years, it has a complex geological and hydrographic history (Lavenu, 1992; Wirrmann et al., 1992; Mourguiart, 2000; Kroll et al. 2012). The present physical configuration of the lake is less than 10K years old and represents a relict form of a water body that formerly covered a larger portion of the Altiplano, the high plain between the Eastern and Western Cordilleras of the Andes (Lavenu, 1992; Wirrmann et al., 1992). By contrast, there have also been times when water levels were lower than at present, and the lake consisted of three isolated sub-basins: the Large Lake (Lake Chucuito) as well as the south-western and south-eastern sub-basins of the present-day Small Lake (Lake Huiñaimarca) (Mourguiart et al., 1986, 1998; Mourguiart, 2000). The endemic fauna of Lake Titicaca may therefore have evolved by myriad complex processes, including intralacustrine diversification within single sub-basins, vicariance among the three main sub-basins that comprise the present-day lake, allopatric divergence and dispersal among water bodies within the broader Altiplano region, colonization from other geographic regions, and reticulate evolutionary processes.

In common with ancient Lakes Baikal, Ohrid, Biwa, and the Caspian Basin, Lake Titicaca possesses an endemic amphipod species complex (we avoid applying the term “species flock” due to its implication of monophyly), although the taxa that have radiated vary among these lakes. *Hyalella* (Talitroidea; Dogielinotidae) is the only freshwater epigean amphipod genus that occurs in South America (ca. 50 described species (Väinölä et al., 2008)), but it also occurs in North America, where a vast number of cryptic species have recently been revealed (Wellborn and Broughton, 2008; Witt et al., 2003, 2006, 2008). South American and North American members of the genus are currently assigned to different subgenera (Bousfield, 1996). The Titicaca complex contains 13 described endemic species, and another non-endemic, wide-spread South American species has also been reported from the lake (Coleman and González, 2006; González and Coleman, 2002; González and Watling, 2003). However, within the lake, the genus remains taxonomically poorly known, and assessments by taxonomists suggest that its actual diversity may be much higher (Bousfield, 1996; González and Watling, 2003). There are some strong morphological parallels between the L. Titicaca *Hyalella* and the distantly-related gammaroidean amphipods in Lake Baikal; both groups contain members that possess embellished spines and other morphological protuberances, but there are also important differences between these lakes, including lake depth, area, and age. Additionally, lower oxygen levels within Lake Titicaca arising from its high elevation have been thought to act as a constraint on the evolution of larger body sizes among its fauna (Peck and Chapelle, 2003).

In this study, we present the first molecular characterization of the Lake Titicaca *Hyalella* species complex, using DNA sequence data from two gene regions, one mitochondrial and one nuclear. Our phylogenetic analysis encompasses specimens collected within the currently connected Small and Large Lakes that comprise Lake Titicaca, water bodies from other high-altitude regions of Bolivia and Peru, and a broad survey of aquatic habitats from across Chile and Argentina. Specifically, we test the hypothesis that the evolution of the complex has involved multiple colonization events (polyphyly) against the null hypothesis of a single radiation and monophyly. This study establishes a phylogenetic framework for further study of the species diversity and diversification modes within this ancient lake species cloud.

1. Methods

2.1 Field collecting

Four hundred nine *Hyalella* specimens from South America are included in this study (Fig S1, Table S2). Amphipods were sampled widely within Lake Titicaca itself, with 135 specimens selected for molecular analysis. Fifty-two specimens originated from 12 sampling locations in the Large Lake (Lake Chucuito), which are concentrated near Copacabana, Bolivia; Puno, Peru; and the Capachica Peninsula, Peru. Eighty-three specimens were derived from 13 sampling sites in the Small Lake (or Lake Huiñaimarca), with samples collected near Huatajata, Tiquina, and Guaqui, Bolivia. Eight specimens were included from Lake Arapa, Peru, which lies in Lake Titicaca's inflow drainage basin and is connected to Lake Titicaca's Large Lake by a river. Other high-altitude habitats in the vicinity of Lake Titicaca were sampled, including localities in the La Paz and Oruro provinces of Bolivia (43 specimens from 13 sites) and the Puno region of

Peru (37 specimens from 8 sites). Both the Lake Titicaca inflow drainage basin and the Desaguadero basin, which contains Lake Titicaca's sole outflow river (R. Desaguadero), were well represented. An additional Peruvian specimen from a high-altitude lake (Fuente del Amazonas, Arequipa region) that represents the origin of the Amazon River was also included and sequenced for both markers. Habitat types included other lakes, ponds, farm dams, ditches, streams, and rivers. The remaining specimens were collected during a broader survey of southern South American sites, including 122 specimens from 41 localities throughout Argentina and 62 specimens from 18 localities across Chile (plus an additional Chilean sequence obtained from GenBank).

Amphipods were collected from littoral and shallow zones using D-frame nets with 250 µm mesh size. Specimens were collected from deeper regions of Lake Titicaca (typically up to 20 m depth with a few samples up to 40 m) by towing a square benthic net with 400 µm mesh behind a motor boat for transects of ca. 20-50 m in length. Whenever possible, the samples were sorted while alive into gross morphospecies based on body shape, dorsal armature, antennal features, and colour. Taxonomic identifications to species level were not attempted here, due to the large diversity of morphological forms encountered as well as prior reports suggesting that a large fraction of total species diversity in the lake likely remains to be described (Bousfield, 1996; González and Watling, 2003). Specimens were preserved in 95% ethanol. When available, five individuals per gross morphospecies from each site within Lake Titicaca were selected for genetic analysis. One to six individuals were processed from other sites. Detailed locality information (Table S2) and sequence data are available through the following project in the Barcode of Life Data Systems (BOLD) online repository (www.boldsystems.org) (Ratnasingham and Hebert, 2007): Amphipod Radiation of Lake Titicaca [TTKK]. (This project will be publically released and GenBank numbers obtained for all sequences upon acceptance of this manuscript.)

2.2 Generating molecular data

Total DNA was extracted from each selected individual by grinding one leg (or 2-3 legs for smaller individuals) in 50 µl of a proteinase-K extraction buffer following the methods outlined in Schwenk et al. (1998). A 680 base pair (bp) fragment (of which 628 bp were used for analysis) of the mitochondrial cytochrome c oxidase subunit I (COI) gene was amplified using the primers LCO1490 and HCO2198 (Folmer et al., 1994). The 50 µl PCR reactions contained 0.5-1 µl of DNA template, 5.0 µl 10x PCR buffer, 0.2 µM of each primer, 2.2 mM $MgCl_2$, 0.2 mM of each dNTP, and 0.4 units of *Taq* DNA polymerase (New England Biolabs). The PCR conditions consisted of 60 sec at 94°C followed by 5 cycles of 60 sec at 94°C, 90 sec at 45°C, 60 sec at 72°C; followed by 35 cycles of 60 sec at 94°C, 90 sec at 51°C, 60 sec at 72°C; followed by 5 min at 72°C. PCR products were gel purified (2% agarose) using the Qiaex kit (Qiagen Inc.). Products were sequenced in one direction using primer LCO1490 on a 3730 automated sequencer (Applied Biosystems). After inspecting the trace files and trimming the primer sequences and any additional poor-quality peaks using MEGA version 5 (Tamura et al., 2011), the COI nucleotide sequences were subsequently aligned using the Clustal module in MEGA, translated with the invertebrate mitochondrial code, and verified to be free of stop codons and indels. (A second alignment was also created for each gene, trimming an additional 20 bp from each end. The phylogenetic pattern was the same, and these results are not reported further.)

Using the COI results as a guide, a sub-set of the DNA extractions was selected for analysis of a nuclear marker, including specimens spanning all of the major genetic groupings detected in the COI data set. We obtained sequences for a fragment from the 5' end of the non-coding, large nuclear ribosomal rRNA gene (28S) from 64 specimens. For 58 of these specimens, sequences were successfully obtained for both markers; for six additional specimens, only 28S sequences were available for analysis, as a high-quality COI sequence was not obtained. Protocols were similar to those used for COI, with the following exceptions. The 28S region was amplified and bidirectionally sequenced using the primers employed by Witt et al. (2006). An annealing of 50°C was used for both the initial PCR amplification and for the cycle sequencing reaction. Following visual inspection and the trimming of all trace files, forward and reverse sequences were combined into a single consensus sequence for each individual. The 28S gene was also aligned using the default settings of Clustal in MEGA5, and the alignment was inspected to verify that that the alignment consisted of large conserved blocks of nucleotides.

2.3 Phylogenetic reconstruction using ML and Bayesian methods

The COI and 28S South American data sets were each augmented by six sequences of North American *Hyalella* (from Jeffery 2015) to serve as outgroups. Jeffery (2015) used mitochondrial and nuclear gene regions to reconstruct phylogenetic hypotheses for a broad sampling of North and South American *Hyalella* specimens, rooted using a more distant outgroup (the related genus *Platorchestia*). The South American *Hyalella* was strongly supported as being a monophyletic group within the monophyletic genus *Hyalella*. Therefore, we elected to use closely related outgroups from the same genus here to resolve the tree for South American *Hyalella*. For each gene, six individuals of North American *Hyalella* from Jeffery (2015) were selected that possessed full-length, high-quality sequences and that were also phylogenetically disparate. These augmented data sets were used for both model testing and phylogenetic reconstruction.

As reticulate evolution has been shown to be common in ancient lake radiations, the COI and 28S gene regions were analyzed separately, including all available individuals for each gene. Duplicate sequences (100% similarity, with sub-sequences considered as duplicates) were first eliminated from each alignment using the online tool ElimDupes (https://hcv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html; last accessed Jan. 19, 2017). Using all nucleotide sites and partial deletion of missing data at the 90% threshold, the best-fit model of nucleotide evolution was selected for each gene according to the Bayesian Information Criterion (BIC) (Posada and Buckley, 2004) and then used to perform a maximum likelihood (ML) phylogenetic analysis in MEGA5. ML phylogenetic analyses were also performed using partial deletion of missing data and gaps at the 90% threshold. Confidence was assessed using 1000 bootstrap pseudoreplicates, and the two single-gene trees were compared.

