Molecular Biogeography of the Amphipod Genus *Hyalella* in North America

By

Matthew J. Hrycyshyn

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Biology

Waterloo, Ontario, Canada, 2015

©Matthew J. Hrycyshyn 2015

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Hyalella azteca (order: amphipoda) was once thought to be one of the most broadly distributed North America freshwater crustaceans, but recent molecular studies have discovered several species belonging to a morphologically cryptic complex. This group consists of over eighty provisional species, most of which are endemic to isolated regions within the Great Basin. Previous research has focused on a limited number of *Hyalella azteca* populations within specific North American regions (the Great Lakes, the Great Basin, Oklahoma and Quebec). The geographic ranges of *Hyalella* species likely exceed the scope previously studied and their characterization requires a broader approach. This study explored the spatial distribution of nine *Hyalella* species from the analysis of five hundred and thirteen North American habitats. The sampling area focused on previously glaciated regions, but also included habitats south of the glacial limit.

The *Hyalella azteca* complex can be separated into two distinct morphological groups on the basis of body size. Previous studies have demonstrated that the large-bodied *Hyalella* species can out-compete their small-bodied counterparts, except in the presence of centrarchids. Size-biased predation of the larger variant allows the small-bodied species to persist in habitats with predatory fish.

The Pleistocene glaciations effect on modern-day distributions of the *Hyalella azteca* species complex is poorly understood. The repeated advancement and retreat of the glaciers in North America has influenced the demography of freshwater taxa. As species were displaced south of the glacial limit, they persisted in ice-free refuges and re-colonized glaciated regions after the last glacial maximum. Prior to this study, it was unknown to what degree that vicariance and post-glacial dispersal have had in shaping the distribution of the *Hyalella* species complex.

In this study, the principal aim was to analyze biogeographical patterns within the *Hyalella* species complex from the spatial distribution of genetic variation. A combination of population genetics and phylogenetic approaches were applied to the cytochrome *c* oxidase 1, 16S, 28S and the first internal transcribed spacer genes; specifically, nested clade analysis, Tajima's D, Fu's F, Ramos-Onsins & Rozas R_2, mismatch distributions, analysis of molecular variance, spatial analysis of molecular variance and isolation by distance curves. These methods were used to infer demographic patterns (isolation by distance, long distance dispersal, range expansions and fragmentation) and refugial origins of *Hyalella* lineages.

This study has helped disentangle the demographic history and evolutionary relationships among a genetically divergent group of cryptic amphipods. Not only has *Hyalella azteca* been profoundly affected by the Pleistocene glaciations, but the distributions of its lineages are being influenced by both current and historical ecological pressures. The ranges of large and small-bodied morphotypes in unglaciated regions have likely been shaped by the presence of centrarchids. Small-bodied *Hyalella* were predominately restricted east of the continental divide and dispersed into the Great Lakes region from Atlantic or Mississippian refugia. Large-bodied *Hyalella* were distributed throughout North America and re-colonized glaciated regions from Mississippian, Missourian and Beringian refugia. The Great Lakes region was colonized by several species of *Hyalella* with habitats possessing sympatric lineages. The degree of spatial and genetic variation within the *Hyalella azteca* species complex demonstrates the need to characterize hidden diversity within North American fauna before it is lost to extinction or climate change.

Acknowledgements

I would like to thank my supervisor Dr. Jonathan D.S. Witt for the extraordinary amount of patience and guidance during the completion of this work. To Dr. John C. Semple, thank you for agreeing to be my co-supervisor. Furthermore, thank you to Dr. Kirsten M. Müller and Dr. Brian Dixon for your advice and participation in my advisory committee.

To my fellow field workers, Susan Usjak, Martin Brummell, Nick Jeffery, Fatima Mitterboeck, David Arnott and Joyce Lee, thank you for helping me during the long drives and even longer sampling days. To my fellow lab mates over the years, Dan Allen, Jessica Leung, Michelle D'souza, Patrick Pathammavong, Evelyn Worthington, thank you for your assistance and friendship. A special thanks to Terin Robinson for your constant support, along with the Dixon lab for your lab space and assistance in cloning the ITS1 rRNA gene (chapter 5). To Dr. Crease, Dr. Brown and Dr. Müller, thank you for helping with the analysis of the ITS1 rRNA gene.

Thank you to my family, my late father Eugene and my mother Nicole, my brother Steven and my sister Kristy. Thank you for your support and encouragement. To Natalie, none of this would have been possible without you.

This research was supported by NSERC (National Science and Engineering Research Council) funding to J.D.S.W., and NSTP (Northern Scientific Training Program) to M.J.H.

Table of Contents

AUTHOR'S DECLARATION ii

Abstract iii

Acknowledgements v

List of Figures xii

List of Tables xx

Chapter 1 1
1.1 General Introduction 1
1.1.1 Understanding organismal variation in the context of evolutionary history 1
1.1.2 How do we organize species? 2
1.1.3 Biological species concept and the Modern Synthesis 4
1.1.4 The genotypic and phylogenetic species concepts 6
1.1.5 Inferring species trees from gene trees 7
1.1.6 Phylogeographic applications of the barcoding gene 9
1.1.7 Phylogeographic inference using multiple markers 12
1.2 The Pleistocene Glaciations 16
1.2.1 Milankovitch astronomical cycles 18
1.2.2 Reconstructing the ecological history of the Pleistocene 19
1.3 Glacial refugia 21
1.4 Characteristics of *Hyalella 'azteca'* 24
1.4.1 Species delineation within *Hyalella 'azteca'* complex 26
1.5 Objectives 34

Chapter 2 38
2.1 INTRODUCTION 38
2.1.1 Size-based predation of *Hyalella 'azteca'* 38
2.1.2 Mitochondrial molecular marker for phylogenetic inference 39
2.1.3 Congruence among genomic and mitochondrial regions 42
2.1.4 Objectives 43
2.2 MATERIALS AND METHODS 45
2.2.1 COI Sequence Naming Convention 45
2.2.2 Field Collection 45
2.2.3 DNA extraction and amplification 46
2.2.4 Model of molecular evolution 48
2.2.5 COI variation among *Hyalella 'azteca'* species 48
2.2.6 Amplification of 16S and 28S genes 49
2.2.7 Tests for congruence 51
2.2.8 Tests for molecular clocks 51
2.3 RESULTS 52
2.3.1 Field Collection of *Hyalella 'azteca'* 52
2.3.2 Sequence alignment and model selection 58

2.3.3 Initial phenetic analysis and species designations 58
2.3.4 COI sequence variation....61
2.3.5 Neighbor-joining analysis of COI sequences 67
2.3.6 16S and 28S sequence variation 68
2.3.7 Maximum Likelihood and Bayesian Analyses 68
2.3.8 Congruence of COI, 16S and 28S genes in *Hyalella 'azteca'* 73
2.3.9 Molecular Clocks....75
2.4 DISCUSSION....74
2.4.1 Spatial distribution of the *Hyalella 'azteca'* species complex....76
2.4.2 Evolutionary history of size morphology 77
2.4.3 The absence of phylogenetic congruence 78

Chapter 3: Phylogeography of Large-Bodied *Hyalella* species....81
3.1 INTRODUCTION 81
3.1.1 The demographic effects of range expansions....81
3.1.2 Introduction to Nested Clade Analysis 84
3.1.3 The debate surrounding nested clade analysis....89
3.1.4 Tests of neutrality and population expansion 94
3.1.5 Analysis of molecular variance....95
3.1.6 Structural analysis of molecular variance....96
3.1.7 Objectives 96
3.2 MATERIALS AND METHODS....100
3.2.1 Partitioning of species 1 haplotypes 100
3.2.2 Nested Clade Analysis....101
3.2.3 Population genetics and mismatch distributions....103
3.2.4 Hierarchical Analysis of Molecular Variance....104
3.3 RESULTS: SPECIES 1....104
3.3.1 Species 1 lineages 104
3.3.2 Phylogeographic analysis of species 1 lineage A 108
3.3.3 Outgroup analysis of species 1 lineage A....113
3.3.4 Nested clade analysis of species 1 lineage A....113
3.3.5 Mismatch distribution and population genetics of COI haplotypes within species 1 lineage A....117
3.3.6 Analysis of molecular variance of COI haplotypes within species 1 lineage A 119
3.3.7 Phylogeographic analysis of species 1 lineage B 122
3.3.8 Outgroup Analysis of species 1 lineage B....122
3.3.9 Nested Clade Analysis of species 1 lineage....125
3.3.10 Mismatch distribution and population genetics of species 1 lineage B....126
3.3.11 Analysis of molecular variance of species 1 lineage B....128
3.3.12 Phylogeography of species 1 lineage C 129
3.3.13 Nested clade analysis of species 1 lineage C....129
3.3.14 Mismatch distribution and population genetics of species 1 lineage C....130
3.3.15 Analysis of molecular variance of species 1 lineage C....130
3.3.16 Phylogeography of species 1 lineage D....133
3.3.17 Outgroup Analysis of species 1 lineage D....133
3.3.18 Nested Clade Analysis of species 1 lineage D....135

3.3.19 Mismatch distribution and population genetics of species 1 lineage D.....135
3.3.20 Nested COI haplotype network of species 1 lineage E.....137
3.3.21 Outgroup analysis of species 1 lineage E 137
3.3.22 Nested Clade Analysis of species 1 lineage E 138
3.3.23 Mismatch distribution and population genetics species 1 lineage E.....140
3.3.24 Analysis of molecular variance of species 1 lineage E.....142
3.3.25 Phylogeography of species 1 lineage F.....143
3.3.26 Outgroup analysis of species 1 lineage F.....144
3.3.27 Nested clade analysis of species 1 lineage F 144
3.3.28 Mismatch distribution and population genetics of species 1 lineage F 146
3.3.29 Analysis of molecular variance for species 1 lineage F.....148
3.3.30 Phylogeography of species 1 lineage G 149
3.3.31 Outgroup analysis of species 1 lineage G 149
3.3.32 Nested clade analysis of species 1 lineage G.....151
3.3.33 Mismatch distribution and population genetics of species 1 lineage G.....153
3.3.34 Analysis of molecular variance for species 1 lineage G 153
3.3.35 Phylogeography of species 1 lineage H 156
3.3.36 Outgroup analysis of species 1 lineage H 156
3.3.37 Nested clade analysis of species 1 lineage H.....157
3.3.38 Mismatch distribution of species 1 lineage H.....161
3.3.39 Analysis of molecular variance for species 1 lineage H 161
3.3.40 Phylogeography of species 1 lineage I.....165
3.3.41 Outgroup analysis of species 1 lineage I.....166
3.3.42 Nested clade analysis of species 1 lineage I 168
3.3.43 Mismatch distribution and population genetics of species 1 lineage I 169
3.3.44 Analysis of molecular variance for species 1 lineage I.....171
3.4 RESULTS: SPECIES 2.....173
3.4.1 Phylogeography of species 2 173
3.4.2 Outgroup analysis of species 2 174
3.4.3 Nested Clade Analysis of COI haplotypes within species 2.....177
3.4.4 Nested clade analysis of clade 5-1 within species 2 180
3.4.5 Nested clade analysis of clade 5-2 within species 2 185
3.4.6 Nested clade analysis of clade 5-3 within species 2 190
3.4.7 Mismatch distribution and population genetics of species 2 194
3.4.8 Analysis of molecular variance of species 2.....194
3.5 RESULTS: SPECIES 7.....197
3.5.1 Phylogeography of species 7 197
3.5.2 Outgroup analysis of species 7 198
3.5.3 Nested clade analysis of species 7 198
3.5.4 Mismatch distribution and population genetics of species 7 206
3.5.5 Analysis of molecular variance of species 7.....208
3.6 RESULTS: SPECIES 8.....209
3.6.1 Phylogeography of species 8 209
3.6.2 Outgroup analysis of species 8 209
3.6.3 Nested clade analysis of species 8 210
3.6.4 Mismatch distribution and population genetics of species 8 211

3.6.5 Analysis of molecular variance of species 8 215
3.7 DISCUSSION 216
3.7.1 Spatial structure among large-bodied *Hyalella* species 217
3.7.2 Genetic patterns within large-bodied *Hyalella* species 220
3.7.3 Allele surfing within expanding ranges 221
3.7.4 Long distance dispersal among large-bodied *Hyalella* species 222
3.7.5 *Hyalella 'azteca'* suture zones 224
3.7.6 Sectoring within species 1 lineage A 225
3.7.7 Overview of species 1 refugia 226
3.7.8 Species 1 lineage B 227
3.7.9 Species 1 lineage C 228
3.7.10 Species 1 lineage D 229
3.7.11 Species 1 lineage E 229
3.7.12 Species 1 lineage F 230
3.7.13 Species 1 lineage G 231
3.7.14 Species 1 lineage H 232
3.7.15 Species 1 lineage I 234
3.7.16 Dispersal of species 2 from southern refugia 234
3.7.17 Species 2 genetic and spatial patterns 236
3.7.18 Repeated expansions within clade 5-2 (species 2) from a Mississippian refugium 237
3.7.19 Species 7 genetic and spatial patterns 238
3.7.20 Species 8 genetic and spatial patterns 239
3.8 CONCLUSIONS 240

Chapter 4: Phylogeography of Small-Bodied *Hyalella* species 242
4.1 INTRODUCTION 242
4.2 MATERIALS AND METHODS 243
4.3 RESULTS: SPECIES 3 244
4.3.1 Overview small-bodied *Hyalella 'azteca'* species 244
4.3.2 Phylogeographic analysis of species 3 244
4.3.3 Outgroup analysis of species 3 245
4.3.4 Nested clade analysis of species 3 248
4.3.5 Mismatch distribution and population genetics of COI haplotypes within species 3 261
4.3.6 Analysis of molecular variance of species 3 263
4.4 RESULTS: SPECIES 4 264
4.4.1 Nested COI haplotype network of species 4 264
4.4.2 Outgroup analysis of species 4 265
4.4.3 Nested clade analysis of species 4 266
4.4.4 Mismatch distribution and population genetics of species 4 268
4.4.5 Analysis of molecular variance of species 4 270
4.5 RESULTS: SPECIES 5 271
4.5.1 Nested COI haplotype network of species 5 271
4.5.2 Outgroup analysis of species 5 272
4.5.3 Nested clade analysis of species 5 273
4.5.4 Mismatch distributions and population genetics of species 5 281

4.5.5 Analysis of molecular variance of species 5 lineage A 283
4.6 RESULTS: SPECIES 6....284
4.6.1 Nested haplotype network of species 6....284
4.6.2 Outgroup analysis of species 6 284
4.6.3 Nested clade analysis of species 6 clade 6-1 285
4.6.4 Nested clade analysis of species 6 clade 5-1 287
4.6.5 Nested clade analysis of species 6 clade 5-2 291
4.6.6 Nested Clade Analysis of species 6 clade 5-3 292
4.6.7 Mismatch distribution and population genetics of species 6 297
4.6.8 Analysis of molecular variance of species 6....297
4.7 RESULTS: SPECIES 9....300
4.7.1 Nested haplotype network of species 9....300
4.7.2 Outgroup analysis of species 9 300
4.7.3 The nested clade analysis of species 9....301
4.7.4 The mismatch distribution and population genetics of species 9 303
4.7.5 Analysis of molecular variance of species 9....303
4.8 DISCUSSION....306
4.8.1 Range expansions of the widely distributed small-bodied species into the Great Lakes region 306
4.8.2 Genetic patterns within the small-bodied species....307
4.8.3 Species 3 genetic and spatial patterns....308
4.8.4 Species 4 genetic and spatial patterns....310
4.8.5 Species 5 genetic and spatial patterns....311
4.8.6 Species 6 genetic and spatial patterns....312
4.8.7 Species 9 genetic and spatial patterns....314
4.9 CONCLUSIONS....314

Chapter 5: Comparison between large and small-bodied *Hyalella* species....315

5.1 INTRODUCTION 315
5.2 MATERIALS AND METHODS....317
5.3 RESULTS 317
5.3.1 Large-bodied *Hyalella* isolation by distance curves....318
5.3.2 Small-bodied *Hyalella* isolation by distance curves....323
5.4 DISCUSSION....329
5.4.1 Comparison of large and small-bodied *Hyalella* distributions 331

Chapter 6: Phylogeographic Analysis of the *Hyalella* First Internal Transcribed Spacer 334
6.1 INTRODUCTION 334
6.1.1 Concerted evolution of multigene families....335
6.1.2 Three-dimensional structure of rRNA sequences and analysis of covariation 336
6.1.3 Objectives 339
6.2 MATERIALS AND METHODS....339
6.2.1 PCR amplification....339
6.2.2 Cloning of ITS1 340

6.2.3 Structural alignment and domain analysis 341
6.2.4 Folding of rRNA structures 342
6.2.5 ITS1 rRNA gene network analysis 344
6.2.6 ITS1 rRNA gene analysis of molecular variance 344
6.3 RESULTS 345
6.3.1 Structure of domain 1 346
6.3.2 Structure of domain II 349
6.3.3 Structure of domain III 349
6.3.4 Structure of domain IV 352
6.3.5 Structure of domain V 352
6.3.6 Structure of domain VI 357
6.3.7 Structure of domain VII 357
6.3.8 Structure of domain VIII 358
6.3.9 Genetic patterns of the ITS1 gene within species 1 and 2 364
6.3.10 Genetic patterns of the ITS1 gene within species 3 370
6.3.11 Genetic patterns of the ITS1 gene within species 4 371
6.3.12 Genetic patterns of the ITS1 gene within species 5 377
6.3.13 Folding structures of the ITS1 gene within species 6 and 9 387
6.3.14 Genetic patterns of the ITS1 gene within species 6 and 9 395
6.3.15 Genetic patterns of the ITS1 gene within species 7 398
6.3.16 Genetic patterns of the ITS1 gene within species 8 401
6.4 DISCUSSION 403
6.4.1 Constrained ITS1 variation within species 1 and 2 404
6.4.2 Isolation-by-distance and gene flow within species 3 405
6.4.3 Potential hybridization among species 5 lineages 406
6.4.4 Incomplete lineage sorting among species 6 and 9 407
6.5 CONCLUSIONS 408

Chapter 7: Comparative Phylogeography of Small and Large-Bodied *Hyalella* Species 409
7.1 Conclusions regarding the comparison of small and large-bodied *Hyalella* species 409

REFERENCES 420

APPENDIX A 454

APPENDIX B 474

APPENDIX C 534

List of Figures

Figure 2.1	**Map showing the distribution of *Hyalella* large-bodied species in North America**	**53**
Figure 2.2	**Map showing the distribution of *Hyalella* small-bodied species in North America**	**54**
Figure 2.3	**Neighbor-joining tree of representative *Hyalella* COI Haplotypes**	**60**
Figure 2.4	**Maximum Likelihood tree of COI haplotypes**	**70**
Figure 2.5	**Maximum Likelihood tree of 16S haplotypes**	**71**
Figure 2.6	**Maximum Likelihood tree of 28S haplotypes**	**72**
Figure 2.7	**Maximum Likelihood tree of concatenated COI/16S/28S Haplotypes**	**74**
Figure 3.1	**Schematic of a Nested Clade Analysis**	**87**
Figure 3.2	**Criterion to break loops in a Nested Clade Analysis**	**93**
Figure 3.3	**COI Neighbor-Joining tree of representative species 1 lineages**	**107**
Figure 3.4	**Map showing the geographical distribution of species 1 lineage A clades 4-1 and 2-6**	**109**
Figure 3.5	**Map showing the geographical distribution of species 1 lineage A clade 4-2**	**112**
Figure 3.6	**Mismatch distribution of species 1 lineage A**	**118**

Figure 3.7	**Map showing the geographical distributions of populations within the 5-group SAMOVA**	**121**
Figure 3.8	**Map showing the geographical distribution of species 1 lineage B**	**124**
Figure 3.9	**Mismatch Distribution of species 1 lineage B**	**127**
Figure 3.10	**Map showing the geographical distribution of species 1 lineage C**	**131**
Figure 3.11	**Mismatch distribution of species 1 lineage C**	**132**
Figure 3.12	**Map showing the spatial distribution of species 1 lineage D**	**134**
Figure 3.13	**Mismatch distribution of species 1 lineage D**	**136**
Figure 3.14	**Map showing the geographical distribution of species 1 lineage E**	**139**
Figure 3.15	**Mismatch distribution of species 1 lineage E**	**141**
Figure 3.16	**Map showing the geographical distribution of species 1 lineage F**	**145**
Figure 3.17	**Mismatch distribution of species 1 lineage F**	**147**
Figure 3.18	**Map showing the geographical distributions of species 1 lineage G**	**150**
Figure 3.19	**Mismatch distribution of species 1 lineage G**	**154**
Figure 3.20	**Map showing the geographical distribution of species 1 lineage H**	**159-160**
Figure 3.21	**Mismatch distribution of species 1 lineage H**	**162**
Figure 3.22	**Map showing the geographical distributions of populations within the 3-group SAMOVA of species 1 lineage H**	**164**
Figure 3.23	**Map showing the geographical distribution of species 1 lineage I**	**167**
Figure 3.24	**Mismatch distribution of species 1 lineage I**	**170**

Figure 3.25 **Map showing the geographical distributions of the groups within the 2-group SAMOVA for species 1 lineage I** **172**

Figure 3.26 **Map showing the distribution of 3-step clades within species 2** **176**

Figure 3.27 **Map showing the geographical distribution of clade 4-1 and 4-3 in species 2** **183**

Figure 3.28 **Map showing geographical distribution of clade 4-2 in species 2** **184**

Figure 3.29 **Map showing geographical distribution of clade 5-2 in species 2** **189**

Figure 3.30 **Map showing geographical distribution of clade 5-3 in species 2** **193**

Figure 3.31 **Mismatch distribution of species 2** **195**

Figure 3.32 **Map showing geographical distribution of species 7 clades 4-1 and 4-2** **204**

Figure 3.33 **Map showing geographical distribution of clade 4-3 in species 7** **205**

Figure 3.34 **Mismatch distribution of species 7 COI haplotypes** **207**

Figure 3.35 **Map showing geographical distribution of clades 4-1, 4-2 and 4-3 of species 8** **212**

Figure 3.36 **Map showing geographical distribution of species 8 clades 4-4, 4-5 and 4-6** **213**

Figure 3.37 **Mismatch distribution of species 8 COI haplotypes** **214**

Figure 4.1 **Map showing the geographical distribution of species 3** **247**

Figure 4.2 **Map showing the geographical distribution of species 3 clade 5-1** **252**

Figure 4.3 **Map showing the geographical distribution of species 3 clade 4-3** **255**

Figure 4.4	**Map showing the distribution of species 3 clade 3-12**	**256**
Figure 4.5	**Map showing the distribution of species 3 clades 3-13, 3-14, 3-16 and 3-17**	**257**
Figure 4.6	**Map showing the distribution of species 3 clade 3-15**	**258**
Figure 4.7	**Mismatch distribution of species 3**	**262**
Figure 4.8	**Map showing distribution of species 4**	**267**
Figure 4.9	**Species 4 lineage A mismatch distribution**	**269**
Figure 4.10	**Map showing distribution of species 5 lineage A**	**276**
Figure 4.11	**Map showing distribution of species 5 lineage B**	**279**
Figure 4.12	**Map showing distribution of species 5 lineages C, D and E**	**280**
Figure 4.13	**Mismatch distributions of species 5**	**282**
Figure 4.14	**Map showing distribution of species 6**	**286**
Figure 4.15	**Map showing distribution of species 6 clade 4-1**	**294**
Figure 4.16	**Map showing distribution of species 6 clades 4-2 and 4-3**	**295**
Figure 4.17	**Map showing distribution of species 6 clades 5-2 and 5-3**	**296**
Figure 4.18	**Mismatch distribution of species 6**	**299**
Figure 4.19	**Map showing distribution of species 9**	**302**
Figure 4.20	**Mismatch distribution of species 9 haplotypes**	**304**
Figure 5.1	**Species 1 lineage A isolation by distance curve**	**320**
Figure 5.2	**Species 1 lineage H isolation by distance curve**	**321**
Figure 5.3	**Species 1 lineage I isolation by distance curve**	**322**

Figure 5.4	**Species 2 isolation by distance curve**	**324**
Figure 5.5	**Species 7 isolation by distance curve**	**325**
Figure 5.6	**Species 3 isolation by distance curve**	**326**
Figure 5.7	**Species 5 lineage A isolation by distance curve**	**327**
Figure 5.8	**Species 6 isolation by distance curve**	**328**
Figure 5.9	**Reduced major axis regressions of the *Hyalella* isolation by distance curves**	**330**
Figure 6.1	**Alignment of *Hyalella* ITS1 according to rRNA structure**	**347**
Figure 6.2	**Structural alignment of domain I in the ITS1 rRNA gene**	**348**
Figure 6.3	**Structural alignment of domain III in the ITS1 rRNA gene**	**350**
Figure 6.4	**Structural alignment of domain IV in the ITS1 rRNA gene**	**353**
Figure 6.5	**Structural alignment of domain V in the ITS1 rRNA gene**	**355**
Figure 6.6	**Structural alignment of domain VI in the ITS1 rRNA gene**	**359**
Figure 6.7	**Structural alignment of domain VII in the ITS1 rRNA gene**	**360**
Figure 6.8	**Structural alignment of domain VIII in the ITS1 rRNA gene**	**362**
Figure 6.9	**Structure (open configuration) of the ITS1 consensus sequence in species 1 and 2**	**366**
Figure 6.10	**Structure (closed configuration) of the ITS1 consensus sequence in species 1 and 2**	**367**
Figure 6.11	**ITS1 rRNA gene network of species 1 and 2**	**368**

Figure 6.12 ITS1 rRNA gene network of species 1 and 2 with respect to faunal provinces 369

Figure 6.13 Structure of the species 3 ITS1 consensus sequence 372

Figure 6.14 Species 3 ITS1 rRNA gene network according to its spatial distribution 373

Figure 6.15 Structure of the species 4 ITS1 consensus sequence 374

Figure 6.16 Species 4 ITS1 rRNA gene network with respect to COI lineages 375

Figure 6.17 Species 4 ITS1 rRNA gene network with respect to faunal province 376

Figure 6.18 Structure of the ITS1 variant 5-1 consensus sequence in species 5 379

Figure 6.19 Structure of the ITS1 variant 5-2 consensus sequence in species 5 380

Figure 6.20 Structure of the ITS1 variant 5-3 consensus sequence in species 5 381

Figure 6.21 Structure of the ITS1 variant 5-4 consensus sequence in species 5 382

Figure 6.22 Structure of the ITS1 variant 5-5 consensus sequence in species 5 383

Figure 6.23 Alternative folding of the ITS1 variant 5-5 consensus sequence in species 5 384

Figure 6.24 Species 5 ITS1 rRNA gene network with respect to ITS1 structural variants 385

Figure 6.25 Species 5 ITS1 rRNA gene network with respect to COI lineage 386

Figure 6.26 Structure of the ITS1 variant 6-1 consensus sequence in species 6 388

Figure 6.27 Structure of the ITS1 variant 6-2 consensus sequence in species 6 389

Figure 6.28	**Structure of the ITS1 variant 6-3 consensus sequence in species 6**	**390**
Figure 6.29	**Structure of the ITS1 variant 6-4 consensus sequence in species 6**	**391**
Figure 6.30	**Structure of the ITS1 variant 6-5 consensus sequence in species 6**	**392**
Figure 6.31	**Structure of the ITS1 variant 6-6 consensus sequence in species 6**	**393**
Figure 6.32	**Structure of the species 9 ITS1 consensus sequence**	**394**
Figure 6.33	**ITS1 rRNA gene network of species 6 and 9 with respect to ITS1 structural variant**	**396**
Figure 6.34	**ITS1 rRNA gene network of species 6 and 9 with respect to faunal province**	**397**
Figure 6.35	**Structure of the species 7 ITS1 consensus sequence**	**399**
Figure 6.36	**ITS1 rRNA gene network of species 7 with respect to faunal province**	**400**
Figure 6.37	**Structure of the species 8 ITS1 consensus sequence**	**402**
Figure C.1	**The nested COI haplotype network of species 1 lineage A**	**534-535**
Figure C.2	**Nested COI Haplotype Network of species 1 lineage B**	**536**
Figure C.3	**Nested COI haplotype network of species 1 lineage C**	**537**
Figure C.4	**Nested statistical parsimony network of COI haplotypes in species 1 lineage D**	**538**
Figure C.5	**Nested COI haplotype network of species 1 lineage E**	**539**
Figure C.6	**Nested COI haplotype network of species 1 lineage F**	**540**
Figure C.7	**Nested COI Haplotype Network of species 1 lineage G**	**541**

Figure C.8 Nested COI haplotype network of species 1 lineage H 542-543

Figure C.9 Nested COI haplotype network of species 1 lineage I 544

Figure C.10 Nested COI haplotype network of species 2 545-548

Figure C.11 Nested COI haplotype network of species 7 549-551

Figure C.12 Nested COI haplotype network of species 8 552-554

Figure C.13 Nested COI haplotype network of species 3 555-558

Figure C.14 Nested COI haplotype network of species 4 559

Figure C.15 Nested COI haplotype network of species 5 lineage A 560

Figure C.16 Nested COI haplotype network of species 5 lineage B 561

Figure C.17 Nested COI haplotype network of species 5 lineages C, D and E 562

Figure C.18 Nested COI haplotype network of species 6 563-566

Figure C.19 Nested COI haplotype network of species 9 567

List of Tables

Table 1.1	**A summary of the historical taxonomy of the *Hyalella azetca* species complex**	**29**
Table 1.2	**Species designations of *Hyalella 'azteca'***	**32**
Table 2.1	**Primer combinations**	**47**
Table 2.2	**Sample sites per province/territory/state**	**55**
Table 2.3	**Sample sites, sample size and diversity of COI sequences for each *Hyalella 'azteca'* species**	**63**
Table 2.4	**Base frequencies of COI sequences by *Hyalella 'azteca'* species**	**63**
Table 2.5	**Average COI nucleotide diversity within each species by codon Position**	**64**
Table 2.6	**Average group mean nucleotide distances among and within *Hyalella 'azteca'* species**	**65**
Table 2.7	**Average group mean amino acid distances among and within *Hyalella 'azteca'* species**	**66**
Table 2.8	**Sequence characteristics of COI, 16S and 28S**	**68**
Table 2.9	**Substitution model parameters estimated for maximum likelihood**	**69**
Table 2.10	**Molecular clock calculations of COI, 16S and 28S**	**75**
Table 3.1	**Population, sample and genetic statistics for each large-bodied species (lineage) of *Hyalella 'azteca'***	**106**
Table 3.2	**Nested Clade Analysis and Inferences of species 1 lineage A**	**114**

Table 3.3 **AMOVA and SAMOVA of COI haplotypes for species 1 lineage A** **120**

Table 3.4 **Nested Clade Analysis of COI haplotypes in species 1 lineage B** **126**

Table 3.5 **AMOVA and SAMOVA of species 1 lineage B** **129**

Table 3.6 **Nested Clade Analysis of COI haplotypes for species 1 lineage D** **135**

Table 3.7 **Nested Clade Analysis of species 1 lineage E** **140**

Table 3.8 **AMOVA and SAMOVA of species 1 lineage E** **143**

Table 3.9 **Nested Clade Analysis of species 1 lineage F** **146**

Table 3.10 **AMOVA and SAMOVA of species 1 lineage F** **148**

Table 3.11 **Nested Clade Analysis of species 1 lineage G** **152**

Table 3.12 **AMOVA and SAMOVA of species 1 lineage G** **155**

Table 3.13 **Nested Clade Analysis of species 1 lineage H** **161**

Table 3.14 **AMOVA and SAMOVA of species 1 lineage H** **163**

Table 3.15 **Nested Clade Analysis of species 1 lineage I** **169**

Table 3.16 **AMOVA and SAMOVA for species 1 lineage I** **171**

Table 3.17 **Nested clade analysis of species 2** **178-179**

Table 3.18 **AMOVA and SAMOVA of species 2** **196**

Table 3.19 **Nested clade analysis of species 7 COI haplotypes** **199**

Table 3.20 **AMOVA and SAMOVA of species 7** **208**

Table 3.21 **Nested clade analysis of species 8 COI haplotypes** **210**

Table 3.22 **AMOVA and SAMOVA of species 8** **215**

Table 3.23 **Summary of nested clade analysis inferences** **221**

Table 4.1	**Population, sample and genetic statistics for each small-bodied species (lineage) of *Hyalella 'azteca'***	**246**
Table 4.2	**Nested clade analysis with inference key for species 3**	**251**
Table 4.3	**AMOVA and SAMOVA of species 3**	**263**
Table 4.4	**Nested clade analysis with inference key of species 4**	**266**
Table 4.5	**AMOVA and SAMOVA of species 4**	**270**
Table 4.6	**Nested clade analysis with inference key of species 5 lineage A**	**273**
Table 4.7	**Nested clade analysis with inference key of species 5 lineage B**	**278**
Table 4.8	**AMOVA and SAMOVA of species 5 lineage A**	**283**
Table 4.9	**Nested clade analysis with inference key of species 6**	**289**
Table 4.10	**AMOVA and SAMOVA of species 6**	**298**
Table 4.11	**Nested clade analysis with inference key of species 9**	**303**
Table 4.12	**AMOVA and SAMOVA of species 9**	**305**
Table 4.13	**Summary of nested clade analysis inferences for small-bodied Lineages**	**308**
Table 5.1	**Summary of *Hyalella* Isolation by distance curves**	**319**
Table 6.1	**Summary of ITS1 sequence variation in *Hyalella 'azteca'* species**	**345**
Table 6.2	**Comparative chi-squared statistics of domain III**	**351**
Table 6.3	**Comparative chi-squared statistics of domain IV**	**354**
Table 6.4	**Comparative chi-squared statistics of domain V**	**356**
Table 6.5	**Comparative chi-squared statistics of domain VI**	**361**

Table 6.6	**Comparative chi-squared statistics of domain VII**	**361**
Table 6.7	**Comparative chi-squared statistics of domain VIII**	**363**
Table 6.8	**AMOVA of the ITS1 gene in *Hyalella 'azteca'***	**365**
Table A.1	**List of locations and sample data**	**454-473**
Table B.1	**Species 1 lineage A Nested Clade Analysis**	**474-477**
Table B.2	**Species 1 lineage B Nested Clade Analysis**	**478-480**
Table B.3	**Species 1 lineage D Nested Clade Analysis**	**481-482**
Table B.4	**Species 1 lineage E Nested Clade Analysis**	**483-484**
Table B.5	**Species 1 lineage F Nested Clade Analysis**	**485-486**
Table B.6	**Species 1 lineage G Nested Clade Analysis**	**487-489**
Table B.7	**Species 1 lineage H Nested Clade Analysis**	**490-492**
Table B.8	**Species 1 lineage I Nested Clade Analysis**	**493-495**
Table B.9	**Species 2 Nested Clade Analysis**	**496-504**
Table B.10	**Species 3 Nested Clade Analysis**	**505-510**
Table B.11	**Species 4 Nested Clade Analysis**	**511-512**
Table B.12	**Species 5 lineage A Nested Clade Analysis**	**513-516**
Table B.13	**Species 6 Nested Clade Analysis**	**517-523**
Table B.14	**Species 7 Nested Clade Analysis**	**524-528**
Table B.15	**Species 8 Nested Clade Analysis**	**529-531**
Table B.16	**Species 9 Nested Clade Analysis**	**532-533**

Chapter 1

General Introduction

1.1.1 Understanding organismal variation in the context of evolutionary history

The Earth's biota is comprised of over 8.7 million species that trace back evolutionary lines over billions of years (Costello *et al.,* 2013). Quantifying the number of current and extinct taxa is a monumental undertaking and only the preliminary step toward understanding the underlying dynamics of speciation, extinction and patterns of variability. Organismal variation is affected by the complex interaction of nuclear gene diversity and the environment.

African lake cichlids are renowned for marked morphological divergence that is a result of recent speciation events in the Great Lakes of the East African Rift Valley. The most notable of these, Lake Victoria, was mostly or completely desiccated until recently (~14,000 years), and now hosts a superflock of diverse lake cichlids species that are distinct in both morphology and niche specialization. Populations in Lake Malawi, on the other hand, were shaped by repeated invasions and hybridization among riverine cichlids (Joyce *et al.,* 2011). Lake Victoria was seeded by the surrounding lake systems that were historically colonized by stepping-stone events out of Lake Kivu (Verheyen *et al.,* 2003). As such, the adaptive radiations (a process of rapid diversification or speciation) that ultimately determined present-day morphological and molecular variation bear the mark of both current (niche specialization) and historical events (timing of migrations). Depending on the species, this becomes a problem of geographical scale. Studying the

endemic Midas cichlid species within crater Lake Apoyo of Nicaragua has the advantage of working within a compact microcosm (Barluenga *et al*., 2006), whereas analyzing pelagic marine invertebrates may require a global initiative. That is not to say that isolated environments can't host amazing degrees of variation, such as evident by explosions in morphological diversity in the endemic species of Lakes Titicaca and Baikal. There is no single explanation for the spatial distribution of molecular or morphological variation; the demographic history of each lineage must be considered in turn. Characterizing the forces that create such extremes in variation is of paramount importance to not only conservation biology, but to the general understanding of evolutionary mechanisms.

1.1.2 How do we organize species?

Categorizing living systems has challenged scientists since Linnaeus proposed a hierarchical relationship among the kinds of organisms. While systematists have wrestled with species concepts for nearly 250 years (de Queiroz, 2005), two of the most significant advancements came with the works of Charles Darwin and again during the Modern Synthesis of the mid 20th century. In the *Origin of Species*, Darwin challenged Natural Theology, arguing that not only were species mutable, but they transformed via evolutionary processes from common ancestry (Mayr, 2001). Central to evolutionary thought is not only the question of how to delineate species, but also to ascertain why biological and ecological variation separates into discrete groups (Mallet, 2007). The method in which we define and delimit species determines not only how we quantify

biodiversity, but also has repercussions in conservation biology as well as studies of ecology and evolution (Frankham *et al.,* 2012).

Darwin has been often criticized for his reliance on a morphological definition of species (Mallet, 1995), though he had already rejected a species definition based on interbreeding because of hybridization between Galapagos finches and a lack of sterility barriers in otherwise good species (groups that maintain their distinctiveness in sympatry). Hybridism, in Darwin's opinion, was a byproduct of evolution and not the defining essence of a species (Mallet, 2010) The causes of speciation eluded Darwin because the nature of inheritance would not be resolved until the rediscovered works of Mendel were incorporated into evolutionary theory during the Modern Synthesis (Turelli *et al.,* 2001). The crowning achievement of Modern Synthetic Theory was the characterization of organismal variation as the natural result of epistatic and pleiotropic interactions among discrete genetic entities (genes) derived from spontaneous mutation and recombination. In addition to reinforcing the significance of natural selection, Modern Synthetic Theory characterized the contribution of genetic drift on speciation. Whether through adaptive or demographic processes, cladogenesis becomes the inevitable result of gradual successive genetic changes among diverging lineages. This was the ultimate rejection of saltationary theory of extreme leaps in evolution across a single or few generations (Bateson, 2002). Though punctuated evolution in the fossil record and adaptive radiations appear rapid on a geological time scale, they still operate through the neo-Darwinian mechanisms.

1.1.3 Biological species concept and the Modern Synthesis

A significant result of the modern synthesis was the adoption of the biological species concept (BSC), popularized by Ernst Mayr (1963), whereby species are defined as actually or potentially interbreeding populations that are reproductively isolated from other similar groups (Sokal & Crovello, 1970). This departed from early attempts to address the species question, which used morphological or typological concepts to organize varieties of organisms into categories (Mayr, 1992). The BSC provided a new operational criterion for the characterization of species based on the reproductive isolation among populations. Once defined by this property, species are no longer considered arbitrary categories, but are treated as fundamental units of biological organization whose members will cohesively evolve (de Queiroz, 2005, de Queiroz 2007). By natural extension of defining species as real biological entities according to an essential quality (reproductive isolation), studies began using the species as units of analysis (Mallet, 2005). Whether species meet with the requirements of essentialism or are merely arbitrary categories (Hart, 2011), their use as a unit of analysis remains a potent tool in the study of ecological variation. Even the subdiscipline of island biogeography began with successfully characterizing the spatial distribution of ecological traits by identifying a logarithmic relationship between species density and land area on the Melanesian islands (MacArthur & Wilson, 1967). Characterizing mechanisms that operate at the species level has been an important endeavor in the study of extinction and speciation rates (Rabosky, 2009), most notably with adaptive radiations (Lieberman, 2012)

A principle component of the biological species concept is that speciation occurs mainly allopatrically, whereby geographically isolated populations develop unique mutations and allelic frequencies diverge on the basis of local adaptive forces. In the event of secondary contact the populations will maintain their reproductive isolation because of their own internal organization. The contribution of sympatric speciation (cladogenesis of a new species within the same geographical region as the ancestor) remains a fertile ground for vigorous debate (Bird *et al*., 2012). A recent study of the plant biota on Lord Howe's islands estimated that 8-13% species had resulted from sympatric speciation (Papadopulus *et al.*, 2011), though it has been criticized for its overreliance on chloroplast DNA (Coyne, 2011). The process of speciation only concerns itself with the evolution of traits that either pleiotropically or epistatically contributes to reproductive isolation (Coyne and Orr, 2004). The biological species concept has been considered the "gold standard" for species concepts (Mallet, 2010), but it suffers from some significant drawbacks. Primarily, the BSC considers species to be real and not just arbitrary categories, with intrinsic reproductive isolation from other groups as an essential quality. Any level of hybridization or introgression among lineages would undermine their classification as a species, though in the recent literature, reproductive barriers are being considered semi-permeable allowing the differentiation of species in spite of introgression (Hausdorf, 2011). Hybridization must be rare on a per individual basis, or the homogenizing influence of gene flow will erode specific characters, but on a species level, hybridization appears quite common (Mallet, 2005). Angiosperms are particularly troublesome for the BSC, because of asexual reproduction and reticulated phylogenetic nets formed from extreme polyploidization and hybridization (Soltis & Soltis, 2009). In

defense of the BSC, there are well documented cases of speciation from the evolution of intrinsic reproductive barriers, such as cyotplasmic incompatibility (Werren, 1997) and chromosomal duplication. Coyne and Orr (2004) argued that such a strict definition loses its pragmatic value and would hinder future biological research. The BSC only provided testable hypotheses for those taxa that sexually reproduce and exist sympatrically or parapatrically with their sister taxon. Otherwise, reproduce isolation must be inferred from introgression or hybridization from migration, historical interbreeding from overlapping dynamic ranges or among populations reared in a laboratory. This can be problematic when dealing with fossils or organisms that cannot be experimentally crossed. Additionally, many "good" species, i.e. populations that retain their diversity in sympatry, would be overlooked with such a strict enforcement of reproductive isolation (Coyne and Orr, 2004).

1.1.4 The genotypic and phylogenetic species concepts

The debate over the essentialism of species remains hotly contested and is likely to continue unabated, but the question remains how do we quantify species? The choice of species concept is critically important for biological conservation and ecological research as the number of current species can vary as much as 50% (Agapow *et al.*, 2004). Mallet (1995) proposed the Genotypic Cluster Species Concept (GSC) as an alternate criterion to infer reproductive isolation without the need for mating trials. In a local area, a species will form a genotypic cluster of unlinked loci with allelic frequencies in Hardy-Weinberg equilibrium, whereas allelic frequencies between species will be in Hardy-Weinberg

disequilibrium with strong heterozygote deficits. The phylogenetic species concept (PSC) defines species as reciprocally monophyletic lineages and does not require those lineages be sympatric. Instead of reproductive isolation from other groups, this concept uses the evolutionary history of traits, or gene trees of molecular markers, to infer species boundaries. The advantages of the GSC and PSC are that they can provide an operational definition to categorize species and side step the argument over the essentialism of species. The phylogenetic species concept is extremely pragmatic as it can delimit allopatric species and requires no *a priori* assumptions about population structure. As such, it has promoted a wealth of biogeographical research in widely distributed taxa (Agapow *et al*., 2004).

1.1.5 Inferring species trees from gene trees

Phylogeography, as a sub-discipline of biogeography, is the study of the spatial distribution of lineages within and among species, and seeks to characterize the relationship between genealogy and geography through time (Avise, 2009). Present day population structures are the result of complex interactions between historical demographic events and current ecological pressures. Mismatch distributions can infer both current and historic range expansions when compared against simulations using coalescent models. The discipline of phylogeography is intrinsically linked to the advances in available molecular markers. DNA sequence data coupled with new phylogenetic inference tools has revolutionized our ability to discriminate species boundaries (Bloomquist *et al.,* 2010) that were previously delineated on the basis of

morphological comparisons (Avise, 2000). Each gene tree represents its own unique history that has been potentially impacted by hybridization, incomplete linage sorting, gene duplication or lateral gene transfer.

Species trees must be cautiously inferred from the sometimes contradictory data of discordant gene trees (Slowinski & Page, 1999). The true genealogy has been especially difficult to obtain in groups that undergo rapid speciation because of short internodes within the phylogeny (Knowles & Klimov, 2011). Historically, the higher support values of multi-gene concatenated alignments were used as justification of the species' trees (Rokas *et al.*, 2003), but more recent work has suggested that these inferences can be easily misinterpreted by incomplete lineage sorting (Kubatko & Degnan, 2007). A solution to resolve discordant gene trees has been to increase the number of independent loci and determine the relative probabilities (likelihoods) of each tree given a species tree. The likelihood scores are calculated using coalescent and nucleotide substitution models and the most probable species tree can be searched using well established Bayesian or maximum likelihood algorithms (Degnan & Salter, 2005; Yang & Rannala, 2010). These Monte Carlo Markov Chain (MCMC) methods are often considered the best choices for species tree estimation (Knowles *et al.*, 2012) if the discordant gene trees are derived from incomplete lineage sorting. Species tree estimation becomes much more problematic when considering hybridization (Meng & Kubatko, 2009) or lateral gene transfer (Chung & Ané, 2011).

1.1.6 Phylogeographic applications of the barcoding gene

Mitochondrial DNA (mtDNA) has been the marker of choice (in animals) since the inception of phylogeography with studies on south-eastern pocket gophers (Avise *et al.*, 1979). As it does not undergo recombination, mtDNA is inherited intact, allowing the individual to be used as an operational taxonomic unit, which circumvents the problem of *a priori* definitions of populations. Furthermore, due to the asexual transmission of mitochondrial markers, their lineages are hierarchical and not subject to the reticulate branching pattern expected in phylogenies of recombining nuclear markers. The combination of these two properties of mtDNA, allows the application of systematics and phylogenetics, normally reserved for interspecies studies, into the realm of population genetics. Prior to this, phylogenetic reconstruction of the genetic architecture was strictly reserved for deep interspecies relationships. Out of 216 original crustacean phylogeographic studies since 1994 the majority use mitochondrial markers and a significant percentage were exclusively mitochondrial (Leese & Held, 2011).

The mitochondrial cytochrome c oxidase subunit 1, referred to as the barcoding gene, has become widely used to delimit animal species in both population genetics and phylogenetic studies (Leese & Held, 2011). Hebert *et al.*, (2003a, 2003b) proposed that a single gene can function like a barcode if sufficient numbers of base pairs are sequenced to have enough combinations of mutations to uniquely indentify every species. As a result of the degeneracy of the amino acid code, the third position of a codon is usually weakly selected, allowing for 1 out of every 3 sites in protein coding genes to accumulate mutations. A gene fragment that is 45 basepairs in length will have 15 such sites, which

can produce over a billion unique combinations, a far greater value than the number of estimated species on the planet (Hebert *et al.,* 2003a). Since a large fraction of Earth's biota remains uncharacterized, a barcoding gene has been used to facilitate attempts to quantify species variation before much of it is lost to extinction (Valentini *et al.,* 2009).

As a consequence of variation in the rates of molecular evolution, different genes can be used to construct phylogenies to analyze both shallow and deep branches. Although any DNA region can be employed, mtDNA provides estimations of population divergence without the confounding influence of recombination, which occurs in the nuclear genome (Avise, 2000). The mitochondrial genome is haploid and inherited uniparentally (maternally in most animals), as such it has an effective population size one quarter that of nuclear genes. As a result, lineage sorting within the mitochondrial genome occurs at four times the rate of the nuclear genome, but rate estimates of mitochondrial evolution have been estimated as high as 10 fold faster than single-copy nuclear DNA (Brown *et al.,* 1979; Osada & Akashi, 2012). Palumbi *et al.,* (2001) used coalescent simulations to derive a "three-times" rule, whereby a nuclear locus among species will reach reciprocal monophyly when the branch lengths of the mitochondrial markers between species are three times longer than the average intraspecific branch lengths within that species. Rates of mitogenomic evolution differ between higher order taxa, with a faster pace in snakes and lizards than birds or turtles, which may be linked to a greater frequency of speciation events (Eo & DeWoody, 2010). In the absence of gene flow, geographically isolated groups will rapidly progress to reciprocally monophyletic states (Hare, 2001). The rapid evolution of mtDNA makes it ideal for phylogeographic studies because it produces a wealth of character information for analysis. The

mitochondrial genome is duplicated hundreds to thousands of times within each cell and despite its large copy number; the prevalence of heteroplasmy (within-individual variation) in animals appears to be rare (Avise, 2009).

Sex biased gene flow is a well recognized limitation of exclusively using mitochondrial DNA (Galtier *et al.,* 2009). Petit and Excoffier (2009) argue that mitochondrial genes make a poor choice of marker for delineating closely allied species, especially if there is recent introgression. Such a marker may have less diagnostic value because the time it takes to reach reciprocal monophyly will be increased and produce discordant gene trees. When two closely related species come into secondary contact, if the colonizing species has high intraspecific gene flow among conspecific populations it will decrease the likelihood of genetic drift fixing introgressed genes. This is potentially important in species undergoing range expansions. The genetic diversity of populations along the leading front are reduced by successive bottlenecks and founder effects, which can cause an allele to 'surf' along the expansion wave resulting in high allele frequencies across a large geographic area (Klopfstein *et al*., 2006). If the expanding population encounters a similar species with imperfect reproductive barriers, then genes with sufficient intraspecific gene flow will have higher allelic frequencies to resist being lost to genetic drift. Otherwise the asymmetric introgression into the expanding species can overwhelm the native alleles and become the 'surfing allele', even if there is no selective advantage (Currat *et al.,* 2008). When applied to organisms with known sex-biased dispersal, the *Myotis* bats (male-biased gene flow) had widespread mtDNA introgression (less diagnostic) and the *Hippolais* birds (female-biased gene flow) lacked notable

mtDNA introgression (more diagnostic). As such, phylogeographic analyses must take into account the interactions among closely related species (Petit & Excoffier, 2009).

Rapid lineage sorting coupled with higher mutation rates (Zink & Barrowclough., 2008), makes mtDNA more sensitive to population subdivision than nuclear DNA. However, any phylogeny constructed solely on the basis of mtDNA may not reflect other regions of the genome nor accurately represent the species phylogeny. The characterization of the spatial distribution of genetic variation is best attempted by showing genealogical concordance between independent nuclear and mitochondrial gene trees. Naturally, this leads back to the original problems with within-species phylogenies that mtDNA circumvented. Nuclear genes within diploid organisms will undergo recombination and potentially produce reticulated phylogenies. Avise (2000) proposed using the inherent noise of nuclear phylogenies as a null expectation, thus significant departures were evidence of a phylogenetic signal.