Phylogenetic hypotheses were also reconstructed for each gene using Bayesian analysis implemented in MrBayes v3.2 (Ronquist et al., 2012). Each gene was analyzed using the best-fit model according to the BIC criterion (with the model parameters estimated by MrBayes), four chains with 10M generations, a sampling and diagnostic frequency of 1000, and a 25% burnin. To ensure that sufficient generations were run, analysis diagnostics were consulted to verify that the average standard deviation of the split frequencies was below 0.01 and that the potential scale reduction factor of each parameter was 1.0. For each gene, a consensus tree of the type “allcompat” was created using MrBayes and then visualized using FigTree V. 1.4.3 (Rambaut 2016). Upon discovering broad concordance in the major clades between the ML and Bayesian inference, only the ML trees are presented here for both genes, with posterior probabilities from the Bayesian analyses also displayed for key nodes.

2.4 Mapping the distributions of major clades

Since the same major clades were resolved in both the COI and 28S trees, the COI tree (which included a much larger sampling of individuals) was used to map the distributions of the major clades. Some haplotypes were distributed both inside and outside of Lake Titicaca (see Results and Fig S2). Therefore, the number of well-supported (≥95% bootstrap support in the COI ML tree) major clades that contained individuals collected from within Lake Titicaca was used to determine a conservative estimate of the minimum number of independent colonization events of the lake.

Pairwise genetic divergences were calculated within and among the major clades for both genes using the Kimura-2-parameter (K2P) (Kimura, 1980) genetic distance measure, to facilitate comparisons with many other works that employ K2P distances when analyzing the standard animal DNA barcode region of COI (as in Hebert et al., 2003).

2.5 Testing the monophyly of the Lake Titicaca Hyalella *using Bayes factors*

Support for the hypothesis of monophyly among Lake Titicaca *Hyalella* was assessed using Bayes factor comparisons in MrBayes v.3.2 (Ronquist et al., 2012), using the method and commands as outlined in Ronquist et al. (2011; also see Kass and Raftery, 1995). For the first test, sequences obtained from individuals collected within Lake Titicaca were constrained to be monophyletic (model 1). When duplicate sequences were present in the original data set, a sequence was assigned as a Lake Titicaca sequence if any individual possessing that sequence was collected from the Lake. This imparts a conservative approach for testing the null hypothesis of monophyly, by considering these cases as dispersal events out of the lake rather than multiple colonizations. The fit of model 1 was contrasted with the fit of model 2, in which a negative constraint was imposed (i.e. sequences from Lake Titicaca do not form a monophyletic group). For both models, each gene was analyzed using the best-fit model of nucleotide substitution according to the BIC criterion (with the model parameters estimated by MrBayes), four chains with 15M generations for COI and 10M generations for 28S, a sampling and diagnostic frequency of 500, and a 25% burnin. We compared the harmonic mean estimates of the likelihood of the two models. We interpreted a log difference of 3–5 as good evidence for the more likely model, and we consider a log difference greater than 5 to be very strong evidence (following Ronquist et al. 2011).

We conducted a second test to assess whether the specimens collected in the broader Lake Titicaca inflow and outflow drainages formed a monophyletic group. This involved repeating the above analysis, for each gene separately, imposing a monophyly constraint upon all of the specimens from Bolivia and Peru and contrasting the model fit with that from the negative constraint model using Bayes factors.

2.6 Constructing a time-calibrated COI tree using Bayesian inference

We explored the likely timing of colonization of Lake Titicaca and diversification within the lake by constructing an ultrametricized, time-calibrated tree using Bayesian inference. The COI gene was selected for this analysis due to the availability of suitable clock calibrations as well as the large sample size of sequences. Initially, the data set was reduced by eliminating sequences exhibiting 99% similarity using the tool ElimDupes (https://hcv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html). Model testing was conducted in MEGA5 on this data subset of 104 sequences, including the 6 outgroup sequences, and the model that exhibited the lowest BIC score was selected.

The Beauti v1.8.1 interface was used to prepare the .xml files for analysis in BEAST v. v1.8.1 (Drummond et al., 2012), with a total of four replicate input files created. The substitution model was set to HKY+G+I, with the base frequencies estimated from the data, and the coalescent (constant size) tree prior was used. A lognormal relaxed clock model (uncorrelated) was selected, with the COI rate of evolution set to 0.0189 nucleotide substitutions per site per million years (Wilke et al., 2009; see their Table 3). Wilke et al. (2009) recalculated the rate derived by Knowlton and Weigt (1998) across seven sister pairs of shrimp spanning the Isthmus of Panama. The selected rate was estimated using an aquatic taxon from the same taxonomic class as *Hyalella*, of similar body size, and using the same model of nucleotide substitution as applied here. Additionally, the selected rate was similar in magnitude to that presented for another crustacean group (0.0161; from Wilke et al., 2009; Schubart et al., 1998). This latter rate was not selected for use here due to the small sample size of sister pairs separated by the vicariance event, and also as a result of the transition to terrestrial life in lineages within this crab group. Nevertheless, this similarity in the rate estimates for shrimp and crabs suggests that these rates can be used for general dating of malacostracan crustaceans over time spans that are similar to those involved in the calibration. As Lake Titicaca is approximately 3 MY old, we would expect a calibration obtained using the closure of the Isthmus of Panama to provide reasonable date estimates.

For each of the four identical input files, Markov chain Monte Carlo (MCMC) analyses were run in BEAST with a MCMC length of 100,000,000 generations, sampling every 10,000, yielding a final sample size of 10,000 trees. The four resulting log files were then assessed using Tracer v. 1.6 (Rambaut et al., 2014) to ensure adequate sampling and convergence. The best two data files were subsequently selected for further analysis; in both cases, all effective sample sizes for parameter estimates were >2000, which suggests sufficient numbers of independent samples. The LogCombiner v1.8.1 tool was used to combine the two best log files, with a burnin of 10M generations, and to combine the two best tree files. A maximum clade credibility tree was created in Tree Annotator v1.8.1, using a burnin of 2000 trees and median node heights.

This tree was visualized and colour coded by locality in FigTree v. 1.4.2. The occupancy of Lake Titicaca was phylogenetically mapped using the parsimony criterion, considering sequences represented by any individuals within Lake Titicaca as being Titicaca sequences. All sequences represented by each tip were considered for assigning the tip as being present in Lake Titicaca or not, as the data set had been reduced to eliminate sequences with high similarity. Estimated ages and 95% confidence intervals were collated for select nodes.

2.7 Estimating Hyalella *species richness in Lake Titicaca*

As the actual species richness has been projected to be far greater than that described to date (Bousfield, 1996; González and Watling, 2003), we provide a preliminary estimate of the number of species in our samples using the molecular evidence. We estimate the number of species present in Lake Titicaca and among our South America *Hyalella* samples using the Barcode Index Number (BIN) system (Ratnasingham and Hebert 2013). BINs are Molecular Operational Taxonomic Units (MOTUs) that are similar to species. The BIN approach uses a seed clustering threshold of 2.2% p-distance to group sequences together and then refines the clusters further, through either splitting or lumping based upon patterns of continuity in genetic distances. The resulting molecular units closely corresponded to species for a variety of well-studied animal groups (Ratnasingham and Hebert 2013). The number of BINs present in Lake Titicaca is presented as a minimum species richness estimate, as deeper waters were not sampled (most samples were collected from waters ≤20 m in depth), and recently-diverged species may be overlooked by this approach.

3. Results

3.1 ML and Bayesian phylogenetic hypotheses

COI sequences were obtained from 409 South American *Hyalella* individuals. The 628 bp alignment consisted of 264 unique haplotypes from South America plus six outgroup sequences from North America. The alignment was unambiguous, as no gaps or stop codons were present, and the sequences were highly conserved at the amino acid level. Pairwise nucleotide sequence divergences within the ingroup, consisting of all South American *Hyalella* sequences, were as high as 28.2% (K2P), while divergences among those individuals that were collected within Lake

Titicaca were as high as 24.7%. Model testing, including the outgroup sequences, indicated that the best-fit model under the BIC criterion was HKY+I+G. The ML phylogram of all COI sequences (Figs 1, S2) included nine major, well-supported clades. Five of these clades (designated A-E) contained individuals collected from Lake Titicaca and were phylogenetically dispersed throughout the broader South American phylogeny (Figs 1, S2). These five lineages were deeply divergent from one another, as average K2P nucleotide sequence divergences between clade pairs ranged from 16.7% to 23.1% (Table 1).

The 28S sequences were obtained from 64 selected South American individuals. The final alignment was 1267 bp long and contained 65 unique sequences, 59 ingroup sequences and 6 outgroups. Constructed using the best-fit model (GTR+I+G), the 28S phylogenetic analysis (Fig S3) recovered the same nine major clades as recovered using COI, although the deeper internal branching pattern differed somewhat (Fig 1). Divergences were lower in the 28S gene, with average pairwise K2P distances among Clades A-E ranging from 1.1-6.4% (Table 1).

For both genes, the Bayesian inference trees were similar to the ML trees, particularly in their support for the nine major clades, and so the support values for both are shown on the same figure for major clades (Fig. 1). For COI, the likelihood of the best state between MCMC runs 1 and 2 was -11,258.34, while for 28S the best state exhibited a likelihood of -4571.94.