1.1.7 Phylogeographic inference using multiple markers

Reciprocal monophyly among lineages, the defining feature of the phylogenetic species concept, is a potent line of evidence to distinguish groups of haplotypes and support species classifications. Though a powerful diagnostic criterion, determining if species represent monophyletic groups can be skewed by an over reliance on a single marker. Every gene has its own independent history and need not be representative of other areas of the genome. As more and more characters achieve reciprocal monophyly, regardless of

their own historical characteristics (mutation rate, linkage disequilibrium, selective sweeps), the confidence of delineating groups of haplotypes into distinct species increases. Avise (2009) devised a scheme consisting of four aspects for assessing genealogical concordance among unlinked loci with respect to their phylogeographic distributions. Aspect I is an intraspecific concept based on reciprocal monophyly within a single molecular marker. When placed in a biogeographic context, aspect I distinguishes phylogeographic units on the basis of concordance among nucleotide positions within a locus. Aspect II extends this to patterns among unlinked loci that confirm the same phylogeographic relationships. Breaks in the mitochondrial phylogeny are often expected to coincide with the historical barriers to gene flow between groups on the gene tree. Kuo & Avise (2005) used coalescent simulations set up to lack genetic barriers to demonstrate that the COI marker can stochastically produce deep splits in the phylogeny that would incorrectly infer cessation of gene flow. This can lead to contradictory phylogeographic groups or the improper invocation of historical events (Pleistocene glaciations) to explain an incorrect phylogeny. If the concept of aspect II is enforced, it is unlikely that a stochastic process will produce identical phylogenies in multiple unlinked markers. Concordant breaks among phylogeographic groups in several gene trees are strong evidence for historic barriers to gene flow and the delineation of phylogeographic units. Aspect III is a multispecies concept that results from a consistent trend among codistributed taxa to produce congruent breaks among unlinked loci. Such phylogeographic systems have been used to infer Pleistocene glacial refugium and to determine the shared historical effects of vicariance and dispersal (Hickerson *et al*., 2010). Areas that underwent cycles of glaciation experienced repeated recolonization by

highly mobile organisms (Alsos *et al.,* 2007). Finally, aspect IV delineates zoogeographical provinces by comparing the codistribution of gene lineages among regions. This highlights areas important for conservation, which has led to insights into the dynamics behind endemism, species persistence and range shifts (Hewitt, 2000; Soltis *et al.,* 2006).

The dates of vicariant events can be employed to determine the age of cladogenic splits of sister-taxa. The well-dated barrier of the Isthmus of Panama has been used to determine the age of the cladogenic divergences among snapping shrimp, genus *Alpheus* (Knowlton *et al.,* 1993). Sister-taxa were identified based on reproductive compatibility, Hardy-Weinberg disequilibrium at allozyme loci and mitochondrial divergences. Though isolation among groups was not simultaneous, Knowlton *et al.* (1993) estimated rates of molecular divergence. A few years later, these the estimates were refined by focusing on sister-taxa restricted to shallow mangroves (Knowlton and Weight, 1998). Caution is required in such studies, as incorrectly identifying sister-taxa or insufficient sampling may introduce significant error in estimates of evolutionary rates. Furthermore, most vicariant events are not instantaneous, which introduces variability in the dates of physical isolation. This method was revisited by Hurt *et al.,* (2009) and they suggested that coalescent-based approaches were better suited to determine COI divergence times because they can account for the variance resulting from ancestral polymorphisms. The closer a cladogenic split among sister taxa is to the vicariant event, they greater the proportion of genetic variance is explained by unsorted ancestral lineages.

Quantifying biodiversity has become a race against time, as the depredation of species diversity has accelerated on a global scale from habitat loss, climate change and invasive species (Thomas *et al.*, 2004; Vié *et al.*, 2009; Harnik *et al.*, 2012). Previous estimates of global species diversity have reached as high as 100 million species (Mora *et al.*, 2011), though current predictions are more conservative, placing the total at 5-8.7 million with 1.5 million currently described (Costello *et al.*, 2013). This figure continues to remain a hotly debated issue, as Hamilton *et al.*, (2013) most recently estimated the number of tropical arthropod species alone at around 6 million. While newly discovered species of birds and terrestrial mammals are not expected to rise significantly, arthropods (from marine, terrestrial and freshwater biomes) represent a staggering proportion of the currently undescribed taxa (Costello *et al.*, 2012). Trends in studies of species diversity have shown a linear increase in the total number of described species since the late 1800's, with only a significant drop in the rate of reporting during the two World Wars (Costello *et al.*, 2012). Interest in the study of systematics has only grown, with more people identifying species than ever before, although with a shift away from traditional taxonomy towards molecular techniques (Costello *et al.*, 2013).

Estimates of species numbers are further confounded by the existence of genetically divergent, but morphologically constrained taxa. Cryptic species or sibling species are groups that are not distinguishable on the basis of morphological characters, but are distinct species based on PSC/BSC. These complexes either possess phylogenetic characteristics not readily apparent to researchers, such as cryptic auditory (Puissant & Sueur, 2010) and chemical cues, or static morphologies enforced through selection (Smith *et al.*, 2011). The factors that result in the formation of cryptic species cannot be

understated, as they can influence groups separated by vast distances in both geography and time. *Cyclestheria hislopi*, a cryptic species complex that is the sister taxon to cladocerans, is distributed throughout three continents (Asia, Australia and South America) with lineages that likely split with the breaking up of Gondwanaland during the Cretaceous around 120 million years ago (Schwentner *et al.*, 2013). Of the ~226,000 described species of eukaryotic marine taxa, only 3% have been barcoded. Thus this group harbors a significant potential for cryptic diversity (Appeltans *et al.*, 2012). Cryptic species are not uniformly distributed across taxonomic groups or biogeographic regions (Bickford *et al.*, 2007), and the Mollusca and Crustacea are notable for this type of diversity (Appeltans *et al.*, 2012). However, when corrected for species richness and study intensity, across 2000 studies the differences in proportions of cryptic taxa among biogeographic regions and most major metazoan taxa disappeared, except for arthropoda, mammalia and amphibia, which had proportions significantly higher than the mean (Pfenninger & Schwenk, 2007). Trontelj & Fiser (2009) criticized this correction as the number of publications is not necessarily indicative of study intensity, and that the included studies are not a random sample. In both cases, arthropods remain a vastly under-reported group with the potential for significant cryptic species diversity.

1.2 The Pleistocene Glaciations

In the Northern Hemisphere, one of the most significant series of vicariant events was the Pleistocene glaciations. Massive ice sheets covered nearly one third of the land mass, destroying both terrestrial and aquatic habitats. The surviving populations persisted in

geographically isolated glacial refugia (Hewitt, 2000). The glacier's effect on species biogeography was profound. Not only were they potent barriers to gene flow, but the interconnected waterways formed from their meltwaters helped shape the current broad codistribution of freshwater North American taxa (Bernatchez & Wilson, 1998). The Pleistocene glaciations began 2.4 million years ago at the onset of the Holocene epoch (Webb & Bartlein, 1992). Four major glacial advances (Nebraskan, Kansan, Illinoian, and Wisconsinan) were originally recognized, but now upwards of six pre-Illinoian glacial advances have been identified (Mickelson & Colgan, 2003). Even in non-glaciated regions, massive climate variability altered the precipitation regimes of many ecological zones. The freezing and melting of sea ice resulted in global fluctuations in sea levels, with a reduction of 100 meters at the Wisconsin maximum. At this time, North America was affected by two major ice sheets. The Cordilleran ice sheet in the West stretched from Alaska, south along the Rocky Mountains, and into northern Washington. The larger Laurentide ice sheet covered Eastern and Central Canada, as well as the Northern USA (Dyke & Prest, 1987). The major dome of the Laurentide ice sheet is estimated to have been five kilometers thick, and it caused substantial down-warping of the Earth's crust in the Hudson Bay area (Pielou, 1991). The principle of uniformitarianism states that the geological processes that occur today are the same as those that operated in the past. Geologists have used this principle to study ice and rock cores to assess historical climatic changes by analyzing the stratified layers (Pielou, 1991). Over the past billion years there have been three glacial ages and we are currently 2 million years into the most recent one.

1.2.1 Milankovitch astronomical cycles

While glacial ages can last millions of years, the current one has been punctuated by a 100,000 year climatic cycle for the past 900,000 years (Raymo, 1997). This period oscillates between glacial periods lasting 60,000 to 90,000 years, and interglacials lasting 40,000 to 10,000 years respectively. The timing of glacial advances and retreats are calibrated using the decay of oxygen isotopes in benthic and planktonic foraminifera from deep sea core samples (Shackleton, 1990). Milankovitch astronomical cycles are climatic oscillations, named after Milutin Milanković, which refer specifically to changes in ice volume in the northern hemisphere resulting from axial tilt and precession of the equinoxes (Ruddiman, 2006). The Earth's axial tilt shifts between 22.1º and 24.5º over a 41,000 year period. The precession of the equinoxes alters the orientation of the Earth's axis as it switches from Polaris to Vega, shifting the date that the Earth is closest to the Sun with a 22,000 year periodicity (Hays *et al.*, 1976). These are coupled with a 105,000 year cycle in orbital eccentricity, whereby the Earth slips between two elliptical orbits, the Aphelion and the Perihelion. The orbital oscillations are determined by the degree of eccentricity, or the Earth's departure from a circular orbit, which results from interactions between planetary bodies. With increased eccentricity, the Earth receives considerably less solar radiation at points furthest from the sun (Aphelion).

The recent development of the 100,000 year climatic periodicity as a consistent pattern appeared less than a million years ago; nearly 2 million years after the Pleistocene glaciations had started (Imbrie *et al.*, 1993). Variations in global ice volumes behave nonlinearly (in a "saw-toothed" pattern), with the slow accumulation of ice followed by

rapid deglaciations. Though the length of the two systems is nearly identical, the degree of variation in ice accumulation caused by changes in the Earth's orbital eccentricity is insufficient to account for the 100,000 year periodicity of the Pleistocene glaciations (Imbrie *et al.,* 1993). The debate in the literature is far from settled, but Imbrie *et al.* (1993) suggests that the system results from a feedback mechanism between ice elevation and ice accumulation rate. As the amount of solar radiation decreases, incomplete summer melting caused the glaciers to form from the annual accumulation of ice and snow. The glaciers themselves act as a positive feedback mechanism. Unlike soil or water, the glacial ice efficiently reflects solar radiation back into space and further decreases the Earth's temperature. Variability in the Earth's reflectivity is referred to as the albedo effect. Subsequently, when the planet's orbit changes back to Perihelion, (resulting in increased solar radiation), the ice begins to melt and exposes the more light-absorbent water and soil, which aides in heating the Earth (Pielou, 1991). Since the last interglacial period, the Sangamon (130 kya), the glaciers have undergone repeated advances, delays and recessions (Martinson *et al.,* 1987). The Wisconsin glaciations reached a peak 20,000 years ago and the northern hemisphere has progressively warmed as the last glaciation transitioned into the most recent interglacial (Clark *et al.,* 2009).

1.2.2 Reconstructing the ecological history of the Pleistocene

Characterizing the ecological history of the Pleistocene and the impact of the glaciations is a complex problem, and to this effort, three principle lines of evidence have been employed: macrofossils, microfossils and geographic ranges of extant species.

Macrofossils are usually hard parts of organisms that occur in great abundance and dated using C-14 radiocarbon techniques. Radiocarbon has increasingly become more accurate (especially with organisms fossilized within the past 20,000 years) with further calibration with tree-ring dating and uranium-thorium dating of corals (Stuvier *et al.,* 1998). Plants tend to leave fossilized organs (leaves, twigs, logs or fruit) that are preserved in sediments, whereas animal fossils usually are composed of tooth, bone or the hard parts of insects (especially beetle carapace). Microfossils can only be viewed through a microscope and one of the more important of these is pollen grains (Pielou, 1991). Pollen settles on lakes and ponds and sinks to the bottom to become part of the sediment or is incorporated into peat bogs. It settles uniformly into stratified layers providing a detailed ecological that can be extracted from vertical core samples. The pollen must be taxonomically identified and its dispersal range must be calibrated on the basis of its buoyancy and the height of the plant (both factors can affect the travel distance). Local topology and ecology must be considered as well, given that pollen will have a far easier time settling on a grassland pond than a forest lake. Other than pollen, foraminifera (Alonso-Garcia *et al.,* 2011) and dinoflagellates (De Schepper *et al.*, 2011) are used in marine paleoecology, whereas ostracods (Frogley *et al.*, 2001) and chironomids (Engels & Cwynar, 2011) are used as freshwater fossils. The very abundant diatoms have been recovered from both types of ecosystems (Sims *et al.,* 2006). The co-distrtibution of species through time can be used to reconstruct accurate maps of the limits of the ice sheets as they advanced and receded throughout the Pleistocene (Jackson *et al.*, 1997). Furthermore, we can address important biogeographical questions regarding

the effects that glaciations had on historic rates of migration, colonization and the formation of present day species ranges.

It is not possible to reconstruct historic biogeographic ranges for species mostly comprised of soft tissue from the fossil record. This presents a unique problem in trying to ascertain a species' history and to characterize both past and present ecological pressures affecting a species' current distribution. Though the direct evidence of the fossil record cannot be used, current phylogeographic techniques have been very effective in studying the effect of the Pleistocene glaciations on both terrestrial and aquatic organisms. Since the turn of the millennium, the number of phylogeographic articles published has increased four-fold (Soltis *et al.*, 2006), and as a discipline it remains a rapidly expanding field. However, the majority of studies have focused on vertebrates, particularly mammals and fish (Venton, 2013).

1.3 Glacial refugia

Freshwater aquatic organisms are restricted to insular habitats and their migratory capacity depends on the topography of rivers and streams. Those that lack the ability to travel across land disperse through interconnected waterways, either actively by swimming or passively by currents or avian vector (Bilton *et al.*, 2001). With each glacial advance, northern fish populations were subjected to severe population reductions and isolation in glacial refugia. The sustained reduction in population size compounded the initial loss of genetic diversity, with haplotype fixation resulting from genetic drift.

Correlations between the patterns of mitochondrial sequence divergence and geographic zones have revealed the historical dispersal routes of refugial fish lineages (Murdoch & Hebert, 1997; Aldenhoven *et al.*, 2010, Witt *et al.*, 2011).

Discerning the refugial origins of taxa from their post-glaciaction distributions can often be a difficult task. In ninespine sticklebacks, the populations in Northwest Territories are in closer proximity to the Beringian refugium, but microsatellite data indicated an upper Mississippian Valley origin for both these and another group of populations in the Great Lakes region. To further complicate the story, relationships among mitochondrial markers were much closer among the Great Lakes and Beringian clades than either was to the Northwest Territories clade (Aldenhoven *et al.*, 2010). Pygmy whitefish have a similarly disjunct, wide distribution that is further confounded by potential dispersal from multiple refugia (Witt *et al.*, 2011). The geographic distribution of freshwater organisms cannot be the sole line of evidence to infer refugial origins, and benefits from taking a holistic approach preferably involving multiple markers.

At least six major glacial refugia have been characterized in North America for freshwater and anadromous species: the Beringian Refuge that stretches from Western Alaska into the northern Yukon; the Nahanni Refuge, which rests in the Mackenzie and Liard river drainages of the unglaciated corridor between Laurentide and Cordilleran ice sheets; the Cascadian Refuge along the Pacific coast; the Missouri-River Valley Refuge south of the ice sheets; the Mississippi River Valley Refuge to the east of the Missouri River; and the Atlantic Refuge near the eastern coast (Pielou, 1991; Bernatchez &

Wilson, 1998; Oswood *et al.*, 2000; Cox & Hebert, 2001; Van Houdt *et al.*, 2003; Stamford & Taylor, 2004; Shafer *et al.*, 2010). Though these represent the major refugia, it does not discount the likely possibility of cryptic refugia (Stewart & Lister, 2001), particularly in the Alexander Archipelago or Haida Gwaii (Shafer *et al.*, 2010). Contact zones between populations dispersing from separate refugia can lead to hot spots of diversity and these can be mistaken as refugia themselves (Petit *et al.*, 2003).

Previous phylogeographic studies have shown a compelling demarcation in freshwater and anadromous fish species (Bernatchez and Wilson, 1998; Harris and Taylor, 2010; Witt *et al.*, 2011). A regression analysis of pairwise genetic divergences among sister-taxa within holarctic fish species was indicative of an inverse relationship between variability and latitude. This gradient ceases at 46° latitude (the Wisconsin glaciations southern limit), at which point the linear relationship levels out (Bernatchez and Wilson, 1998). Fish species have a marked reduction in species diversity and endemism in biogeographical zones north of the glacial limit (McAllister *et al.*, 1986). Though, certainly less studied, similar trends have been identified in freshwater invertebrates (Ishida & Taylor, 2007; Millette *et al.*, 2011). As organisms disperse from the geographically restricted glacial refugia into newly formed freshwater habitats they typically undergo a series of local bottlenecks and founder effects. Only a fraction of the nucleotide diversity will be represented in the colonists, though what variation does exist will be maintained by the rapid expansion of population (Excoffier *et al.*, 2009). In comparison, unglaciated habitats are much older and likely to have deeper divergence

among populations that reside in them. Even the potent dispersal capabilities of some freshwater crustaceans may not have a significant homogenizing effect on neighboring populations. The Monopolization Hypothesis predicts that persistent founder effects followed by rapid local adaptation will prevent gene flow from insurgent populations (De Meester *et al.*, 2002).

1.4 Characteristics of *Hyalella 'azteca'*

The order Amphipoda (class Malacostraca) are crustaceans that occur in marine, fresh and subterranean waters. Most amphipods are restricted to cool environments with an abundance of dissolved oxygen. All amphipods are obligately sexual, and none are capable of producing diapausing eggs (Smith, 2001). Of the four amphipod suborders, only the Gammaridea inhabit North American freshwater systems. They range between 5 and 20 mm in length, and the body is laterally compressed, consisting of a cephalothorax, seven thoracic segments, a segmented abdomen, and a small, terminal telson. There are two pairs of segmented antennae that vary in length. Amphipods have seven pairs of segmented thoracic legs; the first two pairs are called the gnathopods and are capable of grasping, whereas the remaining five pairs are unspecialized pereopods used for walking (Smith, 2001). Amphipods are referred to as sideswimmers, as they roll on their sides and swim using their abdominal pleopods. Their diet consists of diatoms, macrophytes and detritis (Wellborn, 1994b). Their color is highly variable, ranging from light green to brown and occasionally blue, red or purple (Smith, 2001).

The genus *Hyalella* (Dogielinotidae; Hyalellinae) is composed of three subgenera (Bousfield, 1996). *Austrohyalella* and *Mesohyalella* are South American subgenera that include geographically widespread and morphologically variable species. The North American subgenus *Hyalella* includes only a few described species. The most broadly distributed is *Hyalella azteca.* It was originally described by Saussure (1858), using sample from Vera Cruz, Mexico, and was later redescribed by Gonzalez and Watling (2002). Its subfamily, Hyalellinae, has recently been reclassified under the family Dogielinotidae (superfamily Talitroidea) (Serejo, 2004). *Hyalella azteca* was named for its discovery in a well in the Chapultepec Aztec ruins, and has numerous synonyms in the early literature including *H. knickerbockeri, H. inermis* and *H. dentata* (Geisler, 1944). This species is benthic, and was formerly thought to be broadly distributed from the arctic to southern Mexico, as well as on Bermuda and other islands in the Caribbean (Bousfield 1996; Smith, 2001). *Hyalella azteca* has been classified as a cryptic species complex (Witt & Hebert, 2000), however recently two members have been described with the designations *Hyalella spinicauda* and *Hyalella wellborni* (Soucek *et al.,* 2015). At least eight other highly endemic species in the subgenus *Hyalella* have been described from populations in both continental and island habitats. *Hyalella texana* is endemic to the Edward's Plateau, Texas (Stevenson & Peden, 1973), while the recently described *H. longicornus* occurs on a golf course in Utah (Bousfield, 1996). *Hyalella montezuma* is endemic to Montezuma well in Arizona (Cole & Watkins, 1977; Witt *et al.*, 2003). There are two morphologically distinct species endemic to Death Valley National Park, California; *Hyalella sandra* is epigean like most other *Hyalella*, but *Hyalella muerta* is hypogean and eyeless (Baldinger *et al.*, 2000). *Hyalella merispinosa* was recently

described from Pahranagat Valley, Nevada, and is morphologically similar to *H. sandra* (Baldinger, 2004). *Hyalella caribbeana* is endemic to the Caribbean Islands and *H. sapropelica* has been described from Guatemala (Bousfield, 1996). Independent evaluations of the taxonomy *Hyalella 'azteca'* by early systematists resulted in multiple species designations (table 1.1) that were re-evaluated into a single species by Bousfield (1996).

1.4.1 Species delineation within *Hyalella 'azteca'* complex

Early species definitions used exclusively morphological characters and described *Hyalella 'azteca'* as as a member of the Talitridae family and one of the most widely distributed freshwater crustaceans within North America (Saussure 1858). Bulycheva (1957) redefined the family Talitridae as belonging to the superfamily Talitroidea and created new family classifications according to ecological zones; Hyalidae from marine and brackish waters; Hyalellidae for freshwater talitroids and Talitridae was retained for terrestrial or semiterrestrial organisms. Serejo (2004) reclassified the Hyalellidae as family Dogielinotidae on the basis of a cladistic analysis of morphological characters.

Hyalella 'azteca' disperses actively by swimming and passively by way of lake and stream currents. It can also be dispersed passively over relatively short distances by clinging to waterfowl or aquatic mammals (Rosine, 1956; 1962; Swanson, 1984). This pattern is restricted solely to amphipods: the dispersal of cladocerans by waterfowl has also been observed (Green et al., 2002). The dispersal of amphipods is limited by the

length of time a species can survive desiccation in overland flight. For example, *Gammarus lacustris*, another amphipod, can adhere to duck plumage for up to two hours (Figuerola & Green, 2002). Dispersal ability is a widely variable trait in some amphipods groups. Unlike the gammerids, there are several chiltoniids within the Great Artesian Basin of Australia that are frequently endemic to single springs (Murphy *et al.*, 2013).

Inferences using morphological characters suggested that the subgenus *Hyalella* is most closely associated with the northern members of the South American subgenus *Mesohyalella*; this suggests that *Hyelalla* recently invaded North America. Despite its vast geographic range, *Hyalella 'azteca'* is remarkably constrained morphologically with the most common variations being the presence or absence of mucronations on the first two pleon segments (Bousfield, 1996) and the existence of two distinct size-based morphotypes based on size (Wellborn & Bartholf, 2005; Wellborn *et al.*, 2005). The degree of habitat variation between *Hyalella* species of the two morphotypes has been well characterized as a function of fish predation, but it has yet to be determined if size plays a factor in dispersal ability.

A major component that determines current species distributions is the ability to disperse to new habitats. Proglacial habitats were formed from the melt waters of the Pleistocene glaciations and subsequent colonization by *Hyalella 'azteca'* was likely constrained at an ecological level by fish predation (Wellborn & Broughton, 2008). If the adaptive landscape is constrained by ecological forces and results in few adaptive peaks, then it is expected that the environment or specialized niche will select for similar phenotypic traits or species assemblages (Gillespie, 2004). *Hyalella 'azetca'* can be split

into two broad morphological groups characterized by head and body size that experience opposite size-based selective pressures. Small ecomorphs are commonly located in habitats with predatory fish (*Lepomus* sunfish) where larger individuals experience a fivefold higher rate of predation than their smaller counterparts (Wellborn, 1994a).

Table 1.1: A summary of the historical taxonomy of the *Hyalella azetca* species complex

	Early Taxonomy	Bulycheva (1957)	Bousfield (1996)	Serejo (2004)
Superfamily		Talitroidea s.s.	Talitroidea s.s.	Talitroidea s.s.
Family	Talitridae s.l.	Hyalellidae	Hyalellidae	Dogielinotidae
Subfamily		Hyalellinae	Hyalellinae	Hyalellinae
Genus Species	*Amphitoe 'azteca'* (Saussure 1858)	*Amphitoe 'azteca'* (Saussure 1858)	*Hyalella 'azteca'*	*Hyalella 'azteca'*
	Allorchestes knickerbockeria (Bate 1862)	*Allorchestes knickerbockeria* (Bate 1862)		*Hyalella wellborni (Soucek et al., 2015)*
	Hyalella 'azteca' (Saussure 1858)	*Hyalella 'azteca'* (Saussure 1858)		*Hyalella spinicauda (Soucek et al., 2015)*
	Hyalella dentata (Smith 1874)	*Hyalella dentata* (Smith 1874)		
	Hyalella inermis (Smith 1874)	*Hyalella inermis* (Smith 1874)		

The *Hyalella* species complex, as determined by analyses of allozymes and mtDNA sequences, have revealed at least eighty divergent, monophyletic clades (Witt and Hebert, 2000; Witt *et al.*, 2006; Wellborn & Witt in prep, 2015). Seven major mitochondrial clades have been identified in previously glaciated regions of North America, and another seventy-three occur in the Great Basin (south western United States). The seven northern clades frequently form sympatric *Hyalella* assemblages. The majority of sampled habitats (within 5,000 to 10,000 years) in Wisconsin and the Great Lakes had co-occurring members of different clades (Witt & Hebert 2000). These populations of *Hyalella 'azteca'* had profound levels of Hardy-Weinberg disequilibrium and fixed allelic differences at multiple allozyme loci. Clades 1 and 2 are sister taxa, and both occur commonly throughout northern North America in ponds, but also occasionally in lakes and streams. Clade 3 is widely distributed in central Canada, and occurs in both lakes and streams (Major *et al.*, 2013) and recently described and renamed *Hyalella spinicauda* (Soucek *et al.*, 2015). Only a single population of clade 4 has been detected in previously glaciated regions of North America (in New Brunswick), but it is also known from Georgia and Louisiana (Witt & Hebert 2000; Witt unpublished, 2015). Clades 5 and 6 are widely distributed in lakes and streams in the Mississippi River, Great Lakes and Atlantic drainages (Major, *et al.*, 2013). However, a single population of clade 6 has also been discovered in northern Saskatchewan, and this could indicate that it possesses a disjunct distribution (Witt & Hebert 2000; Witt unpublished). Clade 6 was recently described in Soucek *et al.*, (2015) and designated *Hyalella wellborni*. Clade 7 has previously been taxonomically identified as *Hyalella inermis* (Smith, 1875), but later incorporated into *H. 'azteca'* (Weckel, 1907). It occurs in ponds and lakes from the west coast to at least the Great Lakes region, but might also be distributed

further to the east. Wellborn *et al.*, (2005) discovered a previously unidentified clade in Oklahoma and Major *et al.,* (2013) detected populations of this clade in northern Florida.

Though, predominately cryptic in their morphology, *Hyalella 'azteca'* populations can generally be discriminated on the basis of size into large-body and small-body ecomorphs (Wellborn, 1994a; Wellborn *et al.*, 2005). The small-body ecomorph complex consists of Wellborn's (1994a) three small-body lineages in Michigan species: A, B and C, whose mtDNA lineages cluster with Witt & Hebert (2000) Clades 3, 5 and 6, respectively. Wellborn *et al.,* (2005) Oklahoma-small morph was added to the group of small-body ecomorphs and groups with Michigan species C or Clade 6 (Witt & Hebert, 2000), though morphologically it lacks a mucronation on the second pleon. The large-body ecomorph complex consists of Oklahoma and Michigan variants that group with Clade 1 (Witt & Hebert, 2000). Clades 2 and 7 (Witt & Hebert, 2000) were not included in Wellborn's size-based ecological studies (Wellborn and Broughton, 2008), but on the basis of their morphology they fall within the large-body group (Witt unpublished, 2015). Similarly, clades 4 and 9 were not included in those studies, but form well supported monophyletic groups. Though several co-occurring populations of clades 1 and 2 had significant Hardy-Weinberg deviations at three allozyme loci, they are not easily distinguished electrophoretically due to high allelic variability and a lack of fixation for alternate alleles between them (Witt & Hebert, 2000). For the remainder of the thesis, the different clades of *Hyalella 'azteca'* will be designated according to the scheme outlined in table 1.2.

Table 1.2: Species designations of *Hyalella 'azteca'*
The species column contains the designations used in this thesis, while the Witt and Hebert (2000) and Wellborn (2005) columns include the corresponding classifications from their respective studies. The sized-based morphology for each species is listed in the final column.

Species	Witt and Hebert (2000)	Wellborn (2005)	Ecomorphic Type
1	1	Michigan-Large	Large
2	2		Large
3	3	A	Small
4	4		Small
5	5	B	Small
6	6	C	Small
7	7		Large
8	N/A	Oklahoma-Large	Large
9	N/A		Small

The large-body ecomorphs studied in Michigan and Oklahoma are comparable in life-history traits and habitat. Comparisons between *H. 'azteca'* populations in Duck Lake and George Pond, Michigan, have considered the ecological differences between the two size morphs. The resource consumption rate increases with body size, but the smaller ecomorph attains sexual maturity earlier, and reallocates resources from growth to sexual reproduction (Wellborn, 1994a). According to the ecological principle of competitive exclusion, if both ecotypes occupy the same niche, one will eventually exclude the other (Hardin, 1960). However, the presence or absence of the two ecomorphs appears to be linked to predation regimes. The large ecomorph is favored in fishless ponds because it is better able to cope with the predatory invertebrates that dominate these habitats (Wellborn, 1994b). In fishless habitats, the smaller ecomorphs are consistently excluded by the larger *Hyalella* because of a disparity in foraging abilities, resource consumption rates, and resistance to odonate predation (Wellborn, 2002). However, size-biased predation on *H. 'azteca'* in Oklahoma lakes by bluegill sunfish (*Lepomis macrochirus*) specifically favors the survival of the small ecomorph (Wellborn *et al.,* 2005). Dissection of bluegill fish stomachs revealed an overwhelming feeding bias toward larger *H. 'azteca'* (Cooper, 1965; Wellborn,

1994b). Fish express a substantial size-based food preference for large ecomorphs, and allows the smaller ecomorphs to predominate in lakes (Wellborn, 2002; Wellborn *et al*. 2005).

The spatial distributions of the size ecomorphs have been studied in the Great Lakes basin. Two of the large bodied species (1 & 2) occupy smaller ponds and occur less commonly in streams, suggesting that they inhabit a broader range of environments than the small ecomorphs that only inhabit lakes and streams (Witt & Hebert, 2000). In addition, only large bodied species (1, 2, & 7) occur in fishless tundra ponds (Witt unpublished, 2015). The prairie wetland complex in the Missouri Coteau of North Dakota is characterized by shallow, fishless semi-permanent ponds that frequently dry up. Amphipods are highly susceptible to desiccation, and Swanson (1984) observed that *Hyalella 'azteca*' rapidly re-colonized these ponds via waterfowl when the water returned after drying. Fishless ponds represent the habitat type normally colonized by the large ecomorphs. Although the evidence is largely anecdotal, it does suggest that the large ecomorphs possess a greater dispersal capacity than small ecomorphs. This could be a consequence of a better ability to cling imparted by their larger gnathopods. The two ecomorphs represent intermediate points on the freshwater crustacean dispersal ability continuum. One on end, cladocerans (Forró *et al*., 2008) are extremely successful at dispersing, in contrast glacial relicts (e.g. *Mysis, Diporeia*) are limited to deep sections of lakes (Pothoven *et al.*, 2011). Unlike daphnids, *H. 'azteca'* lacks a parthenogenetic stage, and must rely on the stochastic transfer of a fertilized ovigerous female or multiple individuals from both sexes to establish a new population. However, unlike the glacial relicts, they are not confined to cold deep water, and can be dispersed by waterfowl and mammals (Rosine 1956; 1962; Swanson, 1984; Peck, 1975).

1.5 Objectives

The overall objective of my thesis is to characterize the current and historic distribution of the freshwater amphipod species complex *Hyalella 'azteca'* and apply it to the broader study of biogeographic patterns and molecular evolution. My goals are characterized in the following seven objectives.

Characterization of present day species' ranges (chapter 2)

Though populations of *Hyalella 'azteca'* are well studied in the Great Lakes region (Wellborn 1994; 2005; Witt & Hebert 2000), the Great Basin (Witt *et al.,* 2006; 2008) and Oklahoma (Wellborn & Broughton, 2008), the full distribution of the complex across North America has yet to be fully characterized. This work will focus on areas that are presently underrepresented. Even with the current level of study, it is evident that some clades occupy far larger regions and the propensities for sympatric populations among the lineages imply overlapping species' ranges. The variability of *H. 'azteca'* ranges is extremely severe, as high levels of endemism are rampant among the closely situated habitats within the Great Basin (Witt *et al.,* 2006; 2008). In chapter 2, I will provide the geographical distributions of the most widely distributed *Hyalella* species.

Infer historic ranges and refugial origins (chapters 3 and 4)

A primary goal of biogeographic research is to reconstruct the historical dispersal of species from analysis of their present-day distributions. The co-distribution of *Hyalella 'azteca'* lineages

may indicate dispersal from the similar glacial refugia known to be located in Beringia, the Mississippi River and Missouri River drainages, as well as along the Pacific and Atlantic coasts (Pielou, 1991). This type of analysis has been used to determine the refugial origins of North American and European fishes (Bernatchez and Wilson, 1998; Harris and Taylor, 2010; Witt *et al.*, 2011), as well as in freshwater crustaceans (Millette *et al.*, 2011). By using a combination of phylogeographic and population genetics techniques, I inferred past demographic and range expansions and attempted to track the historical dispersal of *Hyalella 'azteca'* after the melting of the Pleistocene glaciations.

Comparative phylogeographic analysis among lineages (chapters 3, 4 and 5)

Once current and historic ranges of the major *H. 'azteca'* clades have been characterized, I will test for phylogeographic concordance among lineages. To date, evidence of dispersal variability remains largely anecdotal. *Hyalella 'azteca'* provides a unique opportunity to examine the relationship between size-linked dispersal capabilities and population structure. Organisms with greater dispersal capabilities should occupy larger ranges, which in turn leads to founder effects and local adaptation, but will have greater opportunity to move between populations. The increased rate of migration among neighboring habitats will mitigate isolation by distance patterns.

Comparative phylogeographic analysis among regions (chapters 3, 4 and 5)

Hyalella 'azteca' lineages with samples sites from glaciated and unglaciated regions provide an opportunity to test for the effects of the Pleistocene glaciations on population structure and nucleotide diversity. Many habitats south of the glacial limit were affected by the Pleistocene glaciations, but largely avoided complete destruction by the advancing ice sheets. As such, many populations of *H. 'azteca'* have potentially persisted continuously in permanent habitats. The vicariance of populations from glacial advances, followed by reduced population sizes and repeated founder effects of post-glacial dispersal should have detectable patterns within the genetic diversity and structure of northern populations.

Nuclear and mitochondrial concordance (chapters 2 and 6)

Strict reliance on a single molecular marker can be problematic when attempting to infer species trees (Knowles, 2009). Genes evolves according to mechanisms that may vary or even be absent in other areas of the genome. Mitochondrial markers are hotly debated in the literature with concerns over their erratic rates of evolution and potential for selective sweeps (Galtier *et al.,* 2009). As such, this study will analyze the first internal transcribed spacer (ITS1) of the rRNA cistron for concordance between mitochondrial and nuclear genes. ITS1 is a commonly used marker in fungi (Schoch *et al.,* 2012) and has been considered a candidate for a barcoding gene in amoebae (Nassonova *et al.,* 2010) and diatoms (Evans *et al.,* 2007). Recently it has been used to support mitochondrial population structure in pygmy whitefish (Witt *et al.,* 2011).

Rate of molecular divergence (chapter 2)

The endemic population of *Hyalella* within Devil's Hole in the Ash Meadows National Wildlife Refuge potentially exhibits extreme divergences rates at mitochondrial loci. This may have resulted from the fixation of nearly neutral alleles in the small populations caused by genetic drift or extreme thermal temperatures accelerating the mutation rate as a consequence of increased metabolic rates (Witt *et al.,* 2008). The most parsimonious split between this population and its sister taxa in Crystal Springs is 60,000 years, but the level of divergence more resembles a split dating back to the Miocene, assuming the crustacean clock calibration of 1.4% sequence divergence per million years (Knowlton and Weigt, 1998). Likelihood ratio tests provided strong evidence for rate heterogeneity, but these may be a consequence of the smaller population sizes endemic to the region. Once the phylogeographic patterns of the *Hyalella* species complex have been identified, they can be used to infer rates of molecular evolution. If there is sufficient concordance among the distributions of the lineages to infer an accurate dating of past vicariant or dispersal events during the Pleistocene glaciations, then a rate of molecular evolution can be estimated.

The broader biogeographical context

This project will attempt to place the historical and present-day distributions and population structure of the *Hyalella 'azteca'* species complex into the broader context of North American freshwater biogeography and cryptic speciation.

Chapter 2

Geographic and molecular characterization of *Hyalella azteca*

2.1 INTRODUCTION

Hyalella 'azteca' is a cryptic species complex of North American freshwater crustaceans. The traditional classification of *H. 'azteca'* on the basis of morphological characters implied a single species broadly distributed throughout North America. However, recent studies using molecular markers has revealed a cryptic species complex of over 100 distinct lineages with sufficient levels of divergence to warrant classification as provisional species (Witt *et al.,* 2006; Wellborn & Broughton, 2008; Witt & Wellborn in prep). The degree of variation in geographic distributions among these lineages appears to be quite heterogeneous. *Hyalella* species within the Great Basin are often isolated to a single habitat, whereas other groups inhabit locations separated by hundreds of kilometers (Witt *et al.,* 2003; Witt *et al.,* 2008). Broad variability in spatial distributions is not uncommon among amphipod species. In the Great Artesian Basin of Australia there are isolated springs frequently inhabited by endemic chiltoniids (Murphy *et al.,* 2013), a pattern that is mirrored in some of the *Hyalella* species detected in the Great Basin (Witt *et al.,* 2006). Alternatively, other amphipod species (e.g. *Gammarus lacustris*) are broadly distributed across one or multiple continents (Väinölä *et al.,* 2008).

2.1.1 Size-based predation of *Hyalella 'azteca'*

A major factor that determines current species distributions is the capacity to migrate to new habitats, but dispersal ability is not the only contributing attribute. Colonization of proglacial habitats by *Hyalella 'azteca'* was likely mediated at an ecological level by fish

predation (Wellborn & Broughton, 2008). If the adaptive landscape is constrained to a few adaptive peaks, then it is expected that analogous environments will select for similar phenotypic traits or species assemblages (Gillespie 2004). *Hyalella 'azetca'* can be split into two broad morphological groups characterized by head and body size that experience different size-based selective pressures. Small ecomorphs are commonly located in habitats with predatory fish (*Lepomis* sunfish) where larger adults experience a fivefold higher rate of predation than smaller juveniles (Wellborn, 1994a). On the other hand, the large morphotypic species in Michigan inhabit fishless sites, whereby smaller *Hyalella* are preferentially consumed by dragonfly nymphs (Wellborn, 1994a). The sizes of *Hyalella* populations sampled in the American Midwest are negatively correlated to the density of fish, specifically cyprinidae minnows and sticklebacks (Anteau *et al.*, 2011). The relationship between fish and *Hyalella* is more complex than a predator-prey relationship, as minnows also consume detritus, which would result in increased competition.

2.1.2 Mitochondrial molecular marker for phylogenetic inference

As outlined in chapter 1, molecular markers can provide a wealth of phylogenetic information, which can be directly applied to biogeographic inquiries. The mitochondrial cytochrome c oxidase subunit 1 gene (COI) also designated as the DNA barcoding gene, has become widely used to infer species boundaries and to test population genetic hypotheses (Leese & Held, 2011). With sufficient length to produce billions of unique combinations (Hebert *et al.* 2003a, 2003b), COI has potentially more variants than the

number of estimated species on the planet. Since a large fraction of Earth's biota remains uncharacterized, the barcoding gene has facilitated attempts to quantify species variation before it is lost to extinction (Valentini *et al.,* 2009). Though COI has been problematic in resolving species boundaries among plants (Hollingsworth *et al.*, 2011) and fungi (Schoch *et al.,* 2012) and is still a hotly debated topic (Barrowclough & Zink, 2009; Edwards & Bencsh, 2009), it remains a potent tool in resolving phylogenetic and population genetic questions concerning crustaceans (Leese & Held, 2011).

The COI gene has consistently been employed to resolve cryptic taxonomic groups that otherwise would have gone unresolved (Bickford *et al.,* 2007). It has been used extensively to explore the species boundaries in several cryptic crustaceans, including cladocerans (Adamowicz *et al.,* 2004; Adamowicz *et al*., 2009), daphnids (Penton *et al.,* 2004), branchiopods (Rowe *et al.,* 2007) and amphipods (Bradford *et al*., 2010; Cristescu & Hebert, 2005; Seidel *et* al, 2009; Westram *et al.,* 2012). Mitochondrial markers offer several advantages over nuclear genes to infer phylogenetic relationships. As a consequence of being haploid and uniparentally inherited they have a lower effective population size (1/4 that of diploid nuclear genes) and circumvent the confounding effects of recombination. Effective population sizes (N_e) are idealized populations of breeding individuals that result in a reduction in population heterozygosity by drift equal to that of the study population. Often this number is less than the total number of individuals in the census population. This is very important because it affects estimates of coalescence times. The rationale of effective population sizes is that not all individuals within a population necessarily contribute equally to the transmission of

genes to the next generation. Fluctuations in population size, sex-biased ratios and overlapping generations can all affect effective population sizes.

On the basis of coalescence simulations, Palumbia *et al.,* (2001) derived a "three-times" rule for estimating monophyly in neutral nuclear locus from the branch lengths in a phylogeny of mitochondrial DNA. As lineages evolve in the absence of recombination, they invariably trace back to a single ancestor. This is referred to as coalescence. Since mitochondrial DNA evolves at a much quicker rate than nuclear DNA, it will form monophyletic groups while the nuclear ones are still paraphyletic or polypheletic. As a consequence of the lower effective population sizes, when the branch length of the mitochondrial DNA is more than three times that of its average intraspecies diversity, the nuclear alleles should also be monophyletic. In other words, this is an estimation of the probability of coalescence at nuclear loci on the basis of mitochondrial branch lengths. The principal advantage of utilizing COI is that its distribution of intraspecific sequence divergences has little overlap with the distribution of interspecific divergences (Hebert *et al.,* 2003a, Lefebure *et al.,* 2006), a trait common among several mitochondrial markers (Avise, 2000) as a result of rapid coalescent times (Zink & Barrowclough, 2008). The presence of COI in nearly all extant lineages allows broad comparisons across disparate taxonomic groups (Ratnasingham & Hebert, 2007). A set of universal primers in combination with its high copy number make COI highly practical for molecular studies, especially for *Hyalella 'azteca'* in which few other markers are currently available.

2.1.3 Congruence among genomic and mitochondrial regions

The 16S rDNA is a short DNA fragment of the large ribosomal subunit and linked to the COI gene as part of the mitochondrial genome. As such, it shares the maternal inheritance and haploid state of other mitochondrial markers. Selective sweeps or strong selection on one area of the mitochondrial genome may have a covarying effect on all linked loci. As a slower evolving gene than COI, 16S will provide a different estimation of the phylogeny (Lefebure *et al.*, 2006). In conjunction with COI, the mitochondrtial 16S rDNA gene has successfully characterized the boundaries of cryptic species, including the palearctic freshwater amphipod *Gammarus fossarum* (Muller, 2000; Westram *et al.*, 2013), the groundwater amphipod *Niphargus rhenorhodanensis* (Lefebure *et al.*, 2006), the desert spring amphipod *Gammarus pecos* (Seidel *et* al, 2009), and several gammarid amphipods in the Black and Caspian seas (Cristescu & Hebert, 2005). As such, the 16S rDNA gene should provide a more conservative estimate of the mitochondrial phylogeny.

A strict reliance on mitochondrial markers may infer an erroneous species tree (Huang *et al.*, 2010). The unique evolutionary mechanisms that drive divergence of mitochondrial genes may make them incongruent with other genomic regions (Ballard & Whitlock, 2004). The nuclear 28S rDNA gene is a fragment of the large ribosomal subunit of the rDNA array and frequently used to support species relationships inferred from mitochondrial markers among crustaceans. Highly conserved regions within the array allow for the use of universal primers, which in conjunction with the higher copy number facilitates DNA amplification. On the other hand, regions of the array (like the 28S rDNA gene and internal transcribed spacers) are sufficiently heterogenous to provide phylogenetic resolution among congenerics (Hillis *et al*., 1991). As a result of its slower

rate of molecular evolution, the 28S rDNA gene will approach monophyly at more gradual rate than COI and 16S rDNA gene, thus making it a more conservative estimate of species boundaries (Witt *et al.,* 2006). Despite different evolutionary rates, functions and location in the genome, Remigio & Hebert (2000) discovered no significant incongruence between 28S and the mitochondrial 16S and COI gene trees in branchiopods. Cristescu & Hebert (2005) successfully employed these three markers to disentangle the speciose gammerids of the Ponto-Caspian region and Murphy *et al.,* (2013) used 28S and COI to delimit 13 chiltoniid amphipod species in the Great Artesian Basin of Australia. Even in cases of significant incongruence among 28S rDNA and mtDNA markers, the major monophyletic groups among congeneric *Gammarus* species have been successfully characterized (Hou *et al.,* 2007). Given the ease at which 28S rDNA gene can be extracted and sequenced from amphipod populations it is an ideal choice to provide an independent line of evidence from the nuclear genome to support the broader phylogenetic relationships among *Hyalella* lineages.

2.1.4 Objectives

This chapter will provide an overview of the geographical distributions and evolutionary relationships among the more widely distributed members of the *Hyalella 'azteca'* species complex. The current geographic distribution of each species is the product of its own demographic history in conjunction with the effects of major biogeographical events. Since *Hyalella 'azteca'* does not have a fossil record, the biogeographical history of each group must be inferred from current molecular diversity and population structure.

This approach will utilize several genetic markers, both nuclear and mitochondrial, to infer genealogical concordance.

The objective of this chapter is to study the evolutionary relationships among the widely distributed North American *Hyalella* species with respect to two mitochondrial (COI and 16S) and one nuclear rDNA gene (28S). The COI gene is tested for a clocklike pattern of sequence evolution. Although the rate of evolution at this locus is highly variable in some species of *Hyalella,* a molecular clock would facilitate estimates of the age of cladogenic splits.

In the following chapters, the *Hyalella 'azteca'* species complex was studied with a focus on glaciated North America. By using a comparative phylogeographic approach, the evolutionary mechanisms that have shaped organismal distributions can be distilled from the co-distribution of extant species among zoogeological provinces (Avise, 2009; Hickerson *et al.*, 2010). Phylogeographical and population genetic analyses were conducted on the basis of a gene survey of the COI barcoding gene (chapters 3 and 4) along with the first internal transcribed spacer of rRNA cistron (chapter 5).

2.2 MATERIALS AND METHODS

2.2.1 COI Sequence Naming Convention

For the remainder of the current and ensuing chapters, COI haplotypes will be assigned designations according to a species and haplotype number (S#H###). A full list of COI haplotypes is included in appendix A.

2.2.2 Field Collection

Hyalella 'azteca' has only been studied in limited scope, thus a broad sampling strategy was employed. With previously published studies focused on the Great Lakes area, Great Basin, Michigan, Oregon, and Oklahoma, sampling regions were selected to provide the greatest amount of coverage to uncharacterized regions. To this effect, sampling strategies were designed according to the North American highway system. The rationale is that by focusing on sites that are accessible near major routes, a much larger area can be covered quickly and cost-effectively. While this will not give in depth coverage of any specific region, the ranges of some of the species of *Hyalella* are expected to be quite large. This bias will be more significant in regions with poor interconnected roadways, as seen in Yukon Territory and the state of Alaska.

In order to test for differences among *Hyalella* populations in glaciated and unglaciated regions, both areas must have sufficient sample sites. In glaciated regions, the areas of interest include the provinces of Quebec, Saskatchewan, Alberta, the Yukon Territory and the state of Alaska. The sample sites within these regions will be added to those previously sampled (but not sequenced) from New Brunswick, Ontario, Manitoba, British Columbia and a few from the Northwest Territories. Within unglaciated regions,

only the southern Great Lakes region, the States of Oregon and Washington and the Great Basin have been previously considered. This project focuses on as much of the unglaciated United States as possible, including Atlantic and Gulf drainages in the southeastern USA, east and west of the Appalachian Mountains and the Mississippi drainages of the Midwest.

Hyalella specimens were collected by running a dip net among the vegetation along the near shore of lakes and ponds, or by scraping along the banks of rivers. In lakes with shallow nearshores, the sediment was disturbed and the dip nets passed through the floating silt to capture benthic specimens. The specimens were sorted on site to ensure that sufficient *Hyalella 'azteca'* were collected, as opposed to the commonly encountered *Gammarus* sp. and *Crangonyx* sp. These classifications were based on morphology and later confirmed in the laboratory under a dissecting microscope and molecular analysis (see below). The amphipods were preserved in 95% ethanol and stored at -20°C.

2.2.3 DNA extraction and amplification

The specimens were identified microscopically and soaked in ultrapure water for at least 12 hours at 4°C prior to tissue extraction. Mitochondrial and genomic DNA was obtained using a 50ul solution containing proteinase K extraction (Sigma-Aldrich Canada Ltd., Oakville, ON, Canada). Tissue samples were collected from the antenna (for larger samples) or legs and digested in proteinase K for 24 hours at 55°C and then placed in a heating block at 98°C for 13 minutes to stop the reaction. Sequences for each individual were amplified using the polymerase chain reaction amplification (table 2.1 for primers).

The 50 ul PCR mixture contained 5 ul of reaction buffer (10x dilution), 0.1 unit of Taq polymerase (Fisher Scientific Canada Co., Ottawa, ON, Canada), 0.4 uM of each primer (Sigma-Aldrich Canada Ltd., Oakville, ON, Canada), 0.2 uM of each dNTP (Fisher Scientific Canada Co., Ottawa, ON, Canada) and 2 ul of the DNA template. The polymerase reaction was conducted on an Eppendorf Mastercycler (Eppendorf Canada Ltd., Mississauga, ON, Canada). The conditions for COI amplification were 94°C for 1 min, 5 cycles of 94°C at 1 min, 45°C at 1 min 15s, 72°C at 1 min, then 36 cycles of 94°C at 1 min, 50°C at 1min 05s, 72°C at 1 min. The reaction finished with a 5 min 15s extension step at 72°C. A series of primer combinations were required to amplify COI (see table 2.1).

Table 2.1: Primer combinations

Marker	Primer	Source	Oligonucleotide sequence
COI	LCO1490	(Folmer *et al.*, 1994)	GGTCAACAAATCATAAAGATATTGG
COI	HCO2198	(Folmer *et al.*, 1994)	TAAACTTCAGGGTGACCAAAAAATCA
COI	CrustDF1	(Radulovici *et al.*, 2009)	GGTCWACAAAYCATAAAGAYATTGG
COI	CrustDR1	(Radulovici *et al.*, 2009)	TAAACYTCAGGRTGACCRAARAAYCA
16S	16STF	(Macdonald *et al.*, 2005)	GGTAWHYTRACYGTGCTAAG
16S	16SBR	(Palumbi *et al.*, 1991)	CCGGTTTGAACTCAGATCATGT
28S	28S-3311F	(Witt *et al.*, 2006)	GGGACTACCCCCTGAATTTAAGCAT
28S	28S-4434R	(Witt *et al.*, 2006)	CCAGCTATCCTGAGGGAAACT

The amplicons were gel extracted and purified with Qiagen QX1 and PE Wash (Qiagen Inc., Mississauga, ON, Canada) and sequenced with the ABI 3130XL capillary sequencer (Applied Biosystems Canada, Streetsville, ON, Canada).

The within species distances of COI sequences partitioned by codon site were analyzed for numts. As a consequence of the degeneracy of the genetic code, the second codon position nearly always results in an amino acid shift, while mutations of the third

position are often neutral. Nuclear copies are extraneous duplicates free of purifying selection that will often have rates of evolution of the first and second codon equal to that of the third position (Buhay, 2009). The COI nucleotide sequences were aligned by eye in MEGA 5.1 (Tamura *et al.*, 2011), and the amino acid sequence was translated using the invertebrate mitochondrial code to check for premature stop codons or unlikely amino acid shifts that could indicate a nuclear copy of the mitochondrial sequence (numt). Such pseudogenes are extra copies bereft of selectional constraints, thus allowing rapid accumulation of mutations that can skew phylogenetic analyses (Buhay, 2009).

2.2.4 Model of molecular evolution

The substitution model, gamma distribution shape parameter (α) and proportion of invariant sites (I) were determined by log likelihood scores of the 88 substitution models using the Akaike Information Criterion in jMODELTEST version 2.1.3 (Posada & Crandall, 1998; Posada, 2008). The degree of difference between the optimal model and the GTR+Γ+I were tested using a likelihood ratio test. As a more general model, GTR can derive simpler time reversible models and is the only option available for some of the subsequent heuristic analyses.

2.2.5 COI variation among *Hyalella 'azteca'* species

The preliminary assignment of *Hyalella 'azteca'* haplotypes to previously published species (referred to as clades in Witt *et al.*, 2003) was determined by a neighbor-joining phylogeny constructed in MEGA 5.1 (Tamura *et al.*, 2011) using the previously determined model of evolution in the section above.

The COI nucleotide diversity (π) and amino-acid diversity were calculated in MEGA 5.1 (Tamura *et al.*, 2011), as well as the base pair percentages and GC content. Since many populations of *Hyalella 'azteca'* co-habit the same localities, the percentage of sites with sympatric species was determined.

2.2.6 Amplification of 16S and 28S rRNA genes

As a consequence of the computational limitations of heuristic methods, a preliminary phenetic analysis was conducted to identify the major mitochondrial groups based on the COI gene. Only a subset of the specimens were haphazardly chosen from those groups and further sequenced for fragments of 16S and 28S rDNA genes (50 individuals). The procedure for amplifying these genes is listed above. The genomic DNA was pulled from the identical extraction tube from which the COI sequences were derived. The PCR conditions were 94°C for 1 min, 41 cycles of 94°C at 1 min, 50°C at 1 min 05s, 72°C at 1 min and the reaction ended with a 5 min 15s extension step at 72°C.

Maximum Likelihood (ML) and Bayesian phylogenies were constructed for all three genes using RAxML version 7.2 (Stamatakis, 2006) and MrBayes version 3.2.0 (Ronquist & Huelsenbeck, 2003), respectively. These analyses are to set up the tests of congruence in the next section and will use the full 16S and 28S sequence data sets as well as the corresponding COI haplotypes. The ML analysis was run using RAxML-HPC version 7.2 on seven cores of an intel i7 processor with both α and I being estimated during the run. The ML tree was estimated using 500 replicates, each starting from a random tree, and confidence limits tested with 1000 nonparametric bootstrap replicates. The Bayesian analysis consisted of two runs with four chains (one hot and three cold)

over 100 million generations. Convergence was determined by analysis of the standard deviations of the split frequencies in Tracer version 1.5. The percentage of discarded burn-in trees were determined by sampling the standard deviations of the most variable splits using Are We There Yet? (Nylander *et al*,. 2008). The ML and Bayesian trees were visualized in FigTree version 1.4 (Rambaut & Drummond, 2009). The posterior distributions of the Bayesian analysis were mapped onto the nodes of the ML tree. Both programs used a GTR+Γ+I model and estimated the gamma distribution shape parameter and proportion of invariant sites. These values were checked against the model parameters of jMODELTEST for significant deviation.

The COI, 16S and 28S sequences (637, 402 and 1360 base pairs, respectively) were concatenated into a single alignment for ML and Bayesian analyses. The alignments were partitioned in RAxML and MrBayes allowing an independent estimation of the GTR+Γ+I model for each region. As above, the ML analysis ran for 500 replicates with confidence limits determined by 1000 pseudoreplicates. The model parameters were estimated independently for each partition in RAxML and MrBayes, which will provide a far better estimate than using the entire concatenated sequence. The Bayesian analysis was run for 40 million generations, convergence was determined in Tracer and the burn-in estimated in Are We There Yet?.

2.2.7 Tests for congruence

Genealogical congruence among markers was tested with the partition homogeneity test in PAUP 4.10b (Swofford, 2003). Each individual gene pair (COI:16S, 16S:28S, and COI:28S) and the overall trio (COI:16S:28S) were tested with 10000 replicates using tree-bisection reconnection branch swapping with taxa added randomly.

2.2.8 Tests for molecular clocks

Each gene tree was tested to see if its rate of molecular evolution behaved in a clock-like fashion using a likelihood ratio test (Huelsenbeck & Bull, 1996). If the null hypothesis of uniform evolutionary rates cannot be rejected, then the mutation rate can be used to estimate the ages of cladogenic splits. The tree with the best likelihood score was imported into PAUP version 4.10b. The tree topology and the model parameters were held constant while the log likelihood scores were calculated for a model with and without an enforced molecular clock (all tips are equidistant from the root of the tree). The parameters for the GTR+Γ+I model were estimated in jMODELTEST and inputted into PAUP and verified for any major deviations from those used in the ML analysis. A χ^2 test was used to determine significance among the two log likelihood scores.

2.3 RESULTS

2.3.1 Field Collection of *Hyalella 'azteca'*

Hyalella 'azteca' specimens were collected from 513 sites across North America (figures 2.1 and 2.2) between the years 2006 and 2011 (see table 2.2) and added to the 2 Nevada populations previously considered (Witt *et al.*, 2006).

In June of 2007, *Hyalella 'azteca'* were collected from Saskatchewan (n=25) and Alberta habitats (n=19). The northern portion of the sampling region included sites along highways 28 and 855 in Alberta and highway 55 from Cold Lake to Tobin Lake, SK. The southern fraction included habitats along Canadian highways 2 and 35, as well as the surrounding areas of Swift Current, SK, Brooks, AB, and Red Deer, AB. The *Hyalella* populations collected from these sites were added to those previously obtained from Lesser Slave Lake, Sturgeon Lake, Jasper National Park and Waterton Lakes National Park, bringing the total number of Alberta sites to 29 (table 2.2).

In August of 2007, *Hyalella* populations in northern Quebec (n = 26) were sampled from habitats along highways 113, 167, 169 and 176, as well as the surrounding region of Lake Timiskaming, ON (n = 3). The most northern point of the sampling region was at the end of highway 167 near Lac Albanel, QC. These sites were added those previously sampled from southern Quebec, resulting in a total of 49 localities.

Figure 2.1: Map showing the distribution of *Hyalella* large-bodied species in North America

The sample sites of four large-bodied *Hyalella 'azteca'* species (1, 2, 7, and 8) are represented by the dots within the map. Each species is colored according to the legend in the lower right.

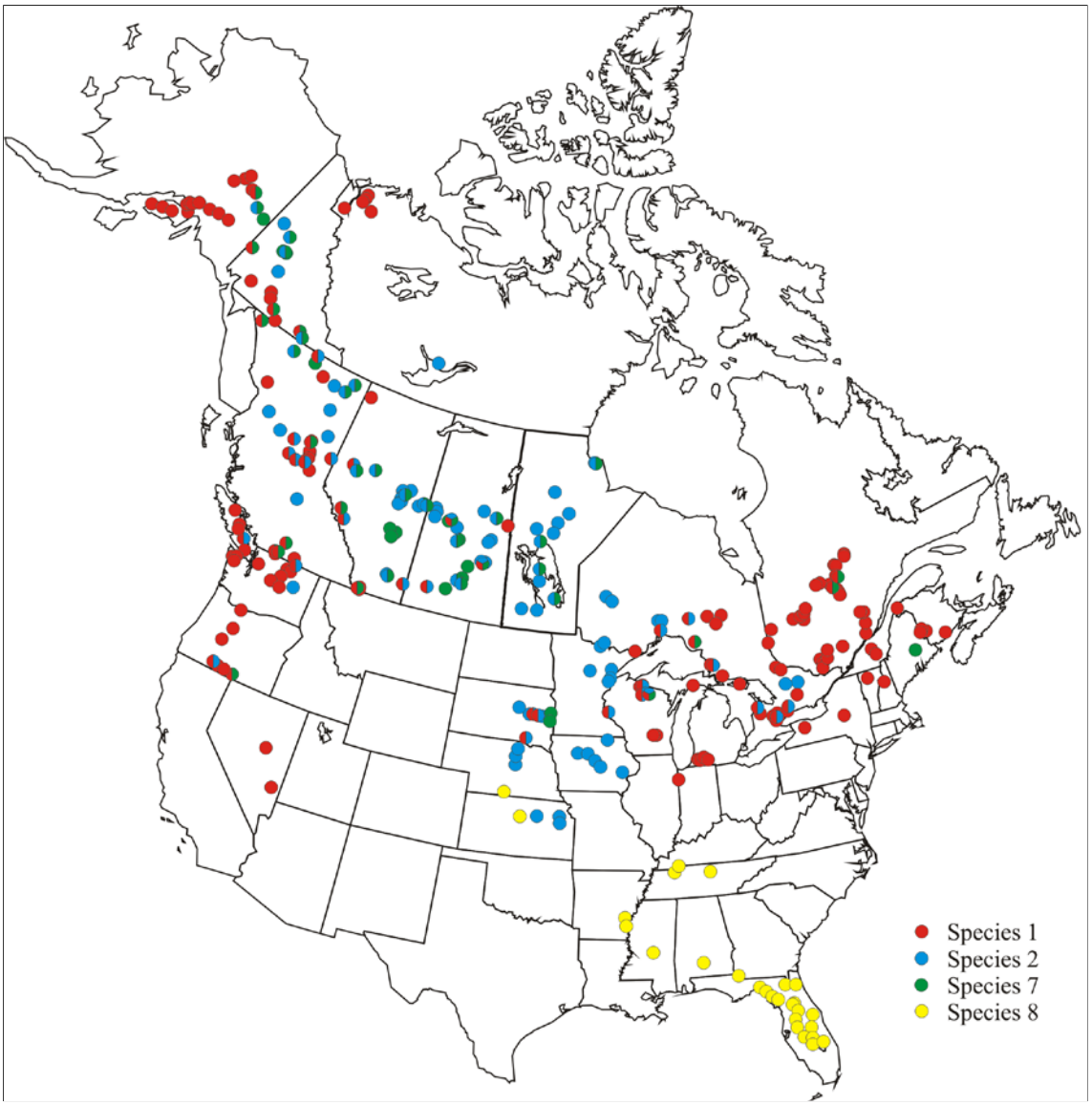


Figure 2.2: Map showing the distribution of *Hyalella* small-bodied species in North America

The sample sites of five small-bodied *Hyalella 'azteca'* species (3, 4, 5, 6 and 9) are represented by the dots within the map. Each species is colored according to the legend in the lower right.

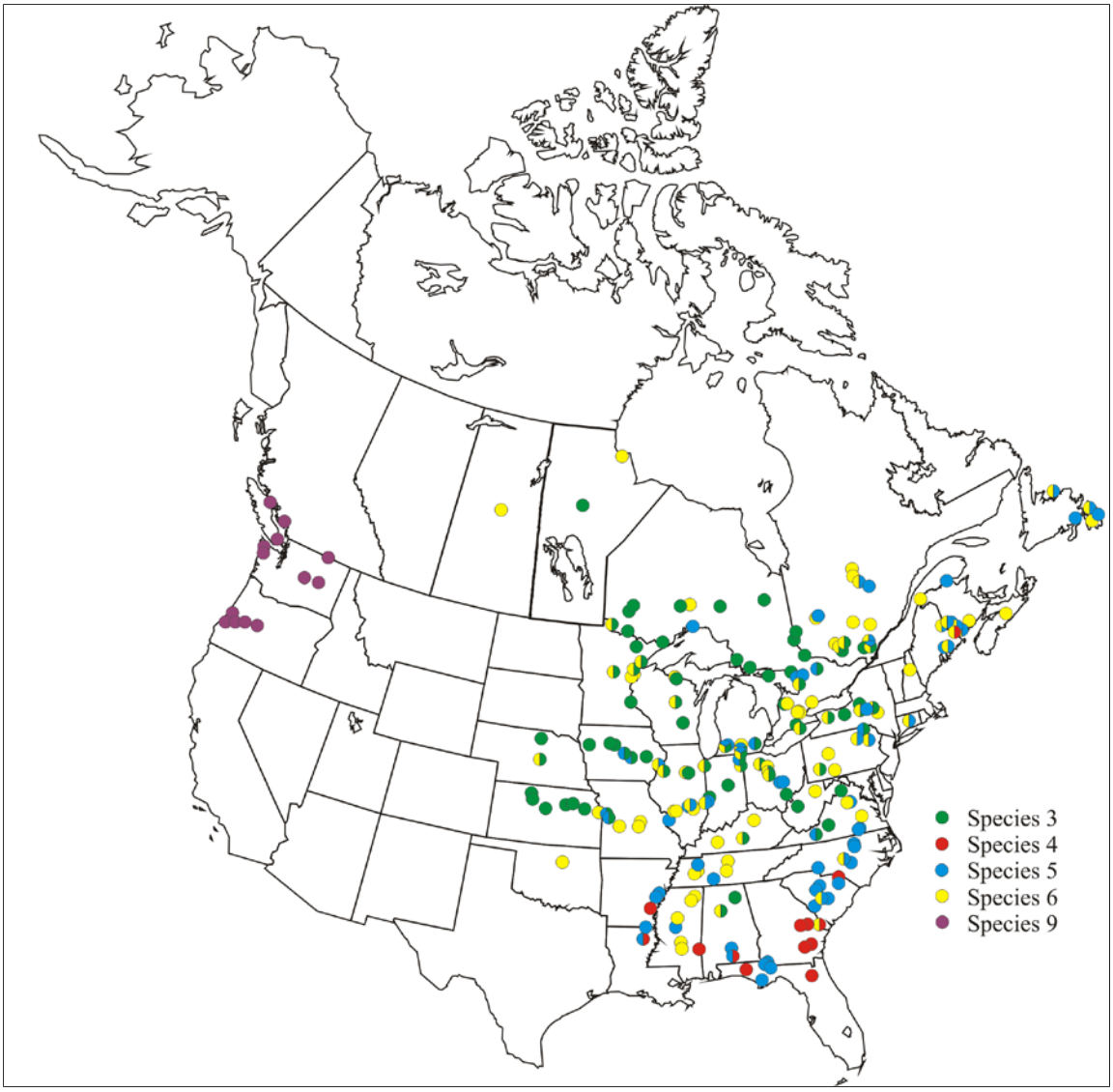


Table 2.2: Sample sites per province/territory/state
The number of sites sampled per state, province or territory is listed. All samples were obtained during the summer, though some regions were sampled multiple times in separate years.

Canadian Provinces and Territories	Samples Sites	Collection Years	American States	Samples Sites	Collection Years
Alberta	29	2006-2007	Alaska	23	2011
British Columbia	44	2005	Alabama	4	2008
Manitoba	22	2006	Arkansas	5	2009
New Brunswick	12	1999	Connecticut	1	2006
Newfoundland	7	2006	Florida	22	2008
Nova Scotia	1	2006	Georgia	8	2008
Northwest Territories	7	2006	Iowa	11	2010
Ontario	49	1996, 2007	Illinois	12	2010
Quebec	46	1999, 2007	Indiana	6	2010
Saskatchewan	25	2006-2007	Kansas	9	2010
Yukon	18	1997, 2011	Kentucky	4	2009
			Louisiana	1	2009
			Maine	5	2006
			Michigan	9	2010
			Minnesota	7	2006
			Missouri	5	2010
			Mississippi	8	2008-2009
			North Carolina	7	2009
			Nebraska	5	2010
			New Hampshire	2	2006
			Nevada	3	2005
			New York	7	2010
			Ohio	7	2010
			Oklahoma	2	2006
			Oregon	19	1999
			Pennsylvania	6	2010
			South Carolina	6	2009
			South Dakota	8	2010
			Tennessee	8	2009
			Virginia	8	2009
			Vermont	1	2006
			Washington	18	2006
			Wisconsin	9	2006
			West Virginia	3	2010

In the summer of 2008, a sampling trip was conducted in the southeastern United States (table 2.2). The northern edge of the sampling region included the coastal plains of Georgia, but yielded few successful sample sites (n = 8) as a consequence of severe drought. *Hyalella 'azteca'* were then collected from localities along interstate highways I-4, I-10 and I-75 in Florida (n = 22), including habitats along the northern edge of Lake Okeechobee. Sites were sampled to include both Atlantic and Gulf drainages along the Florida Peninsula and Panhandle. Additional sites were sampled from drainages of the Appalachicola River in Alabama (n = 4) and various sites Mississippi (n = 8).

In the summer of 2009, a broad approach was employed to sample the regions in the American south-east that were not explored in 2008. The sampling region included (in the order of sampling) Kentucky, Tennessee, Arkansas, Louisiana, South Carolina, North Carolina and Virginia (table 2.2). The *Hyalella* populations within Kentucky (n = 4) were sampled from habitats along the interstate highway I-64. There are three major freshwater ecological zones in Tennessee: Lower Mississippi Basin, Interior Plateau and the Southern Appalachians. An effort was made to sample each region, but similar to the previous sampling efforts of northern Georgia in 2008, sites in eastern Tennessee yielded no *Hyalella* populations despite several localities having submerged vegetation. Sampling the sites within the Lower Mississippi Basin and Interior Plateau yielded 8 *Hyalella* populations. The sample sites in Arkansas were located in the drainages of the Mississippi river (n = 5) and in Louisiana a single *Hyalella* population was sampled from a mangrove swamp. Specimens were collected from habitats in South and North Carolina (n = 6 and n = 7 respectively) along the interstate highways 26, 77 and 95. Sample sites

in Virginia (n = 8) were selected to encompass the Shenandoah Valley and tributaries to the James River and Rappahannock River.

In the summer of 2010 two field collection trips were conducted in the central United States. The first trip sampled from a large region within the American Midwest (table 2.2). Sites within Michigan (n = 9) were sampled along the interstate highway I-94 and in Indiana (n = 6) along I-69. The southern edge of the sampling region included habitats surrounding I-70 in the states of Illinois (n = 8), Missouri (n = 5) and Kansas (n = 9). The western limit included habitats along highway 183 in Nebraska (n = 5). The final sampling regions consisted of sampling sites along U.S. route 14 and 81 in South Dakota (n = 8), U.S. route 20 and interstate highway I-80 in Iowa (n = 11), and along I-74 in Illinois (bringing the total to 12 sites). The second trip covered four states south of the Great Lakes region. *Hyalella* were collected from habitats along US routes 23 and 33 in Ohio (n = 7), I-81 in West Virginia (n = 3), interstate highways I-70 and I-80 in Pennsylvania (n = 6), and the interstate highways I-90, including the Finger Lakes, in New York State (n = 7).

In August of 2011, field collections were obtained in the state of Alaska and Yukon Territory (table 2.2). Two sites in Old Crow, YT had been previously sampled, but the *Hyalella* populations within the Yukon-Mackenzie Delta have been largely uncharacterized. Freshwater habitats along Canadian highways 1 and 2 within the Yukon were sampled, bringing the total number of sample sites to 18. The sampling region within Alaska (n = 23) was largely restricted by the highway system, but two major geographical regions were successfully sampled: Kenai Peninsula and the Alaskan Mountains.

2.3.2 Sequence alignment and model selection

Species 1, 2, 3, 4, 6, 7 and 9 were sequenced with the forward Folmer A and reverse Folmer B primers (table 2.1); species 5 and 8 required multiple primer combinations (Fol A: Crust DR1, Crust DF1: Crust DR1, Crust DF1: Fol B).

Manual alignment of COI sequences revealed no drastic jumps in amino acid evolution or premature stop codons.

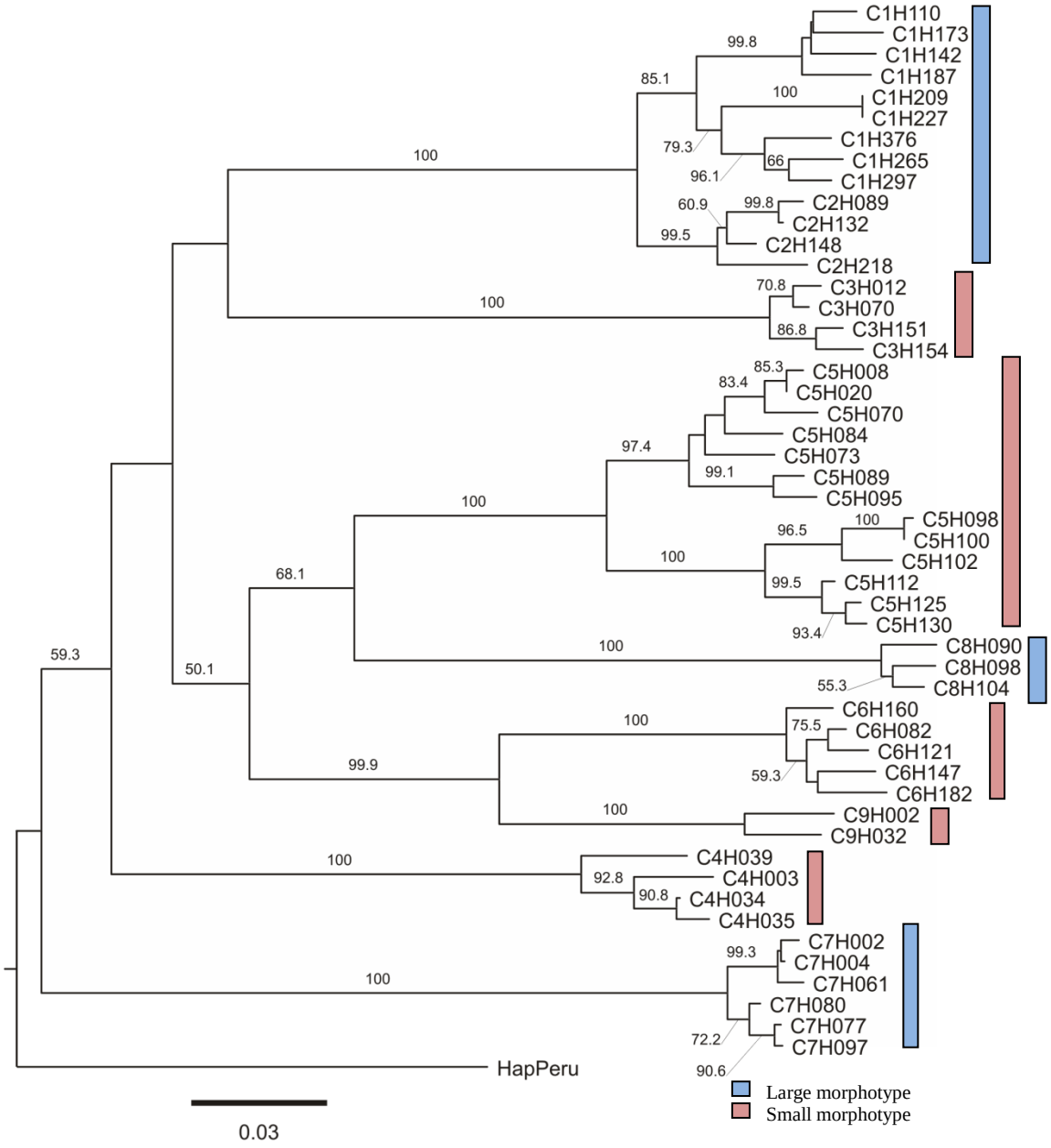
The Tamura-Nei model (1993) for nucleotide substitutions was selected with a gamma distribution shape parameter of $\alpha = 1.32$ and 0.54 proportion of invariant sites as determined by jModeltest version 2.1.3. The Tamura-Nei+Γ+I had the best fit (-lnL 5210.6691), but the general-time-reversible+Γ+I model (GTR) was nearly as good (-lnL 5211.8209). The GTR model is the more general of the two models. A likelihood ratio test indicated no significant difference between the two models (p-value=0.129075) and they have nearly the same number of free parameters (a difference of 1), thus the phylogenies derived from using these two models should have negligible deviations. The Tamura-Nei model was used in the construction of the NJ tree and pairwise distance calculations.

2.3.3 Initial phenetic analysis and species designations

The COI haplotypes were assigned to *Hyalella 'azteca'* species designations on the basis of initial phenetic groupings with sequences from Witt and Hebert (2000) as a reference (figure 2.3). Haplotypes assigned to species 1 through 7 coincide with those previously characterized by Witt *et al.,* (2003). Species 8 haplotypes were consistent with the Oklahoma-Large species considered by Wellborn & Broughton (2008) and species 9 was

a new designation. Species 9 was separated from its sister taxa (species 6) on the basis that it formed a monophyletic group with high bootstrap support and that the branch lengths were greater than the divergence between species 1 and 2. The rational is that species 1 and 2 (Witt & Hebert, 2000) have fixed alloyzme differences at multiple independent loci in a habitat with sympatric populations. This is a strong indicator of reproductive isolation and sufficient justification to delineate them into provisional species. As such, any group that was monophyletic with greater COI divergence from any other such group was assigned a provisional species designation. This resultant list was 9 species (see figure 2.3). Species 1, 2, 7 and 8 are large-bodied morphotypes, while the remainder (3-6, and 9) are small-bodied *Hyalella* species (Wellborn & Broughton, 2008).

Figure 2.3: Neighbor-joining tree of representative *Hyalella* COI haplotypes
The haplotypes within this cladogram are representatives of full neighbor-joining phylogeny that was constructed using the entire COI sequence data set. The listed bootstrap values were calculated on the basis of the complete NJ tree. The outgroup (HapPeru) was selected from a Peruvian freshwater amphipod (*Mesohyalella* sp.). The large and small-bodied species are denoted in blue and red, respectively.



At least 36.5% of the sample sites were sympatric for at least 2 different members of the *Hyalella 'azteca'* complex, with some habitats hosting up to 4 species: Reynolds Lake, MI (species 1, 3, 5 and 6), Lake of the Woods, ON (1, 2, 5, and 6), and Plum Lake, WI (species 1, 2, 3 and 7). Large-bodied species (1, 2, 7 and 8) were sampled from a very broad range throughout North America. Species 1 was the most broadly distributed, with habitats along both the Atlantic and Pacific coasts, especially in glaciated regions (figure 2.1). Species 2 and 7 occupied similar ranges, from the Great Lakes region west to the coast. Species 8 was only collected from unglaciated areas, the most northern point in Nebraska with most of the habitats within the Florida peninsula. The distribution of small-bodied species (3, 4, 5, 6 and 9) were generally more restricted than the large-bodied morphs and predominately sampled from the eastern half of North America (figure 2.2). The populations of species 3, 5 and 6 were co-distributed throughout the Great Lakes region. The range of species 4 populations was predominately restricted to the south-eastern U.S, south of the Appalachian Mountains in both Atlantic and Gulf drainages with the exception of single identified population in New Brunswick (Witt & Hebert, 2000). For biogeographic and population genetics analysis of each species with respect to their spatial distribution, see chapters 3 and 4 for large and small-bodied species, respectively.

2.3.4 COI sequence variation

The 637 bp COI sequence data set (table 2.3) consisted of 2808 *Hyalella 'azteca'* specimens (1294 haplotypes) sampled from 515 sites. The number of sequences per site

ranged from 1 to 23 (Plum Lake, WI) with at least three COI sequences per species successfully amplified from 77% of the populations; the remainder either had too few samples of one of the sympatric species or there was bias in the sampling (see appendix A for full list). For example, Lake Andes, SD, had a 13:1 ratio of species 1 to species 2. The large-bodied *Hyalella 'azteca'* (n=1698) sample size exceeded that for the small-bodied morphotypes (n=1110).

Species 1 had the largest sample size (n=827) with a high nucleotide diversity (π = 0.0459) as well as the greatest number of variable sequence sites (231). Species 2 is closely allied with species 1, but much less diverse (π = 0.0163). Three-hundred and ninety species 3 COI sequences were collected from populations in 84 habitats that were often sympatric (73.8%) with other species (1, 2, 5, 6 and 7). Species 4 had the most limited sample of populations (13 sites) and despite having the fewest collected specimens of the nine species (56), still had a high nucleotide diversity (π = 0.199). Species 5 had the highest diversity among the small-bodied species (π = 0.0427) and the largest number of variable base pair sites (table 2.4). Three-hundred and seventy-six species 6 individuals were collected from 111 habitats, more than any other small-bodied species (table 2.4). Species 7 was the least diverse (π = 0.0109) of all the *Hyalella 'azteca'* species in this study, though not for lack of sample sites (n = 75) or number of samples (n = 255). Species 8 was distinctly separated from other species (only 11.1% sites with two or more *Hyalella* species). One-hundred and two individuals from this species were collected from 27 localities, which had an nucleotide diversity of π = 0.0165. Species 9 was the most isolated and the least diverse (20.0% sympatric sites and π = 0.0120, respectively) of the small-bodied morphotypes.

Table 2.3: Sample sites, sample size and diversity of COI sequences for each *Hyalella 'azteca'* species

This table includes the number of sampled sites and % of sites with sympatric species. The number of COI sequences, COI haplotypes, nucleotide diversity (π), G+C content and number of variable base pair sites are listed. Since many of the sample sites have multiple *Hyalella 'azteca'* species, the total number of sites does not equal the sum of the column.

Species	Number of sites	Sites with sympatric species (%)	Number of COI sequences	Unique Haplotypes	Nucleotide diversity	G+C Content	Variable Base Pairs	Size Morphotype
1	196	45.9	827	375	0.0459	0.462	231	Large
2	133	58.6	514	211	0.0163	0.472	145	Large
3	84	73.8	390	189	0.0139	0.449	164	Small
4	13	46.2	56	34	0.0199	0.428	94	Small
5	78	50.0	235	120	0.0427	0.442	177	Small
6	111	55.0	376	170	0.0179	0.442	155	Small
7	75	66.6	255	83	0.0109	0.409	97	Large
8	27	11.1	102	81	0.0165	0.415	139	Large
9	20	30.0	53	31	0.0120	0.460	61	Small
Total	515	36.5	2808	1294	0.1598	0.447	264	

Both the GC content (table 2.4) and base frequencies (table 2.4) were consistent across all species with the former ranging from 41.2% in Species 7 to 47.2% in species 2.

Table 2.4: Base frequencies of COI sequences by *Hyalella 'azteca'* species

This table lists the average base pair frequencies (A, C, G and T) for each of the *Hyalella 'azteca'* species.

Species	n	Length (bp)	Variable Sites	Base Frequencies			
				A	C	G	T
1	827	637	231	21.4	20.6	25.6	32.4
2	514	637	145	20.9	21.0	26.2	31.9
3	390	637	167	21.7	20.4	24.5	33.3
4	56	637	94	23.0	19.1	23.5	34.4
5	235	637	177	22.4	18.9	25.5	33.2
6	376	637	155	22.3	20.6	24.3	32.8
7	255	637	97	23.2	18.0	23.2	35.6
8	102	637	139	23.3	17.7	23.8	35.2
9	53	637	63	22.2	21.5	24.8	31.5

The average nucleotide diversity (π) and standard deviation for each codon position of the COI gene was calculated (table 2.5) and analyzed for nuclear copies. The third codon position was the most variable and at least an order of magnitude greater in diversity than the highly conserved second position. Species 1 and 5 had higher values of π in the first and third position than the other members of the *Hyalella 'azteca'* complex. The low diversity at the second amino acid position is consistent with genes under selectional constraint. If these were pseudogenes, mutations would be more random, particularly at the second codon position. The manual alignment of COI haplotypes revealed no gaps, premature stop codons or significant leaps in amino acid evolution.

The Tamura-Nei corrected average between-species COI sequence divergences (table 2.6) ranged from 8.4% to 29% (species 1 vs. 2 and species 7 vs. 9, respectively). The between-species amino-acid distances (table 2.7) ranged from 1.1% to 8.0% (species 1 vs. 2 and species 5 vs. 7, respectively). Interestingly, the within-species amino-acid divergence (1.1%) for species 1 was as large as the distance between species 1 and 2.

Table 2.5: Average COI nucleotide diversity within each species by codon position
Each column refers to the position in the codon and lists the average nucleotide diversity (π) and standard deviation as calculated in MEGA 5.1.

	First Position	Second Position	Third Position
Species 1	0.014±0.005	0.006±0.003	0.140±0.016
Species 2	0.006±0.003	0.002±0.001	0.044±0.007
Species 3	0.005±0.002	0.001±0.000	0.038±0.006
Species 4	0.006±0.003	0.001±0.000	0.058±0.008
Species 5	0.014±0.004	0.001±0.000	0.147±0.020
Species 6	0.007±0.003	0.001±0.000	0.051±0.007
Species 7	0.003±0.001	0.000±0.000	0.033±0.007
Species 8	0.001±0.002	0.001±0.000	0.047±0.005
Species 9	0.011±0.004	0.003±0.001	0.024±0.005

Table 2.6 Average group mean COI nucleotide distances among and within *Hyalella 'azteca'* species

The bolded values are the average within species nucleotide distances and the unbolded values are the average between species nucleotide distances using a Tamura-Nei model of nucleotide substitution. The confidence limits were calculated from 500 pseudoreplicates.

	Species 1	Species 2	Species 3	Species 4	Species 5	Species 6	Species 7	Species 8	Species 9
Species 1	**0.049±0.005**								
Species 2	0.084±0.010	**0.017±0.003**							
Species 3	0.225±0.019	0.210±0.020	**0.014±0.002**						
Species 4	0.243±0.021	0.247±0.022	0.236±0.021	**0.021±0.003**					
Species 5	0.235±0.020	0.239±0.021	0.240±0.020	0.210±0.018	**0.046±0.005**				
Species 6	0.250±0.020	0.240±0.020	0.238±0.022	0.237±0.021	0.214±0.019	**0.019±0.003**			
Species 7	0.279±0.022	0.262±0.021	0.245±0.021	0.252±0.021	0.284±0.023	0.270±0.023	**0.011±0.002**		
Species 8	0.259±0.022	0.263±0.023	0.239±0.022	0.230±0.020	0.199±0.017	0.228±0.021	0.280±0.024	**0.017±0.002**	
Species 9	0.235±0.020	0.237±0.021	0.251±0.023	0.248±0.021	0.225±0.019	0.128±0.014	0.290±0.026	0.204±0.019	**0.013±0.002**

Table 2:7 Average group mean COI amino acid distances among and within *Hyalella 'azteca'* species

The bolded values are the average within species amino acid distances and the unbolded values are the average between species amino acid distances using a Poisson model of amino acid substitution. The confidence limits were calculated from 500 pseudoreplicates.

	Species 1	Species 2	Species 3	Species 4	Species 5	Species 6	Species 7	Species 8	Species 9
Species 1	**0.011±0.004**								
Species 2	0.011±0.004	**0.006±0.003**							
Species 3	0.039±0.012	0.042±0.012	**0.003±0.001**						
Species 4	0.043±0.012	0.042±0.012	0.036±0.012	**0.003±0.002**					
Species 5	0.054±0.013	0.058±0.014	0.034±0.011	0.037±0.011	**0.008±0.004**				
Species 6	0.055±0.013	0.058±0.014	0.030±0.011	0.037±0.012	0.035±0.011	**0.003±0.001**			
Species 7	0.078±0.018	0.076±0.018	0.065±0.017	0.051±0.015	0.080±0.019	0.055±0.016	**0.002±0.000**		
Species 8	0.049±0.013	0.051±0.014	0.027±0.010	0.030±0.011	0.012±0.001	0.036±0.012	0.069±0.019	**0.004±0.001**	
Species 9	0.059±0.014	0.063±0.015	0.045±0.014	0.045±0.012	0.044±0.013	0.042±0.013	0.074±0.018	0.039±0.012	**0.007±0.002**

2.3.5 Neighbor-joining analysis of COI sequences

The neighbor-joining analysis (figure 2.1) of the entire COI sequence set resulted in a phylogeny with all species forming monophyletic groups with high bootstrap support (>99) and separated by deep branch lengths. Representative haplotypes from each phylogroup are included in figure 2.1 with bootstrap support derived from the COI phylogeny including all sequences. As a consequence of its high within-species diversity, there were several major phylogroups in species 1 with high bootstrap support. These groups will be further explored in chapter 3 as distinct lineages labeled A to I. In the phylogeny (figure 2.1) the clade represented by haplotypes S1H110, S1H142, S1H173 and S1H187 (99.8 bootstrap support) encompass species 1 lineages A to F, the S1H209/S1H227 clade is lineage G and the remaining clade refers to lineage H and I. Species 6 and 9 were closely allied taxa (12.8% COI divergence). Most species had low intraspecific nucleotide divergence except for species 1 (4.9%) and species 5 (4.6%). Species 2 is sister taxa to species 1, but formed a clear monophyletic group (99.5 bootstrap support). Species 3, 4, 7 and 8 all formed distinct monophyletic groups with little within-species structure. Similar to species 1, species 5 had multiple well supported phylogroups. These lineages (5-A to 5-E) will be fully explored in chapter 4.

2.3.6 16S and 28S sequence variation

The 16S haplotypes have a similar percentage of variable sites (table 2.8) as the COI sequences (16S:40.3% vs. COI:41.4%), but a much lower GC content (only 35.6%). The 28S fragment was considerably more conserved with only 9.6% variable sites and a much higher GC content (56.2%).

Table 2.8: Sequence characteristics of COI, 16S and 28S
This table lists the sequence length, number of variable sites and average base pair frequencies (A, C, G and T) for each of the molecular markers.

Marker	n	Length (bp)	Variable Sites	Base Frequencies			
				A	C	G	T
COI	50	637	264	22.1	19.8	24.9	33.2
16S	50	402	162	34.3	15.6	20.0	30.0
28S	50	1360	131	20.9	25.4	31.8	22.0

2.3.7 Maximum Likelihood and Bayesian Analyses

The GTR+Γ+I model was used in the ML and Bayesian analyses because it is the only option available in RAxML and was the most general of the three models available in MrBayes. Each of these parameters was estimated during the maximum-likelihood run. One of the most significant advantages of RAxML is the ability to estimate different model parameters for distinct segments of a concatenated sequence, specifically for each gene (COI, 16S and 28S).

Table 2.9: Substitution model parameters estimated for maximum likelihood
These parameters were estimated during the maximum-likelihood analyses for each gene in RAxML (COI, 16S and 28S). The parameters include the six rate variables for each nucleotide combination (AC, AG, AT, CG, CT and GT), percentage of invariant sites and α shape parameter of the gamma distribution.

Model Parameters	COI	16S	28S
r (AC)	1.067	0.535	0.655
r (AG)	12.205	4.039	2.198
r (AT)	0.450	0.838	2.167
r (CG)	1.310	0.065	0.477
r (CT)	12.270	6.042	3.300
r (GT)	1.000	1.000	1.000
p-invar	0.532	0.411	0.707
Γ alpha	1.120	1.205	1.872

The log likelihood scores for the best maximum likelihood tree among 500 replicates for COI/16S/28S were -5533.70/-2594.75/-4140.79 (figures 2.3, 2.4 and 2.5, respectively). The COI analysis reached convergence with a 27% burn-in, the 16S at 28% and 28S at 30%. The Bayesian posterior probabilities of each node were calculated from percentage of the total replicates that resolved that node after the burn-in trees were discarded. These values are listed after the bootstraps values on the ML trees (Fig 2.4 2.5 and 2.6).

Figure 2.4: Maximum Likelihood tree of COI haplotypes
This cladogram was constructed using a representative sample of COI haplotypes. Support values were derived from maximum likelihood – 1000 bootstraps/Bayesian posterior probability. Only bootstrap support values greater than 60 were indicated. The outgroup (HapPeru) was selected from a Peruvian freshwater amphipod (*Mesohyalella* sp.). The large and small-bodied species are denoted in blue and red, respectively.

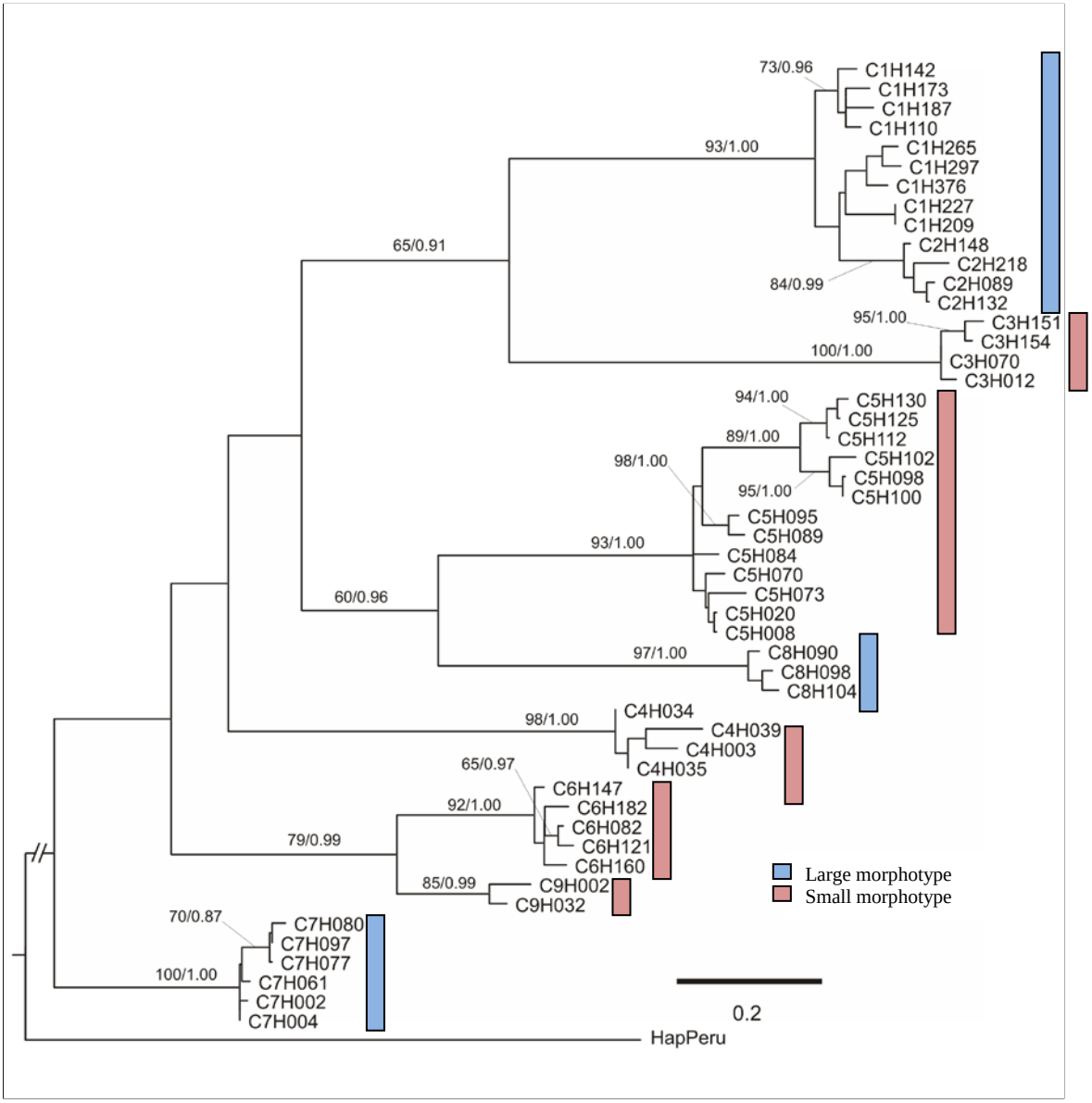


Figure 2.5: Maximum Likelihood tree of 16S haplotypes
This cladogram was constructed using the 16S sequences associated with that COI haplotype. Support values were derived from maximum likelihood – 1000 bootstraps/Bayesian posterior probability. Only bootstrap support values greater than 60 were indicated. The outgroup (HapPeru) was selected from a Peruvian freshwater amphipod (*Mesohyalella* sp.). The large and small-bodied species are denoted in blue and red, respectively.

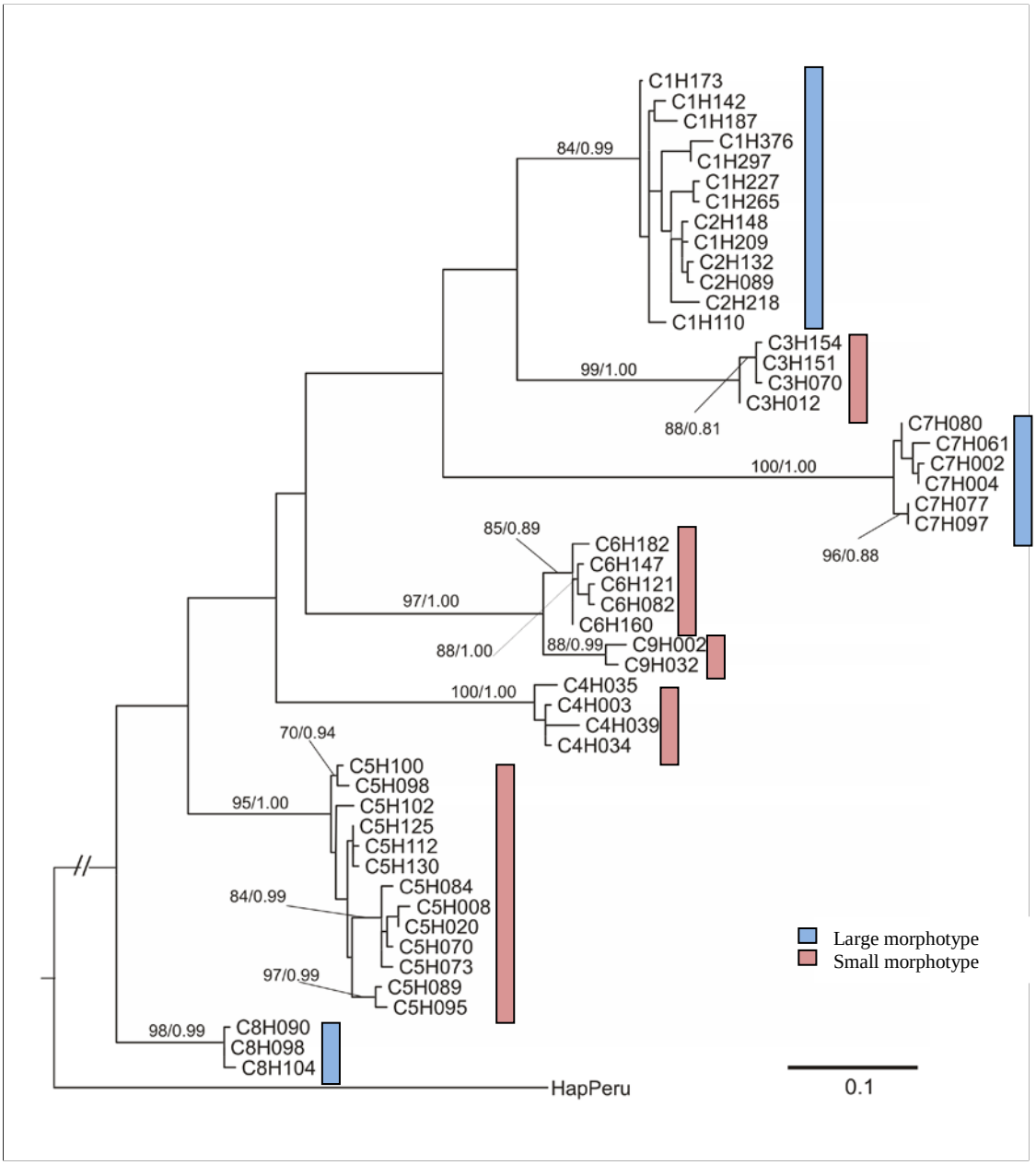
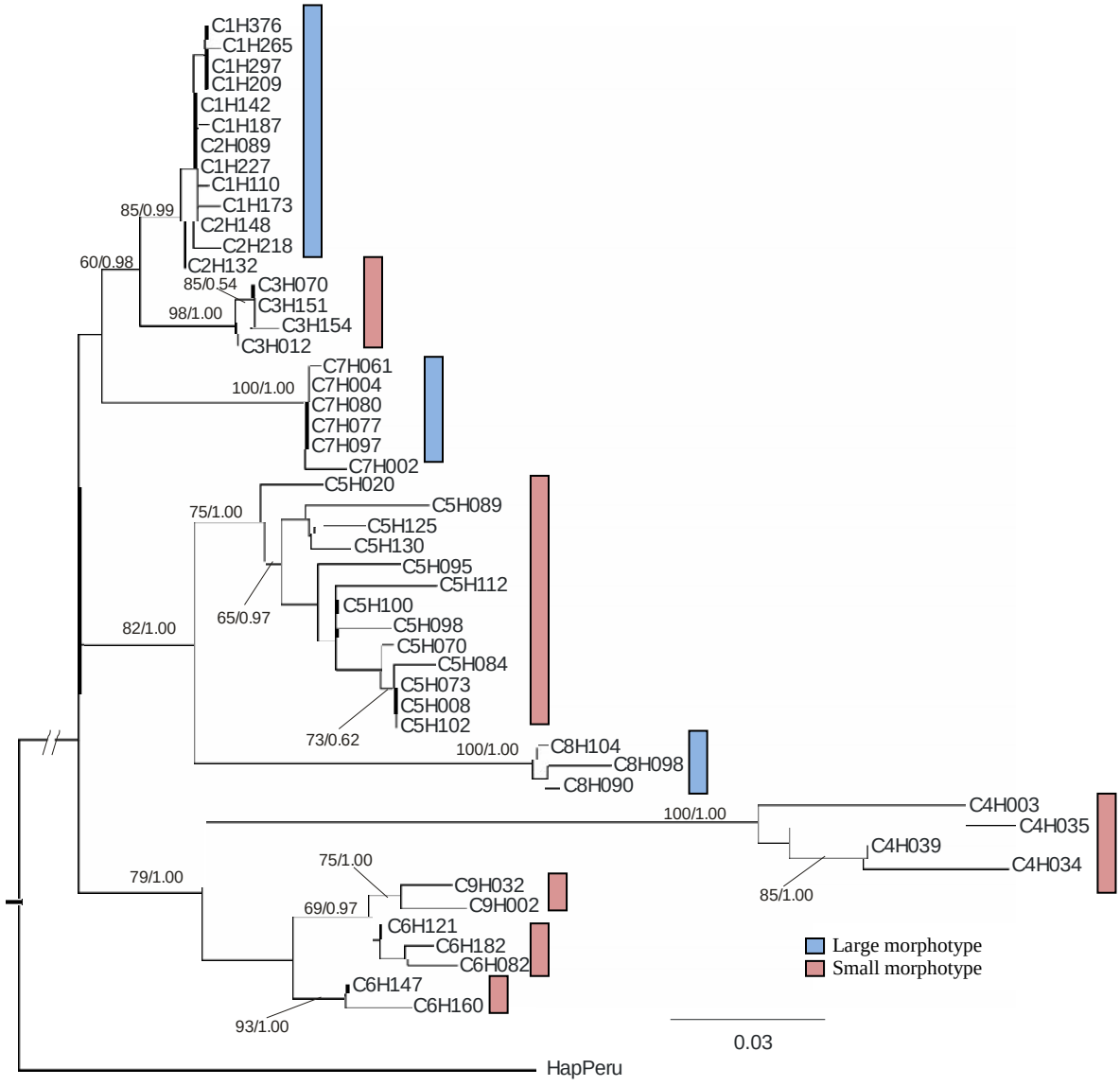


Figure 2.6: Maximum Likelihood tree of 28S haplotypes

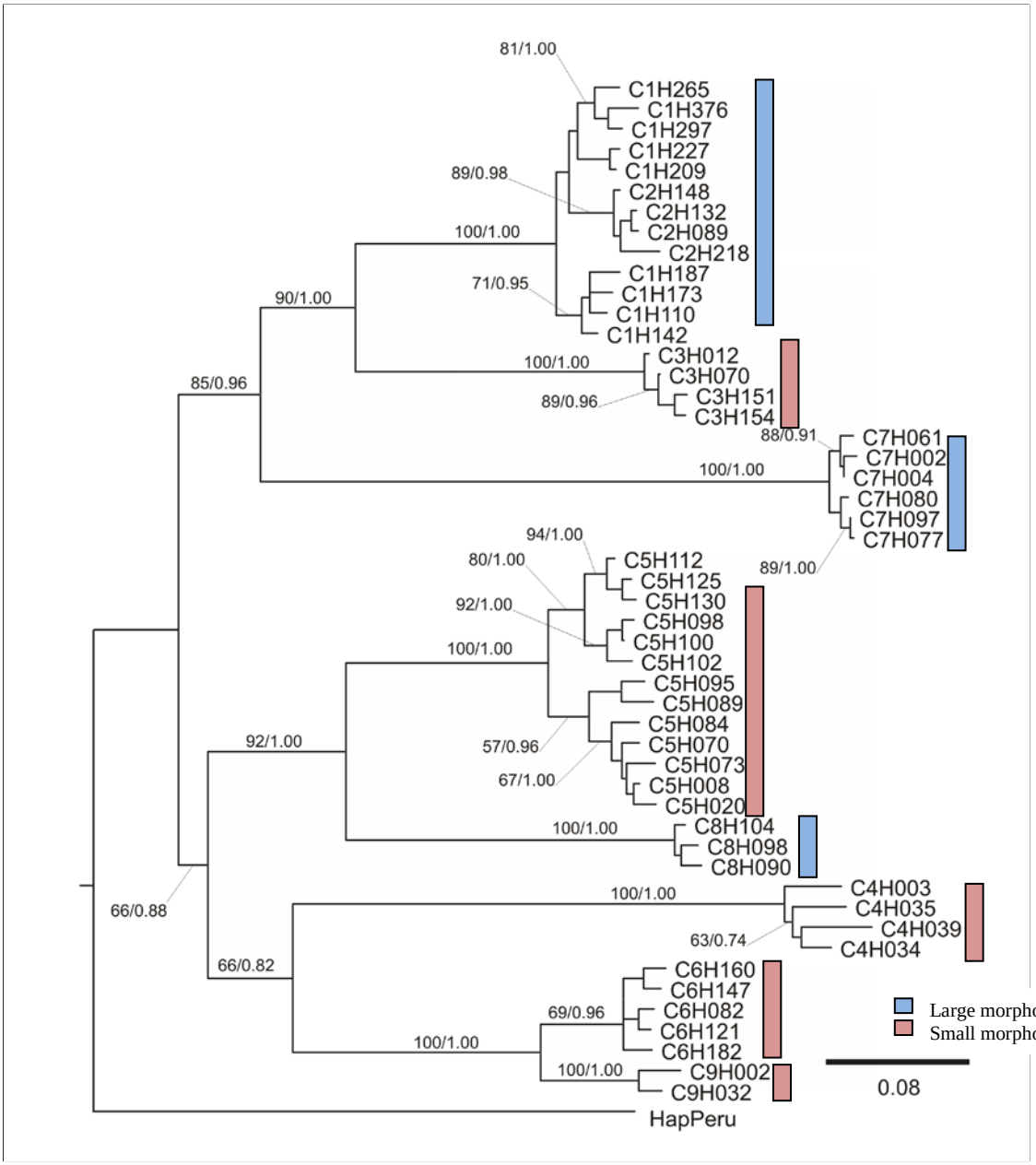
This cladogram was constructed using the 28S sequences associated with that COI haplotypes. Support values were derived from maximum likelihood – 1000 bootstraps/Bayesian posterior probability. Only bootstrap support values greater than 60 were indicated. The outgroup (HapPeru) was selected from a Peruvian freshwater amphipod (*Mesohyalella* sp.). The large and small-bodied species are denoted in blue and red, respectively.