3.2 Geographic distributions of major clades detected within Lake Titicaca

Consisting of 137 specimens, Clade A (red clade in Figs 1-3) was predominantly collected from the Lake Titicaca and Desaguadero drainage basins. It was discovered in both the Large and Small Lakes that together comprise Lake Titicaca, as well as Lake Arapa, diverse habitats from nearby localities in Peru and Bolivia, and a more distant site in northern Chile. The specimens collected from Lake Titicaca were phylogenetically interspersed and closely related to individuals collected outside the lake (Fig S2 in the supplementary information; Fig 3). This clade was genetically heterogeneous; the maximum pairwise COI K2P distance among individuals within Clade A was 11.1%, with a maximum of 8.2% divergence among individuals collected from within Lake Titicaca itself. Maximum parsimony mapping suggests two colonizations of Lake Titicaca within this clade (Fig 3), although a single colonization of the lake combined with multiple cases of dispersal out of the lake could also explain the phylogenetic pattern.

Clade B (green, Figs 1-3) contained 98 individuals and had a broad, primarily non-Titicacan distribution including sites in Chile, Peru, and Argentina, ranging from the province of Córdoba to Tierra del Fuego. One well-supported (99% bootstrap) sub-clade (Fig S2) was detected in northern Chile and in two Peruvian sites near Lake Titicaca; it was represented by a single individual (Specimen ID Pe03-A2) collected from the Large Lake within Lake Titicaca, which was genetically identical to several specimens collected from a nearby river.

Clade C (yellow, Figs 1-3) included 36 individuals and was also widely distributed from Peru to Tierra del Fuego, Argentina. Within Lake Titicaca, this clade was represented by three individuals, each from a different site within the Small Lake. These formed an endemic genetic cluster, with a maximum pairwise COI K2P distance of 1.9% among these three sequences. The Titicaca specimens (Bo21-A5, Bo04-A23, Bo05-A3) were genetically distinct from all members of Clade C collected elsewhere (Fig S2), showing a minimum K2P distance of 7.9% from members of its sister sub-cluster from northern Chile.

Clade D (blue, Figs 1-3) was represented by 63 specimens in our dataset, and its distribution was constrained to the Lake Titicaca drainage basin. Most records were from Lake Titicaca itself, including both the Small and Large Lakes. Just six specimens were collected from riverine localities outside the lake (sites Pe28 and Pe40), which were phylogenetically nested within Clade D in different sub-clusters (Fig S2). Clade D exhibited less genetic diversity than the other major clades (Figs 1, 3, S2), with the maximum pairwise COI K2P distance within this group being 5.5%.

Clade E (black, Figs 1-3), containing 56 individuals, was genetically diverse and widely distributed, with records from Peru to Tierra del Fuego. Clade E contained 11 specimens collected from Lake Titicaca, which were nested with two different sub-clades within Clade E (Figs 3, S2).

3.3 Topology tests using Bayes factor comparisons

The monophyly hypothesis of the Lake Titicaca *Hyalella* was rejected. Upon constraining the sequences derived from Lake Titicaca to be a monophyletic group using the COI data set, the harmonic mean estimate of the marginal likelihood across the two MCMC runs was -12,441.41, in log units. By contrast, the harmonic mean estimate for model 2 (Lake Titicaca sequences do not form a monophyletic group) was -11,424.17. This difference—of 1,017 log units—was interpreted as extremely strong evidence against the monophyly model.

While the average standard deviation of split frequencies was below 0.01 for model 2, this metric hovered around approximately 0.02 for model 1. For COI specifically, the Titicaca constraint model was therefore run again for 50M generations; however, no overall improvement in convergence was observed. Nevertheless, after 50M generations, and a 25% burnin, the resulting harmonic mean across two runs was -12,457.23, yielding the same conclusion as with the 15M-generation runs.

The same conclusion was also drawn from the analysis of the 28S gene sequences. Upon imposing the Lake Titicaca constraint upon the 28S data set, the harmonic mean estimate of the marginal log likelihood across the two MCMC runs was -4838.73, while result for the negative constraint was -4643.22. The constrained model is therefore 121 log units worse than the model containing no constraint. Comparing the single best tree from each model also yields a very large difference, 190 log units. Specifically, with the Lake Titicaca sequences constrained to be monophyletic, the likelihood of the best state across both runs was -4764.08, while the negative constraint (Lake Titicaca not monophyletic) yielded a best state of -4573.70. (The latter value was similar to the single best state obtained in the independent analysis with no constraint imposed: -4571.94; see section 3.1 above.)

Furthermore, the hypothesis that the *Hyalella* inhabiting Bolivia and Peru together form a monophyletic group was also rejected by both markers. Considering COI first, the harmonic mean estimate of the marginal log likelihood was -11,879.23 for the Bolivia + Peru constraint (model 1), while the value was -11,412.37 for the negative constraint (model 2). This difference of 467 was interpreted as extremely strong evidence against the monophyly hypothesis. For 28S, the harmonic mean estimate of the marginal log likelihood was -4860.97 for model 1 and -4640.54 for model 2, a large difference of 220 log units.

3.4 Timing of colonization and diversification

The ultrametricized, time-calibrated COI tree revealed that the five major clades containing specimens collected from Lake Titicaca diverged from one another 12.2-20.6 MYA (Table 2, Fig 3), i.e. before the origin of the lake at approximately 3 MYA. By contrast, in multiple cases, the most recent common ancestors for clades inhabiting Lake Titicaca exhibit shallow date estimates, which post-date the origin of the lake (Table 2, Fig 3). The most recent common ancestors of Clades A and D, which are both genetically diverse and include many individuals that inhabit Lake Titicaca, were dated to 1.67 MY and 0.98 MY, respectively.

Mapping habitat transitions by maximum parsimony suggests a total of seven colonization events of Lake Titicaca (Fig 3). Five of these span distantly-related Clades A-E, with localities from Peru, Bolivia, Chile, and Argentina being widely distributed across the phylogenetic tree. Within two Clades, A and E, two independent colonizations are suggested within each. Within these Clades, the intervening lineages separating Lake Titicaca specimens are located primarily in nearby localities within the inflow and outflow drainage basins of Lake Titicaca. Therefore, a single colonization of Lake Titicaca in each of Clades A and E, combined with multiple cases of dispersal out of the lake, would also generate the phylogenetic patterns of habitat occupancy observed here.

3.5 Species richness estimates using Barcode Index Numbers

Excluding the outgroups, there were 48 BINs among our South American *Hyalella* samples. Twelve BINs occurred in Lake Titicaca. Of these, six were uniquely detected in the lake, while six others were found both in Lake Titicaca and at other sites.

4. Discussion

4.1 Lake Titicaca's amphipods originate from multiple colonizations

This study presents the first molecular phylogenetic investigation of the *Hyalella* amphipod species complex in Lake Titicaca. Our geographic coverage of sites was broad, including both the Large and Small Lakes that comprise Lake Titicaca, water bodies from the surrounding inflow and outflow drainage basins, and other localities throughout both high- and low-altitude regions of southern South America. This sampling program permitted us to investigate whether the lineages present in Lake Titicaca have a monophyletic or polyphyletic origin. Previous studies of ancient lakes "species flocks" have frequently demonstrated non-monophyly, when sampling regimes included sites outside the lake (Table S1). Our results similarly indicate that the *Hyalella* amphipods inhabiting Lake Titicaca are phylogenetically dispersed across the broader South American phylogeny of this genus, occurring within five well-supported and distantly-related lineages that diversified prior to the origin of the lake. Each of these represents at least one independent colonization event, and thus the Lake Titicaca *Hyalella* lineages are considered to have a polyphyletic origin.

The *Hyalella* fauna of Lake Titicaca was previously known to be diverse, with 13 described endemic species, but it has also been recognized that the true diversity is undoubtedly higher (Bousfield, 1996; González and Watling, 2003). While we provide a preliminary estimate of species richness, 12 BINs among our samples, additional investigation is needed to address this question, including further molecular and morphological analysis and consideration of habitat and ecology. As well, more collecting focused on deeper waters in the lake is needed, as most of our samples were collected at depths up to 20 m, and we did not encounter unique deep-water forms reported by others. However, molecular evidence does suggest that the Titicaca clades harbour substantial diversity. Moreover, there appears to be some structure in the molecular data, with sub-lineages present within several major clades, which could represent recently-diverged populations or species. Future analysis focused on species delineation in this species cloud should consider morphology, ecology, and faster-evolving markers; as well, larger portions of the nuclear genome would be helpful for further resolving the phylogenies (e.g. as in Takahashi and Moreno 2015 for Lake Titicaca fishes of the genus *Orestias*) and for exploring the potential role of introgression in diversification.

While the entire *Hyalella* complex is clearly shown to be polyphyletic in origin, the data presented here are also consistent with the concept of evolutionary radiation *within* an ancient lake—in at least two of the five major lineages present in the lake. Two clades exhibit substantial genetic diversity within Lake Titicaca itself (Clade D) or both within the lake and the broader drainage basin (Clade A). Within Clade A, individuals inhabiting Lake Titicaca exhibit up to 8.2% pairwise COI genetic divergence, and Lake Titicaca lineages diverged during the last ~1.7 MY. Meanwhile, Clade D harbours divergences of up to 5.5%, and its deepest divergence within Lake Titicaca dates to approximately 1 MYA, consistent with a scenario of diversification after the formation of the lake. Lake Titicaca may have therefore served as a cradle of diversification in multiple *Hyalella* lineages. Given the complex distributions of many lineages, understanding the role of divergence and dispersal across the broader Altiplano region vs. the role of sympatric processes within the lake would be an interesting avenue for future study.

Interestingly, there are both substantial differences and similarities when comparing our results to those from another Lake Titicaca radiation, the gastropod species flock of the genus

Heleobia (Kroll et al., 2012). Kroll et al. (2012) found that the *Heleobia* of Lake Titicaca and the surrounding Altiplano region form a monophyletic clade that diversified recently, after the origin of the Lake Titicaca. Their results support the conclusion that Lake Titicaca and its surrounding regions have served as a "cradle" of diversification. Our results of deep evolutionary divergences and multiple colonizations within the single genus *Hyalella* contrast with the pattern for *Heleobia*. Nevertheless, both of our studies agree about colonization of Lake Titicaca by some lineages within the last million years. Additionally, both of our studies suggest an apparent burst of lineage diversification at approximately 0.5 MYA (here, based upon visual inspection of lineages A and D in Fig. 3, suggesting a pulse in diversification events), corresponding to the Lake Ballivián period, when the lake level was higher than at present and then subsequently dropped, which may have provoked allopatric divergence events across the basin (Kroll et al., 2012).