2.3.8 Congruence of COI, 16S and 28S genes in *Hyalella 'azteca'*

All three pairwise partition homogeneity tests rejected the null hypothesis of congruence (COI:16S p=0.0001, 16S:28S p=0.0002, and COI:28S p=0.0001) as well as the three-gene concatenated trial (COI:16S:28S p=0.0001). The gene trees displayed varying topologies with the placement of species 7 at the base of the COI tree. Species 6 and 9 were consistently grouped together, as well as species 5 with species 8 and species 1 with species 3. The species 4 branch lengths of the 28S rDNA gene tree were much longer than the other *Hyalella* species. Despite incongruence among markers, the three markers were concatenated and a combined tree was reconstructed with ML and Bayesian methods (Fig 2.7). The posterior distributions of the Bayesian analysis were mapped onto the nodes of the ML tree with the best likelihood score.

Figure 2.7: Maximum Likelihood tree of concatenated COI/16S/28S haplotypes
This cladogram was constructed using the concatenated sequences derived from the COI, 16S and 28S haplotypes. Support values were derived from maximum likelihood – 1000 bootstraps/Bayesian posterior probability. Only bootstrap support values greater than 60 were indicated. The outgroup (HapPeru) was selected from a Peruvian freshwater amphipod (*Mesohyalella* sp.). The large and small-bodied species are denoted in blue and red, respectively.



2.3.9 Molecular Clocks

Tests for molecular clocks on all three genes returned significant deviations from uniform rates of molecular evolution. The best tree determined by the ML analysis was imported into PAUP version 4.10b and the log likelihood scores using a GTR+ Γ+I model and a GTR+ Γ+I with an enforced molecular clock were calculated. The χ^2 test had significant p-values (COI, *p = 5.17E-18*; 16S, *p = 2.2.2E-05*; 28S, *p = 2.753E-11*) for all three markers (table 2.10). The RAxML trees with enforced molecular clocks significantly deviated from the trees without molecular clocks and suggested that none of the genes (COI, 16S, or 28S) evolved at a consistent rate. Otherwise, the age of cladogenic splits could have been estimated from the magnitudes of molecular divergence.

Table 2.10: Molecular clock calculations of COI, 16S and 28S
The log likelihood scores are derived from the comparison RAxML trees with (-lnL clock) and without (-lnL no clock) an enforced molecular clock. The Likelihood Ratio Test (LRT) was twice the difference between the two log likelihood scores. P-values were determined with a χ^2 test using degrees of freedoms (DF) equal to the number of taxa less two.

Molecular Clock	COI	16S	28S
-lnL (clock)	5625.470	2644.857	4212.791
-lnL (no clock)	5531.917	2594.726	4140.773
LRT	187.106	100.261	144.037
DF	49	49	49
χ^2 p-value	5.171E-18	2.202E-05	2.753E-11

2.4 DISCUSSION

Once thought to be North America's most widely distributed amphipod, *Hyalella 'azteca'* is now considered a morphologically cryptic species consisting of over 80 evolutionary significant units. This study's principal outcome is the characterization of the *Hyalella*

'azteca' species complex with regards to the spatial distribution of its molecular diversity. While the vast majority of these members are restricted to endemic springs within the Great Basin (Witt *et al.,* 2006), the potential for broadly distributed *Hyalella 'azteca'* species were hinted in Witt & Hebert (2000) and Wellborn & Broughton (2008). In this study, it is quite evident that some members of the complex occupy vast geographic ranges, especially in previously glaciated regions. Even in the eastern half of nonglaciated North America, spatial patterns consisting of multiple endemic *Hyalella* species (like seen in the Great Basin) never materialized.

2.4.1 Spatial distribution of the *Hyalella 'azteca'* species complex

Species 1, 2, 7 and 8 are the large-bodied members of the *Hyalella 'azteca'* cryptic complex and their geographic ranges far exceed what has been previously characterized (Witt & Hebert, 2000). Species 1, 2 and 7 have all been sampled from the Great Lakes region, but this study demonstrates that their distributions extend into the Mississippi River and Yukon-Mackenzie drainages. Species 8 has only been previously identified in populations in Oklahoma (Wellborn & Broughton, 2008) and represents the only large-bodied species to be exclusively restricted to unglaciated regions. The populations identified in this study indicate a broad distribution throughout the continental United States.

The small-bodied morphotypes tend have more restricted distributions in comparison to those of large-bodied *Hyalella 'azteca'*, although species 8 was predominately detected within the Florida Peninsula. Species 4 was detected in a small region within the southeastern United States, aside from a single population in New

Brunswick. Species 9 (sister taxa to species 6) was the only small-bodied morphotype on the west coast, confined to Oregon, Washington and British Columbia. Species 3, 5 and 6 are all co-distributed through central and eastern United States and Canada.

As indicated by the high degree of sympatric populations, there appears to be a great deal of tolerance among *Hyalella 'azteca'* species for possible competitors. Notably, a significant number of these localities are within the Great Lakes region. This potentially indicates that this area may be a suture zone (Hewitt, 2001), which is a region whereby migrating or colonizing species (or lineages) cause an admixture of populations. This area of secondary contact may result in the hybridization among more closely allied *Hyalella 'azteca'* species. In time these sympatric groups may decline as competition for similar niches drive some populations to extinction. The degree of sympatry within habitats decreases outside the Great Lakes region, with populations in Florida and along west coast being particularly homogenous.

2.4.2 Evolutionary history of size morphology

Size morphology was not a characteristic that aligned with any one monophyletic group in the COI, 16S, 28S or concatenated phylogenies; a pattern that was directly addressed in Wellborn & Broughton (2008). This indicates the potential for it to have switched at multiple instances during the evolution of *Hyalella 'azteca'*. On the basis of the concatenated maximum likelihood and Bayesian phylogenies, either a small-bodied or large-bodied ancestor is equally parsimonious, involving three switches. Regardless of the state of the ancestor to the cladogram, the direct ancestor of species 2, 3 and 7 would also have been large with a switch to small-bodied along the branch to species 3.

Similarly, the root of species 4, 5, 6, 8 and 9 would have been small-bodied with change to large-bodied along the branch to species 8. The third change depends on the ancestral character, if it was small than there was a switch along the branch leading to species 1-2-3-7, otherwise, if it was a large morphotype, then the third switch occurred prior to the species 4-5-6-8-9 clade. This further underscores the cryptic morphology within *Hyalella 'azteca'* species complex. While morphological characters have historically formed the bedrock of taxonomy, spawning several species concepts, this approach does little to elucidate the evolutionary history of these groups. If the primary selective pressures are fish predation and population density, than the heterogeneity of the adaptive landscape could change across both geography and time. As conditions change, it is not unreasonable that a large-bodied morph may be forced to inhabit fish-filled habitats, thus producing increasingly smaller generations as the larger amphipods undergo disproportionate predation. If different morphs were better suited at different times, especially during the restricted habitat availability of the glacial advances, then it is not inconsistent that a morphotype may eventually switch under selective pressure (Wellborn *et al*, 2005). The plasticity of size-based morphology in *Hyalella 'azteca'* remains unexplored. It is unknown whether this is a strictly inherited trait or the product of canalization, whereby the underlying genetic diversity can support a greater diversity of phenotypes, but the environment consistently eliminates them.

2.4.3 The absence of phylogenetic congruence

Heterogeneity in the COI sequence evolutionary rate is well documented in the endemic *Hyalella* populations of the southern Great Basin (Witt *et al*., 2008). These relative rates

have at least a 2.8-fold difference among closely related lineages. This trend continues among these nine *Hyalella* species, with notable incongruence between all three markers (COI, 16S, and 28S). The potential for rate heterogeneity is further supported by the likelihood ratio test, which rejected a clock-like rate of evolution for all three markers.

Despite the incongruence among the markers, the nine *Hyalella* species resolved into monophyletic groups with high posterior distribution and bootstrap support for the majority of gene trees with a few notable deviations. Species 1 and 2 were paraphyletic with respect to the COI and concatenated trees. Witt & Hebert (2000) noted a sympatric population comprised of both species had several fixed allozyme differences as well as loci in Hardy-Weinberg disequilibrium, indicating a high likelihood of reproductive isolation. Furthermore, species 6 and 9 were largely reciprocally monophyletic, except for the more conserved 28S, whereby species 6 was paraphyletic with respect to species 7. Given the slower rate of evolution of 16S and 28S rDNA genes, it is not surprising that COI has progressed towards reciprocal monophyly at a faster rate. Barrowclough and Zink (2008) referred to mtDNA markers as the 'leading' indicators and nuclear DNA as 'lagging' to explain this type of discordant pattern. The stochastic process of lineage sorting can preserve shared alleles among populations that are reproductively isolated, even in the absence of balancing selection. Alternatively, incongruence among mitochondrial and nuclear markers may indicate sex-biased gene flow, selection or hybridization (Palumbi *et al.,* 2001).

A potential limitation of this study is the manner in which the specimens were chosen for further amplification of the 16S and 28S genes. Individuals were selected to

maximize COI variation, which may potentially exacerbate the degree of rate heterogeneity.

A comparative phylogeographic approach will be used in the subsequent chapters to explore patterns in COI diversity and draw a line between genetic and spatial variability. Chapter 3 will deal exclusively with the large-bodied *Hyalella 'azteca'* species, including the methodology and justifications for the biogeographic techniques. Chapter 4 will continue these themes by exploring the spatial distribution of small-bodied morphotypes. The objective is to analyze the current distribution of COI variation in order to ascertain both current and historic demography. This will hopefully allow the reconstruction of the history of these species and placed in the context of the Pleistocene Glaciations. *Hyalella 'azteca'* is a highly varied cryptic species complex, not only in spatial distributions, but in heterogeneity of its molecular markers. This makes it a fascinating organism on which to test biogeographic hypotheses and gain a greater understanding of the evolutionary forces shaping freshwater organisms.

Chapter 3

Phylogeography of Large-Bodied *Hyalella* species

3.1 INTRODUCTION

In this chapter, an in-depth phylogeographic analysis is conducted on the large-bodied species belonging to the *Hyalella 'azteca´* complex. Only recently have these species been characterized and very little is known about their ranges or distributions. This information is required as a critical first step towards understanding the current and historical forces that have shaped the evolutionary relationships within the *Hyalella 'azteca*' species complex. Our aim is to use modern biogeographic techniques to analyze the co-variation of the geographical and genetic data among populations.

3.1.1 The demographic effects of range expansions

The demographic effects of spatial expansion and vicariance have likely had profound genetic consequences, but these processes have not received much attention in the literature (Excoffier & Ray, 2008). The relatively nascent field of colonization genetics has largely focused on admixture zones or the reconstruction of historical distributions from the fossil record (Excoffier *et al.,* 2009). These approaches are problematic with a species complex like *Hyalella 'azteca',* because such zones have not yet been identified, and even if there were fossilized remains, species delineation would be impeded by its morphologically cryptic nature. Recently, the more common methods of reconstructing the demographic history of species have relied on phylogeographic approaches, involving the overlay of patterns of genetic variation (in the form

of haplotype networks or trees) onto species' geographical ranges. By identifying significant associations among genetic and geographical variables, the evolutionary histories of taxa can be uncovered. As a result of the advances in molecular techniques, this approach has evolved from using restriction fragment polymorphisms to the now more widely used mitochondrial markers.

Avise (2000) used mitochondrial haplotype networks to study the underlying causes of present day population structures among co-distributed taxa (green anole lizard and pocket gophers, among others) in the southeastern United States. Phylogroups separated by deep divergent splits are likely derived from allopatric isolation as they develop *de novo* mutations and intrinsic reproductive barriers over time. Extrapolation from the congruent phylogeographic splits among several freshwater and terrestrial taxa hinted that the Appalachian Mountains formed a potent barrier to gene flow (Soltis *et al*., 2006). Similar approaches have been used to identify glacial refugia in the Northwestern USA (Shafer *et al.*, 2010) and Europe. Recent biogeographic research has been accompanied by an explosion in interest regarding refugial populations for insights into species' reactions to climate change (Keppel *et al.*, 2012) and the identification of potential conservation targets (Hampe & Petit, 2005).

Unfortunately, many freshwater invertebrates (including *Hyalella 'azteca'*) leave no fossilized remains and their demographic histories must be determined by applying models of vicariance and spatial expansions to the present-day distribution of genetic variation. Only recently, with the greater acceptance of coalescent-based simulations, have we been able to create more realistic models for these dynamic and often stochastic processes. Working backwards from the tips of a phylogenetic tree, the probability that any two taxa (or haplotypes) will coalesce (two branches on a tree will converge into a single ancestor) in the previous

generation (having the same direct ancestor) is $1/(2N_e)$, where N_e is the effective population size. In a population undergoing demographic expansion, the number of individuals increase rapidly, this in turn decreases the probability of coalescence because of the greater effective population size. This explains two well established trends in growing populations: they have less genetic drift than stationary populations (Kimura & Crow, 1963) and there is an increase in rare alleles within the frequency spectrum (Harpending & Rogers, 2000). Haplotype networks are useful in highlighting these effects as an excess of singletons can result in star shaped branching patterns around central nodes.

While new mutations are likely to go extinct or remain at low frequencies in refugial or stationary populations, rare alleles at the edge of an expansion wave have the potential to propagate and reach extremely high frequencies (Edmonds *et al.*, 2004). This phenomenon is referred to as allele surfing. These alleles can be neutral or even slightly deleterious (Travis *et al*, 2007; Hallatschek & Nelson, 2010), though the sweeping haplotype is more likely to occur among the more frequent alleles (Klopfstein *et al.,* 2006). Simulation studies and experiments with yeast (Hallatschek & Nelson, 2010) have described a variant of this process called sectoring. Instead of a single allele propagating along an expansion front, different alleles can disperse in independent waves to occupy distinct regions. This creates areas of uniform genetic diversity bordered by sudden shifts in allele frequencies. If the organisms migrate through narrow dispersal corridors, such as rivers or land bridges, the propagation wave typically endures a series of bottlenecks. Coalescent events cluster both before and after bottlenecks and can increase the probability of allele surfing, even involving deleterious alleles (Burton & Travis, 2008). Founder populations often possess haplotypes with short coalescence times as small population size promotes allele loss to genetic drift. However, when a new population rapidly

expands, allele frequency fluctuations as a result of drift will be prevented, and coalescence times extended. This genetic pattern will be repeated as the surfing allele propagates to each new habitat along the expansion front. These events should leave distinct observable patterns in the phylogenetic structuring of haplotypes and can be identified by nested clade analyses (NCA), mismatch distributions, Tajima's D and the hierarchical analysis of molecular variance (AMOVA/SAMOVA).

3.1.2 Introduction to Nested Clade Analysis

Nested clade analysis (NCA) was designed by Templeton (1998) to provide a method to test the significance of the inferred geographical associations and enhance the still popular method of overlaying haplotype networks on top of spatial distributions. Without this measure, there is no way to ascertain if an association has sufficient sample sizes or localities to merit further investigation. A series of F-statistics have been the traditional method for such analyses, but since they only take current spatial variation into account, it is difficult to disentangle historical demographic processes from present-day ones. Nested clade analysis utilizes the temporal advantage of haplotype networks and applies the statistical rigor of random permutational contingency tests (Templeton, 2009a). The patterns of significant geographical associations formed from the nested clades are then used to infer possible determinants (Templeton, 1998). A major strength of NCA is that by partitioning the network into nested clades, it can identify when both current and historical effects are at play by the order of events in a nested series. A population expansion inferred at a lower tier would be more recent than one at an upper tier. By application of the inference key, the NCA can identify past fragmentation, long distance

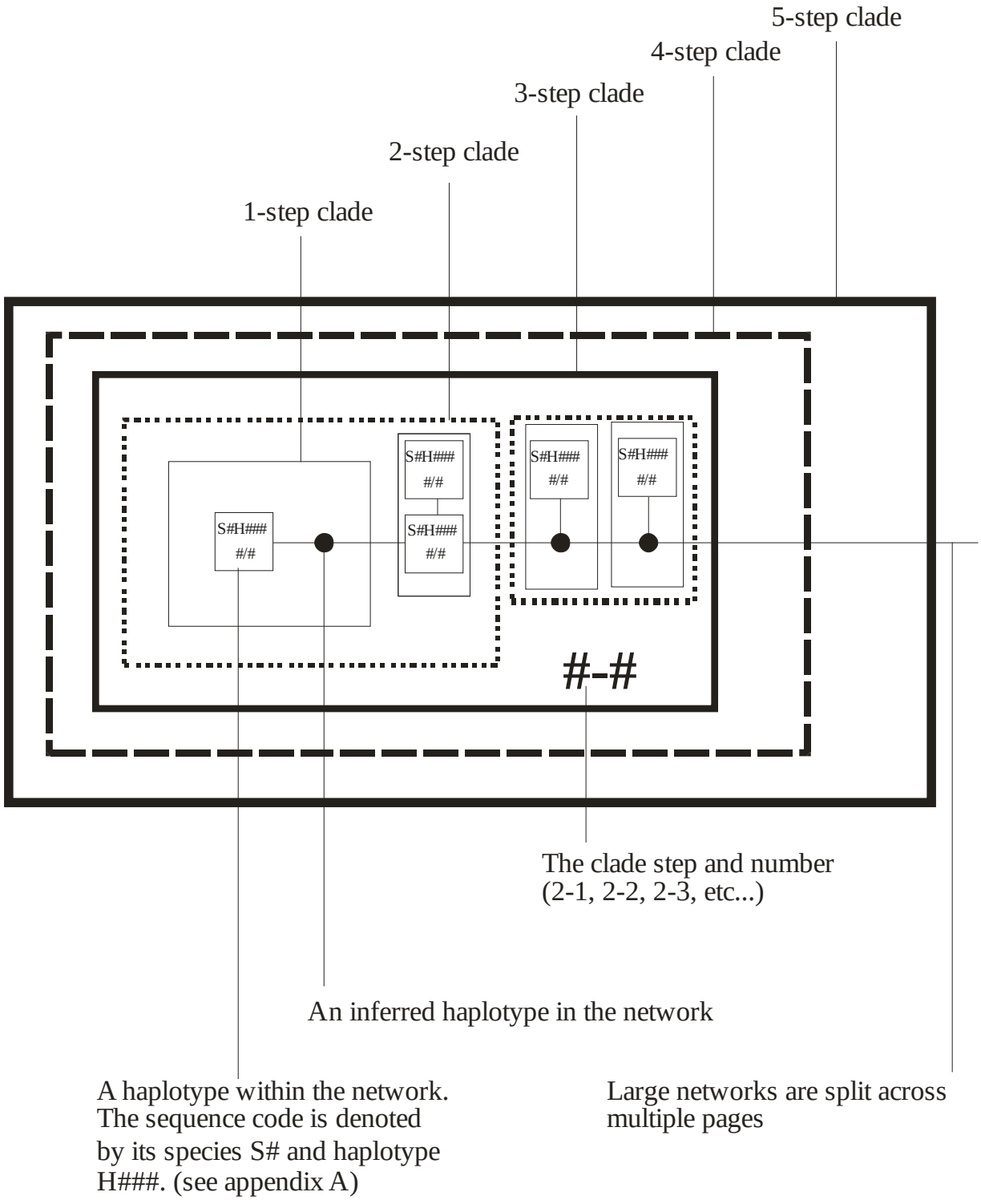
colonization or contiguous range expansion events (Posada *et al.*, 2000). This key was derived from a combination of simulation, model theory and the study of populations that have clearly undergone range expansions into previously uninhabited zones (i.e. previously glaciated regions).

While any network can be nested and used for NCA, the traditional method employed by Templeton & Sing (1993) was statistical parsimony via a program called TCS 1.21 (Clement *et al.*, 2000). At intraspecific levels, they argue that parsimony and likelihood will give similar results. The program TCS employs a parsimony criterion to estimate the probability of a non-parsimonious state, which is the result of hidden mutations derived from multiple nucleotide changes at the same site. These unseen mutations introduce error into the network, though likelihood and Bayesian methods can estimate the number of changes on the basis of probabilistic models. As such, TCS evaluates the limit (usually 5%, though any value can be selected) at which confidence in parsimony deteriorates. Long edges that possess numerous inferred haplotypes would be problematic for the nested clade analysis. The program breaks these branches, which separates the haplotype network into distinct networks. This provides a simple breaking point to delineate COI lineages within a species.

Nested clade analysis is a popular method used to identify the demographic processes that have shaped the distribution of molecular variation (Panchal & Beaumont, 2007). Specifically, NCA provides a framework to recognize clades within a network with sufficient sample sizes to discern geographical associations. The process of nesting the network, as outlined by Templeton (1998), is as follows (refer to figure 3.1). The haplotypes located at the tips (nodes only linked to one other real or inferred haplotype) are combined with their most

immediate interior ancestors into 1-step clades. If multiple tips adjoin the same node, they are all included in that clade. Since all haplotypes must be grouped into 1-step clades, the next step is to group the interior 1-step clades that have no tip haplotypes). To accomplish this, the 1-step clades already formed are ignored and the new tip haplotypes (haplotypes now connected to a single node) are combined with interior nodes. This process of grouping and pruning is repeated until all haplotypes are combined in 1-step clades. The 2-step clades are assigned using the same process, except the 1-step clades act as the new tips and interior nodes. The process is repeated for however many n-step clades that are required for the entire network to be contained within a single clade. After the nesting procedure is complete, every *n*-step clade should be composed of both (*n* minus 1)-step tip clades and (n minus 1)-step interior clades, or in the case of 1-step clades, tip and interior haplotypes. However, some clades at the center of the network may only be composed of interior clades. Understanding the terminology of tip and interior clades and how they relate to each other is crucial to NCA. A tip clade (or haplotype) has only one connection leading to it, whereas an interior clade has two or more edges leading to other nodes. While some interior haplotypes and clades are temporarily considered tips for the process of grouping interior clades, they do not maintain that status for the permutation tests and inference key. In figure 3.1 the box marked as “1-step clade” contains a tip haplotype and an inferred interior haplotype (black dot). This 1-step clade becomes the tip clade for the ensuing 2-step clade, which also has an interior 1-step clade. This process is repeated again as this 2-step clade becomes the tip clade for the 3-step clade in which it is nested.

Figure 3.1: Schematic of a Nested Clade Analysis



The NCA nesting rules were later amended to account for symmetrical clades (Templeton, 1998). In the case of five haplotypes linked in a row, the two outer nodes on each side combine into 1-step clades leaving a single ungrouped haplotype in the middle. In order to determine which 1-step clade the ungrouped haplotype should be grouped with the middle node will be assigned to the 1-step clade with smallest sample size in order to increase the overall power of the NCA. This technique is applied to all tiers of the nesting process.

The program GeoDis applies random permutation tests to geographic distances derived from populations contacting the nested clades in order to identify geographical associations (Posada *et al.*, 2000). For each clade (or haplotypes) within a nested clade it determines the clade distance (average distance of populations to the geographic center of the clade) and the nested clade distance (average distance of populations to the geographic center of the nesting clade).Then it calculates the I-T clade and nested clade distances, which is the difference between the average interior clade (or nested clade) distance and the average tip clade (or nested clade) distance. The distribution of these distances is contrasted against a null hypothesis of no structure. For each statistic, the NCA can return distance values that are either significantly small or large. A clade that is broadly distributed will likely have a significantly large clade distance, alternatively a clade restricted to a single population will have a significantly small clade distance. The I-T measures provide a temporal comparison, because the tip clades are derived and will be younger than the interior clades (Templeton, 2004). Furthermore, these measures often have greater power to reject the null hypothesis.

3.1.3 The debate surrounding nested clade analysis

The literature is currently host to a vigorous debate regarding the efficacy and assumptions of the nested clade analysis. The most salient complaints regarding NCA are that the application of the inference key can lead to false positives (Panchal and Beaumont, 2007; Petit, 2008) and that there is no means to evaluate the veracity of the inferences therein (Beaumont *et al*., 2010). Templeton (2009b) argues that the explicit interpretations of the NCA are grounded in neutral coalescent theory and on the basis of simulations reinforced by 150 positive controls (Templeton, 2008). Much of the argument rests on the comparative strengths and weaknesses of approximate Bayesian computation (Fagundes *et al.,* 2007) vs. NCA, but the ultimate factor for the analysis of the present *Hyalella* data set was computational efficiency. Templeton (2005) analyzed a multilocus data set of over 2000 haplotypes in at least 40 locations, whereas other such methods at the time (approximate bayesian computation) were computationally more difficult with larger data sets. In order to analyze clade 2 for example, the analysis must be able to handle over 130 sites.

Difficulties with false positives derived from the inference key can be mitigated by employing a holistic approach. Any inference derived from the NCA should be supported by corroborating evidence from mismatch distributions and population genetics. A mismatch distribution is a graph plotting the frequency of pairwise distances against the actual pairwise genetic distance. A unimodal peak infers either continuous growth over a long time period or a past explosion in population numbers. In contrast, as alleles are lost within stationary populations, the distribution becomes ragged and multimodal. Harpending *et al.* (1993) developed a raggedness (r) statistic that could reasonably distinguish population expansions from

stationary groups using simulated data. The mismatch distribution's parameters are estimated and used to create a coalescent simulation of a population expansion. If there is a significant deviation between the two curves, then the null hypothesis of population expansion is rejected. For populations with a high level of migration among neighboring demes, a mismatch distribution approaches a demographic expansion (explained above) with a unimodal peak (Excoffier, 2004). If the organisms are highly vagile, then new mutations will be able to migrate out before they can coalesce. This effect would be more pronounced in populations with larger carrying capacities (Excoffier *et al.*, 2009). Organisms with weaker dispersal capabilities will have coalescence events throughout the expansion wave and approach a bimodal mismatch distribution (Excoffier *et al.*, 2009).

Another criticism leveled at NCA is that the methods used to produce networks are varied and can result in different conclusions (Cassens *et al.*, 2005). This can be applied to all network-based approaches and is more problematic for molecular markers undergoing recombination or horizontal gene transfer. Different network methodologies handle the reticulation of the network caused by these effects (or not at all) and can result in widely varied results.

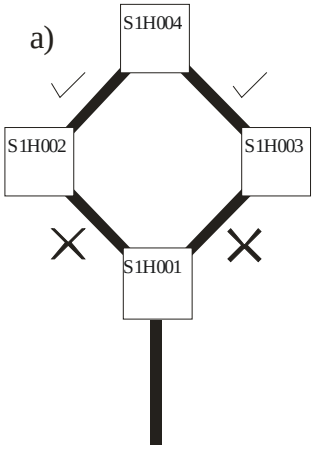
Rate heterogeneity can be problematic in assessing evolutionary relationships, and even molecular markers that evolve in a clock-like manner can be highly variable over short time spans. If a gene has a mutation rate of 10 changes per million years, they can all occur in rapid succession or spread across thousands of years. This effect is likely is more muted at the higher steps of a nested clade analysis, which encompass longer time frames. Given small sample sizes, it is unlikely that many of the 1-step clades will meet the threshold of significant geographical associations. As these smaller nodes are grouped into higher clades, much of that uncertainty will

be lost. Permutation tests for significance at a given tier do not depend on the topology of the lower tiers. For example, the configuration of 1-step clades within a 2-step clade won't matter; all of the haplotypes from each of the nested 1-step clades will be pooled into their respective 2-step clade. Though, some decision points along the inference key depend on geographical patterns identified at lower tiers.

The comparisons of geographic distributions between lower and upper tier clades in a hierarchical series are used within the NCA to discern long distance colonization from past fragmentation. Discerning between these two patterns is important to biogeographic research to ascertain the contribution of dispersal and vicariance in shaping a species distribution. If an upper tier clade possesses a disjunct distribution and the separate groups contain lower tier clades with over-lapping distributions, the inference key of the NCA favors past fragmentation. Of much greater concern is the process used to break reticulated loops, especially if they would influence the structure of higher tiers. On the basis of Templeton *et al.*, (1992) rules of constructing haplotype networks, three criteria have been suggested to resolve reticulation (figure 3.2). The topographic criterion breaks loops in order to maximize the connections to central nodes. These internal haplotypes are likely ancestral and would have had greater time to produce new mutations than tip clades. In a four haplotype loop with an internal node attached to an external by two possible intermediary nodes, this criterion would break the loop between an intermediary and the tip as opposed to between the internal and the intermediary. The frequency criterion is very similar to the method in resolving symmetrical clades, in which loops are broken to maximize sample size. In our four loop example above, breaking the loop between the intermediary and the external still leaves two possible branches to break. If the first intermediary node had fewer samples than the second, then the external node would be grouped with the first

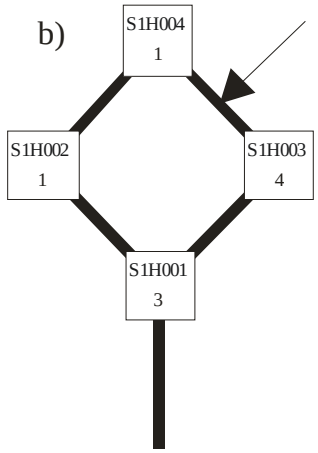
intermediary node. Finally the geography criterion breaks loops in order to group haplotypes that are more geographically proximal as opposed to distant. It is more likely that a new mutation would be locally restricted, especially in species with poor dispersal capabilities. Given the cited ability of *Hyalella 'azteca'* to migrate via avian vectors (Swanson, 1984), the geographical criterion was used only to break loops that the other two could not, in order to prevent underestimation of long distance dispersal events.

Figure 3.2: Criterion to break loops in a Nested Clade Analysis
a) Topographic Criterion, b) Frequency Criterion and c) Geography Criterion.



Topography Criterion
Break loops to maximize internal connections.

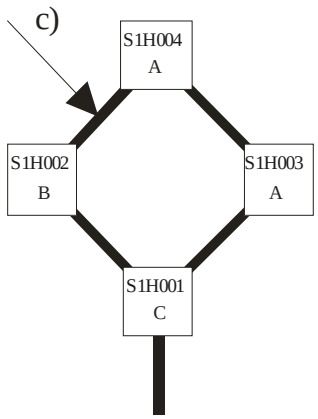
S1H001 is the internal node that connects to the remainder of the network.



Frequency Criterion
Break loops to increase smaller sample sizes.

Numbers below the haplotype names represent the number of copies.

The links between S1H001 and either S1H002 or S1H003 are not eligible to be broken because the topography criterion is applied first.



Geography Criterion
Break loops to group similar sites.

Letters beneath the haplotype names represent populations.

The links between S1H001 and either S1H002 or S1H003 are not eligible to be broken because the topography criterion is applied first.

3.1.4 Tests of neutrality and population expansion

Tajima's D (Tajima, 1989) has been used as a sensitive measure to identify deviations of molecular variation from a null hypothesis of neutrality. Tajima's D is the difference between two other measures that both estimate the effective population sizes; S is the number of segregating sites and π is the average per nucleotide diversity or the average number of nucleotide differences per site. Since both estimate the same parameter, $S = \pi$ and $S-\pi = 0$ becomes the null expectation and deviations can illuminate underlying factors. If the difference is negative it may indicate that the population is undergoing expansion, an extreme bottleneck or purifying selection (Tajima, 1989). Though not strictly a critical value, a score less than -2 is considered significant. Populations undergoing range or demographic expansions will harbor a higher frequency of rare alleles coupled with a lack of intermediary alleles. This pattern inflates the average per nucleotide diversity (or π) and thus the Tajima's D test will return a negative value (Excoffier *et al.*, 2009).

Further tests of neutrality include mismatch distributions using raggedness (Harpending, 1993), Ramos-Onsins & Rozas R_2 (Ramos-Onsins & Rozas, 2002) and Fu's *F* (Fu, 1997). The bonferroni correction for multiple testing was used to determine appropriate *p*-values (Librado & Rozas, 2009). On the basis of a simulation analysis, Ramirez-Soriano *et al.*,(2008) determined that the Fu's *F* and R_2 statistic have the greatest power, especially in cases of large sample size for the former and small sample size for the latter. Given that *Hyalella 'azteca'* is a cryptic species complex and cannot necessarily be identified in the field to ensure adequate collection numbers, the potential for sample size variance among groups is high. R_2 is similar to Tajima's *D* in that both are considered class 1 statistics that rely on the number of singletons as compared to

the number of nucleotide differences. Fu's *F* is a class II statistic, which is based on haplotype distribution and provides an independent measure of deviations from neutrality. The raggedness statistic remains the weakest of these statistics and most subject to power decay with the increase of time since the expansion events. However, it is unique in that it uses a simulated demographic expansion with parameters derived from the data set as the null hypothesis instead of neutrality. Although, Tajima's *D*, Fu's *F* and R_2 are very powerful, they cannot distinguish a demographic expansion from selective sweeps, severe population bottlenecks or genetic hitchhiking. These competing explanations of the molecular variation must be considered against the life history and biogeography of the organism.

3.1.5 Analysis of molecular variance

The analysis of molecular variance (AMOVA) is a technique designed by Excoffier *et al.,* (1992) that partitions the variance components of a data set according to the structure among populations. The haplotype data collected from populations is converted into the Euclidean vector ϕst (an analog of Fst) by combining the frequencies of haplotypes with a matrix of pairwise genetic distances. In an AMOVA, populations are grouped *a priori* in accordance with the hypothesis being tested. The three variance components (within populations, among populations within groups and between groups) are calculated and the resulting values will sum to 100%. If a particular subdivision of populations proposed by a hypothesis is meaningful, then a greater degree of variance will be explained by the between group component. On the other hand, if populations are admixed with little structure, the within populations variance component will be greater. This allows comparisons among different hypotheses, to determine which

subdivisions of populations maximize the between-group variance component. Bernatchez (2001) used AMOVA to discern patterns of fragmentation among brown trout lineages.

3.1.6 Structural analysis of molecular variance

To test for groups that are unbiased by *a priori* assignment, Dupanloup *et al.*, (2002) expanded the capabilities of AMOVA into a structural analysis of molecular variance (SAMOVA) that determines which combination of populations will maximize the percent of variation explained. This technique can be used even if the number of groups is not known by incrementally increasing the number and stopping when a maxima is reached. However, as the populations are increasingly partitioned, the amount of variance explained among groups will artificially increase as a consequence of fewer populations being included in each assemblage. Furthermore, in simulations with low migration rates, significant isolation by distance within groups can mask the population structure and return incorrect partitions. The degree that these separations represent true genetic barriers, and not merely the muddling influence of isolation by distance, must be assessed in conjunction with other approaches (NCA, population genetics).

3.1.7 Objectives

Hyalella 'azteca' presents an excellent opportunity to study the effects of the Pleistocene glaciations and subsequent range expansions on a genetically diverse complex of species. Intraspecific allelic diversity is often greater in populations inhabiting glacial refugia and is expected to gradually decrease with distance. In Nearctic and Palearctic fishes, intraspecific

nucleotide diversity decreases more rapidly north of the glacial limit (Bernatchez & Wilson, 1998). This trend is not universal, as populations inhabiting "suture" zones within glaciated territories have been discovered with greater heterozygosity as a result of the admixture of previously allopatric groups (Petit *et al.*, 2003). Comparative phylogeographic approaches look at the co-distribution of species, but inferences may be clouded by different life histories or dispersal modes. Moreover, relying on single-species studies can be inadequate to infer regional biogeographical histories (Bromilow & Sperling, 2011).

In this chapter, an in-depth analysis of each of the large-bodied members of the *Hyalella 'azteca'* species complex (species 1, 2, 7 and 8) is conducted. The first objective is to construct haplotype networks of the COI sequences to characterize the relationships among them and then to map those haplotypes to their geographical coordinates. This gives an overall survey of the species' ranges and identifies potential areas of structure within the spatial distributions. In particular, the focus is subdivisions that may have resulted from the biogeographical forces of dispersal and vicariance. To further this aim, the haplotype networks are nested according to the rules laid out by Templeton (1998) and subjected to a nested clade analysis. The inferences provided by this technique will identify the regions of the networks that have been shaped by contiguous range expansions, long distance colonization or isolation by distance.

The species or major lineages therein are tested for deviations from neutrality for the purpose of identifying demographic expansions. Population increases happen in conjunction with colonization events, as these populations can be founded by a few invading amphipods, or even a single gravid female (Wellborn & Capps, 2013). Four semi-independent measures are used; raggedness, Tajima's *D*, Fu's *F* and R_2. Fu's *F* works best with large sample sizes, whereas R_2 is

better on smaller groups, which is a great advantage given the high variance in sample sizes. In the case of demographic expansion, Tajima's *D* and Fu's *F* is significantly negative. Additionally, mismatch distributions will be used to provide a direct test of population increases; if the raggedness statistic is significant, then the null hypothesis of demographic expansion can be rejected. If a species is increasing its range, then the mismatch distribution should appear smooth and unimodal, whereas stationary populations tend to be multimodal. A bimodal distribution may indicate multiple expansion waves at different times or extreme bottlenecking in the expansion front.

The objective of using an AMOVA is to test hypotheses regarding the subdivision of species. By comparing different groupings of populations, the degree of variance explained can be compared. The first condition will test populations assigned to faunal provinces as outlined in McAllister *et al.*, (1986), which were designated according to fish endemism and consistent with the topography of major drainages. The faunal provinces are as follows: Yukon-Mackenzie, Cascadian, Mississippian, Hudson Bay, Great Lakes (including the drainages of the St. Lawrence), Northern Appalachian, Central Appalachian and South-Eastern (including the coastal plain of Georgia). The second condition will analyze the variance components explained by separating populations on the basis of glaciated and unglaciated regions. By comparing these two conditions, it can be determined if separating populations on the basis of glaciated and unglaciated regions is a better explainer of the variation than the faunal provinces.

The principal drawback of the AMOVA is the *a priori* assumption of the structure of populations. The structural analysis of molecular variance (SAMOVA) groups the populations in order to maximize the variance component at the between-group tier (as opposed to the within-

group and within-population tiers). The number of groups is set prior to the analysis, but only from 2 to 6 will be tested. The same analysis series was used to structure among populations of the blackfish *Dallia* (Campbell & Lopéz, 2014). The results based on the 2-group SAMOVA will be accepted unless one of the analyses using a greater number of groups meets the following two criteria. First, that there is large increase of between-group variance explained by further partitioning (at least 10%) and that some of those new partitions have at least three populations. As the number of groups increase, single populations are often placed into a new group with only a minor increase in the variance component at the highest tier. Since the goal is to detect meaningful subdivisions within a species' or lineage's distribution of populations, groups consisting of single habitats may not be particularly useful. Eventually a SAMOVA would split off all populations into single groups. The null expectation is that the SAMOVA of a partition of 2 groups will be used unless more groups can explain a significantly greater percentage at the between-group variance component and that the new groups formed have more than three populations. Major subdivisions need not occur at every increase in group number. For example, increasing the number of partitions may only separate groups restricted to single habitat, whereas a significant subdivision of populations may not reveal itself until a greater degree of partitioning. The SAMOVA will provide a point of comparison against the two hypotheses (populations split according to either faunal provinces or glaciated and unglaciated regions) tested in the AMOVA.

Ultimately, the combined information provided by mapping the haplotype networks, NCA, Tajima's D, Fu's F, R_2, raggedness statistic, mismatch distributions, AMOVA and SAMOVA will be used to characterize the competing biogeographical forces that have shaped present day variation. Specifically, the goal is to assess the effect of Pleistocene vicariance and

post-glacial dispersal on the demographic and evolutionary history of the large-bodied members of the *Hyalella 'azteca'* species complex.

3.2 MATERIALS AND METHODS

In this chapter, the term species will refer to the overarching monophyletic groups identified in chapter 2, lineage will refer to the specific subgroups identified within the species delineated by bootstrap support greater than 70% or by the 95% statistical parsimony criterion in TCS 1.21 (Clement *et al.,* 2000), and clade will refer to the n-step clades within the nested clade analysis.

3.2.1 Partitioning of species 1 haplotypes

The identification of haplotype clusters within species 1 must be undertaken with a great deal of caution, because an incorrect *a priori* haplotype structure may introduce unintended biases into subsequent analyses. That being said, it is necessary to discriminate these groups because population structure can override the effects of isolation by distance and muddle attempts to discern the refugial origins of these mitochondrial lineages. Species 1 has a high level of diversity (π=0.045) with deep divergent splits (7.9% COI sequence divergence between lineages A and G). The COI gene has been estimated to diverge at a rate of 1% per million years (Knowlton & Weigt, 1998), which would suggest divergence prior to the Wisconsinan glaciations (~50k years). However, it is suspected that the rates of molecular evolution within the *Hyalella* complex can be greatly accelerated (Witt *et al.*, 2008). Even if two deeply divergent mitochondrial lineages coexisted within the same refugial populations (from admixture of

historically allopatric groups), inclusion of both groups in a mismatch distribution could produce a multimodal distribution that would override any signal from a demographic expansion. Phylogenetic structure must be taken into account. Statistical parsimony provides an inherent mechanism to delineate such groups, as any haplotypes that do not meet its 95% confidence limit are parsed into separate networks. This is the point at which some nucleotide sites are likely to have experienced multiple nucleotide changes. Disjoint networks parsed in this manner will be separated by at least ten inferred mutational steps in the 637 base pair COI fragment. We assume is that a divergence of that magnitude is sufficient to place the cladogenic split prior to the most recent glacial retreat. Even if the networks were forced together by ignoring the parsimony criterion, the nested clade analysis will treat them as independent groups until a very high clade-step. A similar process was employed by Bernatchez (2001) on distinct lineages of brown trout and had little effect on the mismatch distribution and NCA's ability to discern common effects and admixture among groups. Therefore, the species 1 lineages were either split into their monophyletic groups identified in a neighbor-joining tree (using a Tamura-Nei model of evolution with Γ=0.23) with bootstrap support (using 1000 replicates) greater than 70% (see figure 3.3) or by the 95% statistically parsimony criterion employed by TCS 1.21 (Clement *et al.,* 2000).

3.2.2 Nested Clade Analysis

Mitochondrial COI sequences from chapter 2 were aligned by eye and exported from MEGA 6.06 (Tamura *et al.,* 2013). Statistical parsimony haplotype networks were created in TCS 1.21 (Clement *et al.,* 2000) and reticulated loops were broken according the topographical, then

frequency criteria as outlined previously. Groups were nested according to the rules devised by Templeton (1998) and Templeton and Singh (1993), including the frequency-based solution to resolve symmetrical clades. The NCA assumes that interior clades are ancestral to tip clades, but this analysis can be enhanced by determining the root of the network. This allows the designation of clades as ancestral that otherwise might be considered derived, especially if the higher nested clades coalesce into a network with no internal nodes. For example if the network forms into 4-step clades, without the root, neither can be considered interior on this basis alone. The determination of the root was conducted by comparing the neighbor-joining and parsimony phylogenies derived from the comparison of COI haplotypes to a closely related outgroup with high bootstrap support to ensure it is not an actual member of the ingroup. The neighbor-joining tree was constructed in PAUP 4.10 (Swofford, 2003) using a model selected by Akaike information criteria (Posada & Buckley, 2004) in jmodeltest 2.6. The heuristic parsimony search was conducted in PAUP 4.10b (Swofford, 2003) use tree bi-section re-connection branch swapping if sample sizes are small ($n<100$) or nearest-neighbor interchange. The parsimony analysis employed a starting tree obtained by neighbor-joining (using Kimura-2 parameter model of evolution) followed by 20 search replicates starting from random trees. The NJ and parsimony analyses were compared to determine among which clade within the haplotype network represents the most likely root.

The nested groups along with their geographical positions (decimal degrees) were imported into GeoDis 2.6 (Posada *et al.*, 2000) and patterns of geographical associations were analyzed using the most recent inference key. GeoDis determines the clade distance *Dc* (the average distance of a haplotype from the geographical center of its clade), the nested clade distance *Dn* (the average distance of a haplotype to the geographical center of its nested clade),

and the difference between the interior and tip clades for both the clade and nested clade distances ($Dc_{(I\text{-}T)}$ and $Dn_{(I\text{-}T)}$ respectively). The distances were re-calculated after each of the 10000 random permutation tests and compared against a null expectation of no geographical associations. Large or small *Dc*, *Dn* or *I-T* contrasts were deemed significant at a threshold of 0.05 after the application of the Dunn-Sidak correction for multiple tests. Maps were generated by hand using CorelDraw X3.

3.2.3 Population genetics and mismatch distributions

The mtDNA sequence alignments were imported into DNAsp to assess deviations from neutrality and test for potential demographic expansions. Tajima's *D* and Fu's *F* were calculated and tested for significance against 1000 coalescent simulations based on the θ ($4N\mu$) per gene value estimated by applying a poisson distribution to the mutations along a lineage (Ramos-Onsins & Rozas, 2002). Simulations can also be derived from the number of segregating sites (Watterson estimation) and though this changes the critical values, they were only included if they altered the determination of significance. To assess the potential for population expansion, Tau was estimated for the mismatch distributions and then used to derive a simulated expected curve for a population undergoing a constant rate of expansion using coalescent models. The raggedness statistic (Harpending, 1993) and R_2 (Ramos-Onsins and Rozas, 2002) were calculated and tested for significance against 1000 coalescent simulations. The mismatch distribution was exported and the curve was generated in Excel 2010.

3.2.4 Hierarchical Analysis of Molecular Variance

Sequence alignments were imported into Arlequin 3.5 (Excoffier & Lischer, 2010) along with an arlequin file of population groups structured on the basis of faunal provinces (McAllister *et al*,, 1986) or glaciated and unglaciated regions. The structural analysis of molecular variance was conducted using the SAMOVA program (Dupanloup *et al*., 2002) for two groups and then incrementally repeated up to six. Each new tier was checked to determine if the assemblages of populations had at least 3 members or if the new groups only cleaved off single habitats.

3.3 RESULTS: SPECIES 1

This chapter focuses on the COI sequence data set derived from the large-bodied members (species 1, 2, 7 and 8) of the *Hyalella 'azteca'* species complex. As outlined in chapter 2, species 1 was collected from 184 localities and the COI gene was sequenced for 827 individuals with a nucleotide diversity of 0.0459. Its geographic distribution stretches from the Great Lakes region to Alaska, and was predominately located in glaciated regions.