4.2 Limited evidence for divergence between sub-basins

Allopatric divergence between lake sub-basins is one of several possible mechanisms for *in situ* diversification in ancient lakes. Lake Titicaca would be expected to be an ideal venue for this process, resulting from the presence of two distinct present-day lakes: the larger Lago Chucuito, with mean and maximum depths of 135 m and 284 m, and the smaller and much shallower Lago Huiñaimarca, with a mean depth of 9 m and a maximum of just 41 m (Wirrmann, 1992) yet containing two historically-isolated sub-basins (Mourguiart, 2000). The Large and Small Lakes are currently connected by the Tiquina Strait, which is just 850 m wide and a maximum of 21 m deep (Wirrmann, 1992). Lake Titicaca has been subject to substantial fluctuations in water levels, with the lake formerly covering a much larger region of the Altiplano (see Kroll et al., 2012 for an overview). By contrast, there have also been periods of greater isolation between the basins than at present (Mourguiart et al., 1986, 1998; Mourguiart, 2000). Moreover, there are additional peripheral lakes, particularly Lake Arapa, which is currently connected to Lake Titicaca by a river. In addition to physical separation, there was increased salinity in some lakes within the broader paleolake system during arid periods (Servant-Vildary and Mello e Sousa, 1993), which may also have contributed to lineage divergence and to complex patterns of extirpation and recolonization throughout the system.

Despite these favourable conditions for allopatric divergence between sub-basins, there is only limited evidence for this within Lake Titicaca itself among the lineages considered here. Three of the five major clades collected from Lake Titicaca occur in both and Small and Large Lakes, while the two clades detected in just one sub-basin were represented by very few specimens; a single specimen of clade B was collected from the Large Lake, and 3 individuals of clade C occurred in the Small Lake. Close genetic relationships between the sub-basins were detected within the major clades as well, but a more in-depth examination of phylogeographic patterns and population genetics within these clades is required. It is possible that inter-basin divergence did play a role in the past, such as during drier periods, but that subsequent dispersal at times of higher water levels has eroded this signature. Moreover, deeper-water populations may exhibit a different pattern and more marked geographic structure within Lake Titicaca. Some morphological species are specific to deeper waters (Dejoux, 1992), and thus the two sub-basins may be effectively isolated for such species, while truly deep-water specialists would be expected to occur in the deeper Large Lake only. The relative roles of allopatric and sympatric divergence within Lake Titicaca and its broader drainage basins require further investigation, and deeper-water habitats should be targeted for investigation.

4.3 Complex phylogeographic patterns revealed

We observed complex phylogeographic patterns in this species group. While vicariance among Altiplano water bodies likely played a role in the evolution of this complex, the phylogenetically nested positions of many non-Titicaca individuals within the Titicaca clades—together with shared sequences distributed inside and outside the lake—strongly suggest numerous independent cases of secondary dispersal out of Lake Titicaca. In some cases, this may be linked to the lake's sole outflow (Río Desaguadero) (Wirrmann, 1992); thus, it is not surprising that the Titicaca lineages present in the Small Lake would also occur in some connected regions of Bolivia, into which this river flows. Dispersal out of the Large Lake and into surrounding regions of Peru could be explained by upstream migration as well as by dispersal via waterfowl, which has been observed in the genus *Hyalella* (Swanson, 1984). It is also possible that some of the patterns were caused by multiple movements both out and into the lake.

Shallow-water lineages may be more likely to disperse out of Lake Titicaca into other habitats, and thus our focus on littoral and shallow-water (≤20 m depth) sampling has likely not provided a full account of the genus's true diversity and endemicity in the lake. Future work might demonstrate that deeper-water specialists have a higher degree of endemism to Lake Titicaca, resulting from a lack of dispersal opportunities and an absence of any similar deep-water habitats in the surrounding region. Interestingly, we discovered a divergent, apparently endemic mitochondrial lineage within Lake Arapa in somewhat deeper sample (15 m depth; site Pe35). Similarly, in Lake Baikal, shallow-water and bay regions contain more cosmopolitan or widespread species, whereas deeper-water species are more often endemic to the lake itself (Kozhova and Izmest'eva, 1998).

"Escapes" (Neilson and Stepien, 2009) of endemic lineages that evolved *in situ* within ancient lakes are common. This is especially the case for the brackish water Ponto-Caspian fauna. There have been natural invasions of marine environments by this euryhaline fauna (e.g. Audzijonyte et al., 2008; Cristescu and Hebert, 2002). Such escapes may in fact contribute to diversity within the ancient lake itself, as lineages may diverge once outside the lake and then transition back into it. These complex scenarios can increase genetic variance and set up the conditions for the evolution of novel phenotypes and reticulate evolutionary processes (reviewed in Cristescu et al., 2010). Investigating hybridization in detail was beyond the scope of the present work, but is an important topic for future investigation. We observed concordance between both genetic regions in the identities of the major clades, suggesting their long reproductive isolation, but topological differences among and within these major clades. For example, Lake Arapa (sampling site Pe35) displayed a unique mitochondrial lineage, but the one individual from this cluster that was sequenced for the nuclear marker was identical to two individuals from Lake Titicaca. Although based upon few individuals, this interesting pattern mirrors the finding of unidirectional hybridization—resulting in nuclear gene flow while maintaining mitochondrial divergence—among diaptomid crustaceans in the Malili ancient lake system in Sulawesi, Indonesia (Vaillant et al., 2013). While hybridization is an interesting hypothesis that could explain the genetic patterns we observed for Lake Arapa, 28S evolves slowly compared to COI (see Table 1), and thus the slow pace of molecular evolution could also explain the lack of divergence in the nuclear marker. Therefore, future studies should include faster-evolving regions of the nuclear genome to investigate hybridization in this system. The phylogenetic results presented here indicate that this polyphyletic *Hyalella* species cloud will represent a rich study system for the exploration of biodiversity and the geographic and genetic modes of radiation.

4.4 Amphipods in ancient lakes: repeated genesis of diversity

The radiation in Lake Titicaca has some important parallels with the large Baikalian amphipod radiation of 375 species (Kamaltynov, 1999), despite involving disparate evolutionary lineages. The Baikalian endemic amphipod radiation (primarily of the families Acanthogammaridae and Eulimnogammaridae) is also polyphyletic, with two major lineages involved (Macdonald et al., 2005), although there has been limited phylogenetic research from regions surrounding the lake. However, although the Baikal amphipods originated from multiple colonizations, it has also exhibited substantial *in situ* diversification and endemism. Another parallel between the amphipod radiations of Lakes Baikal and Titicaca is the extensive evolution of armature, which is unusual among freshwater amphipods. This observation is suggestive of coevolution with predators (e.g. Vermeij and Covich, 1978; West and Cohen, 1996). A recurring theme in the study of ancient lakes faunas is that intensive biological interactions modulate morphological evolution.

Endemic radiations of amphipods have occurred repeatedly in temperate ancient lakes of the world (Martens and Schön, 1999; Cristescu and Hebert, 2005; Cristescu et al. 2010). In addition to endemic species of Lakes Titicaca (Dejoux, 1992; González and Coleman, 2002; González and Watling, 2003) and Baikal (Kamaltynov 1999; Macdonald et al. 2005), a substantial radiation of ca. 100 species has occurred within the Ponto-Caspian basin (Cristescu and Hebert, 2005), while the younger Lake Ohrid is home to ca. 7 endemic species of amphipod (Karaman, 1985). Variation among lakes in the size of the radiations likely results from a combination of factors: lake age, lake size, hydrographic stability, and physicochemical parameters and history of fluctuation; intrinsic biological traits promoting or inhibiting diversification as well as biological interactions; and contingent or random effects. For example, the low oxygen concentration in Lake Titicaca is thought to constrain the evolution of larger body sizes, e.g. in comparison to Baikalian amphipods (Peck and Chapelle, 2003), and perhaps limits the range of niches that they occupy. Also, the diversity of other lineages present in a lake may either promote or inhibit further niche diversification. For example, there are endemic amphipods in Lake Baikal that are brood parasites upon other amphipods, as well as species that form associations with sponges (Kozhova and Izmest'eva, 1998). Thus, both the biotic and abiotic environments are expected to impact the scale of an ancient lake radiation.

The fact that multiple evolutionarily independent lineages of amphipods have radiated in several ancient lakes suggests an inherent propensity for diversification in these settings, perhaps attributable to a biological trait. Brooding has been proposed as being associated with ancient lakes radiations in a variety of different organisms, including crustaceans (Cohen and Johnston, 1987; Martens, 1997; Martens and Schön, 1999) and gastropods (Michel, 1994, 2000). Brooding and direct development are predicted to be associated with lower dispersal rates and thus greater potential for (micro)allopatric divergence and local adaptation, possibly leading to elevated speciation rates. Greater speciation rates could lead to higher net diversification in stable environments, where local extinction is expected to be lower than in other freshwaters. Despite the elegance of this hypothesis, a thorough analysis must be conducted prior to concluding the existence of a link between a specific trait and diversification. Often, ancient lakes studies have tended to focus on radiations and have largely ignored cases of colonization without radiation. However, including such "non-radiations" (as in Seehausen, 2006) is important for providing a full assessment of the associations between traits such as brooding and diversification.