3.3.1 Species 1 lineages

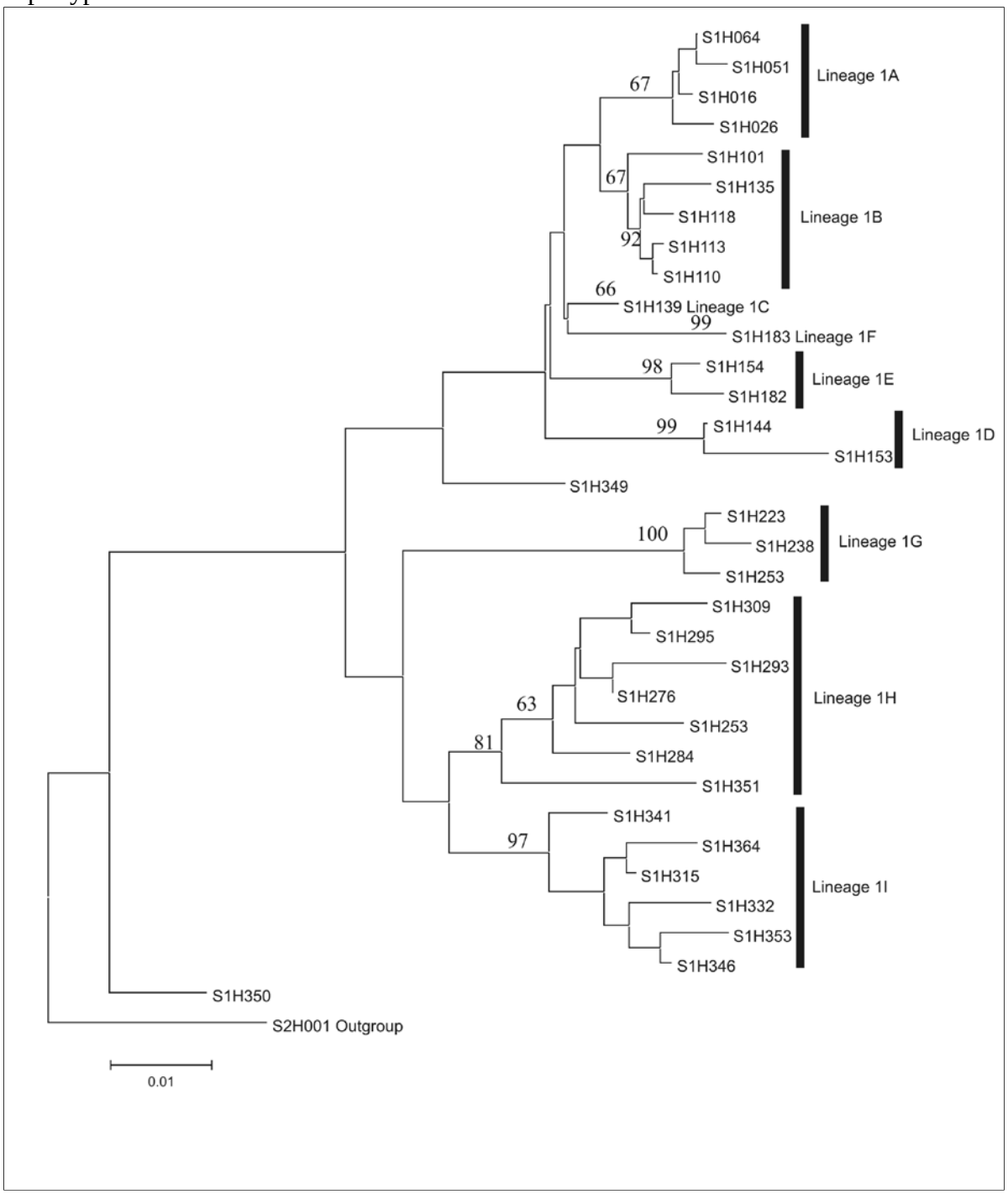
The major lineages within species 1 were delineated on the basis of monophyletic clades with high bootstrap support in a neighbor-joining tree or by the parsimony criterion employed by TCS 1.21. In these networks, the default 95% setting was used, resulting in the separation of any groups that were connected by ten or more inferred haplotypes. These groups were then checked for congruence against a neighbor-joining tree constructed with a Tamura-Nei model of

evolution (see chapter 2 for justification). Nine major lineages within species 1 were identified (A to I) as well as three distinct ungrouped haplotypes: S1H139, S1H349 and S1H350 (figure 3.3). The placement of haplotype S1H101 was clearly problematic for the neighbor-joining analysis, and eroded some of the bootstrap support defining lineages A and B. This haplotype was within six inferred mutational steps of lineage B haplotypes. Inclusion of haplotype S1H101 in lineage A would force a network beyond the limits of the parsimony criterion. Similar conflicts in the tree (separation of species 1 lineages C from F and the exclusion of haplotype S1H351 from lineage H) were also resolved on the basis of their respective haplotype networks. Lineage H possessed the highest nucleotide diversity (π=0.01458), whereas the remaining groups had within-lineage genetic distances of less than 1% (table 3.1).

Table 3.1: Population, sample and genetic statistics for each large-bodied species (lineage) of *Hyalella 'azteca'*

Species (Lineage)	Number of populations	Number of sequnces	Number of haplotypes	Nucleotide diversity	Raggedness	Tau	Tajima's D (D_T)	R_2	Fu's F
1 (A)	60	231	98	0.00711	0.0214 (*p=0.154*)	4.154	-2.25 (*p=0.011*)	0.0235 (*p=0.004*)	-127.68 *(p<0.001)*
1 (B)	25	64	37	0.00120	0.0038 (*p=0.001*)	5.039	-0.99 (*p=0.165*)	0.0711 (*p=0.172*)	-18.34 *(p<0.001)*
1 (C)	12	23	5	0.00093	0.1420 (*p=0.186*)	0.593	-1.65 (*p=0.018*)	0.0751 (*p<0.001*)	-3.75 *(p<0.001)*
1 (D)	8	21	12	0.00673	0.0990 (*p=0.116*)	1.535	-0.86 (*p=0.227*)	0.0988 (*p=0175*)	-4.24 (*p=0.027*)
1 (E)	25	66	28	0.00669	0.0180 (*p=0.060*)	3.601	-1.89 (*p=0.008*)	0.0421 (*p=0.014*)	-15.58 *(p<0.001)*
1 (F)	18	55	23	0.00468	0.0241 (*p=0.055*)	2.98	-1.84 (*p=0.011*)	0.0457 (*p=0.014*)	-15.38 *(p<0.001)*
1 (G)	28	109	48	0.00574	0.0334 (*p=0.219*)	3.657	-2.10 (*p=0.002*)	0.0314 (*p=0.007*)	-41.95 *(p<0.001)*
1 (H)	25	103	41	0.01458	0.0061 (*p=0.008*)	6.562	-1.29 (*p=0.073*)	0.0554 (*p=0.074*)	-36.67 (*p<0.001*)
1 (I)	27	129	59	0.00924	0.0064 (*p=0.001*)	4.477	-2.02 (*p=0.003*)	0.0335 (*p=0.006*)	-46.50 (*p<0.001*)
1 (ALL)	184	827	354	0.04590	0.0013 (*p=0.050*)	15.476	-0.855 (*p>0.10*)	0.0628 (*p=0.477*)	-351.98 *(p<0.001)*
2	123	514	211	0.01660	0.0017 (*p<0.001*)	5.445	-1.79 (*p=0.005*)	0.0352 (*p=0.018*)	-262.58 *(p<0.001)*
7	69	255	100	0.01090	0.0058 (*p=0.002*)	2.234	-1.79 (*p=0.010*)	0.0343 (*p=0.019*)	-63.07 (*p* <0.001)
8	27	102	81	0.01651	0.0028 (*p=0.062*)	6.752	-2.14 (*p=0.001*)	0.0347 (*p<0.001*)	-83.98 (*p<0.001)*

Figure 3.3: COI Neighbor-Joining tree of representative species 1 lineages.
The bootstrap support values were calculated on a separate neighbor-joining tree containing all species 1 haplotypes and rooted with a species 2 outgroup (S2H001). A haplotype was haphazardly selected from each of the 3-step clades defined in their respective lineage's COI haplotype networks.



3.3.2 Phylogeographic analysis of species 1 lineage A

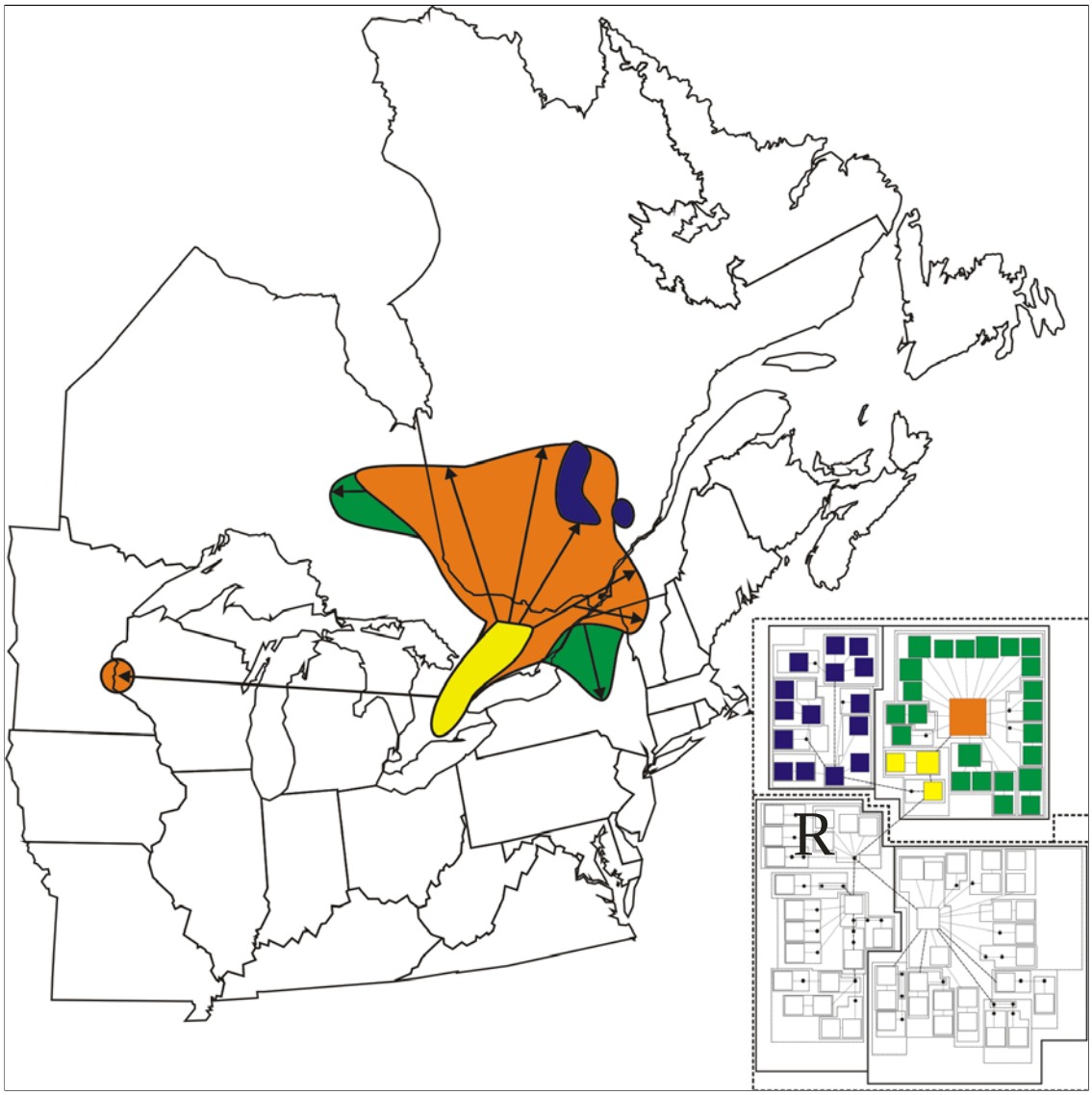
Lineage A was predominantly collected from habitats in the Great Lakes region, along the Saint Lawrence River as well and the Atlantic drainages of eastern Canada. It had the largest sample size of the lineages within species 1, with 231 individuals being sequenced (98 haplotypes) from 60 freshwater habitats and a total nucleotide diversity of 0.0071 (table 3.1).

The COI haplotype network (appendix C figure C.1) was constructed from ninety-eight haplotypes (S1H001-S1H098) that were nested into fifty-three 1-step clades (1-1 to 1-53), fourteen 2-step clades (2-1 to 2-14), four 3-step clades (3-1 to 3-4), two 4-step clades and the total cladogram was a 5-step clade. The members of clade 4-1 (figure 3.5) were predominately sampled from habitats in eastern Ontario and southern Quebec with additional sites in Connecticut, New York and Wisconsin. Clade 4-1 consists of two 3-step clades: a tip clade (3-1) and an interior clade (3-2).

Clade 3-1 haplotypes (figure 3.4) were distributed across nine habitats (QC001, QC002, QC003, QC006, QC007, QC008, QC012, QC013, QC021) in northern Quebec (purple region on the map in figure 3.5) and two habitats south of Lac St. Jean (QC023, QC038). It was the smallest of the 3-step clades, containing sixteen haplotypes. Only haplotypes S1H004 and S1H009 were collected from multiple habitats and both were interior nodes within clades 1-2 and 1-3, respectively. Central haplotypes in a network tend to be older and more widely distributed.

Figure 3.4: Map showing the geographical distribution of species 1 lineage A clades 4-1 and 2-6

The haplotypes of clades 4-1 and 2-6 have geographic distributions represented on the map (left) and their corresponding network positions are indicated on the insert (right). Clade 4-1 is composed of the tip clade 3-1 (purple) and interior clade 3-2 (yellow + orange + green). The root is denoted by (R). The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades.

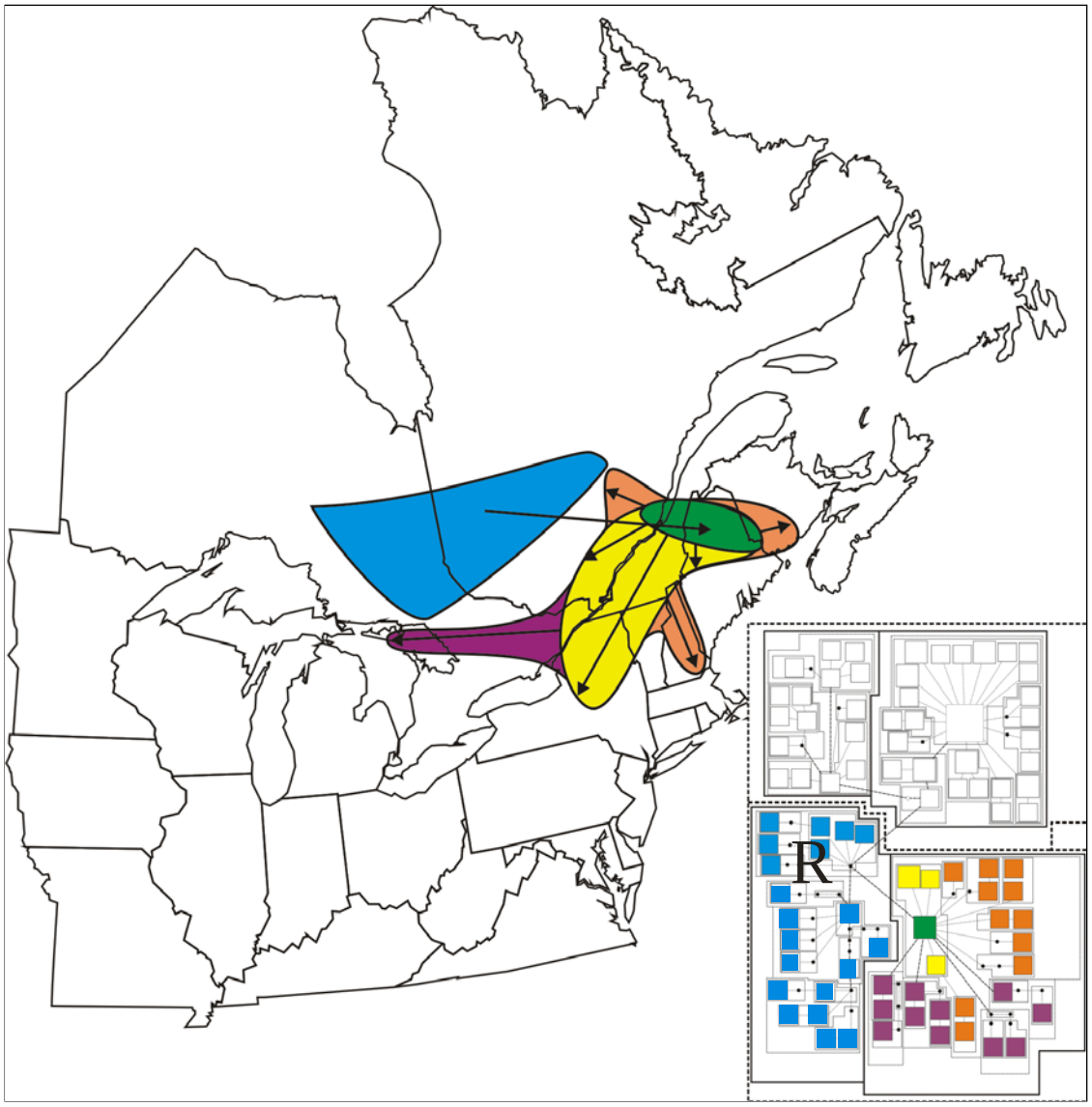


Clade 3-2 (appendix C, figure C.1) was a broadly distributed group (yellow, orange and green colored distributions on the map in figure 3.5) and occurred in habitats that were located in the Great Lakes region up to the western bank of the St. Lawrence River, with a few amphipods collected from sites across its eastern and southern drainages. The clade was comprised of twenty-seven unique haplotypes (S1H017-S1H043) nested into ten 1-step clades (1-8 to 1-17) and two 2-step clades (2-4, 2-5). The interior haplotype (S1H026) was collected from 15 localities and formed a star phylogeny with 16 edges in the network. Furthermore, the distribution of this central haplotype encompassed the collective ranges of the other haplotypes in clades 1-8 and 2-4, which is a pattern consistent with isolation by distance.

Clade 4-2 possessed haplotypes derived from individuals collected from the St. Lawrence River drainage, the Finger Lakes in New York and a single site north of Lake Huron (figure 3.5). It consisted of two 3-step clades: the interior clade 3-3 and tip clade 3-4. Clade 3-3 contained the symmetrical clade 2-6, which could have grouped with clade 3-2, 3-3 or 3-4, but was sorted according to sample size (figure 3.5). According to the outgroup analysis, the root of the network was located within clade 3-3, which in turn makes clade 4-2 interior to clade 4-1. The populations possessing clade 3-3 individuals were detected in localities in northern Ontario and Quebec (the blue distribution in fig. 3.4 and fig. 3.5). Members of clade 3-4 were broadly distributed in habitats along the St. Lawrence River (fig. 3.5). The center of the clade was composed of a star phylogeny around haplotype S1H054 with 22 edges, but was only collected from two locations, one in south-eastern Quebec and the other in southern New Brunswick (green in figure 3.6). The geographical distributions of the nested clades (1-33, 2-10 and 3-4) have progressively larger ranges (yellow, orange and purple colored regions in figure 3.5, respectively). This pattern was consistent with a contiguous range expansion and quite different

from the pattern possessed by clade 2-4. The interior haplotypes of clade 3-4 have a limited distribution, whereas within clade 2-4 there was a clear indication of allele surfing by the interior haplotype (S1H026). As clade 3-4 expanded, there was sufficient time between colonization events to accrue *de novo* mutations in the haplotypes of the dispersing individuals.

Figure 3.5: Map showing the geographical distribution of species 1 lineage A clade 4-2
The haplotypes of clades 4-2 have geographic distributions represented on the map (left) and their corresponding network positions are indicated on the insert (right). Clade 4-1 is composed of tip clade 3-4 (green + yellow + orange + purple) and interior clade 3-3 (blue). The root (R) is marked on the insert within clade 3-3. The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades.



3.3.3 Outgroup analysis of species 1 lineage A

The root of the haplotype network was determined by comparing the roots of phylogenetic trees derived from the neighbor-joining and parsimony analyses using haplotype S1H209 as the outgroup. This sequence was chosen on the basis of the phylogeny. It was a member of lineage 1G, which forms a distinct monophyletic group from the other lineages with high bootstrap support (fig. 3.3). The model of molecular evolution provided by jmodeltest was the TIM1 model (Posada, 2008) with a proportion of invariant sites (I = 0.622). The parsimony heuristic searches were run with nearest neighbor interchange, but as a consequence of the number of taxa it ran out of memory before reaching completion. As such, the 427 000 stored trees were condensed into a 50% majority rule tree (not shown). The root of both the neighbor-joining and parsimony trees was unambiguously located within clade 1-21, between haplotype S1H050 (sampled from Fukishima Lake, Ontario) and clade 1-18 (appendix C, figure C.1).

3.3.4 Nested clade analysis of species 1 lineage A

The nested clade analysis indicated 11 significant geographical associations (table 3.2). Application of the inference key returned 9 cases that indicate restricted gene flow with isolation by distance (either IBD or LDD in table 3.2), one contiguous range expansion and one range expansion with long distance colonization (table 3.2).

Table 3.2: Nested Clade Analysis and Inferences of species 1 lineage A
Only the clades with significant geographical associations are included, refer to appendix C figure C.1 for cladogram. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance, CRE: Contiguous Range Expansion, RE(C): Range Expansion with Long Distance Colonization, LDC: Long Distance Colonization and LDD: Isolation by Distance with Long Distance Dispersal. See appendix B for full NCA table of clade (Dc) and nested clade (Dn) distances, as well as p-values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key	Conclusion
5-1	1-2-3-4-NO:	IBD
4-1	1-2-3-5-6-7-YES:	LDD
3-1	1-2-3-5-6-(TooFew)-7-YES:	IBD or CRE or LDC
3-2	1-11-12-NO:	CRE
2-4	1-2-3-4-NO:	IBD
1-8	1-2-5-6-15-NO:	LDD
4-2	1-2-3-4-NO:	IBD
3-3	No Inference	No Inference
2-7	1-19-20-21-YES:	RE(C)
3-4	1-2-3-4-NO:	IBD
2-10	1-2-3-4-NO:	IBD
1-33	1-2-3-4-NO:	IBD

The nested clade analysis of the total cladogram (clade 5-1) denoted isolation by distance as a result of a significantly small clade and nested clade distances (Dc=311.07, $p<0.001$; Dn=328.63, $p<0.001$ respectively) in the tip clade 4-1.

The NCA of clade 4-1 indicated restricted gene flow in conjunction with potential long distance dispersal. This was a direct result of the tip clade 3-1 having a significantly small clade distance and a significantly large nested clade distance (Dc=80.06, $p<0.001$; Dn=389.72, $p=0.0174$ respectively). When the significance of Dc and Dn values are reversed like this, the inference key indicates that the clade was locally restricted but disjunct from the geographic center of clade 4-1.

The NCA of clade 3-1 returned significant geographical associations among the tip clades 2-1 and 2-2, which possess haplotypes collected from locations surrounding Lac Chibougamau (QC013) and Lac St. Jean (QC021), respectively (figure 3.5). The inference key could not distinguish between isolation by distance, contiguous range expansion and long distance colonization (table 3.2). Furthermore, as a consequence of the overlapping distributions of the clades, it could not discern short distance from long distance dispersal. Tip clade 2-1 has significantly small clade and nested clade distances (*Dc*=45.31, *p=0.008; Dn*=174.85, *p=0.008* respectively*)* consistent with isolation by distance. Instead, the tip clade 2-2 has significantly small clade and significantly large nested clade distances (*Dc*=9.25, *p<0.001; Dn*=174.85, *p<0.001* respectively*)*, which the inference key interprets as long distance colonization.

The NCA of clade 3-2 indicated a significantly large nested clade distance (*Dn*=302.61, *p=0.0091*) for clade 2-4 and a significantly small clade distance in the interior clade 2-5 (*Dc*= 155.12, *p=0.0190*). This means that the average distance of the haplotypes within tip clade 2-4 were significantly distant from the average geographical center of all haplotypes within clade 3-2 (which pools together both clade 2-5 and 2-4 haplotypes). This fact, in conjunction with the spatially restricted interior clade 2-5 (as dictated by the small clade distance) was a good indicator of a contiguous range expansion. This inference was supported by mapping the geographical range of the clades (fig. 3.5). The interior clade 2-5 possesses haplotypes collected from habitats in southwestern Ontario (the yellow region on the map in figure 3.5), whereas the haplotypes of the tip clades occupy a much broader region throughout northern Ontario and Quebec, New York and a single isolated population south of Lake Superior (the orange and green regions on the map in fig. 3.5). Nested within clade 3-2 were clades 1-8 and 2-4, which the inference key indicated had genetic patterns consistent with by isolation by distance. In clade 1-

8, the tip haplotype S1H020 had a significantly small clade distance (Dc=4.28, *p<0.001*). Similarly, within clade 2-4, the tip clades 1-13, 1-14 and 1-15 had significantly small clade distances (Dc= 23.01, *p<0.001*, Dc= 5.81, *p=0.002* and Dc= 79.12, *p=0.001*, respectively), as well as a significantly large difference in clade distance ($I\text{-}T_{(Dc)}$=258.86, *p=<0.001*).

The inference key indicated a pattern of isolation by distance with respect to clade 4-2 because the tip clade 3-4 had both significantly small clade and nested clade distances (Dc=239.70, *p=0.000;* Dn=345.60, *p=0.006* respectively*)* and the difference in geographic distance between the internal and tip clades were significantly large (I-T Dc=-99.72, *p=0.001*, I-T Dn=-138.63, *p=0.000*).

Within clade 3-3, only clade 2-7 had a significant geographical association, because its tip clade (1-26) had a sufficient sample size in one habitat to get a significantly small clade distance (Dc=0, *p=0.002*). Any haplotype detected at a single location with sufficient sample size will likely give a significantly small clade distance. In the absence of other significant geographical associations within that clade, the inference key will always return an answer of restricted gene flow with isolation by distance.

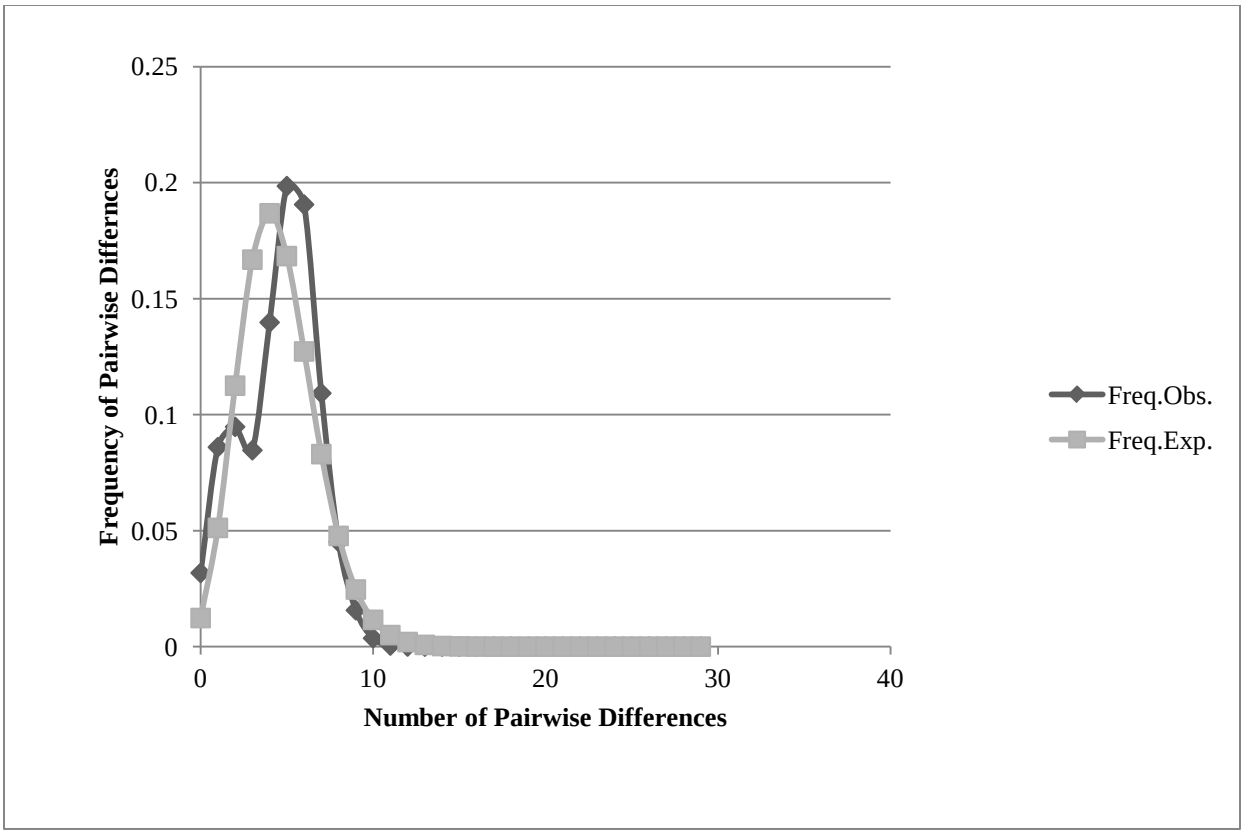
All of the clades with significant geographical associations within clade 3-4 (1-33 and 2-10) indicated restricted gene flow with isolation by distance as a consequence of tip clades with significantly small clade distances ($Dc_{(C1H066)}$=0, *p=0.001;* $Dc_{(1\text{-}45)}$=0, *p<0.001*, respectively).

3.3.5 Mismatch distribution and population genetics of species 1 lineage A

The mismatch distribution had bimodal peaks (fig. 3.6) and was plotted against a simulated population undergoing constant expansion based on tau=4.154. The raggedness statistic (table 3.1) was not significant (r=0.0215, $p = 0.141$) and could not reject the null hypothesis of a demographic expansion.

All of the tests for neutrality indicated significant deviations when compared against 1000 coalescent simulations. The excess of rare alleles resulted in a significantly negative Tajima's D (-2.25, $p=0.011$) and a significantly negative Fu's F (F = -127.68, $p<0.001$). Ramos-Onsins and Rozas R_2 statistic (R_2=0.0235, $p=0.004$) was significant, indicating an excess of singletons per sequence (U) as compared to the average nucleotide difference (k) divided by two. The rationale was that in a population undergoing a sudden increase in size, the value of U should equal $k/2$. In this scenario, R_2=0. This is consistent with a high haplotype diversity mixed with low nucleotide divergences, which coincides with a demographic expansion.

Figure 3.6: Mismatch distribution of species 1 lineage A
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with tau = 4.154.



3.3.6 Analysis of molecular variance of COI haplotypes within species 1 lineage A

The AMOVA partitioned on the basis of faunal provinces had four groups of populations (Central Appalachian: 1 population, Northern Appalachian: 7 populations, Great Lakes: 33 populations and Hudson Bay: 19 populations) and the between-group variance component was 12.59% (table 3.3). The second AMOVA subdivided the populations with respect to glaciated and unglaciated regions, but had only a single population located in ice-free regions. This resulted in a negative variance component, which is equivalent to 0.

The SAMOVA was conducted using two to six groups of populations (table 3.3). The 5-group analysis (V(BG) = 45.44) was deemed optimal because it had more than a 10 point increase in the between-group variance component over analyses using fewer groups (table 3.3) and the 6-group partition only placed a single population into the additional group. The *Fst* was determined using this partition (*Fst* = 0.639, $p<0.001$). The 5-group SAMOVA consisted of three major assemblages with two groups containing a single population. The first group had 26 populations and occupied an area north of the Great Lakes (green in fig. 3.7). The second group was an assemblage of 15 populations along the St. Lawrence River and east into Maine and New Brunswick (yellow in fig. 3.7). The distribution of populations in groups 1 and 2 coincides with clade 3-2 (fig. 3.4) and clade 3-4 (fig. 3.5), respectively. As described in the NCA, these regions represent species 1 lineage A spatial structure formed from the Post-Pleistocene dispersal of *Hyalella* into the Great Lakes region. The third major group consists of 17 populations and was split into two zones in northern Quebec and northern Ontario (blue in fig. 3.7). Given the degree of similarity between these two regions, it may be an indication of gene flow. Since the mode of long distance dispersal among *Hyalella* is via avian vectors, it would be expected that long

distance dispersal would occur along a North-South corridor, as opposed to East-West. However, the NCA traces back an ancestral root of lineage A to populations west of the Great Lakes.

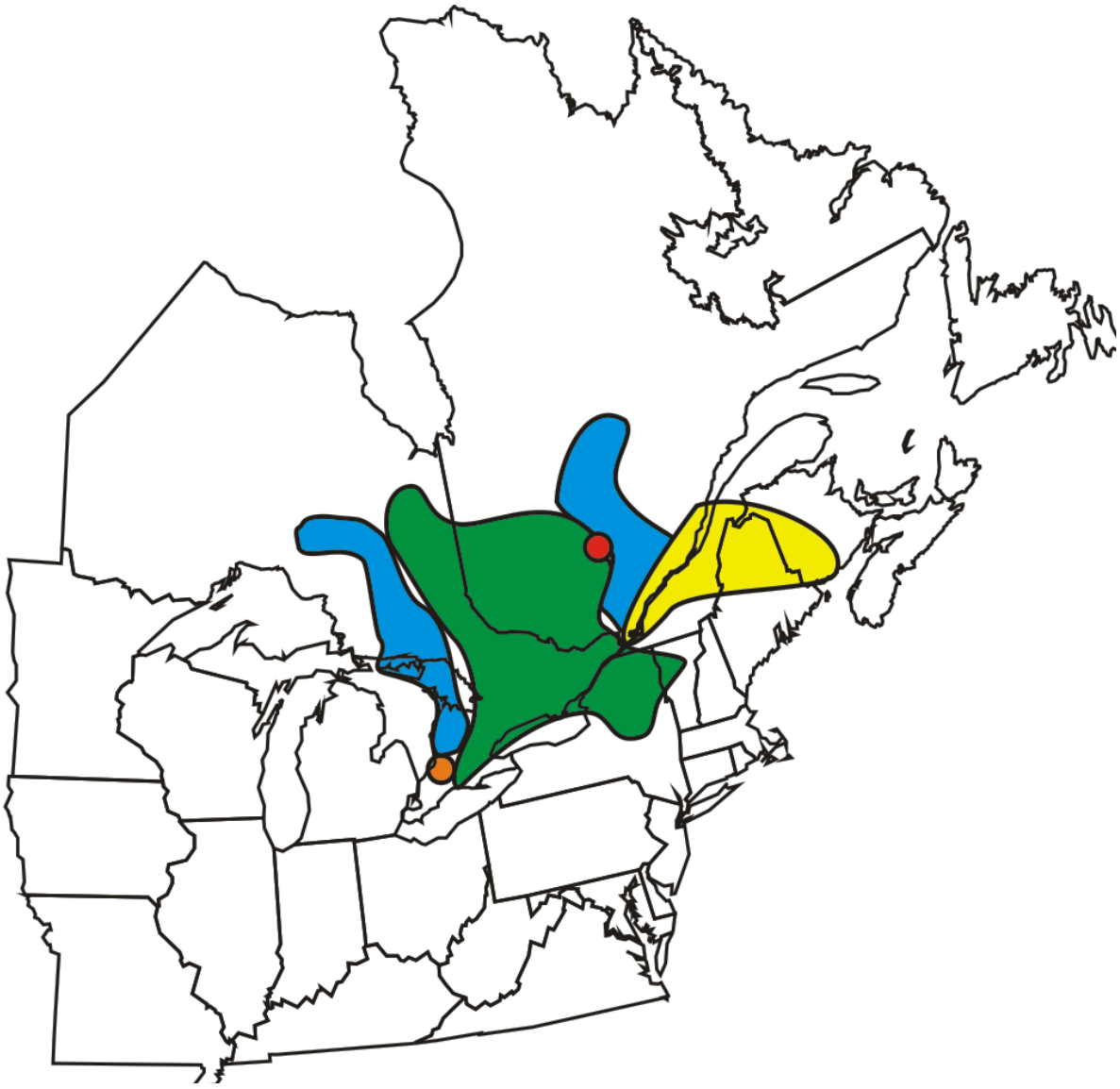
Table 3.3: AMOVA and SAMOVA of COI haplotypes for species 1 lineage A
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. The 5-group partition in the SAMOVA (bolded) was the best supported (see text). * $p<0.01$ ** $p<0.001$ NS = Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	12.59**	47.23**	40.18**
Unglaciated vs. Glaciated	-13.24^{NS}	65.26**	47.98**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	28.58**	36.94**	34.48**
3	29.00**	36.62**	34.38**
4	30.21**	34.97**	34.82**
5	**45.44****	**18.43****	**36.13****
6	47.00**	16.94**	36.07**

Figure 3.7: Map showing the geographical distributions of populations within the 5-group SAMOVA

The groups are based on the 5-group SAMOVA with three major population assemblages (blue, green and yellow) and two groups containing single populations (orange and red).



3.3.7 Phylogeographic analysis of species 1 lineage B

Sixty-four species 1 lineage B amphipods (table 3.1) were collected from 25 habitats in the Great Lakes region, northern Quebec and the Fairford River in Manitoba. Ten of these locations (especially around the Great Lakes) also contained other species 1 lineages. In the neighbor-joining analysis, the lineage B COI sequences (S1H101 to S1H135) formed a monophyletic cluster (figure 3.3) with tepid bootstrap support (67).

The statistical parsimony network (appendix C, fig. C.2) of COI haplotypes formed a 5-step clade (5-2) constructed from 37 unique haplotypes (S1H099 to S1H135): seventeen 1-step clades (1-54 to 1-70), seven 2-step clades (2-16 to 2-22), five 3-step clades (3-5 to 3-9) and two 4-step clades (4-3 and 4-4). Clade 3-5 (haplotypes S1H099 to S1H110) was represented by the haplotype S1H101 in the neighbor-joining phylogeny (figure 3.3) and was separated from the rest of the network by 6 inferred mutational steps (appendix C, fig. C.2). While inclusion of this haplotype weakened the bootstrap support for lineage B in the neighbor-joining phylogeny (92 to 67), it did not break the limit set by the parsimony criterion.

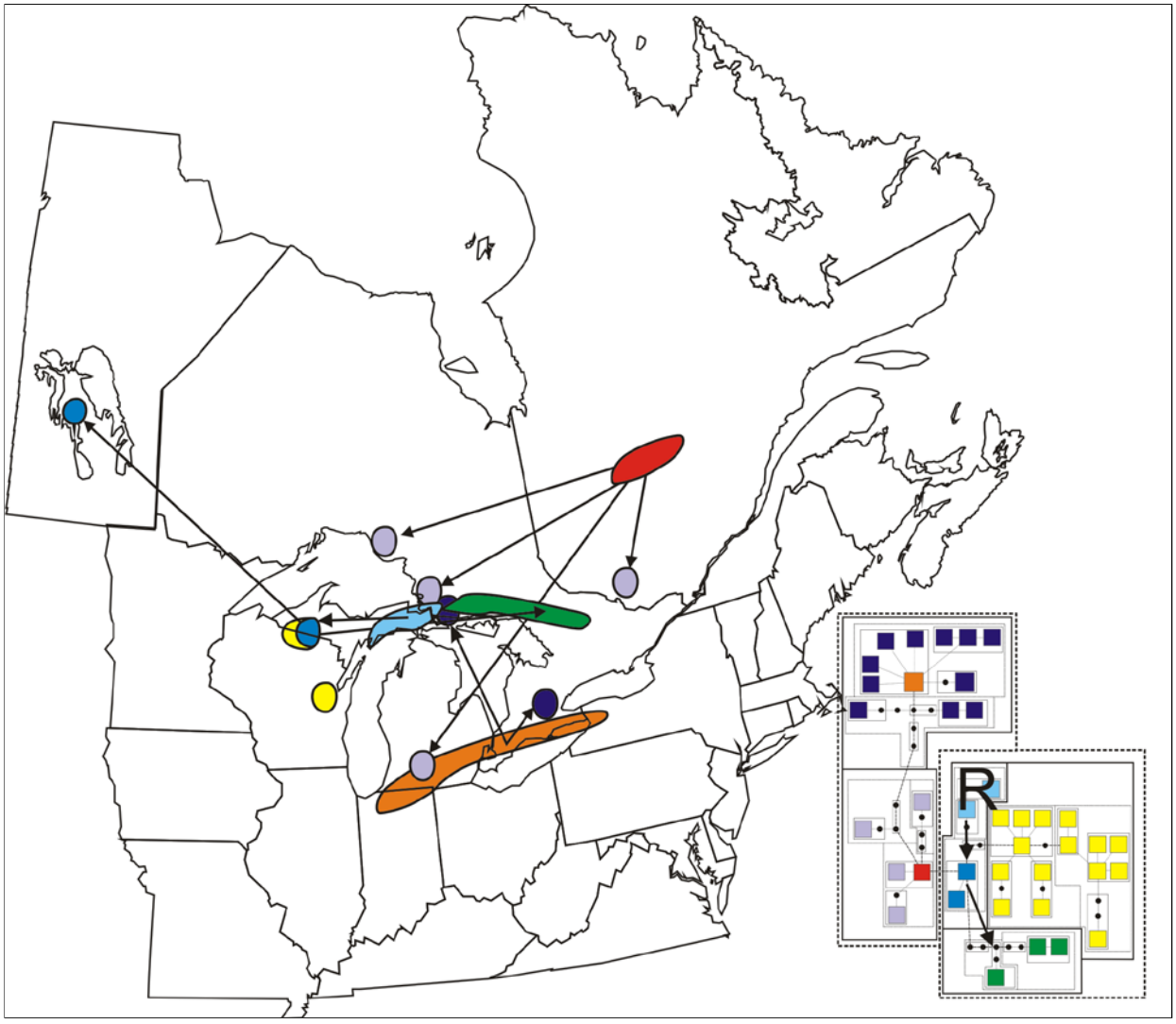
3.3.8 Outgroup analysis of species 1 lineage B

The HKY model of molecular evolution (Hasegawa *et al.,* 1985) with invariant sites (I=0.653) and gamma distributed rate variation (α=0.858) was selected as the optimal model on the basis of the Akaike information criterion. The resulting neighbor-joining tree placed the root between haplotypes S1H118 and S1H119 (appendix C, fig. C.2). The parsimony tree was derived from a heuristic search using tree bisection reconnection branch swapping. Although the search

replicates were started from both a neighbor-joining tree and 20 random trees, the haplotypes of clades 3-6 and 3-7 formed a polytomy with the root at the base of the 50%-majority consensus tree. Both the NJ and MP trees were estimated on the basis of lineage 1G haplotype (S1H209) as the outgroup.

When lineage B haplotypes were mapped (fig. 3.8), most of the 3-step clades formed disjunct groups of populations. Clade 3-5 has a central haplotype (S1H103) that was broadly distributed in the southern Great Lakes region (dark purple and orange distributions in fig. 3.8) and tip clades that were collected from disparate localities. The only instances where individuals were detected outside of the internal haplotype's (S1H103) distribution were the Eramosa River, ON (ON013), Speed River, ON (ON040) and a pond north of Lake Huron, ON (fig. 3.8). Clade 3-6 had an interior haplotype detected in northern Quebec (fig. 3.8) and derived haplotypes that were detected in populations throughout the Great Lakes region (fig. 3.8). Clade 4-4, which consists of clades 3-7 to 3-9, was largely restricted to Minnesota populations, although some tip clades were located in Manitoba and Ontario (blue, green and yellow distributions in fig. 3.8).

Figure 3.8: Map showing the geographical distribution of species 1 lineage B
The haplotypes of lineage B have geographic distributions represented on the map (left) and their corresponding network positions are indicated on the insert (right). The root of the network (R) connected to an edge within clade 2-20. The distributions of clades 2-19, 2-20 and 3-9 are color coded light blue, blue and green, respectively. Haplotype S1H103 (orange) is nested into clade 3-5 (orange + dark purple). Haplotype S1H12 (red) is nested within clade 3-6 (red + light purple). The distributions of clades 2-22 and 2-23 are color coded yellow. The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades.



3.3.9 Nested Clade Analysis of species 1 lineage B

The nested clade analysis returned 6 significant geographical associations (table 3.4), of which the inference key indicated four patterns of isolation by distance, one contiguous range expansion and one incomplete geographical sampling.

The NCA of the total cladogram (clade 5-2) indicated a contiguous range expansion, because the interior clade 4-4 had a significantly small clade distance (Dc = 307.85, $p<0.001$). When applied to clades 4-3 and 4-4, the inference key returned a pattern of isolation by distance because the tip clades 3-5 and 3-8 had significantly small clades distances respectively ($Dc_{(3-5)}$ = 283.09, $p<0.001$, $Dc_{(3-8)}$ = 126.87, $p<0.001$). Clade 3-8 had a significant geographical association, but there was insufficient geographical sampling to make an inference. The NCA of clade 2-15 indicated isolation by distance because the tip clades 1-53 and 1-55 have significantly small nested clade distances (Dc = 52.74, $p<0.001$ and Dn = 0, $p<0.001$, respectively). Similarly, isolation by distance was also inferred for clade 2-23, but as a result of the interior clade 1-68 having a significantly small nested clade distance (Dn = 10.91. $p=0.008$).

Table 3.4: Nested Clade Analysis of COI haplotypes in species 1 lineage B
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. CRE: Contiguous Range Expansion, IBD: Isolation by Distance and IGS: Inadequate Geographical Sampling. See appendix B for full NCA table with Dn and Dc values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

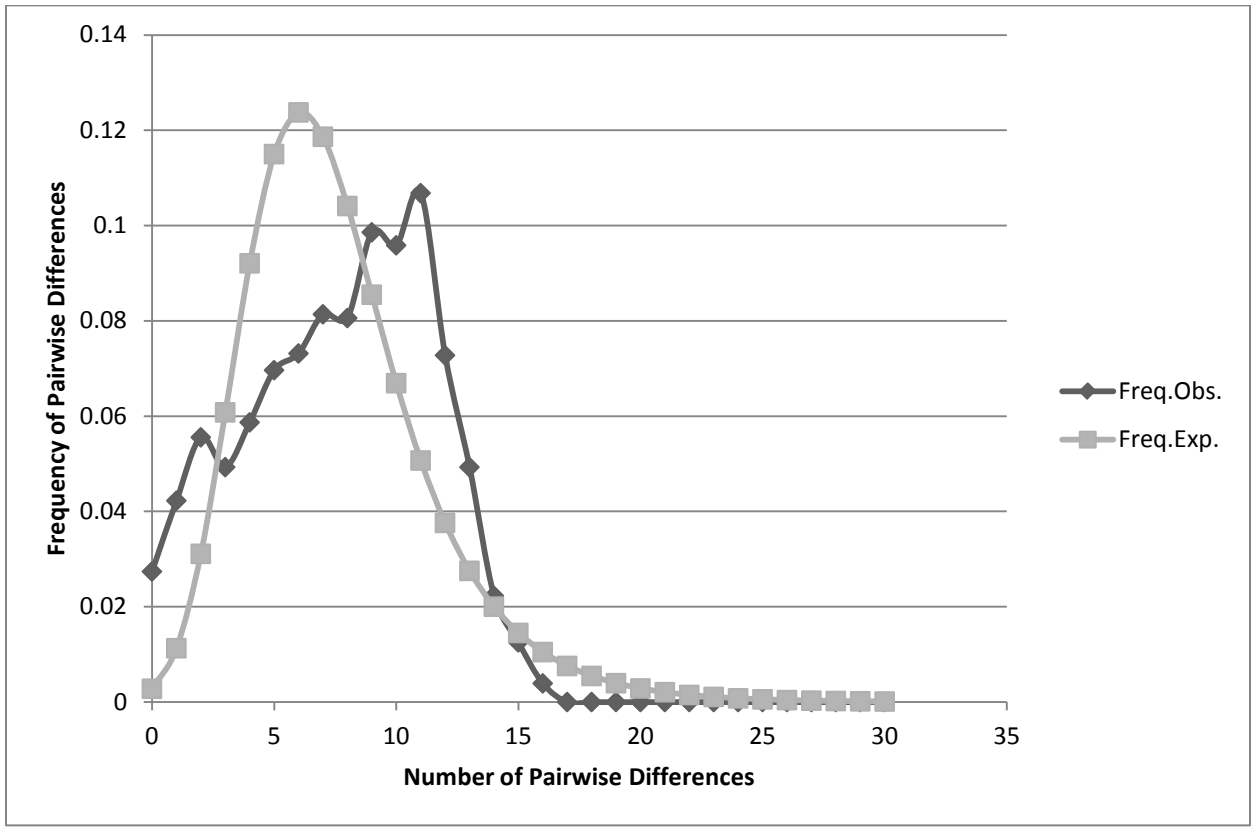
Clade	Inference Key	Conclusion
5-2	1-2-11-12-NO:	CRE
4-3	1-19-20-2-3-4-NO:	IBD
3-5	No Inference	No Inference
2-15	1-2-3-4-NO:	IBD
4-4	1-2-3-4-NO:	IBD
3-8	1-19-20-NO:	IGS
2-23	1-2-11-17-4-NO:	IBD

3.3.10 Mismatch distribution and population genetics of species 1 lineage B

The mismatch distribution was multimodal and lacked a smooth peak (fig. 3.9). The raggedness statistic (table 3.1) rejected the null hypothesis (r =0.004, *p=0.001*) of a demographic expansion. The expected curve of a population undergoing a constant population expansion was estimated setting Tau at 5.039.

Neither Tajima's *D* ($D_{T=}$ -0.994, *p=0.165*) nor R_2 (R_2=0.0711, *p=0.172*) indicted departures from neutrality and only Fu's *F* (F_F=-18.34, *p<0.001*) was significantly negative, which was not consistent with a lineage undergoing demographic expansion.

Figure 3.9: Mismatch Distribution of species 1 lineage B
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with $\tau = 5.039$.



3.3.11 Analysis of molecular variance of species 1 lineage B

The populations were partitioned into groups according to three faunal provinces (Great Lakes: 15 populations; Hudson Bay: 4 populations; Mississippi: 6 populations), which resulted in an AMOVA (table 3.5) with a between-group variance component of 22.66%. The AMOVA partitioned into glaciated and unglaciated regions (unglaciated: 6 populations, glaciated: 19 populations) was not significant (table 3.5) and the between-group variance component was only 19.6%.

The SAMOVA of COI haplotypes was initially conducted using 2-groups and compared to analyses incorporating a greater degree of partitioning. Adding more groups did not result in a major increase of the between-group variance component above the initial 2-group SAMOVA (V(BG)=52.57) and the additional partitions were composed of single populations. The 2-group assemblages of populations were consistent with the clades identified in the haplotype network (figure 3.9), with the first group being composed of 9 populations from clade 3-5 (orange and purple distributions in figure 3.8) and the second group containing 16 populations (red, light purple, blue, green yellow and pink distributions in figure 3.8) belonging to clades 3-6 through 3-9. The between-group variance component of the 2-group SAMOVA was significant ($p<0.001$) and explains over twice the variation of either the faunal provinces (V_{BG}=22.66) or unglaciated and glaciated (V_{BG}=19.6) partitions tested in the AMOVA (table 3.5). The *Fst* value was determined using the 2-group partition (*Fst* = 0.890, $p<0.001$).