4.5 Concluding remarks

Our molecular data on the amphipods of Lake Titicaca and other regions of southern South America revealed at least five independent colonizations of the lake, at least two of which appear to have undergone subsequent diversification. This pattern suggests that the Lake Titicaca environment promoted the diversification of amphipods. Moreover, our results indicate a complex relationship between the fauna of the lake and its surrounding regions, with multiple apparent cases of dispersal out of the lake, as well as possible cases of dispersal back into the lake secondarily. As well, it is possible that historical vicariance among basins within Lake Titicaca played a role, and diversification dynamics across the broader drainage basins should also be considered. Despite this complex pattern, our results suggest that a portion of the lineage diversity originated inside Lake Titicaca, perhaps representing intralacustrine diversification.

The amphipods of Lake Titicaca have evolved interesting morphological features such as embellished dorsal armature. Given that the other South American *Hyalella* clades and a diverse array of deeply divergent lineages in North America (Witt and Hebert, 2000; Witt et al., 2006) lack such features, it appears that that this trait evolved relatively rapidly within Lake Titicaca, perhaps in response to predation. The apparent relatively rapid pace of morphological evolution in the genus in Lake Titicaca contrasts sharply with the broader pattern of morphological evolution in *Hyalella* within North and South America: morphological stasis and substantial cryptic species-level diversity (Wellborn and Broughton, 2008; Witt et al., 2003, 2006, 2008; Witt and Hebert, 2000). If it is correct that the dorsal armature evolved in response to sustained predation in Lake Titicaca, then this would represent another example of biological interactions governing important evolutionary processes in ancient lake faunas.

Acknowledgements

We are very grateful to Agustín Bachmann, Santiago Echaniz, Alicia Vignatti, Juan Paggi, Alfonso Giudici, Eliana Ibáñez, Elizabeth Escalante, the Family Escalante-Astudillo, Elisa Pistán, Cinthia Ramos, Claudia Zepita, Veronika Sacherová, and Bohumír Janský for providing specimens, assisting with field collecting, and/or assisting with field logistics. Angela Holliss of the Advanced Analysis Centre at the University of Guelph conducted the DNA sequencing. We thank Austin Duong for assistance with proof-reading and referencing and Tzitziki Loeza-Quintana for valuable input on the Bayesian analyses. We also thank two anonymous reviewers for their thoughtful and constructive comments. This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (2012-327509) to JDSW, an NSERC postdoctoral fellowship and NSERC Discovery Grant (386591-2010) to SJA, NSERC Undergraduate Student Research Awards (USRA) to DCA and JWM, and a grant to PR from the Research Direction of the Catholic University of Temuco (funding for development of Limnology Project DGI-DCA 2007-01, Project MECESUP UCT 0804, with valuable support from MI).

References

Audzijonyte, A., Daneliya, M.E., Mugue, N., Väinölä, R., 2008. Phylogeny of *Paramysis* (Crustacea: Mysida) and the origin of Ponto-Caspian endemic diversity: Resolving power from nuclear protein-coding genes. *Mol. Phylogenet. Evol.* **46**, 738-759. (doi:10.1016/j.ympev.2007.11.009)

Bousfield, E.L., 1996. A contribution to the reclassification of neotropical freshwater hyalellid amphipods (Crustacea: Gammaridea, Talitroidea). *Bolletino del Museo Civico di Storia Naturale di Verona* **20**, 175-224.

Brooks, J.L., 1950a. Speciation in Ancient Lakes. *Q. Rev. Biol.* **25**, 30-60.

Brooks, J.L., 1950b. Speciation in Ancient Lakes (Concluded). *Q. Rev. Biol.* **25**, 131-176.

Cohen, A.S., Johnston, M.R., 1987. Speciation in brooding and poorly dispersing lacustrine organisms. *Palaios* **2**, 426-435.

Coleman, C.O., González, E.R., 2006. New hyalellids (Crustacea: Amphipoda, Hyalellidae) from Lake Titicaca. *Org. Divers. Evol.* **6**, 218-219. (doi: 10.1016/j.ode.2005.11.001)

Cristescu, M.E.A., Hebert, P.D.N., 2002. Phylogeny and adaptive radiation in the Onychopoda (Crustacea, Cladocera): evidence from multiple gene sequences. *J. Evol. Biol.* **15**, 838-849.

Cristescu, M.E.A., Hebert, P.D.N., 2005. The "Crustacean Seas" – an evolutionary perspective on the Ponto-Caspian peracarids. *Can. J. Fish. Aquat. Sci.* **62**, 505-517. (doi: 10.1139/F04-210)

Cristescu, M.E., Adamowicz, S.J., Vaillant, J.J., Haffner, D.G., 2010. Ancient lakes revisited: from the ecology to the genetics of speciation. *Mol. Ecol.* **19**, 4837-4851. (doi: 10.1111/j.1365-294X.2010.04832.x)

Dejoux, C., 1992. The Amphipoda. In *Lake Titicaca. A Synthesis of Limnological Knowledge* (eds. C. Dejoux, A. Iltis), pp. 346-356. Dordrecht: Kluwer Academic.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol*. **29**, 1969-1973.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech.* **3**, 294-299.

Fryer, G., 1991. Comparative aspects of adaptive radiation and speciation in Lake Baikal and the great rift lakes of Africa. *Hydrobiologia* **211**, 137-146.

González, E.R., Coleman, C.O., 2002. *Hyalella armata* (Crustacea, Amphipoda, Hyalellidae) and the description of a related new species from Lake Titicaca. *Org. Divers. Evol.* **2,** 271-273.

González, E.R., Watling, L., 2003. Two new species of *Hyalella* from Lake Titicaca, and redescriptions of four others in the genus (Crustacea: Amphipoda). *Hydrobiologia* **497**, 181-204.

Hebert, P.D.N., Cywinska, A., Ball, S.L., deWaard, J.R., 2003. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond.* B . **270**, 313-321.

Herder, F., Nolte, A.W., Pfaender, J., Schwarzer, J., Hadiaty, R.K., Schliewen, U.K., 2006. Adaptive radiation and hybridization in Wallace's Dreamponds: evidence from sailfin silversides in the Malili Lakes of Sulawesi. *Proc. R. Soc. B.* **273**, 2209-2217.

Jeffery, N.W., 2015. Genome Size Diversity and Evolution in the Crustacea. PhD Thesis. University of Guelph, Canada. (Available from: https://atrium.lib.uoguelph.ca/xmlui/handle/10214/9216)

Kamaltynov, R.M., 1999. On the higher classification of Lake Baikal amphipods. *Crustaceana* **72**, 933–944.

Karaman, G.S., 1985. Contribution to the knowledge of the Amphipoda 151. *Gammarus salemaai* new species from Lake Ohrid (Fam. Gammaridae). *Fragm. Balcan.* **12**, 155-168.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 11–120.

Knowlton, N., Weigt, L.A., 1998. New dates and new rates for divergence across the Isthmus of Panama. *Proc. R. Soc. Lond. B.* **265**, 2257-2263.

Kozhova, O.M., Izmest'eva, L.R. (eds.), 1998. *Lake Baikal: Evolution and Biodiversity*. Leiden: Backhuys Publishers.

Kroll, O., Hershler, R., Albrecht, C., Terrazas, E.M., Apaza, R., Fuentealba, C., Wolff, C., Wilke, T., 2012. The endemic gastropod fauna of Lake Titicaca: correlation between molecular evolution and hydrographic history. *Ecol. Evol.* **2**, 1517-1530.

Lavenu, A., 1992. Formation and geological evolution. In *Lake Titicaca. A Synthesis of Limnological Knowledge* (eds. C. Dejoux, A. Iltis), pp. 3-15. Dordrecht: Kluwer Academic.

Macdonald, K.S., Yampolsky, L., Duffy, J.E., 2005. Molecular and morphological evolution of the amphipod radiation of Lake Baikal. *Mol. Phylogenet. Evol.* **35**, 323-343.

Martens, K., 1997. Speciation in ancient lakes. *TREE* **12**, 177-182.

Martens, K., Schön, I., 1999. Crustacean biodiversity in ancient lakes: a review. *Crustaceana* **72**, 899-910.

Martin, C.H., Cutler, J.S., Friel, J.P., Dening Touokong C., Coop, G., Wainwright, P.C., 2015. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution* **69**, 1406-1422.

Michel, E., 1994. Why snails radiate: a review of gastropod evolution in long-lived lakes, both recent and fossil. *Arch. Hydrobiol. Beih. Ergebn. Limnol.* **44**, 285-317.

Michel, E., 2000. Phylogeny of a gastropod species flock: exploring speciation in Lake Tanganyika in a molecular framework. *Adv. Ecol. Res.* **31**, 275-302.

Mourguiart, P., 2000. Historical changes in the environment of Lake Titicaca: evidence from ostracod ecology and evolution. *Adv. Ecol. Res.* **31**, 497-520.

Mourguiart, P., Carbonel, P., Peypouquet, J.-P., Wirrmann, D., Vargas, C., 1986. Late Quaternary palaeohydrology of Lake Huinaymarca (Bolivia). Scenarios based on ostracods fauna. *Hydrobiologia* **143,** 191-197.

Mourguiart, P., Corrège, T., Wirrmann, D., Argollo, J., Montenegro, M.E., Pourchet, M., Carbonel, P., 1998. Holocene palaeohydrology of Lake Titicaca estimated from an ostracod-based transfer function. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **143,** 51-72.

Neilson, M.E., Stepien, C.A., 2009. Escape from the Ponto-Caspian: Evolution and biogeography of an endemic goby species flock (Benthophilinae: Gobiidae: Teleostei). *Mol. Phylogenet. Evol.***52**, 84-102. (doi:10.1016/j.ympev.2008.12.023)

Peck, L.S., Chapelle, G. 2003. Reduced oxygen at high altitude limits maximum size. *Proc. R. Soc. B* **270**, (Suppl 2), S166-S167. (DOI: 10.1098/rsbl.2003.0054)

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53,** 793-808.