Table 3.5: AMOVA and SAMOVA of species 1 lineage B
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. The 2-group partition in the SAMOVA (bolded) was the best supported (see text). * $p<0.01$, ** $p<0.001$, NS= Not Significant.

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	22.66**	63.81**	13.53**
Unglaciated vs. Glaciated	19.6^{NS}	66.99**	13.41**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	**52.67****	**36.31****	**11.02****
3	56.70**	30.98**	12.32**
4	59.96**	27.49**	12.55**
5	61.92**	25.25**	12.83**
6	65.07**	22.32**	12.62**

3.3.12 Phylogeography of species 1 lineage C

Twenty-three species 1 lineage C individuals were collected from 12 populations broadly distributed across North America, including disparate localities in Alaska and South Dakota. The COI nucleotide diversity (table 3.1) was extremely low (π=0.0009) compared to other species 1 lineages.

The COI statistical parsimony network (appendix C, fig. C.3) for species 1 lineage C was entirely contained within a single 1-step clade (1-73). The interior haplotype (S1H139) was collected 16 times from 9 habitats and its geographical distribution encompassed the entire range of the nested clade (appendix C, fig. C.3).

3.3.13 Nested clade analysis of species 1 lineage C

The nested clade analysis conducted on clade 1-73 did not return any significant geographical associations.

3.3.14 Mismatch distribution and population genetics of species 1 lineage C

The mismatch distribution produced a unimodal peak (fig. 3.11), and the raggedness statistic (table 3.1) could not reject the null hypothesis (r=0.1420, $p=0.186$) of a demographic expansion. The expected frequency distribution was simulated from a population undergoing constant population expansion with $\tau = 0.593$.

Tajima's D ($D_=$ -1.65, $p=0.018$), Ramos-Onsins and Rozas R_2 statistic (R_2=0.0711, $p=0.172$) and Fu's F (F=-3.75, $p<0.001$) were all significant and indicated departures from neutrality consistent with a demographic expansion.

3.3.15 Analysis of molecular variance of species 1 lineage C

While species 1 lineage C *Hyalella* were detected in multiple faunal provinces, the AMOVA/SAMOVA was not conducted due to small sample size in the populations. The F_{ST} determined from the AMOVA was not significant (F_{ST} = 0.403, $p=0.173$).

Figure 3.10: Map showing the geographical distribution of species 1 lineage C
The dots represent locations of the populations containing lineage C haplotypes (appendix C fig. C.3).

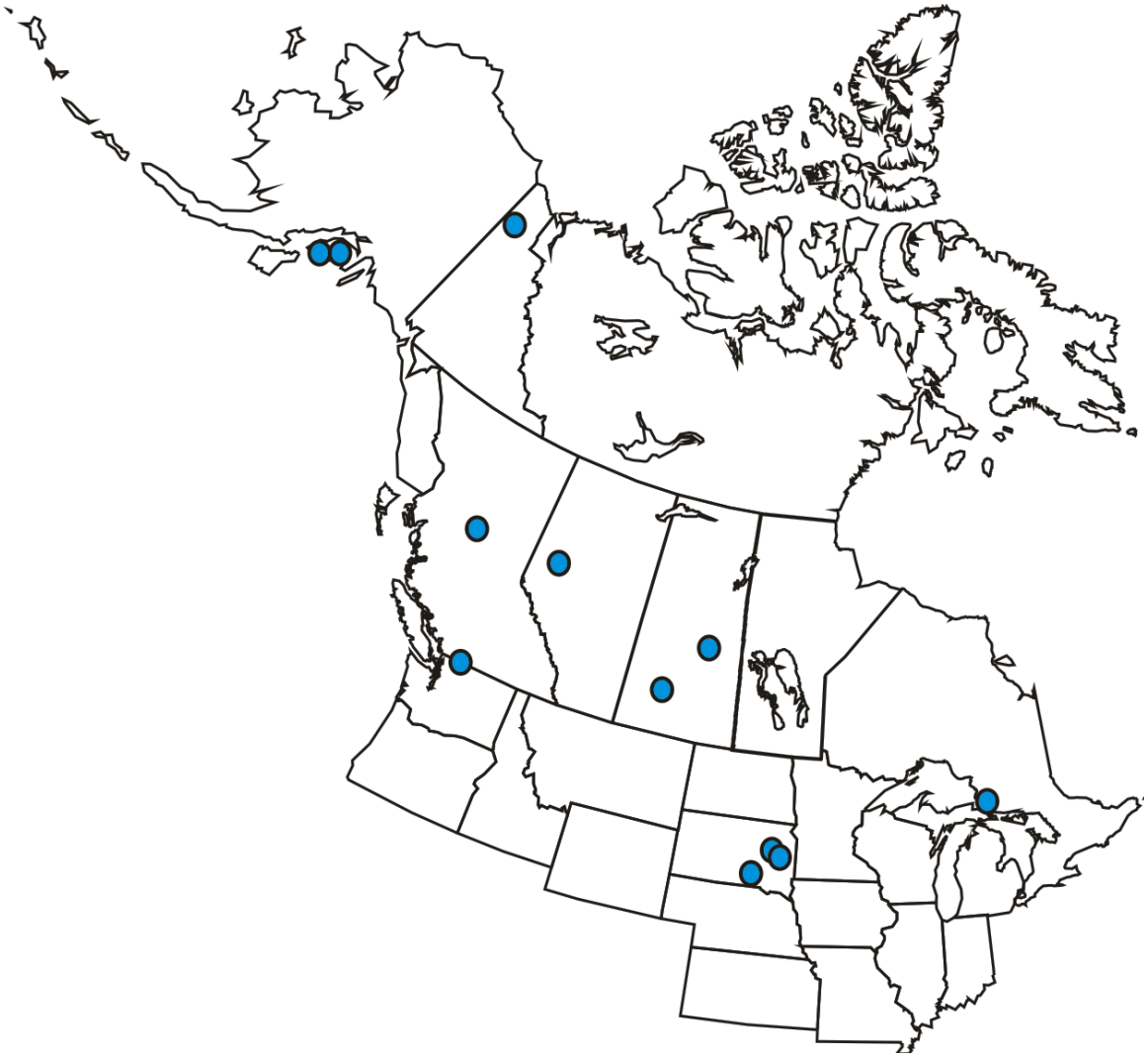
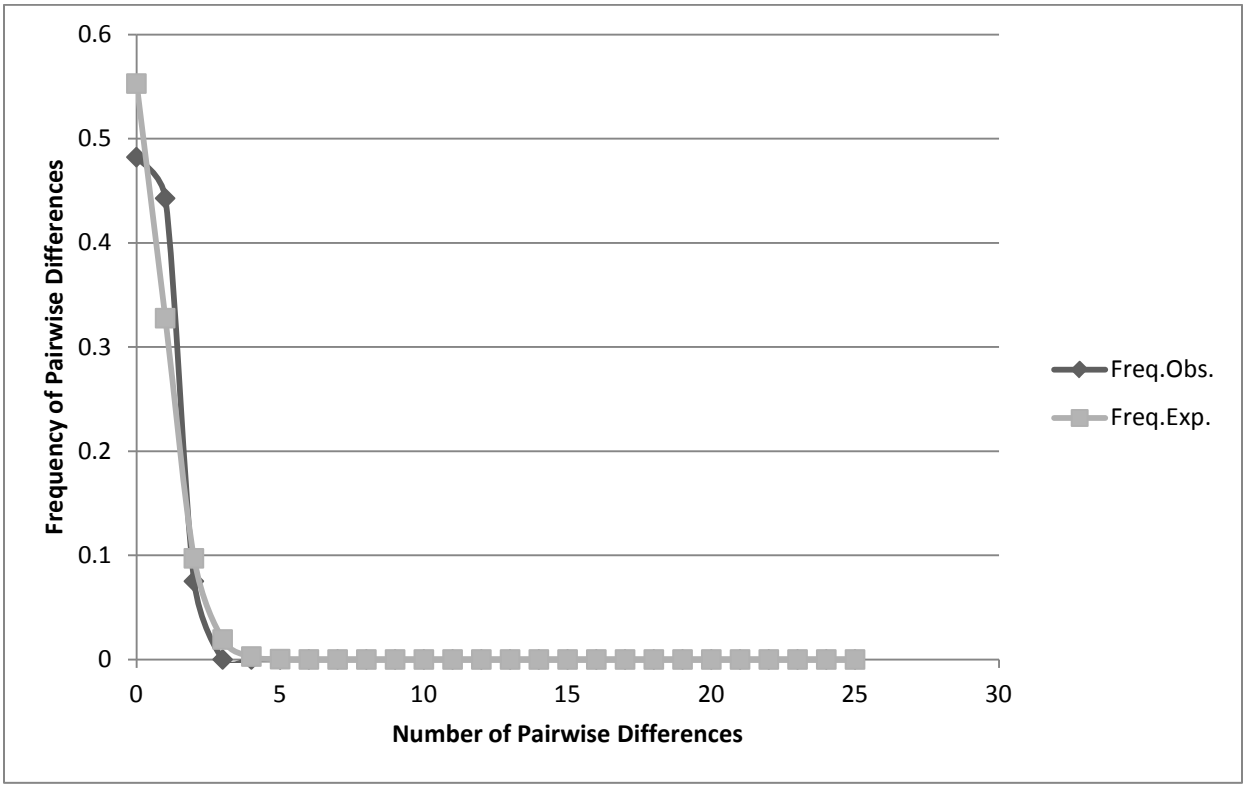


Figure 3.11: Mismatch distribution of species 1 lineage C
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with $\tau = 5.039$.



3.3.16 Phylogeography of species 1 lineage D

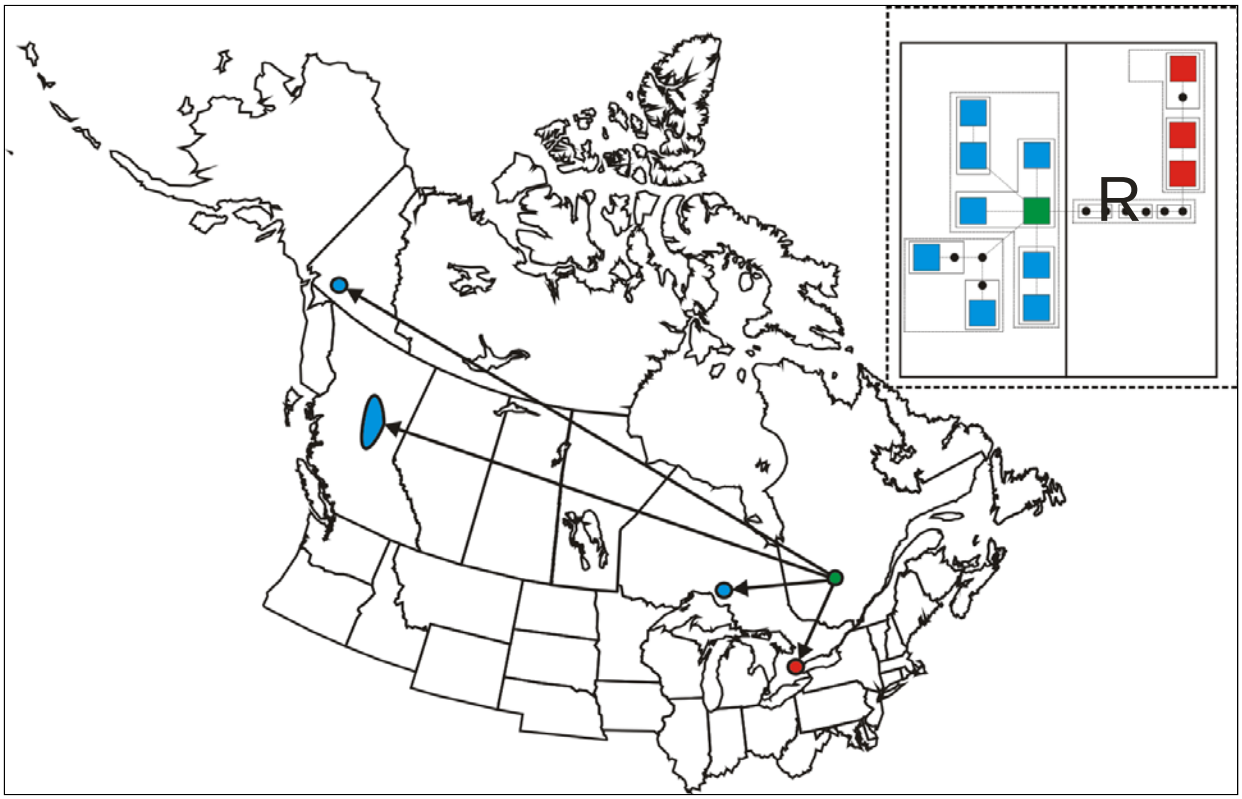
Twenty-one *Hyalella 'azteca'* specimens assigned to lineage D were collected from 8 localities (table 3.1), some of which were separated by extreme distances. Three of the populations were discovered in Ontario (ON003, ON019) and Quebec (QC038), while the remaining sites were situated in British Columbia and the Yukon. The COI sequences formed a monophyletic clade in the neighbor-joining tree (fig. 3.3) with high bootstrap support.

The statistical parsimony network (appendix C fig. C.4) of COI sequences was constructed from twelve haplotypes (S1H142 to S1H153) that nested into six 1-step clades (1-75 to 1-80), three 2-step clades (2-25 to 2-27), two 3-step clades (3-10 and 3-11) and a single 4-step clade (4-5). Clade 3-10 was detected in northern British Columbia, Yukon and Quebec (green and blue in fig. 3.12), whereas clade 3-11 was restricted to a single habitat (red in fig. 3.12) in southern Ontario (ON034).

3.3.17 Outgroup Analysis of species 1 lineage D

The HKY model of molecular evolution with a proportion of invariant sites (I=0.843) was selected on the basis of the Akaike information criterion. The neighbor-joining and maximum parsimony trees were constructed using species 1 haplotype S1H209 as the outgroup (lineage 1G), and unambiguously placed the root of the network along the edge between clades 2-25 and 2-27.

Figure 3.12: Map showing the spatial distribution of species 1 lineage D
Lineage D haplotypes have geographic distributions represented on the map (left) and their corresponding network positions are indicated on the insert (right). Haplotype S1H144 (green) was nested into clade 3-10 (green + blue) and clade 3-11 was coded red. The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades. The root is denoted by (R)



3.3.18 Nested Clade Analysis of species 1 lineage D

The NCA returned two significant geographical associations (table 3.6). The inference key indicated that the distribution of clade 2-25 was shaped by restricted gene flow with some long distance dispersal because the interior clade (1-75) had a significantly large clade distance (Dc=1671.03, $p<0.001$). Clade 1-75 was detected in sites in British Columbia, Ontario and Quebec, whereas the tip clades 1-74 along with 1-76 were only located in British Columbia. The inference key for clade 3-10 also indicated restricted gene flow with isolation by distance because the tip clade 2-26 had a significantly small clade distance (Dc=0, $p<0.001$).

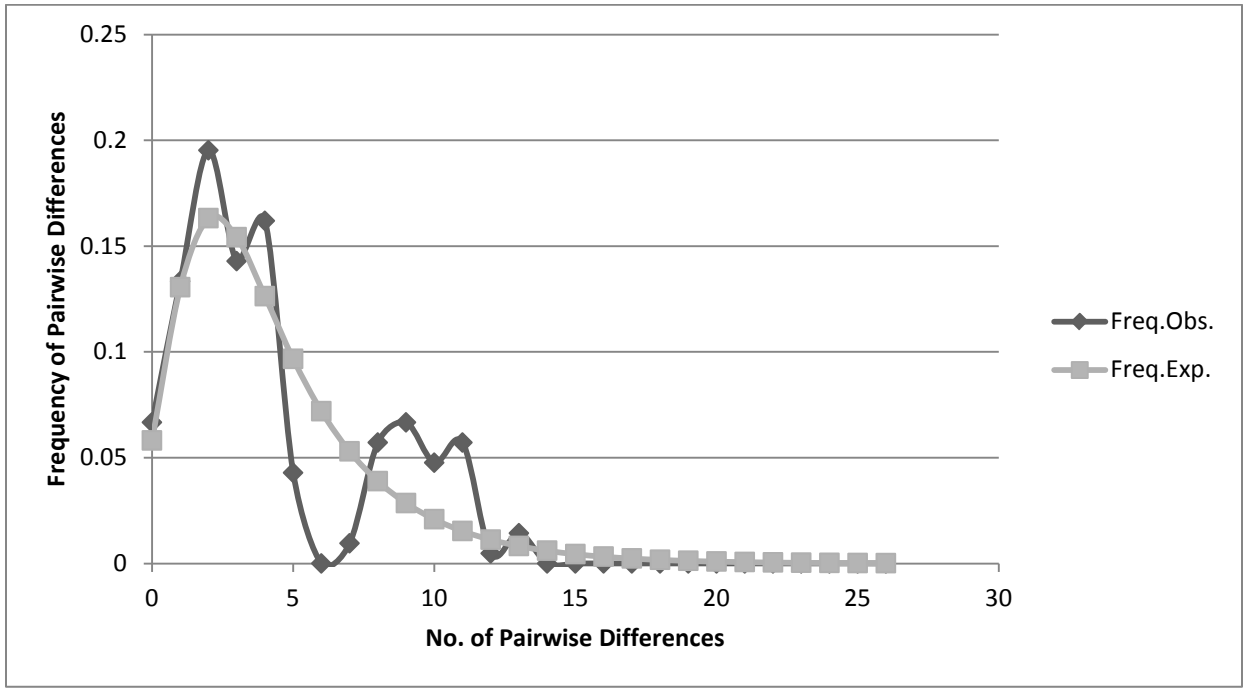
Table 3.6: Nested Clade Analysis of COI haplotypes for species 1 lineage D
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance and LDD: Long Distance Dispersal. See appendix B for full NCA table. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key	Conclusion
4-5	No Inference	No Inference
3-10	1-2-3-5-7-YES:	IBD with LDD
2-25	1-2-3-4-NO:	IBD

3.3.19 Mismatch distribution and population genetics of species 1 lineage E

The mismatch distribution (fig. 3.13) of COI haplotypes consists of two bimodal peaks. The raggedness statistic (table 3.1) failed to reject the null hypothesis (r=0.033; $p=0.143$) of demographic expansion. The expected frequency curve was calculated on the basis of $\tau = 1.535$.

Figure 3.13: Mismatch distribution of species 1 lineage D
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with $\tau = 1.535$.



None of the tests for departures from neutrality were significant; Tajima's *D* (*D*=-0.860; p=0.183), Ramos-Onsins and Rozas R_2 statistic (R_2=0.0988, *p=0.0988*) and Fu's *F* (*F*=-4.24, *p=0.027*). Fu's (1997) analysis of the *F* statistic indicates that a conventional p-value of 0.05 was equivalent to a p-value of 0.02 when using coalescent simulations.

Sixty-six species 1 lineage E individuals representing 28 haplotypes were detected in 25 locations (Appendix A) around the Great Lakes region and northern Quebec. These haplotypes formed a monophyletic group in the neighbor-joining analysis (fig. 3.3) with high bootstrap support and low nucleotide diversity (π=0.0067). Only one of the 25 populations (QC033) did not possess other *Hyalella 'azteca'* species and nineteen of them were co-distributed with lineage A (Appendix A).

3.3.20 Nested COI haplotype network of species 1 lineage E

The haplotype network (appendix C, fig C.5) consisted of 28 unique COI sequences nested into eighteen 1-step clades (1-181 to 1-98), five 2-step clades (2-28 to 2-32), two 3-step clades (3-12 and 3-13), and a single 4-step clade (4-6).

3.3.21 Outgroup analysis of species 1 lineage E

For the neighbor-joining analysis, a Tamura-Nei model of evolution with a proportion of invariant sites (I=0.654) was selected on the basis of the Akaike information criterion. The parsimony analysis was conducted with tree bisection reconnection branch swapping starting from both a neighbor-joining tree (using a Tamura-Nei model) and 20 random trees. Both

analyses placed the root within clade 2-32, along the edge between clades 1-96 and 1-98 (figure 3.18). The species 1 lineage G haplotype (S1H209) was used as the outgroup.

3.3.22 Nested Clade Analysis of species 1 lineage E

The NCA (figure 3.18) returned three significant geographical associations (clades 1-18, 2-29 and 3-12). Clade 1-18 (blue in figure 3.14) had a broadly distributed central haplotype S1H154 (9 habitats) with 16 connections. The rest of the haplotypes within this clade either co-occurred in populations with this haplotype or were detected in individuals within its range. The inference key (table 3.7) indicated a pattern of restricted gene flow with isolation by distance as the tip clades were locally restricted. This was also evident with clade 3-12, except that the isolation by distance was inferred in conjunction with either short or long distance dispersal. This outcome was a consequence of clade 2-30 having a small clade distance coupled with a large nested clade distance (Dc = 61.23, $p<0.001$, Dn = 212.70, $p<0.001$ respectively). Clade 2-29 possessed a disjunct distribution, occurring in northern Quebec and a habitat north of Lake Huron (figure 3.14). The inference key indicated that this clade had a pattern of isolation by distance because the tip clade (1-91) was restricted to three populations in northern Quebec that were located in close proximity.

Figure 3.14: Map showing the geographical distribution of species 1 lineage E
The haplotypes of species 1 lineage E possess geographic distributions represented on the map (right) and their corresponding network positions are indicated on the insert (left). There are two 3-step clades: the tip clade 3-12 (blue) and interior clade 3-13 (yellow). The root is denoted by (R).

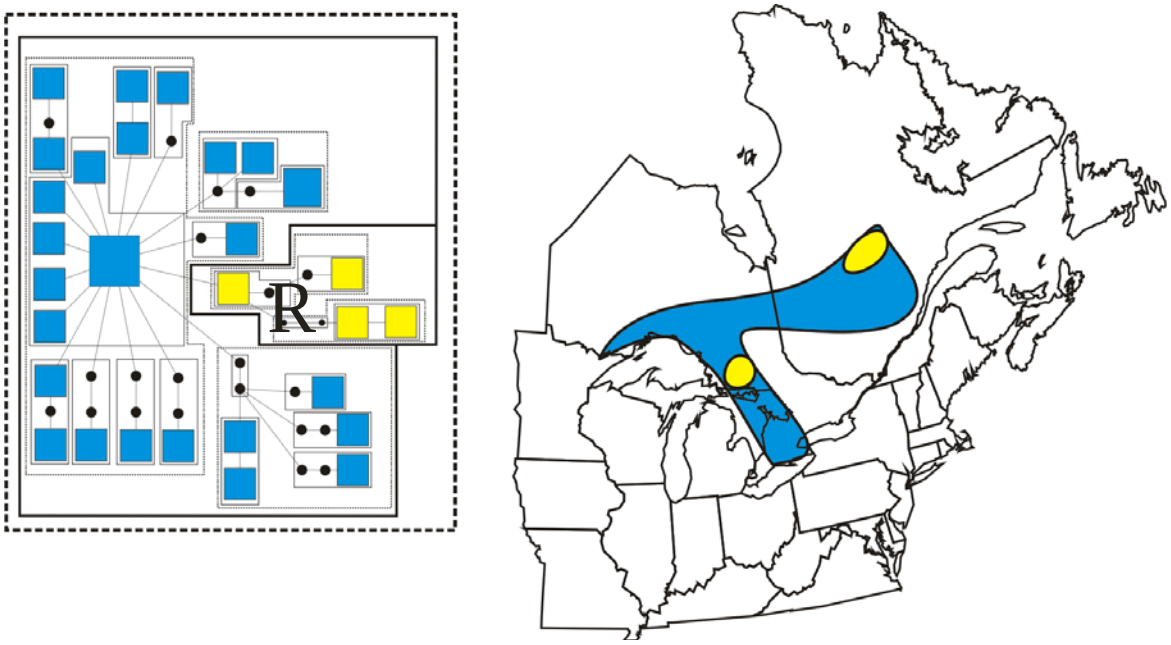


Table 3.7: Nested Clade Analysis of species 1 lineage E
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance, SDD: Short Distance Dispersal and LDD: Long Distance Dispersal. See appendix B for full NCA table. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

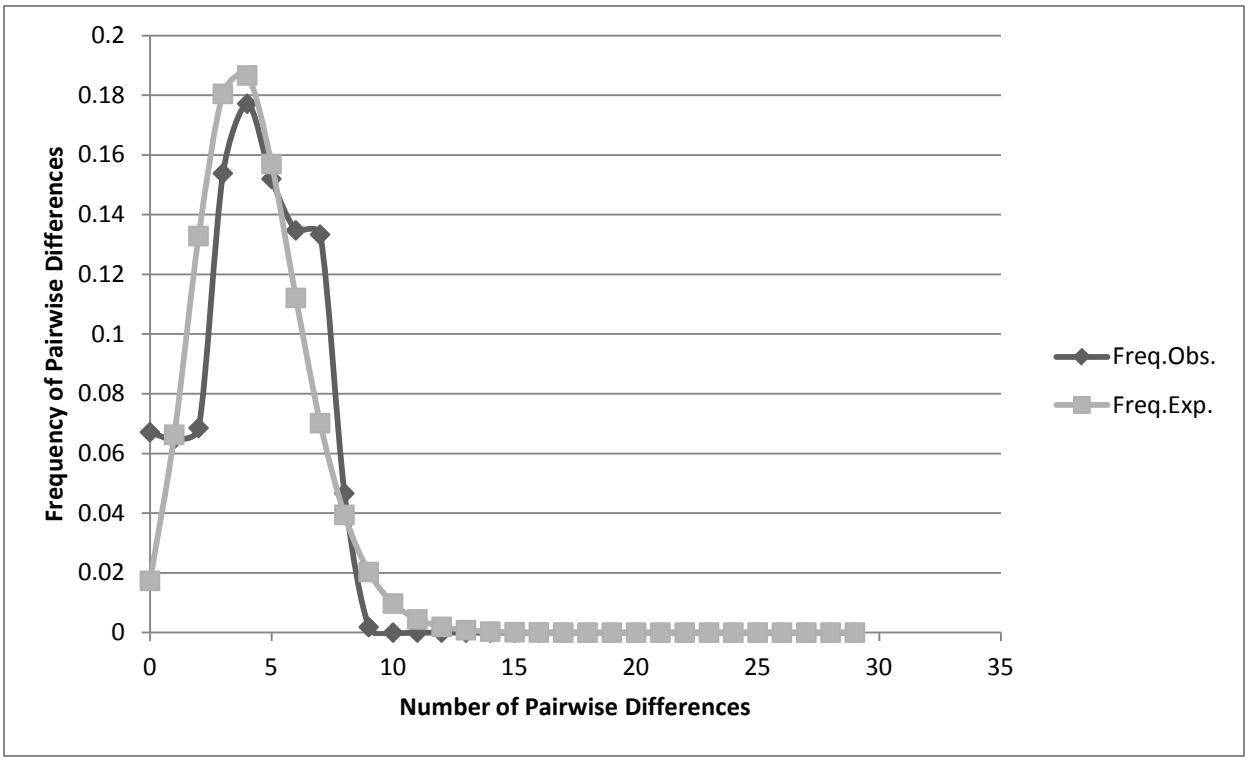
Clade	Inference Key	Conclusion
4-6	No Inference	No Inference
3-12	1-2-3-5-6-7-8-NO:	IBD with either SDD or LDD
2-25	1-19-20-2-3-4-NO:	IBD
1-81	1-2-3-4-NO:	IBD

3.3.23 Mismatch distribution and population genetics of species 1 lineage E

The mismatch distribution (fig. 3.15) was unimodal and failed to reject the null hypothesis (r=0.0183; $p=0.07$) of a demographic expansion. The expected distribution curve was determined from a simulated population undergoing constant population expansion with τ = 3.601.

Tajima's D (D=-1.89, $p=0.001$) indicated a significant departure from neutrality and the negative score was the result of an excess of rare alleles. Similarly, Ramos-Onsins and Rozas R_2 statistic (R_2=0.0421 $p=0.014$) and Fu's F ($_F$=-15.58 $p<0.001$) indicated a similar deviation from neutrality on the basis of excess singletons and alleles, respectively.

Figure 3.15: Mismatch distribution of species 1 lineage E
The mismatch distribution of COI haplotypes was plotted against an expected curve of a population undergoing constant demographic expansion with $\tau = 3.601$.



3.3.24 Analysis of molecular variance of species 1 lineage E

The AMOVA (table 3.8) partitioned by faunal provinces resulted in two groups (Great Lakes: 14 populations, Hudson Bay: 11 populations) and the between-group variance component (V_{BG}=13.30) was significant (*p<0.001*). Grouping the populations on the basis of glaciated and unglaciated regions was not possible due to a lack of individuals detected in ice-free areas.

The SAMOVA (table 3.8) was conducted with increasing numbers of groups. The 4-group setup (V_{BG} = 52.35) was the first to have at least two assemblages of 3 or more populations (Group 1: 19 populations, Group 2: 3 populations, Group 3: 2 populations and Group 4: 1 population) and had the highest between group variance component (52.35%).The second group included three populations that coincide with clade 2-30 (fig. 3.14) that were located within the watersheds of Lake Ontario. This partition accounted for a large percentage of the variance, in contrast to the amount explained by differences among populations within the groups (16.42%). This suggests that the 19 populations within the larger partition were homogenized and while partitioning the populations according faunal province was significant (table 3.8), it explained far less of the variation than indicated by the SAMOVA.

Table 3.8: AMOVA and SAMOVA of species 1 lineage E
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. The 4-group partition in the SAMOVA (bolded) was the best supported (see text). * p<0.01, ** p<0.001, ---- Not Applicable

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	13.30**	49.21**	37.49*
Unglaciated vs. Glaciated	----	59.99**	40.01**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	34.18**	37.29**	28.05**
3	45.21**	27.31**	27.48**
4	**52.35****	**16.42****	**31.23****
5	38.01**	28.32**	33.67**
6	51.62**	14.59**	33.78**

3.3.25 Phylogeography of species 1 lineage F

Fifty-five species 1 lineage F individuals (18 haplotypes) were collected from 16 sites. Thirteen of these populations were located within a region stretching from the southern end of the Canadian Rockies to the state of Alaska. The remaining five populations were distinctly separated from the main cluster, with two in Northwest Territories and three distributed across central Canada. The COI sequences of lineage F formed a monophyletic group with high bootstrap support (98) in the neighbor-joining analysis (figure 3.3) with a shallow nucleotide diversity (π=0.00468, table 3.1).

The 24 unique haplotypes (S1H183 to S1H205) were nested into twelve 1-step clades (1-99 to 1-110), three 2-step clades (2-33 to 2-35) and one 3-step clade (3-14). The network was centered around the interior haplotype S1H183, which was detected in 4 populations.

3.3.26 Outgroup analysis of species 1 lineage F

The HKY model was selected on the basis of the Akaike information criterion. The neighbor-joining and parsimony analyses (starting from either a neighbor-joining tree or 20 random trees) indicated a root between haplotypes S1H199 and S1H205 in clade 1-110 (appendix C, fig. C.6). The species 1 lineage G haplotype (S1H209) was used as the outgroup.

3.3.27 Nested clade analysis of species 1 lineage F

The nested clade analysis (table 3.9) returned three significant geographical associations (clades 1-105, 2-33 and 3-14). Clade 1-105 was discovered in three habitats in Alaska (AK001, AK009 and AK017) and the inference key of the NCA indicated a contiguous range expansion. The interior haplotype S1H192 was restricted to a single habitat (AK001) and had a significantly small clade distance (Dc = 0, *p=0.0014*) and the tip haplotype S1H193 had a significantly large nested clade distance (Dn = 54.12, *p=0.0099*). The inference key returned two possible outcomes for clade 2-33 as a consequence of clade 1-101 having a significantly small clade distance and a significantly large nested clade distance (Dc = 36.68, *p<0.001*, Dn = 1216.59, *p=0.003* respectively); either a contiguous range expansion or past fragmentation of a larger range (fig. 3.16). The NCA for the total cladogram (clade 3-14) indicated restricted gene flow with isolation by distance (table 3.9) as a consequence of the significantly large I-T (interior minus tip) nested clade distance ($I\text{-}T_{Dn}$ = 806.46, *p = 0.0057*).

Figure 3.16: Map showing the geographical distribution of species 1 lineage F
The haplotypes of species 1 lineage F possess geographic distributions represented on the map (right) and their corresponding network positions are indicated on the insert (left). The central haplotype S1H183 as well as its co-distributed tip clades (1-100 to 1-105) are depicted in green. Haplotypes S1H185, S1H199 and clade 1-107 share a habitat (yellow + purple). The distributions of the remaining haplotypes in clade 2-34 are colored blue and clade 2-35 is orange. The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades. The root is marked on the insert as (R) and by a dotted line on the map.

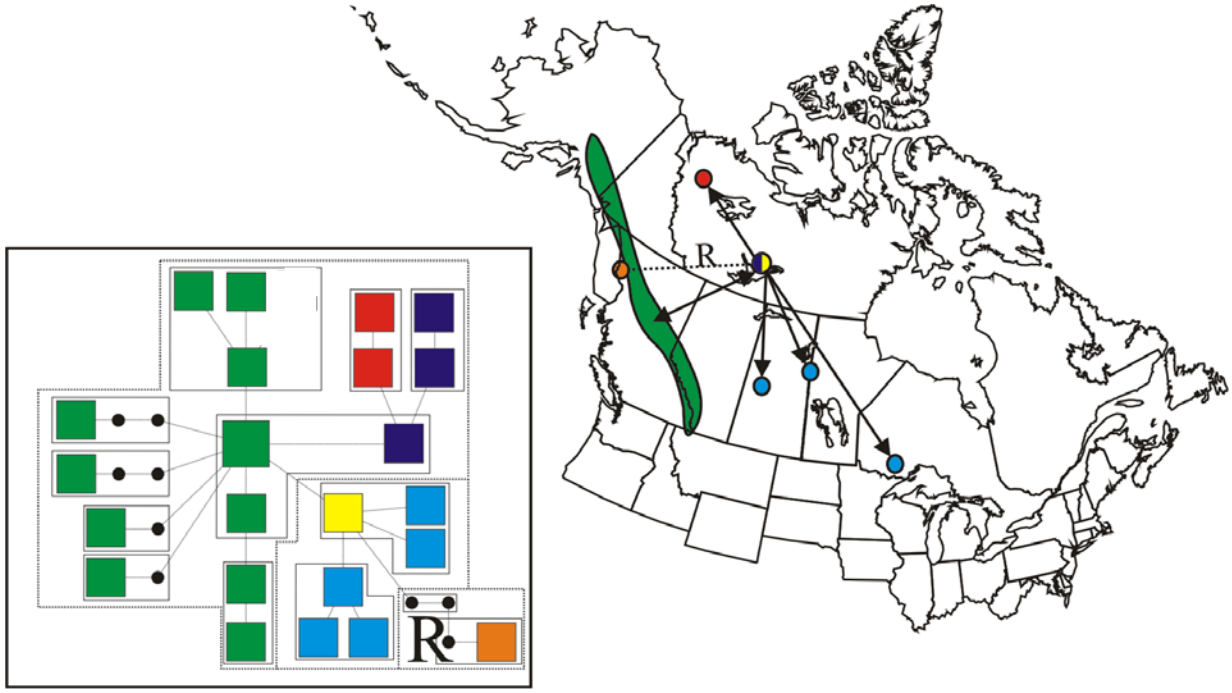


Table 3.9: Nested Clade Analysis of species 1 lineage F
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance, RE: Range Expansion, PF: Past Fragmentation and LDD: Long Distance Dispersal. See appendix B for full NCA table of Dc and Dn values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

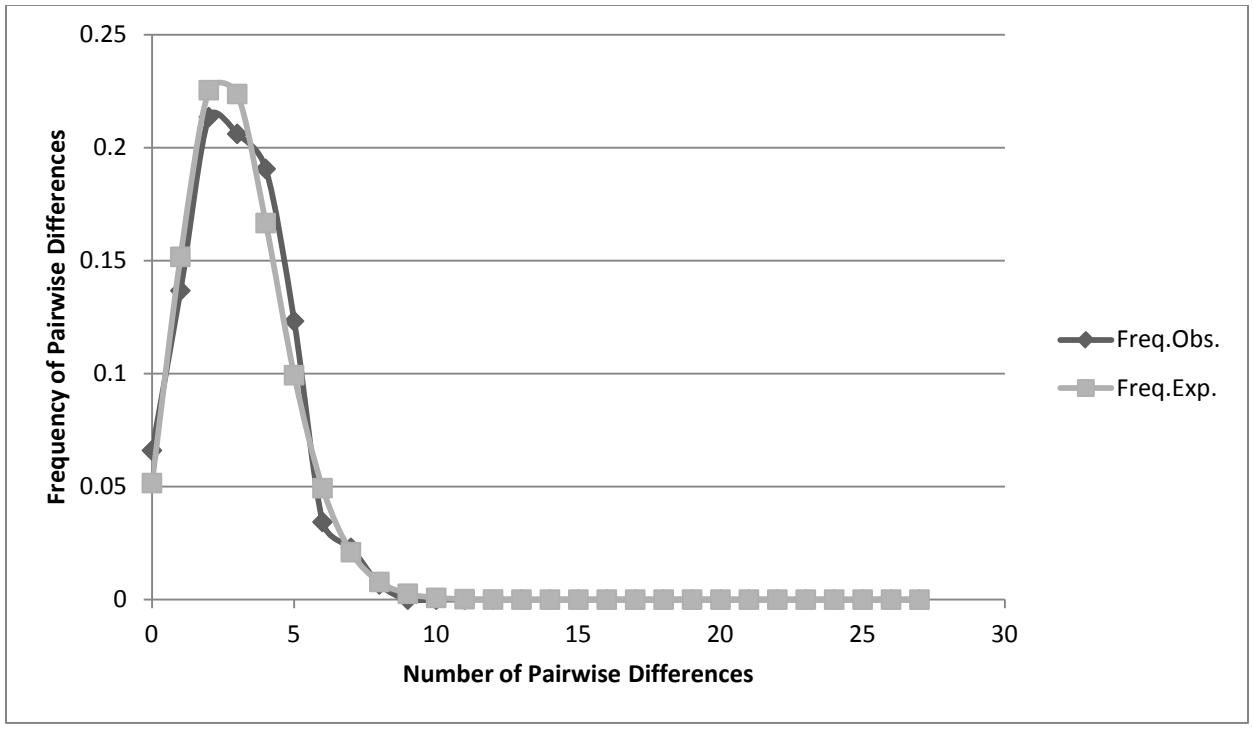
Clade	Inference Key	Conclusion
3-14	1-19-20-2-11-17-4-NO:	IBD
2-33	1-2-11-12-13-YES:	RE or PF
1-105	1-2-11-12-NO:	CRE

3.3.28 Mismatch distribution and population genetics of species 1 lineage F

The mismatch distribution (fig. 3.17) had a unimodal peak and the raggedness statistic (table 3.1) did not reject the null hypothesis of population expansion (r=0.024, *p=0.052*). The expected distribution curve of a population undergoing a constant population expansion was estimated from $\tau = 2.980$.

Deviations from neutrality were detected in Tajima's *D* (D_T=-1.84, *p=0.003*), Ramos-Onsins and Rozas R_2 statistic (R_2=0.0457, *p=0.014*) and Fu's *F* (*F*=-15.38, *p=0.000*). These measures were all significant (table 3.1) and consistent with a population undergoing demographic expansion.

Figure 3.17: Mismatch distribution of species 1 lineage F
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with $\tau = 2.980$



3.3.29 Analysis of molecular variance for species 1 lineage F

The AMOVA partitioned by faunal provinces possessed 3 groups (Hudson Bay: 4 populations, Cascadia: 5 populations and Yukon-Mackenzie: 9 populations), but the between-group and among population within groups variance components were not significant (table 3.10). An AMOVA subdivided into glaciated and unglaciated regions (table 3.10) was not possible because no populations were discovered in ice-free areas.

Table 3.10: AMOVA and SAMOVA of species 1 lineage F
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * $p<0.01$, ** $p<0.001$, NS – Not Significant, ---- Not Applicable

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	2.55^{NS}	41.46^{NS}	55.99*
Unglaciated vs. Glaciated	----	----	----

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	41.36^{NS}	24.53**	34.11**
3	35.38*	26.73**	37.89**
4	29.49**	21.26**	49.25**
5	37.73**	21.74^{NS}	50.38**
6	30.80**	16.46^{NS}	50.43**

The SAMOVA (table 3.10) consisting of 4 groups was the only setup in which both the Among-Populations (V_{AP} = 21.26) and Between-Group (V_{BG} = 29.49) variance components were significant (*$p<0.001$*). At this level only two of the assemblages had three or more populations, an Alaskan group (AK001, AK009 and AK017) and a non-Alaskan group (11 populations). Lake Palmer (BC040) and two Saskatchewan locations (SK001 and SK009) were partitioned into the last two groups, respectively. The haplotypes of the Alaskan group were consistent with clade 1-105 and were detected in populations with little haplotype diversity. The

Hyalella collected from samples sites AK001 (n = 5) and AK007 (n = 6) were fixed for a single haplotype each (S1H192 and S1H195, respectively).

3.3.30 Phylogeography of species 1 lineage G

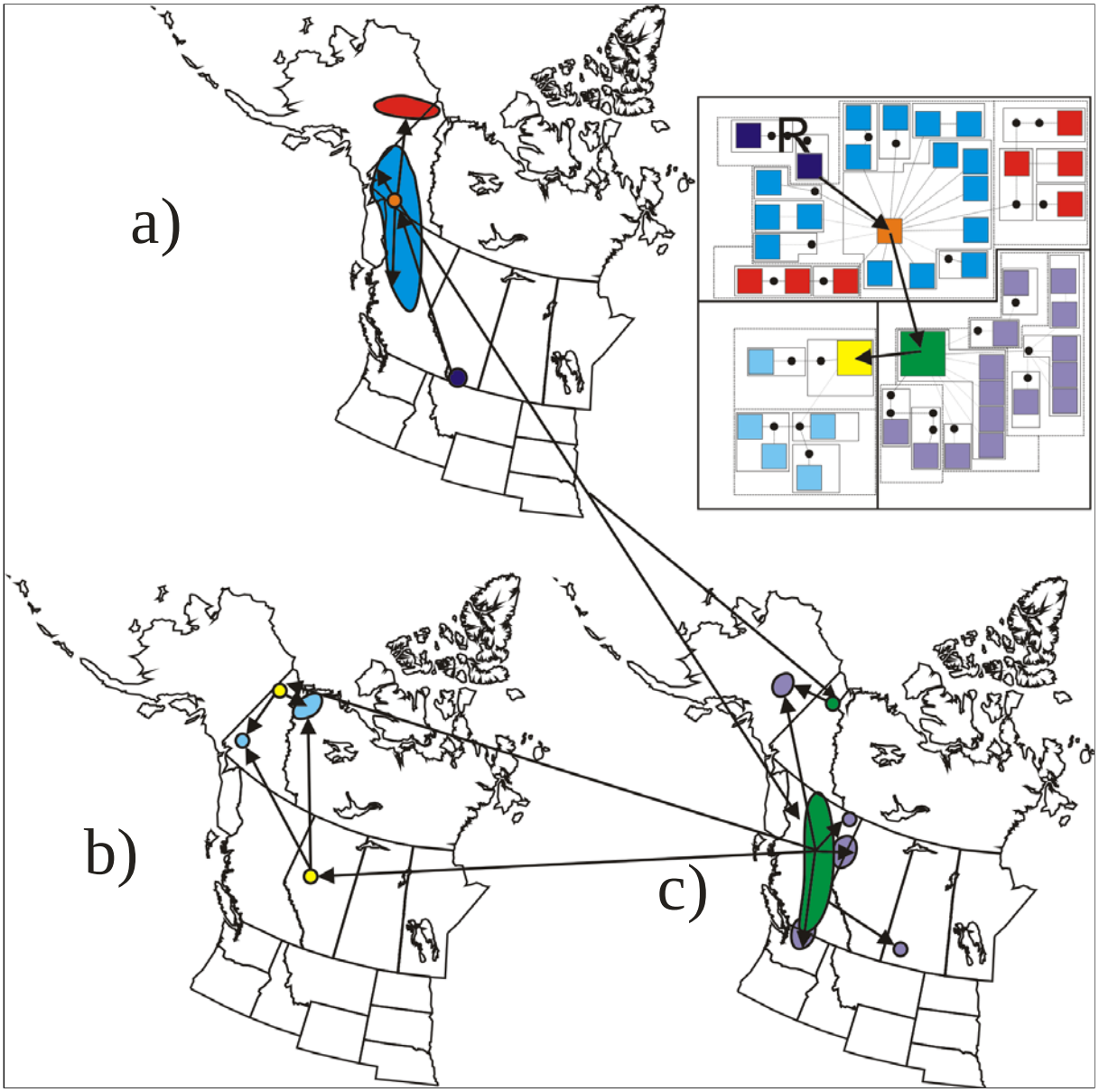
One-hundred and nine species 1 lineage G individuals were sampled (representing 48 haplotypes) from 28 sites throughout western Canada. In the neighbor-joining analysis of all species 1 lineages, it formed a monophyletic group (figure 3.3) with high bootstrap support (100) and low nucleotide diversity (π=0.0057).

The COI statistical parsimony network (appendix C, fig. C.7) was constructed using 48 unique haplotypes and nested into ten 2-step clades (2-36 to 2-45), three 3-step clades (3-15 to 3-17) and a single 4-step clade (4-7).

3.3.31 Outgroup analysis of species 1 lineage G

The Tamura-Nei model with a proportion of invariant sites (I=0.784) was selected on the basis of the Akaike information criterion. The neighbor-joining and parsimony analyses unambiguously placed the root along the edge connecting haplotypes S1H207 to S1H208 between clades 1-111 and 1-112. Species 1 lineage A haplotype S1H003 was used as an outgroup.

Figure 3.18: Map showing the geographical distributions of species 1 lineage G
The geographic distributions of species 1 lineage G haplotypes are represented on the maps and their corresponding network positions are indicated on the insert (upper right): a) clade 3-15 (orange + blue + red + purple), b) clade 3-16 (green and light purple) and c) clade 3-17 (yellow + light blue). The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades.



3.3.32 Nested clade analysis of species 1 lineage G

The NCA indicated 6 significant geographical associations (table 3.11). The total cladogram (4-7) consists of a tip clade 3-17 and two interior clades (3-15 and 3-16). Both interior clades had significantly small clade distances (Dc(3-15) = 571.82, *p = 0.005*; Dc(3-16) = 536.21, *p<0.001*), which the inference key identifies as a pattern of contiguous range expansion in clade 4-7. All three of these 3-step nested clades also possessed significant geographical associations.

The difference between interior and tip nested clade distances in clade 3-15 was significantly small (I-T Dn = -359.66, *p=0.010*), as such the inference key returned multiple explanations: long distance colonization, extinction of intermediary regions of a larger past range, past range expansion with subsequent fragmentation or past fragmentation with subsequent range expansion. The inference key returns these results when the tip clades are geographically separated from the distribution of interior haplotypes by areas that have been sampled but contain no haplotypes specific to either the tip or interior clades. When the tip and interior clades form disjunct distributions, additional information would be required to resolve the competing inferences. Nested within clade 3-15, clade 1-127 had insufficient geographical sampling to return an inference, but the clade 2-40 NCA indicated restricted gene flow with isolation by distance. This inference resulted from a tip clade (1-128) having a significantly small clade distance, in addition to a significantly large difference in clade distances (Dc = 0, *p=0.001*, I-T Dc = 825.09, *p=0.001*, respectively).

Clade 3-16 was characterized by restricted gene flow with isolation by distance because its interior clade (2-40) had significantly large clade and nested clade distances (Dc = 763.22, *p=0.003,* Dn = 688.89, *p=0.009,* respectively) along with a tip clade (2-43) with significantly small clade and nested clade distances (Dc = 42.45, *p<0.001,* Dc = 121.85, *p<0.001,* respectively).

The NCA of clade 3-17 indicated restricted gene flow with isolation by distance as a consequence of its interior clade (2-44) having significantly large clade and nested clade distances (Dc = 761.49, *p=0.006,* Dn = 758.41, *p=0.001* respectively) along with the tip clade (2-45) having significantly small clade and nested clade distances (Dc = 18.85, *p=0.002,* Dc = 281.36, *p=0.001,* respectively).

Table 3.11: Nested Clade Analysis of species 1 lineage G
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance, CRE: Contiguous Range Expansion, and IGS: Incomplete Geographical Sampling. See appendix B for full NCA table of Dc and Dn values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key	Conclusion
4-7	1-2-11-12-NO:	CRE
3-15	1-2-11-17-4-NO:	IBD
3-16	1-2-3-4-NO:	IBD
2-40	1-2-3-4-NO:	IBD
1-127	1-19-20-2-11-NO:	IGS
3-17	1-2-3-4-NO:	IBD

3.3.33 Mismatch distribution and population genetics of species 1 lineage G

The mismatch distribution (fig. 3.19) was unimodal and the raggedness statistic (r=0.0334, *p=0.219*) failed to reject the null hypothesis of demographic expansion. The expected distribution curve of a population undergoing a constant population expansion was estimated with tau set to 3.657.

Tajima's *D* (*D* =-2.103, *p=0.002*), Ramos-Onsins and Rozas R_2 statistic (R_2 = 0.0314, *p=0.007*) and Fu's *F* (*F* = -41.95, *p=0.000*) all indicate significant departures from neutrality (table 3.1), which was consistent with a pattern of demographic expansion.

3.3.34 Analysis of molecular variance for species 1 lineage G

The AMOVA partitioned by faunal provinces (table 3.11) resulted in two groups (Cascadia: 11 populations, Yukon-Mackenzie: 17 populations), however the variance explained by the between-group component was small (V_{BG} = 5.96, *p = 0.002*). An AMOVA partitioned into glaciated and unglaciated regions was not possible because populations with lineage G individuals were not detected in ice free regions.

Figure 3.19: Mismatch distribution of species 1 lineage G
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with $\tau = 3.657$.

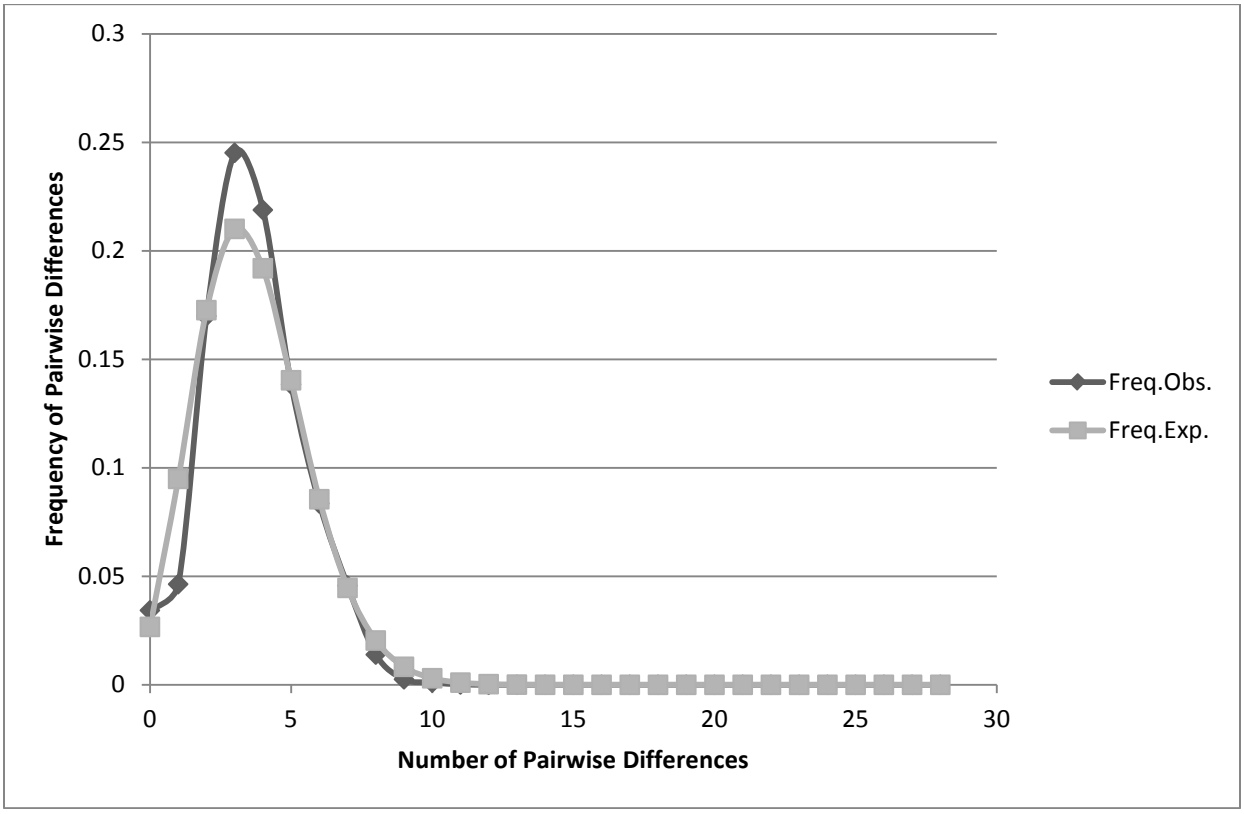


Table 3.12: AMOVA and SAMOVA of species 1 lineage G
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. The 3-group partition in the SAMOVA (bolded) was the best supported (see text). * p<0.01, ** p<0.001, ---- Not Applicable

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	5.96*	47.72**	46.31**
Unglaciated vs. Glaciated	----	----	----

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	31.50**	34.55**	33.95**
3	**27.97****	**35.48****	**36.55****
4	24.82**	31.16**	44.01**
5	25.09**	36.29**	38.62**
6	24.41**	31.00**	44.59**

* p<0.01, ** p<0.001, NS = Not Significant, ---- Not Applicable

The SAMOVA was conducted using 2 to 6 groups and was significant at all levels (table 3.12). The initial 2-group setup had the highest between-group variance component (table 3.12), however only one of the groups had more than three populations. The 3-group partition had at least 2 assemblages with three populations: a 3-population group in the Northwest Territories (NT001, NT003 and NT006), a single Yukon population group (YT001) and the remaining 24 populations in the third assemblage. The SAMOVA setups with more assemblages had a between-group variance component that explained a lower amount of variance (table 3.12) and did not result in any additional subdivisions of three or more populations. The haplotypes in the three-population assemblage were consistent with clade 2-45, although the samples sites had low sample sizes.

3.3.35 Phylogeography of species 1 lineage H

One-hundred and three individuals assigned to lineage H (41 haplotypes) were sampled from 21 habitats predominately located south of the Cordilleran glacial limit as well as within two populations from Nevada considered by Witt *et al.*, (2006). These COI haplotypes formed a monophyletic group in the neighbor-joining analysis of all species 1 haplotypes (figure 3.3) with low bootstrap support (63). The inclusion of haplotype S1H351 from Chuchi Lake (BC017) increased the bootstrap support (81), but was greater than ten mutational steps from the nearest sequence. Even with haplotype S1H351 excluded, lineage H had the highest nucleotide diversity (π=0.0158) of any species 1 lineage (table 3.1).

The COI statistical parsimony network (appendix C, fig. C.8) was constructed using forty-one unique haplotypes (S1H253 to S1H293) that were nested into twenty 1-step clades (1-141 to 1-160), ten 2-step clades (2-46 to 2-55), six 3-step clades (3-18 to 3-23), two 4-step clades (4-8 and 4-9) and one 5-step clade (5-3). The network was highly diverse and many of the tip clades were separated by interior clades composed predominated by inferred mutational steps (see clade 3-19 in figure 3.27).

3.3.36 Outgroup analysis of species 1 lineage H

The root of the network was determined on the basis of the neighbor-joining and parsimony phylogenetic analyses using species 1 lineage G as an outgroup (haplotype S1H209). The model of molecular evolution selected was HKY with a gamma distributed rate variation (α=0.158). The maximum parsimony analysis was conducted using nearest neighbor interchange branch

swapping starting from both a neighbor-joining tree (using a HKY parameter model) as well as 20 replicates starting from random trees. As a result, the root from both analyses was unambiguously placed within clade 3-19 among the inferred haplotypes between the 2-step clades 2-47, 2-48, 2-49 and 2-56 (appendix C, fig. C.8).

3.3.37 Nested clade analysis of species 1 lineage H

The nested clade analysis returned 6 significant geographical associations. Three of which (clades 2-46, 3-20 and 5-3) the inference key indicated the potential for range expansions in conjunction with long distance colonization (table 3.12).

The NCA for the total cladogram (clade 5-3) indicated a range expansion as a consequence of significantly small interior clade (4-8) and nested clade distances (Dc = 326.62, $p<0.001$, Dn = 367.68, $p<0.001$ respectively). The inference key also returned an explanation of long distance colonization resulting from a tip clade (4-9) having a significantly large nested clade distance (Dn = 625.77, $p=0.001$).

Clade 4-8 included four of the 3-step clades (3-18 to 3-21; 23 populations, 72 haplotypes) and the NCA (table 3.12) indicated isolation by distance as result of a tip clade (3-20) having significantly small clade and nested clade distances (Dc = 115.35, $p=0.004$, Dn = 140.82, $p=0.002$, respectively). Nested within clade 4-9 were two 3-step clades with significant geographical associations. Clades 3-19 (7 populations; 16 haplotypes) and 3-20 (5 populations; 21 haplotypes) both occupied locations along the Canadian-US border in southern British Columbia and were sympatric in only a single habitat (BC024), although clade 3-19 was also

detected in populations in Oregon (OR006, OR008, OR010 and OR015). The inference key returned multiple explanations in either case. The genetic patterns in clade 3-19 suggested either allopatric fragmentation or isolation by distance because the I-T clade distance was significantly large (I-T D_n = 58.66, *p=0.002*), the tip clade (2-48) had a significantly small clade distance (D_c = 52.95, *p<0.001*), and the geographic region between clades 2-48 and 2-49 was devoid of clade 3-19 individuals. Since the two 2-step clades were separated by a long edge of inferred haplotypes, the NCA favors allopatric fragmentation. The NCA indicated that the distribution of clade 3-20 was shaped by a range expansion with long distance dispersal, because the tip clade 2-51 had significantly small clade and large nested clade distances (D_c=0, *p=0.010*, D_n=147.43, *p<0.001*, respectively).

Nested within clade 3-18 were 2-step clades (2-46 and 2-47) with significant geographical associations. Clade 2-46 (8 populations, 27 haplotypes) individuals were collected from localities in southern British Columbia (yellow region on the map in figure 3.28), as well as two additional populations in the northern end of the province (BC048 and BC050). The two disparate locations formed a tip clade (1-141) with a significantly large nested clade distance (D_n=653.52. *p=0.005*) and in conjunction with an interior clade (1-142) with significantly small clade distance (D_c=133.90. *p=0.011*), the inference key returned an explanation of range expansion with long distance colonization. The NCA of clade 2-47 (6 populations, 10 haplotypes) indicated restricted gene flow with isolation by distance among populations in the states of Washington and Oregon (red and orange in figure 3.28) because the I-T (interior minus tip) clade distance was significantly small (I-T D_c=-267.41, *p=0.017*).

Figure 3.20: Map showing the geographical distribution of species 1 lineage H
The geographic distributions of species 1 lineage H haplotypes are represented on the maps and their corresponding network positions are indicated on the insert (lower): a) clade 3-18 (orange + yellow+ green), b) clades 3-19 to 3-21 (blue + light purple + pink) and clade 4-9 (dark purple + magenta). The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades. The root (R) is noted on both the network and as dotted lines on map.

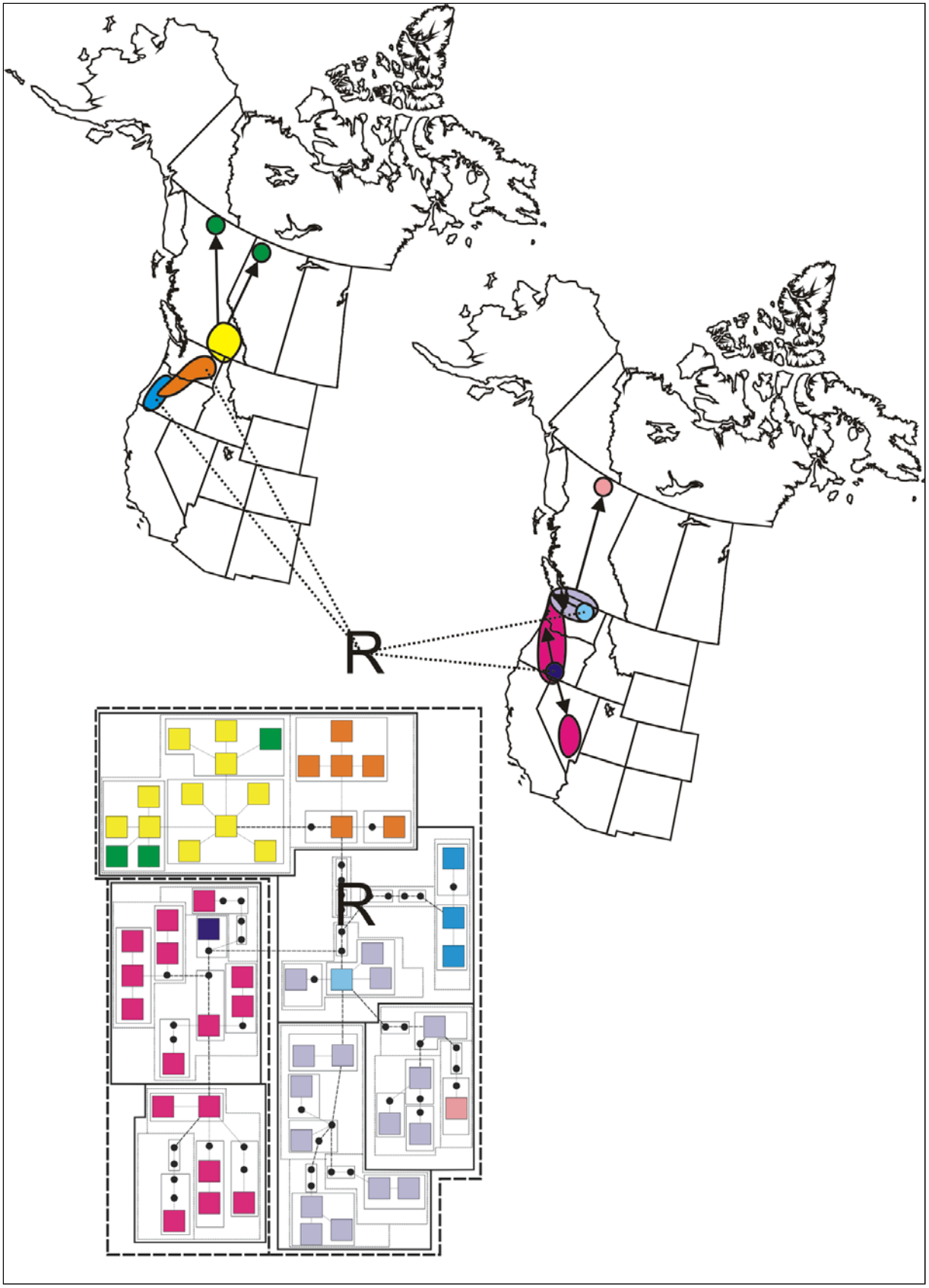

R
R

Table 3.13: Nested Clade Analysis of species 1 lineage H
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance, RE: Range Expansion, CRE: Contiguous Range Expansion, AF: Allopatric Fragmentation, LDD: Long Distance Dispersal, RE(C): Range Expansion with Long Distance Colonization, PF: Past Fragmentation or IGS: Incomplete Geographical Sampling. See appendix B for full NCA table of Dc, Dn and p values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key	Conclusion
5-3	1-2-11-12-13-YES:	RE(C)
4-8	1-2-3-4-NO:	IBD
3-18	No Inference	No Inference
2-46	1-2-11-12-13-YES:	RE(C)
2-47	1-2-3-4-NO:	IBD
3-19	1-19-20-2-3-4-9-10-NO:	PF or IBD
3-20	1-2-3-5-6-7-YES:	IBD with LDD

3.3.38 Mismatch distribution of species 1 lineage H

The mismatch distribution (fig. 3.21) was not unimodal, although a large observed peak was offset from the expected frequency curve derived from a simulated population undergoing a constant demographic expansion with tau set at 6.562. The raggedness statistic (table 3.1) rejected the null hypothesis of population expansion (r=0.0061, *p=0.008*).

3.3.39 Analysis of molecular variance for species 1 lineage H

The AMOVA partitioned by faunal provinces (table 3.13) resulted in 3 groups of populations (Cascadia: 9 populations, West Coast: 15 populations, Yukon-Mackenzie: 2 populations) and the between-group component explained 13.14% of the variance, although it was not significant (*p = 0.0137*). An AMOVA subdivided into glaciated and unglaciated regions (glaciated: 11 populations, unglaciated: 15 populations) also had a non-significant between-group variance component (V_{BG} = 14.24, *p = 0.0117*).

Figure 3.21: Mismatch distribution of species 1 lineage H
The mismatch distribution of COI haplotypes plotted against an expected frequency curve of a population undergoing constant demographic expansion with $\tau = 6.562$.

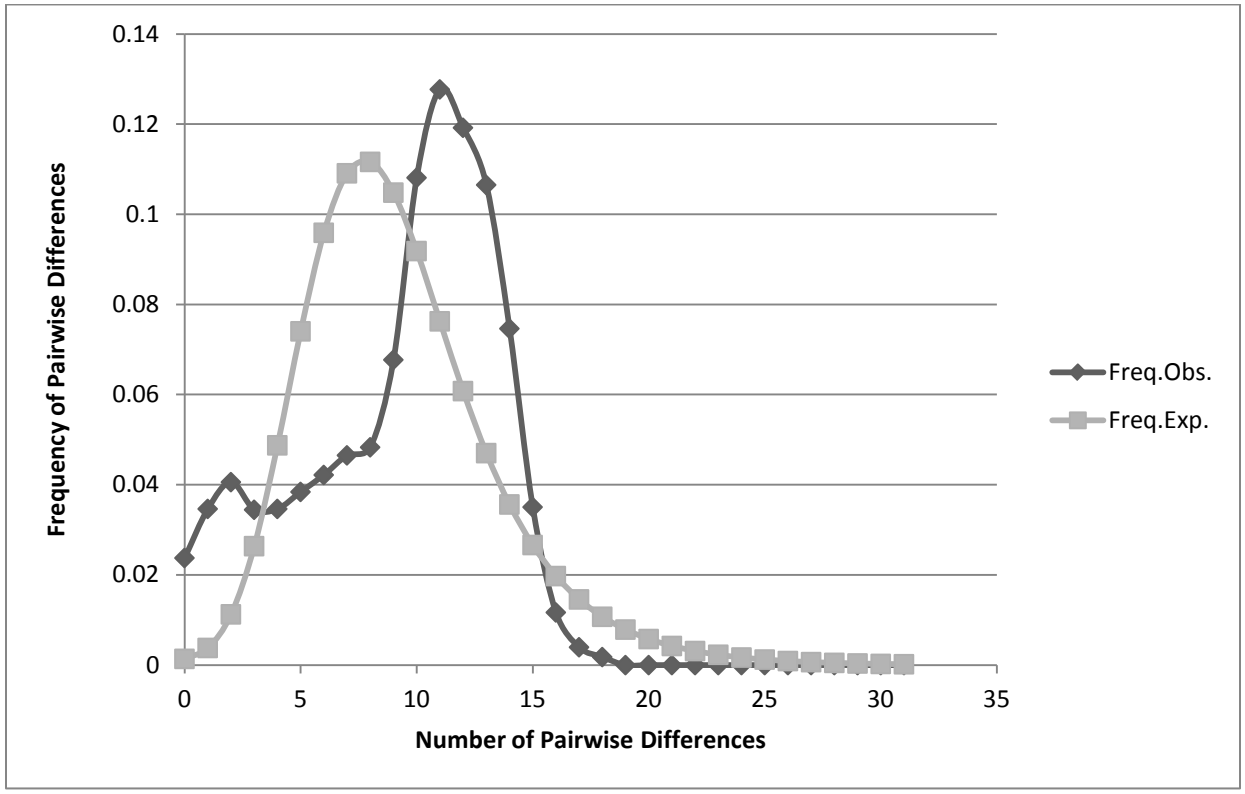


Table 3.14: AMOVA and SAMOVA of species 1 lineage H

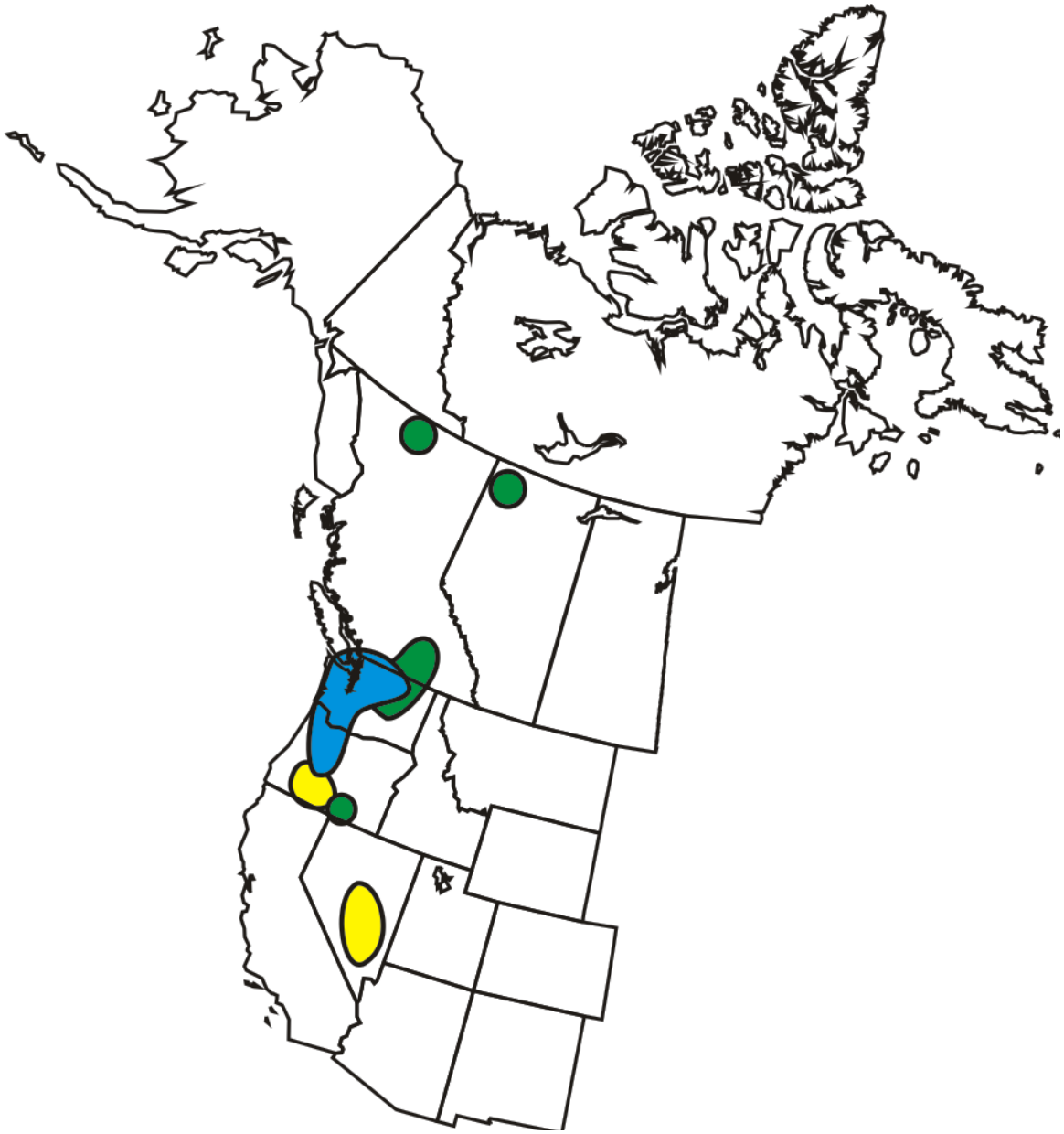
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. The 3-group partition in the SAMOVA (bolded) was the best supported (see text). * p<0.01, ** p<0.001, NS = Not Significant, ---- Not Applicable

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	13.14^{NS}	54.46**	32.40**
Unglaciated vs. Glaciated	16.85^{NS}	51.74**	31.41**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	28.51**	42.03**	29.46**
3	**46.17****	**24.75****	**29.08****
4	46.07**	23.84**	30.09**
5	55.35**	14.58**	30.07**
6	55.77**	11.78**	32.45**

The 3-group SAMOVA (table 3.14) split the populations into a Cascadian group (11 populations; blue in fig. 3.21), a southern group (7 populations; yellow in figure 3.30) and an eastern group (8 populations; green in figure 3.30). These assemblages coincide with the major clades within the COI haplotype network (fig. 3.22); the Cascadia group was consistent with clades 3-19 to 3-21, the southern group matches clade 4-8 and the eastern group with clade 3-18. The SAMOVA at higher levels did not result in any further subdivisions consisting of three or more populations and the 2-group condition had a notably lower between-group variance component (table 3.14).

Figure 3.22: Map showing the geographical distributions of populations within the 3-group SAMOVA of species 1 lineage H
The SAMOVA split lineage H into 3 assemblages of populations: the Cascadian group (blue), southern group (yellow) and eastern group (green).



3.3.40 Phylogeography of species 1 lineage I

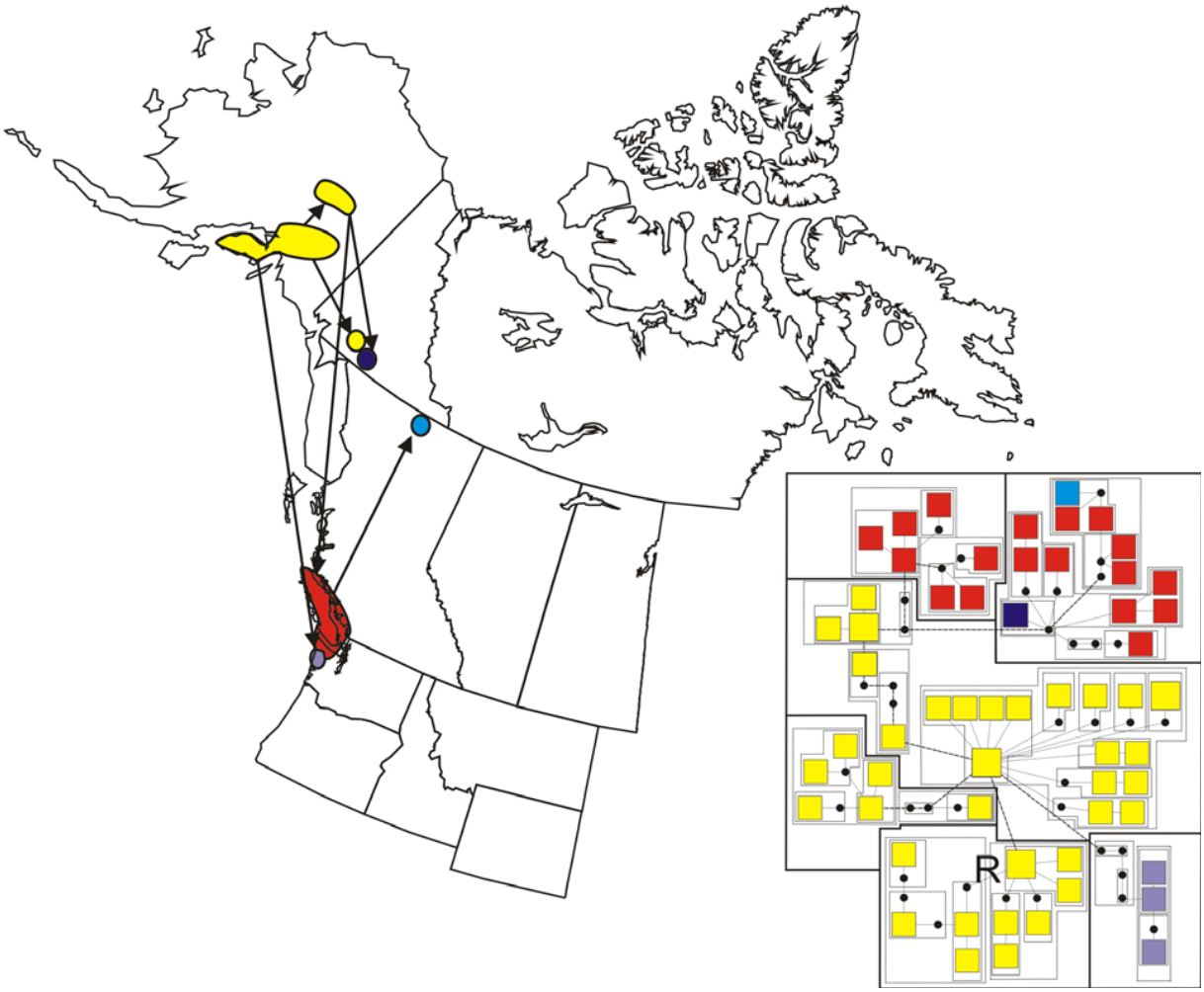
One-hundred and twenty-nine individuals (representing 59 haplotypes) were sampled from 27 habitats (table 3.1). The COI sequences formed a monophyletic group with strong bootstrap support (97) within the neighbor-joining phylogeny that included all species 1 lineages (figure 3.3). When compared to other species 1 lineages, the nucleotide diversity (π=0.00924) was the second highest (table 3.1). The spatial distribution of lineage I individuals consisted of two disjunct regions; a northern Alaskan area along the Kenai Peninsula and a second group of populations restricted to western Washington State and Vancouver Island, BC.

The statistical parsimony network (appendix C, fig. C.9) was constructed using fifty-nine unique COI haplotypes (S1H316 to S1H374) that nested into thirty 1-step clades (1-172 to 1-201), thirteen 2-step clades (2-61 to 2-73), six 3-step clades (3-24 to 3-29) and a single 4-step clade (4-9). The central clade 3-24 was collected from sites throughout Alaska and the Yukon, but was predominately detected in the Kenai Peninsula (figure 3.32). Clade 3-25 was restricted to a single habitat in Washington State (WA015). Clades 3-26 and 3-27 were codistributed with clade 3-24 along the Kenai Peninsula. Clades 3-28 and 3-29 were located throughout Vancouver Island and Washington State (figure 3.23), although there was a single tip haplotype (S1H373 in clade 1-200; see appendix C, fig. C.9) collected from northern British Columbia (BC048). Interestingly, the haplotypes connecting these two 3-step clades in the network were only detected in northern Alaskan populations (fig. 3.23).

3.3.41 Outgroup analysis of species 1 lineage I

The general time reversible model with a proportion of invariant sites (I=0.299) and a gamma distribution (α=1.241) for rate heterogeneity was selected based on Akaike information criterion. The parsimony analysis, using the nearest neighbor interchange branch swapping algorithm, ran out of memory, as such a 50% majority rule consensus tree was constructed using the 486,000 saved trees. Both the neighbor-joining and parsimony analyses unambiguously placed the root of the network between haplotypes S1H347 and S1H348 in clade 3-26 (appendix C, C.9). The species 1 lineage G haplotype S1H209 was used as an outgroup.

Figure 3.23: Map showing the geographical distribution of species 1 lineage I
The geographical distributions of species 1 lineage I *Hyalella* coincide with the haplotypes shown in the insert (appendix C, fig. C.9). Clade 3-24, 3-26 and 3-27 have distributions represented on the map by the color yellow. The region containing clade 3-2 is denoted by light purple. Clades 3-28 and 3-29 are represented by the color red, except for the isolated habitats containing individuals with haplotypes S1H362 and S1H373, which are colored dark purple and blue, respectively. The root is denoted by (R). The arrows on the map trace a path from the distributions of the central haplotypes towards the distributions of the derived clades.



3.3.42 Nested clade analysis of species 1 lineage I

The nested clade analysis returned 5 significant geographical associations (table 3.14). When the inference key was applied to the total cladogram (4-9), it indicated isolation by distance in conjunction with long distance dispersal. The interior clade 3-24 had significantly small clade and nested clade distances along with tip clades (3-25 and 3-28) that had significantly small clade and large nested clade distances ($Dn_{(3\text{-}25)}$=0.000, *p<0.001*, $Dn_{(3\text{-}28)}$=0.730, *p=0.0033*, respectively). The distributions of the clades with reversals in significance formed a congruent geographic subset located south of the Cordilleran ice sheet, which may indicate fragmentation rather than long distance dispersal.

The NCA of clades 3-26 and 3-28 returned inferences of isolation by distance as a result of the small clade distances in the tip clades 2-66 and 2-70, respectively ($Dc_{(2\text{-}66)}$=2.76, *p<0.001*, $Dc_{(2\text{-}70)}$=0, *p=0.012*).The clade 2-63 NCA indicated isolation by distance in conjunction with long distance colonization, a larger range in the past coupled with subsequent extinction within intermediate areas, or past range expansion. This was the result of tip clades (1-176 and 1-180) having significantly small clade distances ($Dc_{(1\text{-}176)}$=0, *p<0.001*, $Dc_{(1\text{-}180)}$=0, *p<0.001*) along with a tip clade (1-182) with a significantly large nested clade distance (Dn=1954.76, *p<0.001*). The NCA of clade 1-175 (nested within clade 2-63) indicated isolation by distance because the tip haplotypes S1H324 and S1H325 had significantly small clade distances ($Dc_{(S1H324)}$=0, *p=0.001*, $Dc_{(S1H325)}$=0, *p=0.009*).

Table 3.15: Nested Clade Analysis of species 1 lineage I
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IBD: Isolation by Distance, and LDD: Isolation by Distance with Long Distance Dispersal. See appendix B for full NCA table of Dc, Dn and p-values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

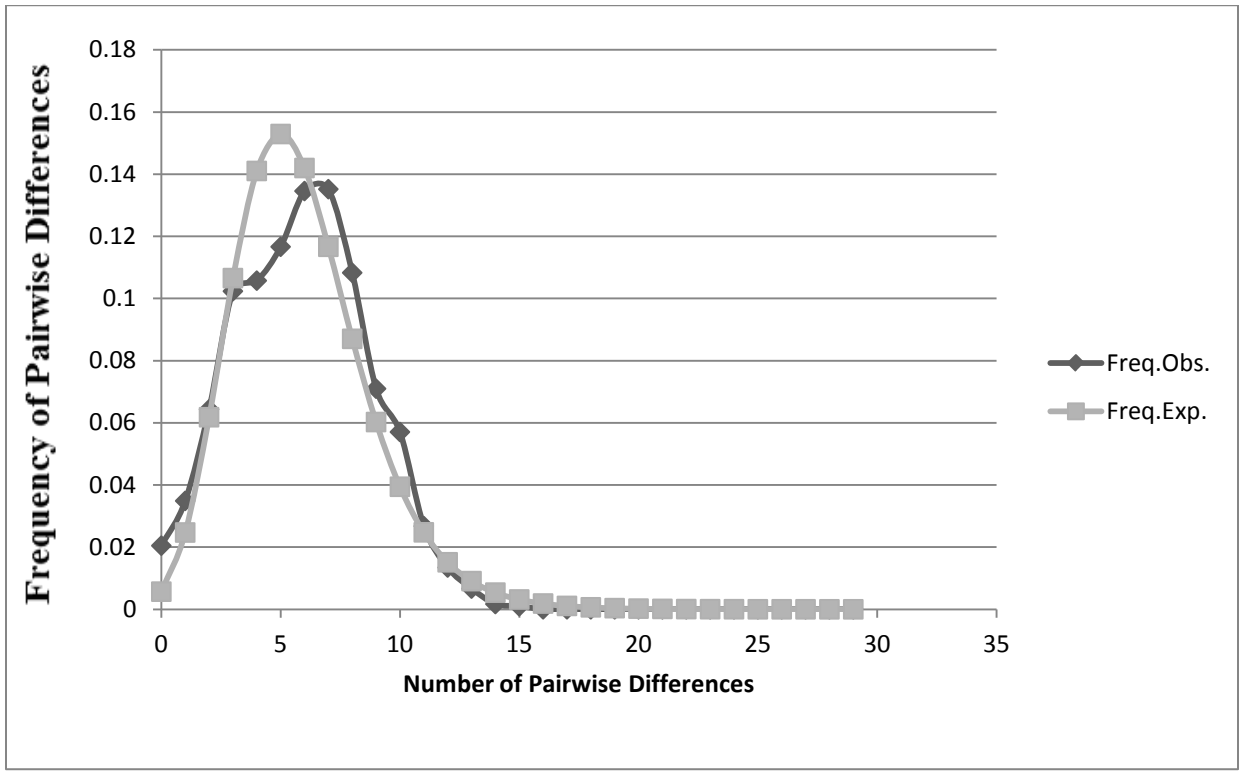
Clade	Inference Key	Conclusion
4-10	1-2-3-5-6-7-YES:	LDD
3-24	No Inference	No Inference
2-63	1-2-3-5-6-13-YES:	see text
1-175	1-2-3-4-NO:	IBD
3-26	1-2-3-4-NO:	IBD
3-28	1-2-3-4-NO:	IBD

3.3.43 Mismatch distribution and population genetics of species 1 lineage I

The mismatch distribution of COI sequences was not completely unimodal (fig. 3.24) and the raggedness statistic (table 3.1) rejected the null hypothesis of a demographic expansion (r=0.0064, *p=0.001*). The expected frequency curve was estimated from a population undergoing a constant population expansion with tau set at 4.477.

Tajima's *D* was negative and identified a significant departure from neutrality (*D*= -2.021, *p=0.003*), as a result of excess rare alleles. The Ramos-Onsins and Rozas R_2 statistic (R_2= 0.033, *p=0.006*) and Fu's *F* (*F*= -46.60, *p<0.001*) also indicated deviations from neutrality consistent with a lineage undergoing demographic expansion.

Figure 3.24: Mismatch distribution of species 1 lineage I
The mismatch distribution of COI haplotypes was plotted against an expected frequency curve of a population undergoing constant demographic expansion with $\tau = 4.477$.



3.3.44 Analysis of molecular variance for species 1 lineage I

The AMOVA partitioned by faunal provinces split the populations into 2 groups (Cascadia: 8 populations; Yukon-Mackenzie: 19 populations). The between-group component explained 21.32% of the variance (*p<0.001*). AMOVA partitioned into unglaciated and glaciated regions could not be conducted, because none of the populations were from ice-free regions.

Table 3.16: AMOVA and SAMOVA for species 1 lineage I
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. The 2-group partition in the SAMOVA (bolded) was the best supported (see text). * p<0.01, ** p<0.001, ---- Not Applicable

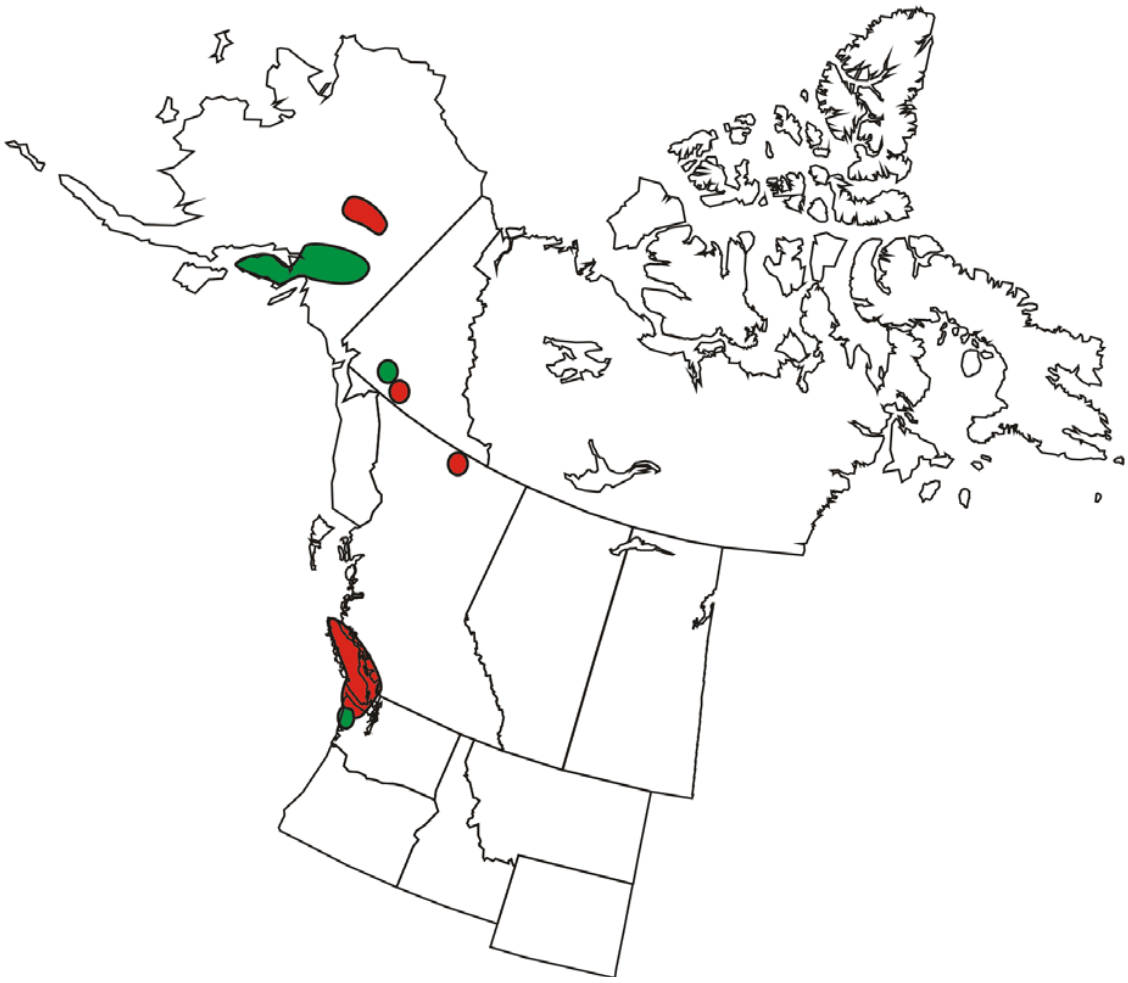
AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	21.32**	61.02**	17.66**
Unglaciated vs. Glaciated	-----	-----	-----

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	**41.30****	**42.99****	**15.71****
3	44.96**	39.18**	15.86**
4	45.60**	38.06**	16.34**
5	43.20**	40.46**	16.34**
6	43.61**	39.07**	17.32**

The SAMOVA divided the populations into 2 major assemblages (group 1: 17 populations; group 2: 10 populations). Increasing the number of subdivisions (V_{BG}=41.3, *p<0.001*) did not explain a greater proportion of the between-group variance component (table 3.15) or result in additional 3+ population assemblages. The 2-group SAMOVA divided the populations into a Beringian group (figure 3.25) and a second assemblage whose range spanned both the Cascadian and Yukon-Mackenzie faunal provinces (figure 3.25).

Figure 3.25: Map showing the geographical distributions of the groups within the 2-group SAMOVA for species 1 lineage I

The groups are based on the 2-group SAMOVA (table 3.15) with two major population assemblages. The first partition consists of 17 populations and its distribution is shown on the map in green. The second group possesses 10 populations is denoted by red.



3.4 RESULTS: SPECIES 2

3.4.1 Phylogeography of species 2

Hyalella 'azteca' assigned to species 2 were collected from 123 sites (table 3.1) throughout glaciated and ice-free regions of North America. This species occurred over a region spanning more than 8000 km, stretching from Alaska to Washington State, and had an eastern limit demarcated by a line between the Great Lakes region and Nebraska. Over 58% of the habitats possessed at least one additional species of *Hyalella* (appendix A). Twenty-five of the populations within the Great Lakes region contained both species 1 and 2, whereas throughout central Canada, 26 habitats possessed both species 2 and 7. Five-hundred and fourteen individuals were sequenced at the COI locus (representing 211 haplotypes), which had a nucleotide diversity of $\pi = 0.0163$ (table 3.1).

The statistical parsimony COI haplotype network (appendix C, fig. C.10) of species 2 was the largest among the large-bodied morphotypes. While species 1 had more unique COI haplotypes (375 vs. 221) they were delineated into 9 separate lineages, whereas all species 2 haplotypes were separated by edges in the network that did not violate the parsimony criterion limit. Despite the size of the cladogram, there was a distinct lack of reticulation within the network. At the lower tiers there were very few loops that required breaking and the upper tier clades were sorting into three major groups. The two-hundred and twenty COI haplotypes were nested into one-hundred and two 1-step clades (1-1 to 1-102), thirty-seven 2-step clades (2-1 to 2-37), sixteen 3-step clades (3-1 to 3-16), seven 4-step clades (4-1 to 4-7), three 5-step clades (5-1 to 5-3) and the total cladogram was composed of 6 steps.

The structure of the COI network consists of three 5-step clades connected through interior nodes (clades 3-11 and 4-5) with few representatives (appendix C, fig. C.10b). The largest group of haplotypes was within clade 5-1 (appendix C, fig. C.10a), which was composed of four 4-step clades (4-1 to 4-4) that were connected to the second cluster of three 3-step clades (3-8, 3-9 and 3-11) through clades 3-5 and 3-10 (appendix C, fig. C.10a and fig. C.10b, respectively). A third group of haplotypes included four 3-step clades (3-13 to 3-16) and was connected through clade 3-12 (appendix C, fig. C.10c).

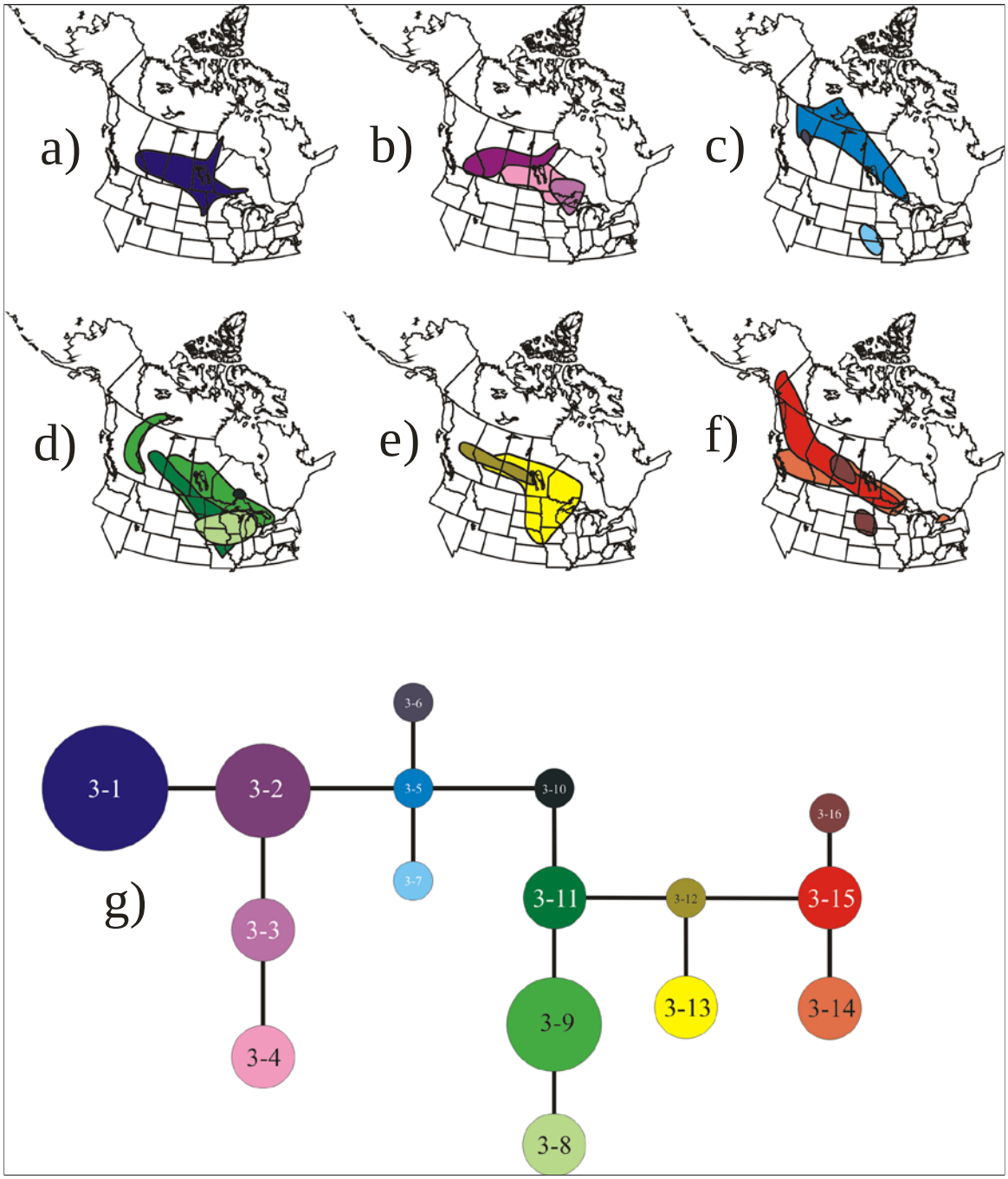
The species 2 network must be treated with some caution at the higher levels of the nested clade analysis. While many of the 4-step clades (4-1, 4-2, 4-4, 4-6 and 4-7) form well defined haplotype groups with clear spatial structure, they connect to each other in the network through a central core that is mainly inferred haplotypes with low sample size (appendix C, fig. C.10a and fig. C.10b). This area of the network represents missing data or the ancestral haplotypes has gone extinct.

3.4.2 Outgroup analysis of species 2

The general time reversible model of sequence evolution with a proportion of invariant sites (I=0.650) and a gamma distributed (α=1.44) rate variation was selected on the basis of the Akaike information criterion. The root of the network was determined by comparing the neighbor-joining and parsimony phylogenies using a species 1 haplotype (S1H209) as the outgroup. The neighbor-joining analysis placed the root along an edge between haplotypes S2H180 and S2H199, a connection that does not exist in the statistical parsimony network. The

root of the maximum parsimony phylogeny is between haplotypes S2191 and S2H199, which is within clade 3-2 in the haplotype network. As such, there were three potential 3-step clades in which the ancestral haplotype may reside (3-2, 3-3 or 3-5). As a result of the network structure, clades 3-2 and 3-5 were always interior clades and only the interior/tip status of clade 3-3 along with its nesting clade (4-4) were altered by the placement of the root. The nested clade analyses considering the affected clades will be repeated using clades 3-3 and 4-1 as interior as well as tip clades.

Figure 3.26: Map showing the distribution of 3-step clades within species 2
This figure includes maps of the following clade distributions: a) clade 3-1, b) clades 3-2, 3-3 and 3-4, c) clades 3-5, 3-6 and 3-7, d) clades 3-8, 3-9, 3-10 and 3-11, e) clades 3-12 and 3-13, and f) clades 3-14, 3-15 and 3-16. The species 2 cladogram is condensed to show only 3-step clades and the diameters of the nodes are relative to their sample sizes (bottom insert). The labels of each node refer to their 3-step clade in fig. 3.35 and their colors are associated with the corresponding distributions represented on the maps.



3.4.3 Nested Clade Analysis of COI haplotypes within species 2

The species 2 nested clade analysis (NCA) returned 34 significant geographical associations (table 3.17), which 20 indicated potential patterns of restricted gene flow with isolation by distance. The inference key also indicated patterns of contiguous range expansion for 10 of the clades. The remaining 6 associations included 2 that the NCA could not differentiate between isolation by distance and range expansion with long distance colonization, 3 patterns of allopatric fragmentation (2 of which that could also indicate long distance colonization) and one inconclusive outcome. The clades with a pattern of restricted gene flow with isolation by distance made up the large majority of significant geographical associations; six of the seven 1-step clades (1-40, 1-42, 1-48, 1-64, 1-75 and 1-88), three of the eight 2-step clades (2-1, 2-17, and 2-25), four of seven of the 3-step clades (3-1, 3-3, 3-4 and 3-8) and the remaining four significant associations among the 4- to 6-step clades (4-1,4-4, 4-7, 5-1 and 6-1). The clades indicating contiguous range expansions were well represented in the 2-step clades (2-2, 2-7, 2-16, 2-21 and 2-28) and throughout the higher level clades (3-15, 4-2, 5-2 and 5-3).

The NCA of the total cladogram indicated isolation by distance with long distance dispersal because the tip clade (5-3) had a significantly large nested clade distance (Dn=1216.86, *p=0.003*) and the clades with large nested clade distances (or the haplotypes therein) did not form two or more geographically concordant subsets. If they did, the inference key would favor allopatric fragmentation (Templeton, 2004).

Table 3.17: Nested clade analysis of species 2
Only the species 2 clades with significant geographical associations are included. The ‘Clades’ column is staggered and includes some clades with no inference to show the hierarchical structure. The inference key shows the decision-tree and the conclusions are denoted as follows: AF: Allopatric Fragmentation, CRE: Contiguous Range Expansion, IBD: Isolation by Distance, IO: inconclusive outcome, LDC: Long Distance Colonization, LDD: Long Distance Dispersal, and PRE: Past Range Expansion with subsequent extinction. See appendix B for full NCA table of Dc, Dn and p-values.