Rambaut, A., 2016. FigTree V. 1.4.3. Available from: http://tree.bio.ed.ac.uk/software/figtree/ (accessed January 21, 2017)

Rambaut, A., Suchard, M.A., Xie, D., Drummond, A.J., 2014. Tracer v1.6. Available from: http://beast.bio.ed.ac.uk/Tracer (accessed February 2, 2017)

Ratnasingham, S., Hebert, P.D.N., 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364. (DOI: 10.1111/j.1471-8286.2006.01678.x)

Ratnasingham, S., Hebert, P.D.N., 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS One* 8, e66213. (doi:10.1371/journal.pone.0066213)

Ronquist, F., Huelsenbeck, J., Teslenko, M., 2011. MrBayes version 3.2 Manual: Tutorials and Model Summaries. Nov. 5, 2011. Available from: http://mrbayes.sourceforge.net/ (accessed Jan. 1, 2017)

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539-42. (doi: 10.1093/sysbio/sys029)

Schön, I., Martens, K., 2004. Adaptive, pre-adaptive and non-adaptive components of radiations in ancient lakes: a review. *Org. Divers. Evol.* **4**, 137-156. (doi:10.1016/j.ode.2004.03.001)

Schubart, C. D., Diesel, R., Hedges, S.B., 1998. Rapid evolution to terrestrial life in Jamaican crabs. *Nature* **393,** 363-365.

Schwenk, K., Sand, A., Boersma, M., Brehm, M., Mader, E., Offerhaus, D., Spaak, P., 1998. Genetic markers, genealogies and biogeographic patterns in the Cladocera. *Aquat. Ecol.* **32**, 37–51.

Seehausen, O., 2006. African cichlid fish: a model system in adaptive radiation research. *Proc. R. Soc. B* **273**, 1987-1998.

Servant-Vildary, S., Mello e Sousa, S.H., 1993. Palaeohydrology of the Quaternary saline Lake Ballivian (southern Bolivian Altiplano) based on diatom studies. *Int. J. Salt Lake Res.* **2,** 69-85.

Swanson, G.A., 1984. Dissemination of amphipods by waterfowl. *J. Wildlife Manage.* **48**, 988-991. (DOI: 10.2307/3801453)

Takahashi, T., and Moreno, E. 2015 A RAD-based phylogenetics for *Orestias* fishes from Lake Titicaca. *Mol. Phylogenet. Evol.* 93, 307-317.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731-2739.

Vaillant, J.J., Bock, D.G., Haffner, G.D., Cristescu, M.E., 2013. Speciation patterns and processes in the zooplankton of the ancient lakes of Sulawesi Island, Indonesia. *Ecol. Evol.* **3**, 3083-3094.

Väinölä, R., Witt, J.D.S., Grabowski, M., Bradbury, J.H., Jazdzewski, K., Sket, B., 2008. Global diversity of amphipods (Amphipoda; Crustacea) in freshwater. *Hydrobiologia* **595**, 241-255. (doi: 10.1007/s10750-007-9020-6)

Vermeij, G.J., Covich, A.P., 1978. Coevolution of freshwater gastropods and their predators. *Am. Nat.* **112**, 833-843.

Wellborn, G.A., Broughton, R.E., 2008. Diversification on an ecologically constrained adaptive landscape. *Mol. Ecol.* **17**, 2927-2936. (doi: 10.1111/j.1365-294X.2008.03805.x)

West, K., Cohen, A., 1996. Shell microstructure of gastropods from Lake Tanganyika, Africa: adaptation, convergent evolution and escalation. *Evolution* **50**, 672-681.

Wilke, T., Schultheiß, R., Albrecht, C., 2009. As time goes by: A simple fool's guide to molecular clock approaches in invertebrates. *Amer. Malac. Bull.* **27**, 25-45.

Wirrmann, D., 1992. Morphology and bathymetry. In *Lake Titicaca. A Synthesis of Limnological Knowledge* (eds. C. Dejoux, A. Iltis), pp. 16-22. Dordrecht: Kluwer Academic.

Wirrmann, D., Ybert, J.-P., Mourguiart, P., 1992. Paleohydrology: A 20,000 years paleohydrological record from Lake Titicaca. In *Lake Titicaca. A Synthesis of*

Limnological Knowledge (eds. C. Dejoux, A. Iltis), pp. 40-48. Dordrecht: Kluwer Academic.

Witt, J.D.S., Hebert, P.D.N., 2000. Cryptic species diversity and evolution in the amphipod genus *Hyalella* within central glaciated North America: a molecular phylogenetic approach. *Can. J. Fish. Aquat. Sci.* **57**, 687-698.

Witt, J.D.S., Blinn, D.W., Hebert, P.D.N., 2003. The recent evolutionary origin of the phenotypically novel amphipod *Hyalella montezuma* offers an ecological explanation for morphological stasis in a closely allied species complex: *Mol. Ecol.* **12**, 405-413. (doi: 10.1046lj.1365-294X.2003.O l728.x)

Witt, J.D.S., Threloff, D.L., Hebert, P.D.N., 2006. DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: Implications for desert spring conservation. *Mol. Ecol.* **15**, 3073-3082.

Witt, J.D.S., Threloff, D.L., Hebert, P.D.N., 2008. Genetic zoogeography of the *Hyalella azteca* species complex in the Great Basin: Rapid rates of molecular diversification in desert springs. Geol. S. Am. S. **439**, 103-114. (doi: 10.1130/2008.2439(05))

Table 1. Average K2P genetic distances between the five major clades that contained individuals inhabiting Lake Titicaca. Values for the COI gene are presented below the diagonal, and values for the 28S gene are above the diagonal. These clades are estimated to have diverged 12.2-20.6 MYA (Table 2, Fig 3).

	Clade A	**Clade B**	**Clade C**	**Clade D**	**Clade E**
Clade A	--	0.016	0.064	0.017	0.021
Clade B	0.167	--	0.057	0.013	0.014
Clade C	0.208	0.203	--	0.056	0.054
Clade D	0.183	0.171	0.224	--	0.011
Clade E	0.212	0.196	0.231	0.197	--

Table 2. Age estimates and confidence intervals for select nodes obtained from Bayesian analysis of COI sequences. Node labels refer to the rounded age estimates presented in Fig 3.

Node Label (Fig 3)	Age Estimate (MY)	(95% Confidence Interval)
20.6	20.633	(15.615-27.086)
14.6	14.646	(11.643-18.121)
13.3	13.343	(10.434-16.453)
12.2	12.160	(9.447-15.381)
1.67	1.668	(1.210-2.178)
0.98	0.977	(0.609-1.500)

Figure Legends

Fig. 1. Phylogenetic hypotheses for the *Hyalella* of South America based upon the A) COI and B) 28S genes. For each gene, the topology is shown according to ML analysis, with the node support values based upon 1000 pseudoreplicates of the ML analysis displayed first, followed by the posterior probabilities from Bayesian analysis of the same dataset. An asterisk (*) denotes that a particular node is missing from that analysis. Major clades are collapsed for visualization, while the full trees including specimen codes are found in the supplementary information, with individuals collected from Lake Titicaca highlighted (Fig. S2: COI tree; Fig. S3: 28S tree). The coloured clades were those containing individuals detected within Lake Titicaca (A-red, B-green, C-yellow, D-blue, E-black), while the grey clades were only detected from other sites. The distributions of major clades that included representatives from Lake Titicaca are shown in Fig. 2. The same major clades were recovered using both genes, given the available data: 409 individuals from South America for COI and 64 individuals for 28S (excluding outgroups).

Fig. 2. Map of Lake Titicaca, which consists of the Small Lake (Huiñaimarca) and Large Lake (Chucuito), linked by a strait, as well as its surrounding inflow (Titicaca) and outflow (Desaguadero) drainage basins. The distributions of the five major South American *Hyalella* clades (Fig. 1) that contain representatives within Lake Titicaca are shown (A-red, B-green, C-yellow, D-blue, E-black). Some points are offset for visualization; GPS co-ordinates are available

in supplementary Table S2. In addition to the mapped localities, Clade A was also collected in northern Chile, and clades B, C, and E were widely distributed in Chile and Argentina.

Fig. 3. Time-calibrated tree for the *Hyalella* of South America, based upon Bayesian analysis of COI sequences. The dataset was first reduced to exclude sequences more than 99% similar to others. Magenta terminal branches denote lineages collected from inside Lake Titicaca; turquoise is used to mark tips that represent individuals found both inside Lake Titicaca and surrounding regions; black branches indicates lineages collected exclusively outside Lake Titicaca; and dark blue is used for the six outgroup sequences of *Hyalella* from North America. Habitat transitions are mapped onto the phylogeny according to the maximum parsimony criterion; branches leading to the most recent common ancestor of Lake Titicaca clades are not coloured, as the timing of lake colonization is uncertain. The seven reconstructed colonizations of Lake Titicaca are marked with circles at the nodes. Age estimates (in MY) are provided for select nodes; confidence intervals for these are provided in Table 2. The full tree including tip labels is available as Fig. S4.

Supplementary files

Table S1. Review of the phylogenetic pattern and extent of sampling for a selection of ancient lakes "species flocks" that have been characterized using molecular phylogenetic methods.

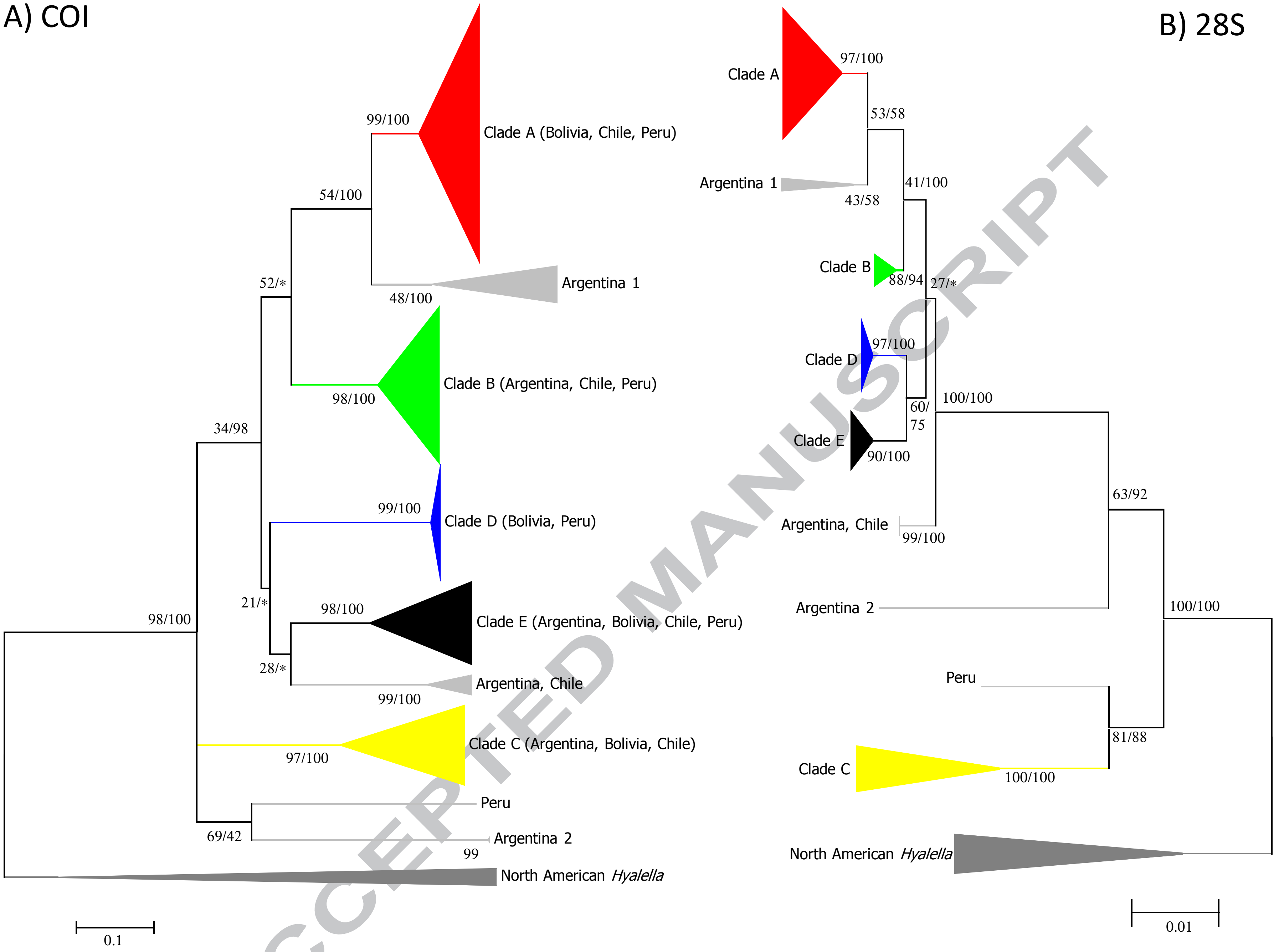
Table S2. List of *Hyalella* specimens from Bolivia, Peru, Argentina, and Chile that were included in this study, including detailed locality data and GenBank Accession numbers for sequences. (XXX – GenBank accessions to be populated upon acceptance of this manuscript.)

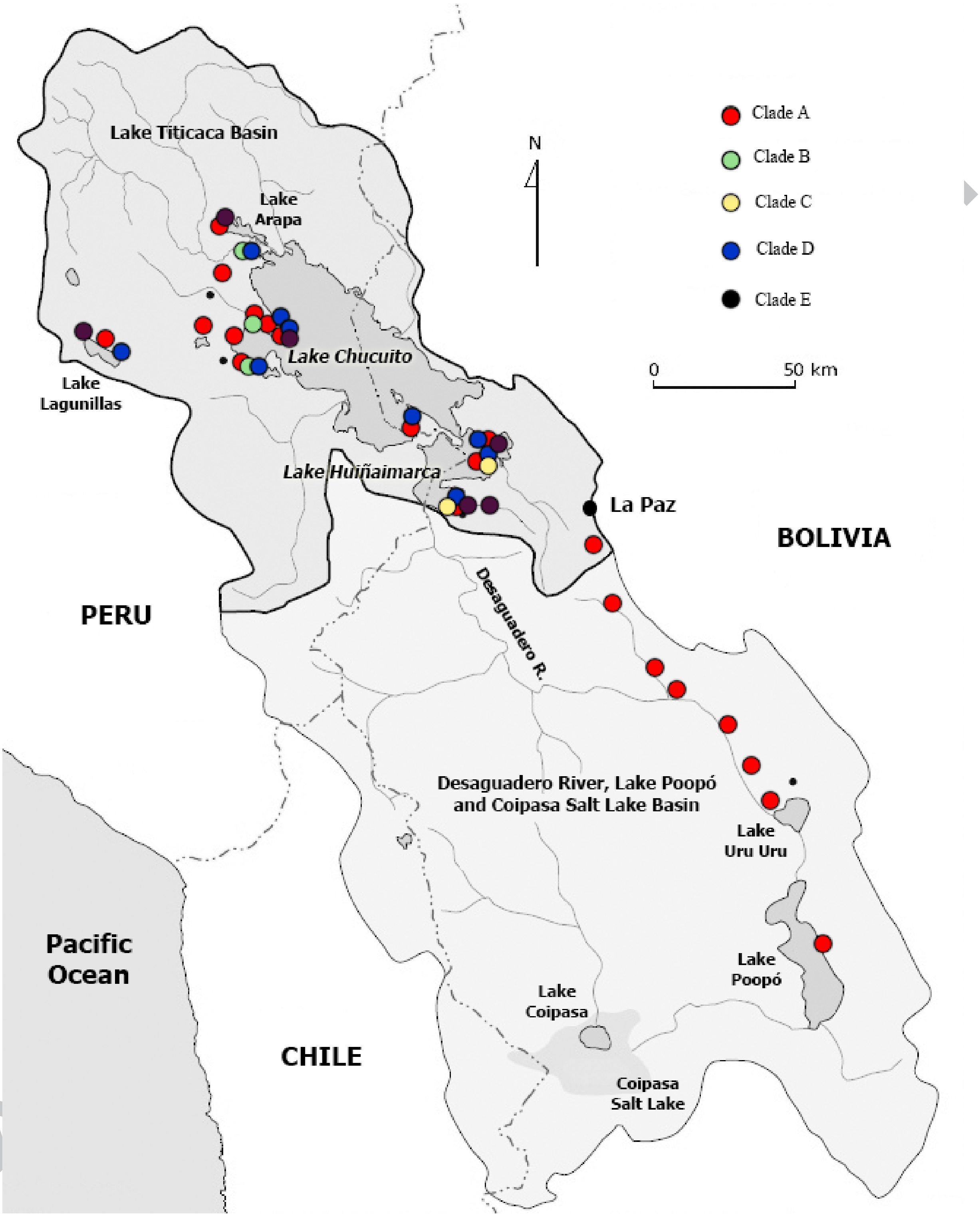
Fig S1. Map showing sampling locations for all *Hyalella* specimens included in this study. A) Sampling sites across southern South America; B) Detailed view of sites sampled from the Lake Titicaca region. GPS co-ordinates are available in Table S2 and online through BOLD (www.boldsystems.org) (project TTKK: Amphipod Radiation of Lake Titicaca).

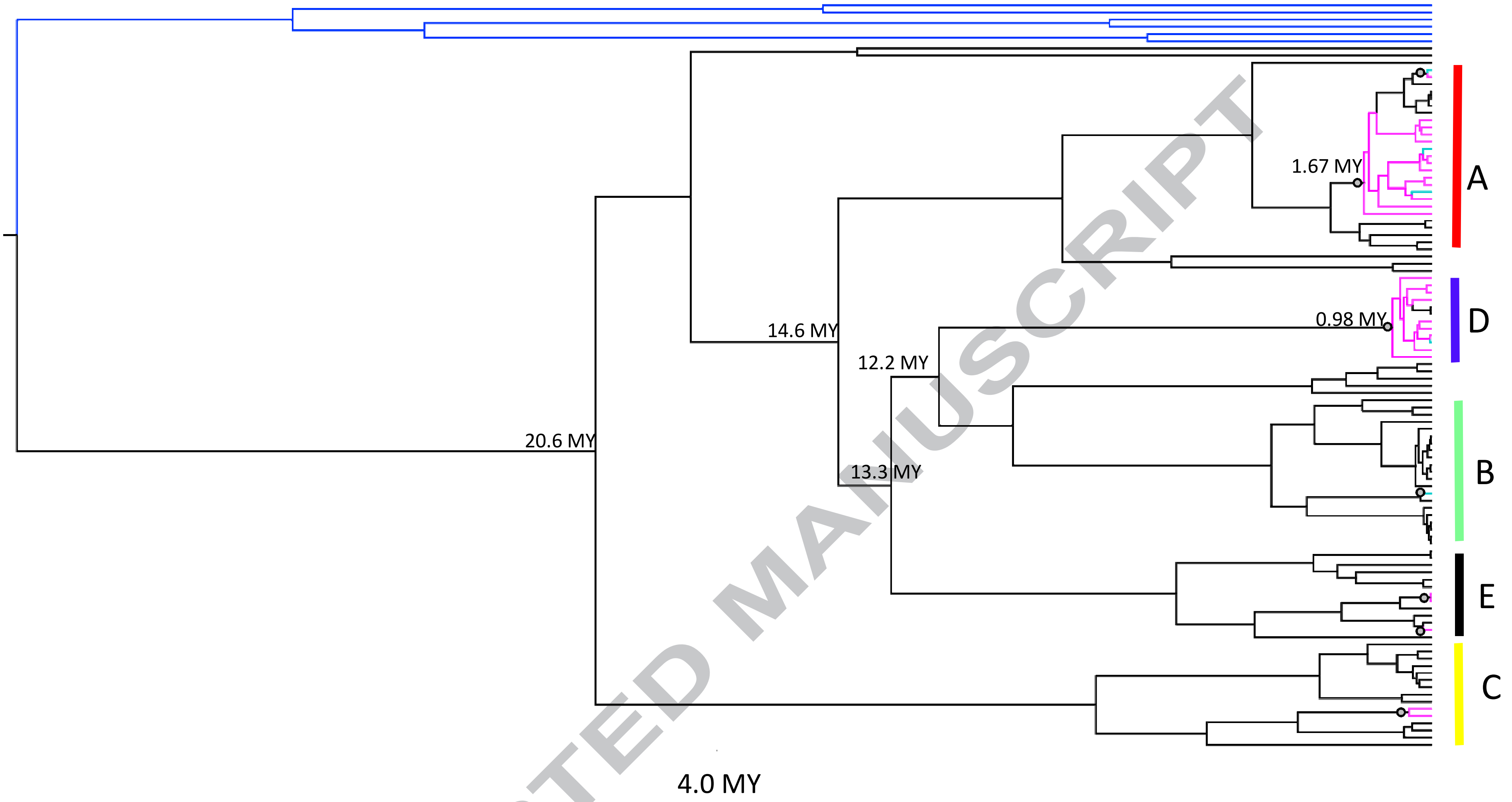
Fig S2. Full COI ML tree of *Hyalella* specimens, matching the tree with collapsed clades presented in Fig. 1A. The sequence data set was reduced to include only unique haplotypes prior to analysis. The tips are labelled with the Process ID (a sequence identifier assigned by BOLD), followed by the Specimen ID, with the beginning of each Sample ID consisting of a country code (Arg-Argentina, Bo-Bolivia, Ch-Chile, Pe-Peru) followed by a site number (sites were numbered within each country), followed by a hyphen and then a code to distinguish individuals from the same site. Five clades (A-E) contain specimens collected from Lake Titicaca. The clade designations are shown using both colour and letters (A-red, B-green, C-yellow, D-blue, E-black), matching the clades as marked in Figs 1 and 2. Full locality data for all specimens are provided in Table S2. Bootstrap values are based upon 1000 pseudoreplicates; values ≥70% are displayed.

Fig S3. Full 28S ML tree of *Hyalella*, matching the tree with collapsed clades presented in Fig. 1B. The tips are labelled with the Specimen IDs. For the ingroup sequences, the Specimen IDs consist of the country code and site code, with a designation for the individual specimen shown after the hyphen. Bootstrap values are based upon 1000 pseudoreplicates. Cases of identical sequences are reflected in the tip labels. Specimens collected from Lake Titicaca are marked by colouring the font of the tip label with the clade designation as well as by adding an asterisk. No individuals from Lake Titicaca from Clade C were sequenced for 28S.

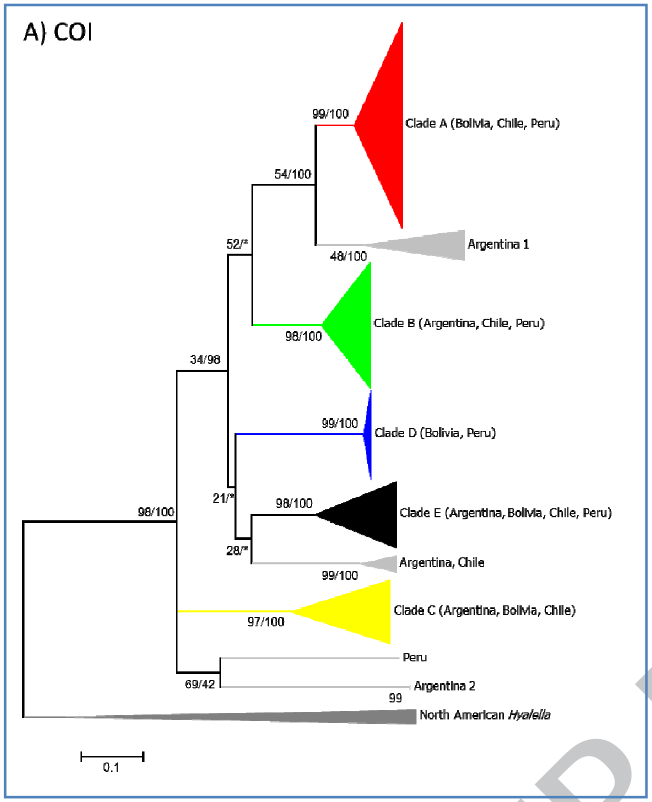
Fig. S4. Full time-calibrated COI tree including tip labels, matching Fig. 3 in the main manuscript. The sequence dataset was reduced to exclude those sequences more than 99% similar to others prior to analysis. The colour coding of branches reflects these duplicate sequences. Magenta terminal branches denote lineages collected from inside Lake Titicaca; turquoise is used to mark tips that represent individuals found both inside Lake Titicaca and surrounding regions; and black branches indicates lineages collected exclusively outside Lake Titicaca. The six outgroup sequences are *Hyalella* from North America. Habitat transitions are mapped onto the phylogeny according to the maximum parsimony criterion. Branches leading to the most recent common ancestor of Lake Titicaca clades are not coloured, as the timing of lake colonization is uncertain.

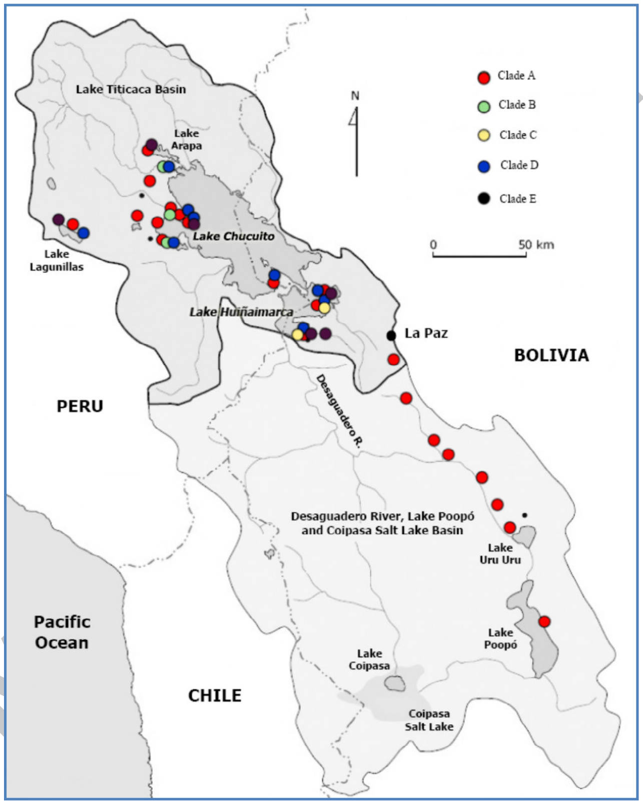

A) COI
99/100
Clade A (Bolivia, Chile, Peru)
54/100
52/*
Argentina 1
48/100
Clade B (Argentina, Chile, Peru)
98/100
34/98
99/100
Clade D (Bolivia, Peru)
21/*
98/100
Clade E (Argentina, Bolivia, Chile, Peru)
98/100
28/*
Argentina, Chile
99/100
Clade C (Argentina, Bolivia, Chile)
97/100
Peru
69/42
Argentina 2
99
North American *Hyalella*
0.1
B) 28S
Clade A
97/100
53/58
Argentina 1
43/58
41/100
Clade B
88/94
27/*
Clade D
97/100
60/
75
Clade E
90/100
100/100
Argentina, Chile
99/100
63/92
Argentina 2
100/100
Peru
81/88
Clade C
100/100
North American *Hyalella*
0.01


Lake Titicaca Basin
Lake Arapa
Lake Chucuito
Lake Lagunillas
Lake Huiñaimarca
La Paz
BOLIVIA
PERU
Desaguadero R.
Desaguadero River, Lake Poopó and Coipasa Salt Lake Basin
Lake Uru Uru
Lake Poopó
Lake Coipasa
Coipasa Salt Lake
Pacific Ocean
CHILE
N
0
50 km
● Clade A
● Clade B
● Clade C
● Clade D
● Clade E


1.67 MY
0.98 MY
14.6 MY
12.2 MY
13.3 MY
20.6 MY
4.0 MY
A
D
B
E
C

Lake Titicaca *Hyalella* Have Polyphyletic Origin


A) COI
99/100
Clade A (Bolivia, Chile, Peru)
54/100
52/*
Argentina 1
48/100
Clade B (Argentina, Chile, Peru)
98/100
34/98
99/100
Clade D (Bolivia, Peru)
21/*
98/100
Clade E (Argentina, Bolivia, Chile, Peru)
98/100
28/*
Argentina, Chile
99/100
Clade C (Argentina, Bolivia, Chile)
97/100
Peru
69/42
Argentina 2
99
North American *Hyalella*
0.1



Non-monophyly:

Specimens collected from Lake Titicaca were dispersed across 5 different clades (A-E) from southern South America, which diversified before the origin of the lake.

Complex biogeography:

Hyalella have colonized Lake Titicaca at least 5 times independently. There is also evidence of secondary dispersal of largely endemic sub-clades (A,D) out of the lake into surrounding regions.

- The *Hyalella* amphipod species cloud of ancient Lake Titicaca is polyphyletic.
- Lake Titicaca was colonized at least 5 times independently.
- Evolutionary radiation occurred within Lake Titicaca in two primarily endemic clades.
- The dispersal history is complex, including migrations out of the lake.