Clade	Inference Key	Conclusion
6-1	1-2-3-5-6-7-YES:	IBD with LDD
5-1	1-2-3-5-6-7-YES:	IBD with LDD
4-1	1-2-3-4-NO:	IBD
3-1	1-2-3-4-NO:	IBD
2-1	1-2-3-4-NO:	IBD
2-2	1-2-11-12-NO:	CRE
4-2	1-2-11-12-NO:	CRE
3-3	1-2-3-4-NO:	IBD
2-7	1-2-3-4-NO:	IBD
3-4	1-2-3-4-NO:	IBD
2-9	No Inference	No Inference
1-26	1-2-11-17-4-NO:	IBD
4-3	1-19-20-2-3-4-9-NO:	AF
5-2	1-2-11-12-NO:	CRE
4-4	1-2-11-12-NO:	CRE
3-8	1-2-3-4-NO:	IBD
2-15	No Inference	No Inference
1-40	1-2-3-4-NO:	IBD
2-16	1-2-11-12-NO:	CRE
1-42	1-2-3-4-NO:	IBD
3-9	1-2-3-5-6-13-YES:	LDD or AF
2-17	1-2-3-4-NO:	IBD
1-48	1-2-3-4-NO:	IBD
2-21	1-19-20-2-11-12-NO:	CRE
2-22	No Inference	No Inference
1-64	1-2-3-5-6-too few-7-8	IBD or IBD with LDD
3-11	1-2-11-12-NO:	CRE
2-25	1-2-3-4-NO:	IBD
5-3	1-2-11-12-NO:	CRE
4-6	No Inference	No Inference
3-13	No Inference	No Inference
2-28	1-2-3-5-15-NO:	PF and/or LDC
1-75	1-2-3-5-6-7-YES:	IBD with LDD
4-7	1-2-3-5-6-7-YES:	IBD with LDD
3-15	1-2-11-12-NO:	CRE
2-33	1-2-3-4-NO:	IBD
1-88	1-2-3-4-NO:	IBD
1-89	1-2-11-12-NO:	CRE
3-16	1-19-20-2-11-12-13-YES:	LDC or AF

3.4.4 Nested clade analysis of clade 5-1 within species 2

Clade 5-1 was collected from habits principally located in central Canada and the northern United States (fig. 3.27 and fig. 3.28). The maximum parsimony analysis placed the root along the edge joining clades 4-2 and 4-3. Phylogenetic analyses enforce bifurcating branching patterns and place roots on haplotypes assumed to be extinct. The statistical parsimony network in contrast, can allow for the persistence of ancestral haplotypes. As determined in the outgroup analysis (see section 3.4.2), the root was placed along the branch connecting clades 2-4 and 2-5 (appendix C, fig. C.10). The ancestral node could be one of several inferred haplotypes along this edge that spans the demarcating line between clades 4-1 and 4-2. Clade 4-2 was considered an interior clade if it contained the ancestral root. In this case, all the 4-step clades within clade 5-1 would be interior nodes and the inference key can make no determination. However, if clade 4-2 was a tip clade, as would be the case if clade 4-1 possessed the root, then the inference key would return a pattern of restricted gene flow with isolation by distance in conjunction with long distance dispersal. The interior clade 4-3 had significantly large clade and nested clade distances (Dc = 1308.35, $p<0.001$, Dn = 1287.04, $p<0.001$, respectively) and the tip clade 4-2 had a significantly small clade distance (Dc = 612.52, $p<0.001$). However, this inference must be treated with extreme caution given the ambiguity of the root.

The NCA of clade 4-1 indicated isolation by distance because a tip clade (3-1) had significantly small clade and nested clade distances (Dc =506.76, $p=0.007$, Dn = 508.43, $p=0.003$), along with the interior clade (3-2) having significantly large clade and nested clade distances (Dc = 760.41, $p=0.019$, Dn = 788.22, $p=0.004$). The majority of sample sites within the clade 3-2 distribution were located in central Canada near the geographic center of the nested

clade 4-1 (as determined by the NCA), which explains this clade's lower clade and nested clade distances. The clade 3-1 NCA indicated either isolation by distance or allopatric fragmentation because the tip clade (2-1) had significantly small clade and nested distances (Dc = 397.35, *p=0.003,* Dn = 411.95, *p=0.004,* respectively), the interior clade (2-2) had a large nested clade distance (Dn = 888.88, *p<0.001),* and the I-T (interior minus tip) nested clade distance was significantly large ($I\text{-}T_{(Dn)}$=476.92, *p<0.001*). This indicates that the interior clades' distributions were broad when compared against the distributions of the tip clades. If the I-T nested clade distance was significantly small, then the inference key would have favored long distance colonization over fragmentation. The distribution of clade 3-1 was disjunct, with populations in central Canada and the Great Lakes regions. The area between these two sections has been sampled and other species 2 clades were isolated, which may suggest that more sampling could result in additional clade 3-1 populations.

The NCA of clade 4-2 indicated pattern consistent with a contiguous range expansion, because the interior clade 3-3 had a significantly small clade distance (Dc = 385.30, *p=0.006*). The interior clade (3-3) was mainly distributed throughout the Great Lakes region, while the tip clade (3-4) occupied a range in central Canada (fig. 3.28). Nested within clade 4-2 were 3 significant geographical associations among the lower tier clades (3-3, 3-4 and 2-7). The NCA of clade 3-3 indicated restricted gene flow with isolation by distance because the interior clade 2-5 had a significantly large nested clade distance (Dn = 555.79, *p=0.012*), in addition to a tip clade (2-6) with significantly small clade and nested distances (Dc = 235.00, *p=0.016,* Dn=336.90, *p=0.002*, respectively). Similar patterns were suggested by the clade 2-7 NCA, because the tip clade (1-23) had significantly small clade distances ($Dc_{(1\text{-}23)}$ = 0, *p<0.010*). The inference key for clade 3-4 returned two possible outcomes: isolation by distance and allopatric fragmentation.

While the tip clade (2-9) had a significantly small clade distance ($Dc_{(2\text{-}9)}$ = 294.00, *p<0.001)*, the difference between interior and tip nested clade distances was significantly large ($I\text{-}T_{Dn}$ = 394.28, *p=0.001)*. As noted in clade 3-1 (see above), the comparison of interior and tip clades favors fragmentation over long distance colonization when the difference in average nested clade distances was large. The non-overlapping distributions of the 3-step clades (fig. 3.28) further support allopatric fragmentation, except that the intervening region was not devoid of other species 2 haplotypes. Interestingly, the location of the gap in the distribution of clade 3-4 was consistent with the one in clade 3-1.

The NCA of clade 4-3 indicated allopatric fragmentation. This was a result of the tip clade (3-7) having a significantly small clade distance (Dc = 219.71, *p<0.001*) in conjunction with a significantly large difference between the interior (3-5 and 3-6) and tip clade distances ($I\text{-}T_{(Dc)}$ = 10001.75, *p<0.001*). The distributions of the tip clades were limited and non-overlapping with the interior clades (fig. 3.27). Clades 3-5 and 3-6 were collected from extremely disparate habitats (British Columbia, Ontario and South Dakota), whereas clade 3-7 was detected in a much more restricted group of populations in the central USA (Kansas, Nebraska and South Dakota). Furthermore, this inference was at a high clade level and the non-overlapping clades were connected by a larger than average number steps (7 inferred haplotypes between clades 3-5 and 3-7). Templeton (2004) considers disjunct distributions at higher tiers to be more indicative of fragmentation, while these patterns at lower tiers were more suggestive of recent long distance colonization. Particularly, where the derived clades within each distinct region were locally restricted and possess distributions that do not overlap with those of the other nested clades.

Figure 3.27: Map showing the geographical distribution of clade 4-1 and 4-3 in species 2
The species 2 clades 4-1 and 4-3 possess geographic distributions represented on the map (left) along with their corresponding network structure as indicated on the insert (right). This map shows the distributions of clade 3-1 (blue + green + red), clade 3-2 (yellow), clade 3-5 (dark purple), clade 3-6 (light purple) and clade 3-7 (orange). The root of the network is denoted by (R) on the basis of the maximum parsimony analysis. The arrows highlight the network connections that stem from the root pointing from the ancestral character to the derived haplotypes.

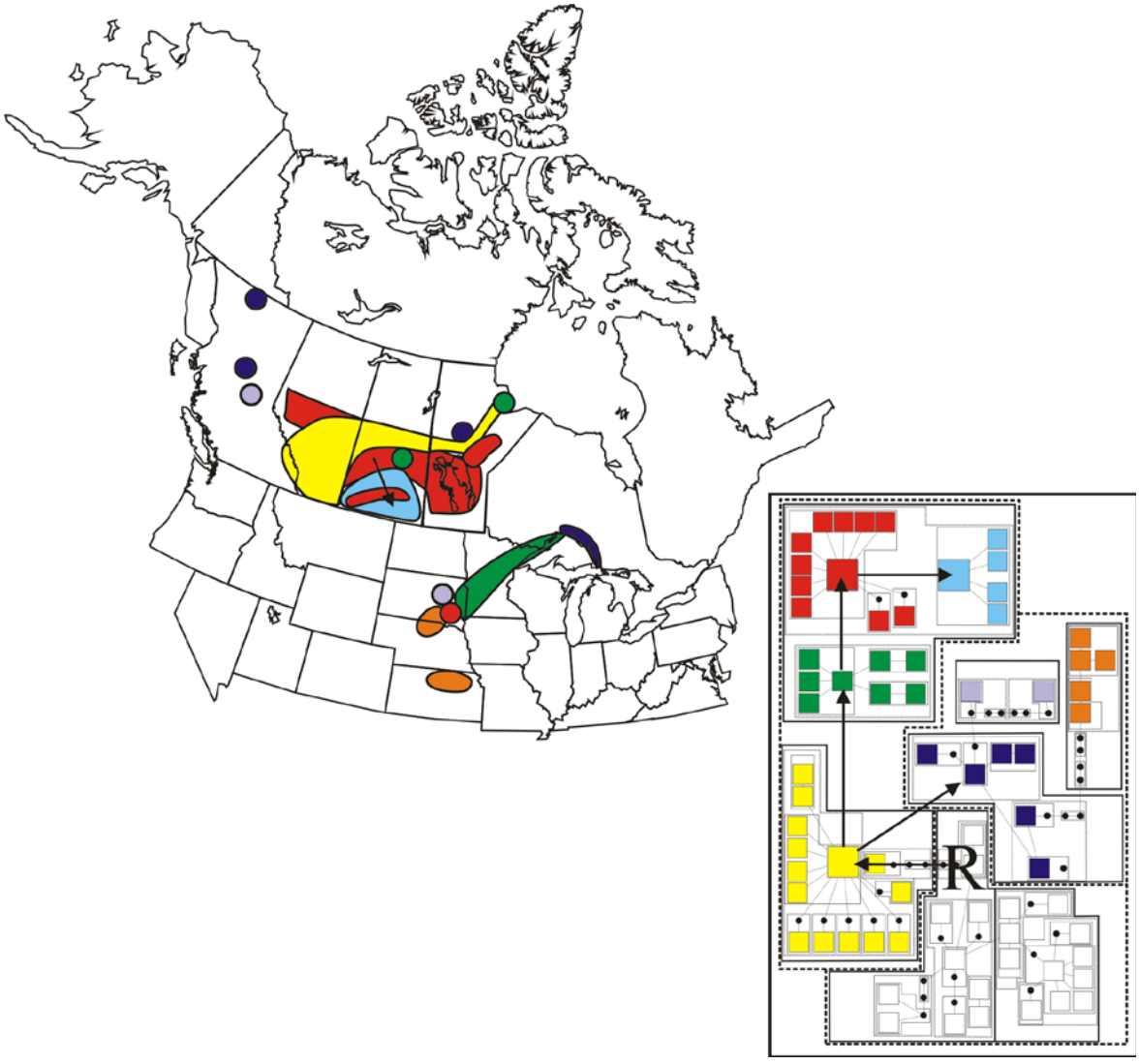
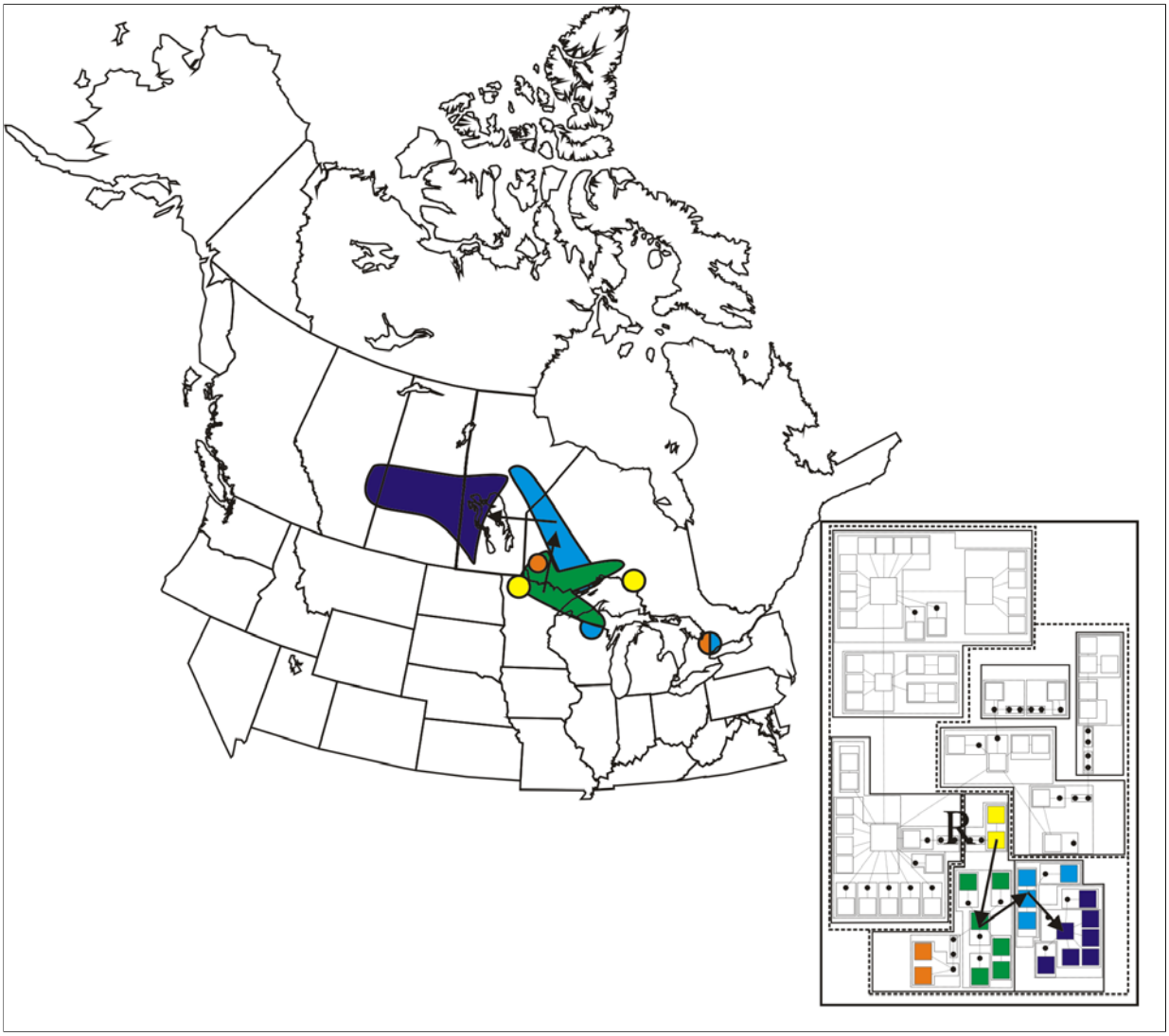


Figure 3.28: Map showing geographical distribution of clade 4-2 in species 2
Clade 4-2 possess a geographic distribution represented on the map (left) and its corresponding network structure as indicated on the insert (right). This map shows the distributions of clade 3-3 (green + orange + yellow) and clade 3-4 (blue + purple). The root of the network is denoted by (R) on the basis of the maximum parsimony analysis. The arrows highlight the network connections that stem from the root pointing from the ancestral to the derived haplotypes.



3.4.5 Nested clade analysis of clade 5-2 within species 2

The clade 5-2 NCA indicated a clear pattern consistent with contiguous range expansion. The interior clade 4-5 possessed significantly small clade and nested clade distances (Dc = 737.00, *p<0.001,* Dn = 832.87, *p<0.001,* respectively), in addition to the tip clade 4-4, which had significantly large clade and nested clade distances (Dc = 1119.23, *p=0.003,* Dn = 1120.55, *p=0.002,* respectively). Clade 4-5 was located within the Missouri River drainages and in central Canada (fig. 3.29) with a range that likely extended throughout North Dakota, although habitats within this state were not sampled. As indicated by the large clade and nested clade distances, the distribution of tip clade 4-4 was very broad, stretching from the Yukon Territory to the Great Lakes region (fig. 3.29). This was consistent with a post-Pleistocene range expansion of species 2 emanating from habitats along the Missouri River. An additional 11 clades with significant geographical associations (clades 1-40, 1-42, 1-48, 1-64, 2-16, 2-17, 2-21, 2-25, 3-8, 3-9 and 4-4) were nested within clade 5-2.

The inference key returned a pattern of restricted gene flow with isolation by distance for clade 4-4. The tip clade 3-8 had significantly small clade and nested clade distances (Dc = 468.50, *p<0.001,* Dn = 823.43, *p<0.001,* respectively) in conjunction with an interior clade (3-9) that had a significantly large nested clade distance (Dc = 1270.00, *p<0.001*). The difference between the clade and nested clade distances was also significantly large ($I\text{-}T_{Dn}$=-640.07, *p<0.001*), in addition to the interior and tip clades possessing overlapping distributions in the Missouri River basin. If the interior and tip clade distributions were separate, the inference key would have indicated a pattern consistent with allopatric fragmentation as a possible outcome.

Interestingly, the spatial range of clade 3-9 was split between two disparate regions: north-western Canada and the Missouri River drainage (fig. 3.29).

Two significant 3-step clades (3-8 and 3-9) were nested within clade 4-4. The NCA for clade 3-8 indicated isolation by distance because the tip clade (2-15) possessed significantly small clade and nested clade distances (Dc = 300.45, *p=0.006,* Dn = 328.87, *p<0.001*), while the interior clade (2-16) had significantly large clade and nested clade distances (Dc = 713.01, *p=0.001*, Dn = 714.72, *p<0.001*, respectively). The NCA of clade 3-9 detected two tip clades with significantly small clade distances ($Dc_{(2\text{-}20)}$ = 1270.00, *p<0.001*, $Dc_{(2\text{-}21)}$ = 415.00, *p<0.001*) and two tip clades 2-19 and 2-22 with significantly reversed clade and nested clade distances; the clade distances were significantly small ($Dc_{(2\text{-}19)}$ = 0, *p<0.001;* $Dc_{(2\text{-}22)}$ = 310.61, *p<0.001*), while the nested clade distances were significantly large ($Dn_{(2\text{-}19)}$ = 1689.33, *p=0.005;* $Dn_{(2\text{-}22)}$ = 1457.84, *p=0.006*). The distributions of these clades were at the extreme ends of clade 3-9's spatial range. Clade 2-19 was restricted to a single habitat in British Columbia (BC010) and clade 2-21 occupied localities within the Great Lakes region. The inference key favors long distance colonization if it is biologically realistic for the organism to disperse in this fashion and if the there is evidence of population expansion by other measures (i.e. mismatch distributions).

Nested within clade 3-8 were three lower tier clades (2-16, 1-42 and 1-40) with significant geographical associations. The NCA for clade 2-16 indicated a pattern consistent with contiguous range expansion because the interior clade 1-42 had a significantly small clade distance (Dc=217.60, *p=0.009*). This clade occupied habitats along the western side of the Missouri River in Nebraska and South Dakota, while the tip clade (1-43) was located at a single site in northern Ontario (ON021). The clade 1-42 NCA indicated isolation by distance as a result

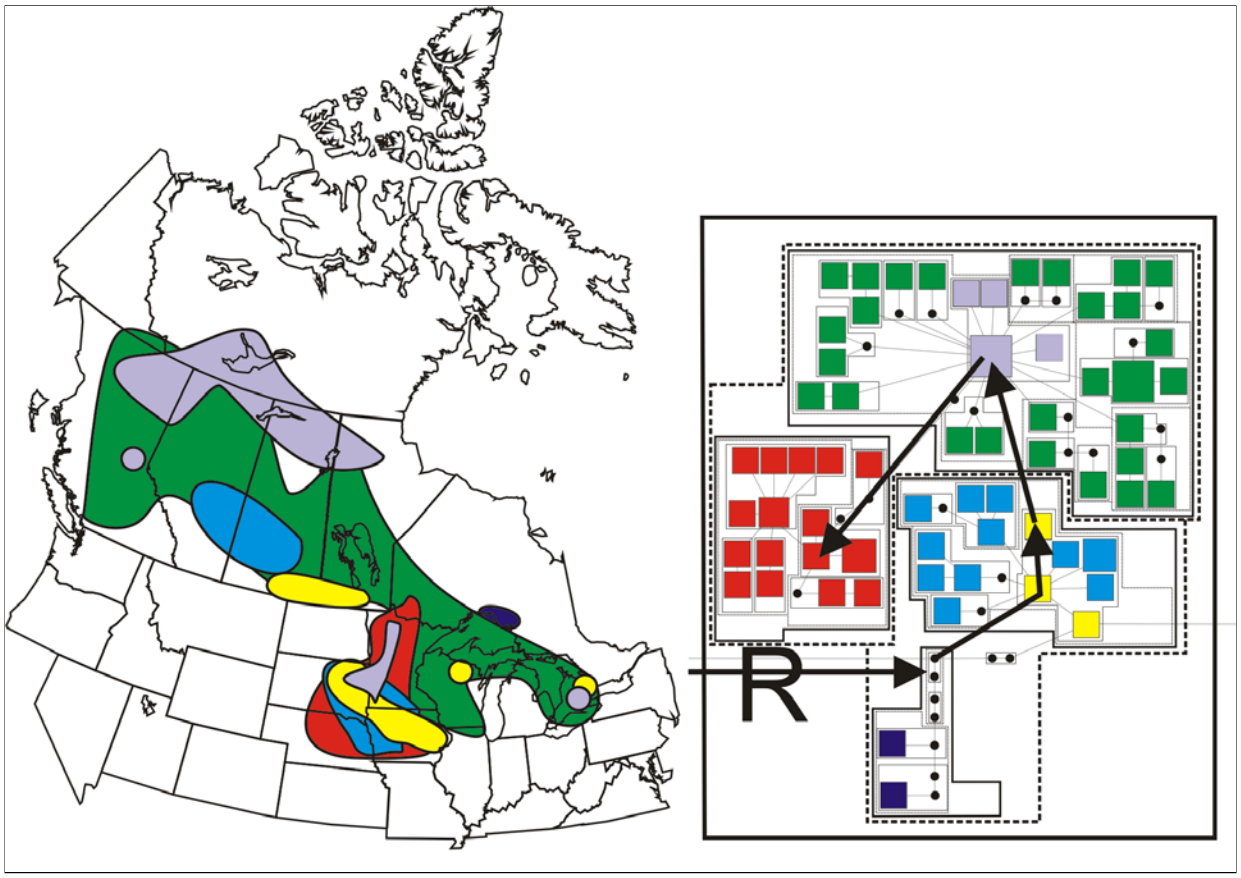
of the tip haplotype S2H038 having a significantly small clade distance (Dc=133.47, *p=0.013*). The third notable clade (1-40) had a central haplotype (S2H050) that was detected in 10 separate localities throughout the Missouri River system. The tip haplotype S2H045 had significantly small clade and nested clade distances (Dc=0, *p=0.003,* Dn=156.63, *p=0.001,* respectively), in addition to a significantly large difference between interior and tip clade distances ($I\text{-}T_{Dc}$=243.63, *p<0.001*). The NCA interprets broadly distributed interior haplotypes in conjunction with locally restricted tip haplotypes as a pattern of restricted gene flow with isolation by distance.

Clade 3-9 had 4 lower tier clades (2-17, 1-48, 2-21, and 1-64) that the NCA detected significant geographical associations. The inference key returned an outcome of isolation by distance for clade 2-17 because a tip clade (1-52) had significantly small clade and nested clade distances (Dc = 32.50, *p<0.001,* Dn = 168.89, *p=0.001*, respectively). This was not surprising given the vast range of the interior clade (1-48), which was detected in disparate regions; a group in the Missouri River basin and another in northern Canada. The NCA of clade 1-48 also indicated isolation by distance because of two tip haplotypes with significantly small clade distances ($Dc_{(S2H001)}$ = 0, *p=0.003*, $Dc_{(S2H016)}$ = 198.88, *p<0.001*). Given the absence of clade 1-48 haplotypes throughout central Canada, this may be the result of a long distance colonization event; however the sample size of S2H002 was insufficient to make that determination. The NCA for clade 2-21 indicated a contiguous range expansion because the interior clade (1-60) had a significantly small clade distance (Dc=166.48, *p=0.001*) and a more limited distribution than the tip clades. While this clade occupied habitats throughout Saskatchewan, the tip clades (1-61 and 1-62) were broadly distributed throughout Manitoba and western Ontario. Clade 1-64 had two interior haplotypes with significant geographical associations. Haplotype S2H025 had a

significantly large clade distance (Dc=404.07, *p<0.001*), in addition to haplotype S2H027 possessing significantly small clade and nested clade distances (Dc=14.34, *p<0.001,* Dn=182.76, *p<0.001*). The difference between interior and tip average clade distances was significantly large ($I\text{-}T_{Dc}$=116.01, *p<0.001*). As such, the inference key returned an outcome of restricted gene flow with some long distance dispersal for clade 1-64. The interior haplotype (S2H025) was detected in localities within both Wisconsin and Southeastern Ontario; however haplotypes S2H027 and S2H028 were only detected in the latter region.

The only significant geographical association within clade 4-5 was clade 2-25. The interior 1-step clades were distributed north of the Missouri River as well as within habitats in the Great Lakes region (fig. 3.29, the distribution marked in yellow). On the other hand, tip clade 1-71 was locally restricted to the region north of the Missouri River only. The NCA indicated isolation by distance because the tip clade 1-71 had significantly small clade and nested clade distances (Dc=136.38, *p=0.016*, Dn=290.60, *p=0.007*, respectively).

Figure 3.29: Map showing geographical distribution of clade 5-2 in species 2
Clade 5-2 haplotypes possess a geographic distribution represented on the map (left) and their corresponding network positions are indicated on the insert (right). The distribution of clade 1-70 is marked in yellow and clade 3-11 by yellow and blue. Clade 1-48 (distribution marked by light purple) was interior to clade 3-9 (light purple + green). The distribution of clade 3-8 is given in red. The root is denoted by (R). The arrows highlight the network connections that stem from the root pointing from the ancestral character to the derived haplotypes.



3.4.6 Nested clade analysis of clade 5-3 within species 2

The clade 5-2 NCA indicated a pattern consistent with contiguous range expansion. The interior clade (4-6) had significantly small clade and nested clade distances (Dc= 681.59, *p<0.001,* Dn= 817.50, *p=0.005,* respectively), in addition to the tip clade (4-7) having a significantly large clade distance (Dc= 1204.83, *p=0.006*). Clade 4-6 occupied habitats throughout central Canada, the western Great Lakes region and the Missouri River drainage. On the other hand, tip clade 4-7 possessed a range that included Alaska, Yukon, Northwest Territories in addition to central Canada and a limited representation in South Dakota (fig. 3.40).

Clade 5-3 had 8 lower tiered clades with significant geographical associations, 2 of which were within clade 4-6 (2-28 and 1-75) and the remainder nested in clade 4-7 (1-88, 1-89, 2-33, 3-15, 3-16 and 4-7). The NCA of clade 2-28 indicated a contiguous range expansion because the interior clade 1-75 had significantly small clade and nested clade distances (Dc= 0, *p<0.001,* Dn= 208.22, *p=0.005,* respectively*),* additionally the tip clade 1-77 possessed a significantly large nested clade distance (Dn= 967.56, *p=0.003*). The interior clade 1-76 occupied a region encompassing southern Manitoba, Kansas and Nebraska (fig. 3.30, the distribution marked green), alternatively the tip clade was located in habitats within the Missouri River drainage, Minnesota and central Canada (fig. 3.30). The remaining significant geographical association within clade 4-6 (clade 1-75) was nested within clade 2-28. The NCA returned a pattern of restricted gene flow with isolation by distance, with the potential for long distance dispersal. The tip haplotype S2H079 had a significantly small clade distance (Dc=38.24, *p=0.002*) and the tip haplotype S2H076 possessed significantly small clade and large nested clade distances (Dc= 0, *p<0.001,* Dn= 769.99, *p<0.001,* respectively). These haplotypes were geographically restricted

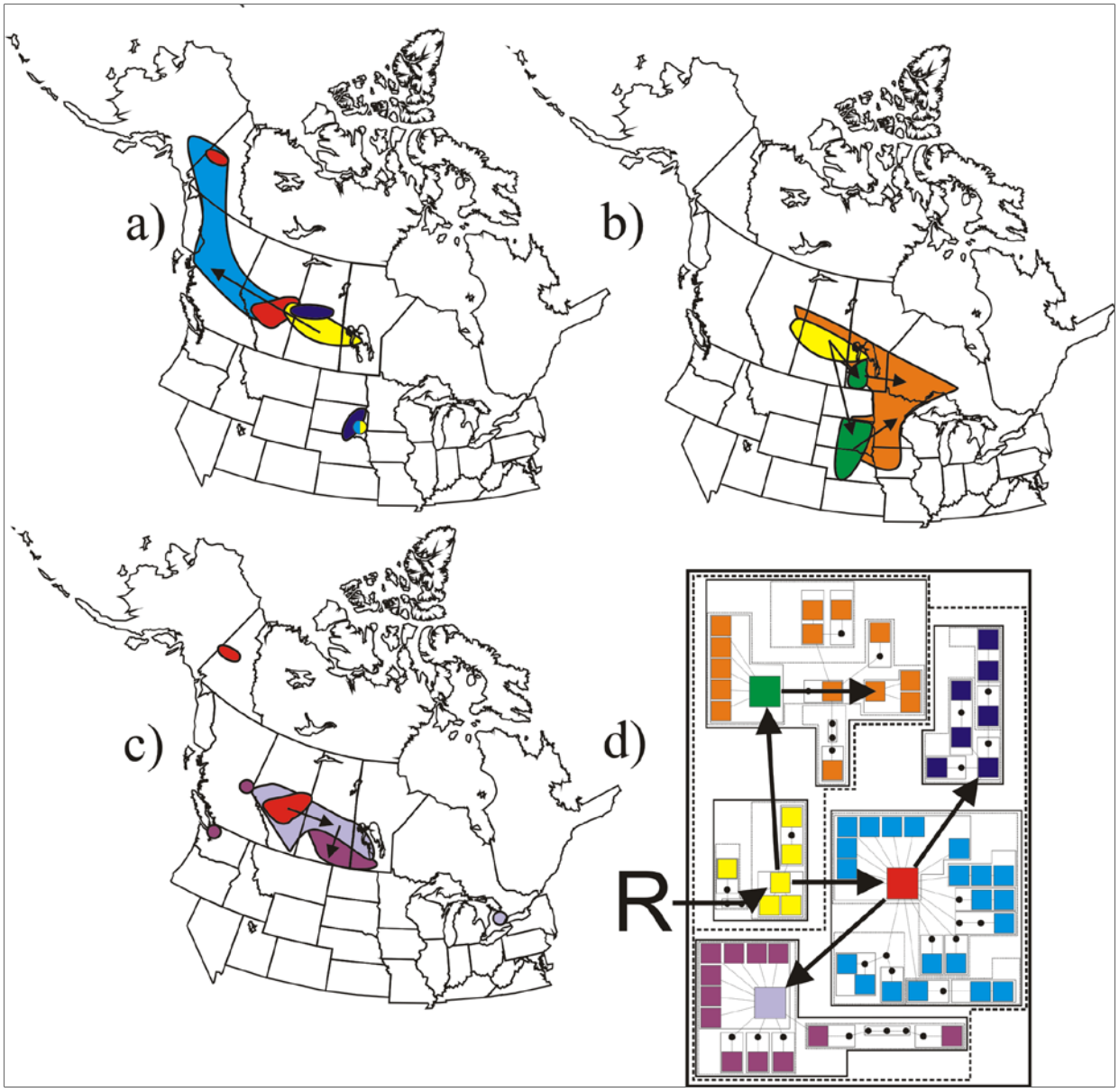
to their respective regions; haplotype S2H079 was detected in three habitats in South Dakota and haplotype S2H076 was isolated from a single lake in Ontario.

The clade 4-7 NCA indicated either isolation by distance or allopatric fragmentation. The interior clade 3-15 possessed a distribution throughout north-western Canada, stretching from Alberta to the State of Alaska (fig. 3.30a). The tip clades, 3-14 and 3-16, had more restricted ranges within central Canada and the Missouri River basin (fig. 3.30a and fig 3.30c). The tip clade 3-14 had significantly small clade and nested clade distances (Dc= 581.67, *p<0.001,* Dn= 618.36, *p<0.001*), in addition to tip clade 3-16 possessing a significantly small nested clade distance (Dn=553.78, *p=0.006*). In this case, the inference key favored allopatric fragmentation because and interior clade 3-15 had a significantly large nested clade distance (Dn=1460.96, *p<0.001*), and the distribution of the tip clades were predominately non-overlapping with the interior clades; with the exception of the populations possessing the interior haplotype S2H216, which were located within the region occupied by the tip clade 3-14 (fig 3.30c). This presents two scenarios: a) clade 4-7 dispersed throughout western Canada followed by fragmentation into eastern and western groups, or b) clade 4-7 initially colonized central Canada, then clade 3-15 colonized westward while the tip clades dispersed eastward, through short distance dispersal events resulting in a pattern of isolation by distance.

Nested within clade 4-7 were 5 lower tier clades with significant geographical associations (clade 3-15 and the hierarchically nested clades 1-88, 1-89 and 2-33, along with clade 3-16). The clade 3-15 NCA returned a pattern of contiguous range expansion because the interior clade 2-33 had significantly small clade and nested clade distances (Dc=866.20, *p=0.002,* Dn=1252.74, *p=0.002,* respectively). The interior clade distribution encompassed a

region that included Alaska, Yukon, British Columbia, Alberta and Saskatchewan; a spatial range that overlaps with that of the tip clades. The tip clade clade 2-34 was detected into habitats in Wisconsin and the Yukon Territory, which accounts for its high clade and nested clade distances. Clade 2-33 had five significantly small tip clade distances: 1-89 (Dc = 73.85, *p<0.001*), 1-90 (Dc = 0, *p<0.003*), 1-91 (Dc = 0, *p<0.001*), 1-92 (Dc = 0, *p<0.001*) and 1-93 (Dc = 0, *p=0.003*). The NCA of clade 1-88 indicated isolation by distance because the tip haplotype S2H116 possessed a significantly small clade distance (Dc = 0, *p<0.001*) and the difference between interior and tip clade distances was significantly large ($I\text{-}T_{Dc}$ = 2213.40, *p < 0.001*). On the other hand, the clade 1-89 NCA suggested a contiguous range expansion because the interior haplotype S2H097 had a significantly small clade distance (Dc = 43.69, *p=0.012*) and the difference between clade distances was significantly small ($I\text{-}T_{Dc}$ = 31.20, *p=0.012*). The clade 3-16 NCA indicated either long distance colonization or allopatric fragmentation because the interior clade 2-36 had significantly small clade and nested clade distances (Dc=133.13, *p=0.001*, Dn=532.03, *p<0.001*, respectively), along with the tip clade 2-37 having a significantly small clade and large nested clade distances (Dc=71.98, *p<0.001*, Dn=881.13, *p<0.001*, respectively). Additionally, the difference in nested clade distances was significantly small ($I\text{-}T_{Dn}$=-349.09, *p<0.001*). The distributions of the interior and tip clades were located in central Canada and South Dakota, respectively.

Figure 3.30: Map showing geographical distribution of clade 5-3 in species 2
Species 2 clades nested within clade 5-3 possess geographic distributions represented by the colored regions on the maps: a) Clade 3-12 (yellow), clade 3-15 (red + blue) and clade 3-16 (dark purple), b) clade 3-12 (yellow) and clade 3-13 (green + orange), c) haplotype S2H126 (red) and clade 3-14 (light purple + magenta), d) the colored haplotypes of the clade 5-3 network correspond to the distributions denoted on the maps. The root is denoted by (R). The arrows highlight the network connections that stem from the root pointing from the ancestral character to the derived haplotypes.



3.4.7 Mismatch distribution and populations genetics of species 2

The mismatch distribution (fig. 3.21) was bimodal and the raggedness statistic (table 3.1) rejected the null hypothesis of demographic expansion (r = 0.0017, *p<0.001*). The expected frequency curve was estimated from a coalescent simulation undergoing constant population growth with tau = 5.445.

The three tests for deviations from neutrality all returned significant results (table 3.1). Tajima's *D* was negative, denoting an excess of rare alleles (*D*=-1.790, *p=0.005*), Ramos-Onsins and Rozas R_2 statistic indicated an excess of singletons (R_2=0.0335, *p = 0.006*) and Fu's *F* indicated an excess of alleles (*F*=-262.58, *p<0.001*). These measures were consistent with a lineage undergoing demographic expansion.

3.4.8 Analysis of molecular variance of species 2

The 123 populations were partitioned on the basis of faunal provinces and arranged into 5 groups: Cascadia (9 populations), Great Lakes (13 populations), Hudson Bay (56 populations), Mississippi (23 populations) and Yukon-Mackenzie (21 populations). The between group component of the faunal provinces AMOVA (table 3.18) explained 13.80% (p<*0.001*) of the variance. The AMOVA partitioned into ice-free (23 populations) and glaciated (100 populations) regions had a between-group component that only explained 10.17 % (*p<0.001*) of the variance.

Figure 3.31: Mismatch distribution of species 2
The mismatch distribution of COI haplotypes plotted against an expected frequency curve of a population undergoing constant demographic expansion with $\tau = 5.445$.

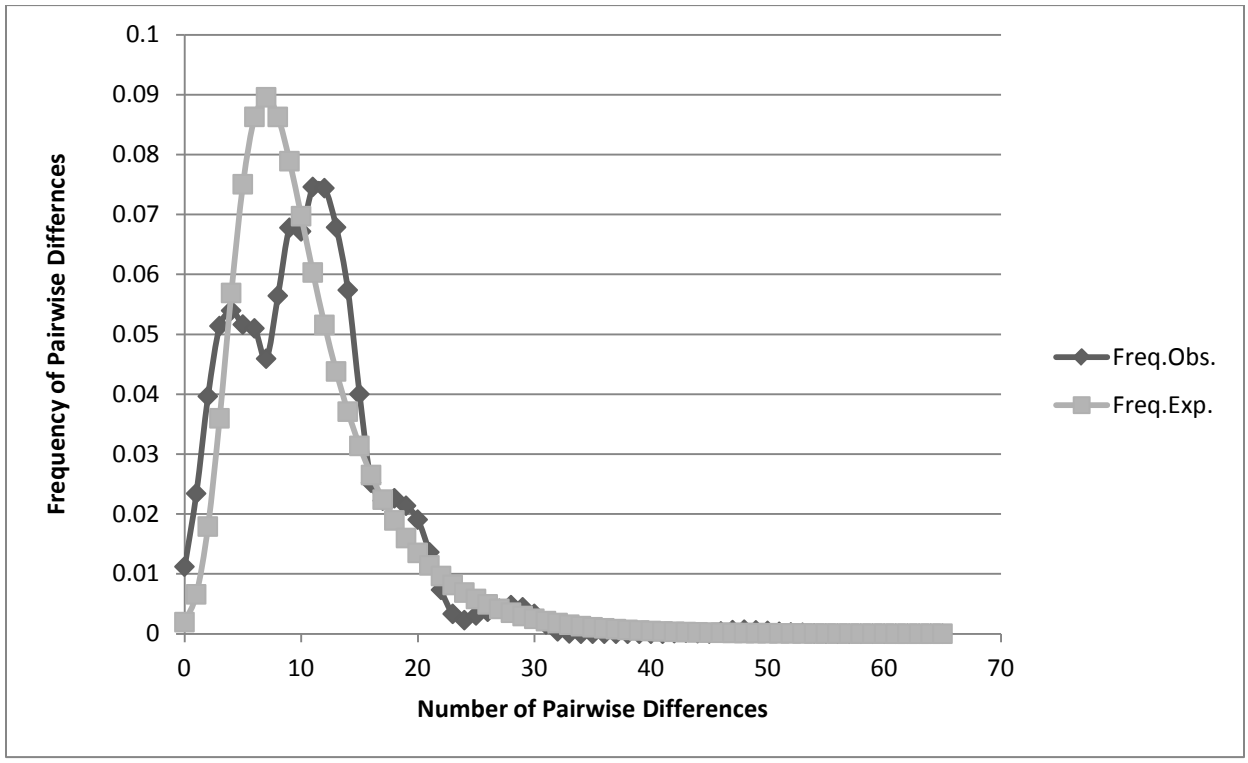


Table 3.18: AMOVA and SAMOVA of species 2
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * p<0.01, ** p<0.001, ---- Not Applicable

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	13.80**	32.83**	53.37**
Unglaciated vs. Glaciated	10.17**	37.60**	52.22**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	19.52**	34.44**	46.03**
3	21.60**	31.88**	46.52**
4	**34.58****	**28.19****	**37.23****
5	31.26**	29.59**	39.15**
6	21.23**	31.73**	47.04**

The 4-group SAMOVA for species 2 maximized the between-group variance component (V_{BG} = 34.58%, *p<0.001*) as compared to the other conditions (table 3.18). This analysis partitioned the populations into four groups with two assemblages having at least three populations (group 1:118 populations, group 2: 3 populations, group 3: 1 population, group 4: 1 population). The second group included clade 3-3 populations within southern. The two 1-population groups included isolated habitats from Iowa (IA009) and Yukon Territory (YT012). The remaining 118 populations were included in the largest partition.

3.5 RESULTS: SPECIES 7

3.5.1 Phylogeography of species 7

Species 7 haplotypes were collected from 69 populations broadly distributed throughout central and western Canada, but the majority were sampled from sites within Alberta and Saskatchewan. Two-hundred and fifty-five individuals were sequenced, resulting in 100 COI haplotypes. The nucleotide diversity (π=0.0109) was relatively shallow compared to the other large-bodied members of the *Hyalella 'azteca'* complex (table 3.1). Notably, haplotype S7H002 nearly accounted for one fifth of the sampled sequences and was collected from 21 distinct populations throughout species 7 range.

The statistical parsimony network was constructed (fig. 3.42) using one-hundred COI haplotypes, which were nested into forty-one 1-step clades (1-1 to 1-41), seventeen 2-step clades (2-1 to 2-17), six 3-step clades (3-1 to 3-6), three 4-step clades (4-1 to 4-3) and the total cladogram consisted of 5 steps. Clades 4-1 and 4-2 were separated from clade 4-3 by seven inferred haplotypes, but remained within the limit of the parsimony criterion.

3.5.2 Outgroup analysis of species 7

The TPM3uf model with a proportion of invariant sites (I=0.555) and a gamma distributed (α=0.966) rate variation was selected on the basis of the Akaike information criterion. Both the neighbor-joining and parsimony analyses placed the root along the edge between clades 4-2 and 4-3. Therefore, there was some ambiguity regarding the NCA of the total cladogram, depending if clade 4-3 was an interior or tip clade. A species 1 lineage G haplotype was used as the outgroup (S1H209).

3.5.3 Nested clade analysis of species 7

The nested clade analysis of species 7 returned 14 significant geographical associations (table 3.19). Patterns of isolation by distance were identified in six clades, contiguous range expansions in two clades and long distance colonization with respect to four clades. The remaining two significant geographical associations had inconclusive outcomes.

Table 3.19: Nested clade analysis of species 7 COI haplotypes
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IO: Inconclusive Outcome, IBD: Isolation by Distance, CRE: Contiguous Range Expansion, LDC: Long Distance Colonization, PF: Past Fragmentation, PRE: Past Range Expansion or RE: Range Expansion. See appendix B for Dn, Dc and p-values.

Clade	Inference Key	Conclusion
5-1	1-2-3-4-NO:	IBD
4-1	1-2-3-4-NO:	IBD
3-1	1-2-11-12-13-YES:	LDD
2-2	1-2-3-5-6-15-NO:	PF and/or LDC
3-2	No inference	No inference
2-3	1-2-3-5-6-15-NO:	LDC, PF or PRE
2-4	1-2-3-4-NO:	IBD
2-5	No inference	No inference
1-14	1-2-11-17-NO:	IO
2-8	1-2-11-17-NO:	IO
1-22	1-2-11-12-NO:	CRE
3-4	No inference	No inference
2-10	1-2-3-5-6-15-NO:	PF and/or LDC
1-25	1-2-3-4-NO:	IBD
2-11	1-2-11-12-NO:	CRE
1-30	1-2-3-4-NO:	IBD
4-2	1-2-3-4-NO:	IBD
3-5	1-2-3-5-15-16-18-NO:	PF, RE or IBD

The NCA for the total cladogram was conducted twice with clade 4-3 as a tip clade and again as an interior clade to account for the ambiguity of the root. The outcome was isolation by distance for the former and isolation by distance with long distance dispersal for the latter. Switching the tip/interior status of clade 4-3 had little effect on the Dc, Dn and p-values (only the inference key was affected), thus the values from the first analysis were included below. Clade 4-1 had significantly small clade and nested clade distances (Dc=534.50, $p<0.001$, Dn=623.10, $p<0.001$, respectively). Clade 4-2 possessed a significantly large nested clade distance (Dn=1268.17, $p=0.002$), whereas clade 4-3 had significantly large clade and nested clade distances (Dc=1298.47, $p=0.012$, Dn=1391.67, $p=0.002$, respectively).

The NCA for clade 4-1 indicated isolation by distance because the tip clade (3-1) had significantly small clade and nested clade distances (Dc=445.64, $p<0.001$, Dn=445.92, $p<0.001$, respectively) and the interior clade (3-2) possessed significantly large clade and nested clade distances (Dc=732.72, $p<0.001$, Dn=727.97, $p<0.001$, respectively). The distributions of these clades were predominately overlapping and occupy a region that encompasses most of central Canada (appendix C, fig. C.12). Disparate clade 4-1 populations were detected in British Columbia, Manitoba and Washington State, but the majority of sites were located in Saskatchewan and Alberta. Nested within clade 4-1 were four lower tier clades with significant geographical associations. Clade 2-2 was nested within clade 3-1 and the NCA for both indicated long distance colonization. Clade 2-2 had significantly small difference in nested clade distances ($I\text{-}T_{Dn}$ = -208.00, $p=0.001$), in conjunction with a tip clade (1-5) that possessed a significantly small clade and large nested clade distance (Dc=369.72, $p=0.007$, Dn=372.30, $p=0.008$, respectively). Clade 1-5 was detected in only a single habitat in northern Manitoba (MB005) and this pattern suggests either past fragmentation of a larger range or long distance colonization.

The inference key suggests the latter conclusion was more likely if it occurred at a lower tier and long distance dispersal was a realistic possibility for the organism. Similarly, clade 3-1 possessed an interior clade (2-2) with significantly small clade and nested clade distances (Dc=440.50, *p=0.019*, Dn=441.13, *p=0.022*, respectively), as well as tip clade (2-1) with a significantly large clade distance (Dn=1092.01, *p=0.006*). The interior clade 2-2 was distributed predominately in central Canada and the tip clade 2-1 was restricted to a single habitat in the Yukon (YT018). The remaining two clades (2-3 and 2-4) with significant geographical associations were nested within clade 3-2. The clade 2-3 NCA indicated past fragmentation or long distance colonization because the interior clades (1-9, 1-10) both had significantly small clade and large nested clade distances ($Dc_{(1-9)}$=0, *p=0.002*, $Dn_{(1-9)}$=1056.04, *p=0.018*, $Dc_{(1-10)}$=0, *p<0.001*, $Dn_{(1-10)}$=1056.04, *p=0.012*). Interior clades 1-8 and 1-9 were detected in the Liard River system in northern British Columbia, while the tip clade 1-10 occupied a region in southern Saskatchewan. The NCA of clade 2-4 indicated restricted gene flow with isolation by distance because the difference between interior and tip clade and nested clade distances was significantly large ($I\text{-}T_{Dc}$=704.85, *p=0.013*, $I\text{-}T_{Dn}$=493.02, *p=0.015*). The distribution of the interior clade 1-12 was located predominately in central Canada, particularly haplotype S7H039, which was detected in nine separate populations.

The NCA for clade 4-2 indicated either isolation by distance or allopatric fragmentation because the tip clade (3-4) had significantly small clade and nested clade distances (Dc=332.57, *p<0.001*, Dn=912.85, *p=0.019*, respectively), whereas the I-T (interior minus tip) clade and nested clade distances ($I\text{-}T_{Dc}$=517.42, *p<0.001*, $I\text{-}T_{Dn}$=268.12, *p=0.022*) were significantly large. The inference key favors allopatric fragmentation if the interior and tip clades have non-overlapping distributions and the intervening area was sampled, yet devoid of populations.

Interior clade 3-3 occupied a region in the Yukon and Alaska, while tip clade 3-4 was detected in habitats throughout central and western Canada (fig. 3.32). While British Columbia was sampled (and predominately lacking in clade 4-2 haplotypes), northern Alberta and Northwest Territories remain largely uncharacterized.

Nested within clade 4-2 were 6 clades with significant geographical associations (clades 1-22, 1-25, 1-30, 2-8, 2-10 and 2-11). The clade 2-8 NCA had a tip clade with significantly large nested clade distance, but that alone was insufficient to derive an inference. Nested within this clade was clade 1-22, which the inference key indicated a genetic pattern consistent with a contiguous range expansion, because the interior haplotype S7H059 had significantly small clade and nested clade distances (Dc=377.44, *p=0.002,* Dn=519.30, *p=0.005,* respectively). The range of this haplotype encompassed five habitats located in central Canada, while the tip haplotypes expanded the clade's range to include habitats within British Columbia (fig. 3.32).

There were four clades nested within clade 3-4 with significant geographical associations. The 1-step clades (1-25, 1-30) NCA returned a pattern of isolation by distance because of tip haplotypes with significantly small clade distances ($Dc_{(S7H060)}$=0, *p<0.001*, $Dc_{(S7H071)}$=0, *p<0.001,* respectively). The inference of clade 1-30 was further supported by significantly large I-T clade distance ($I\text{-}T_{Dc}$ = 142.42, *p<0.001*). The clade 2-10 NCA indicated either past fragmentation or long distance colonization because tip clade 1-26 had significantly small clade and large nested clade distances (Dc = 0, *p<0.*001, Dn = 447.61, *p = 0.010,* respectively). This tip clade was detected within a single population located in Squaunga Lake, YT, while the remaining populations of clade 2-10 were distributed throughout central and northern Yukon. The clade 2-11 NCA indicated a contiguous range expansion because the interior clade (1-29) possessed significantly small clade and nested clade distances (Dc=184.66, *p<0.001,*

Dn=203.38, *p=0.005*, respectively). While the interior clade was restricted to a region within Alaska, the tip clade range's encompasses localities in Yukon and northern British Columbia (fig. 3.32).

The only significant geographical association indicated within clade 4-3 was for clade 3-5. The inference key returned three possible outcomes: Past fragmentation of a larger range with subsequent extinction, past range expansion or long distance colonization. The interior haplotypes of clade 3-5 were located throughout western Canada with the tip clades being distributed throughout eastern Canada (fig. 3.33). The tip clade 2-14 was the only group located in northern Quebec and possessed a significantly small clade and large nested clade distance (Dc=162.53, *p<0.001*, Dn=2080.03, *p=0.004*) and the difference among interior and tip clade distances was significantly large ($I\text{-}T_{Dc}$ = 918.13, *p<0.001*). This clade may have dispersed into this region through the Lake Bonneville system with subsequent extinction of the intermediate areas, or could be the product of a long distance dispersal event via avian vectors.

Figure 3.32: Map showing geographical distribution of species 7 clades 4-1 and 4-2
Clades 4-1 and 4-2 haplotypes possess geographic distributions represented on the map (top) and their corresponding network positions are indicated on the insert (bottom). The regions occupied by each 3-step clade within clade 4-1 are marked on the map as follows: 3-1 (green) and 3-2 (light purple). The 3-step clades within clade 4-2 are marked on the map as follows: 3-3 (orange + blue + magenta) and 3-4 (dark purple). The root is denoted by (R).

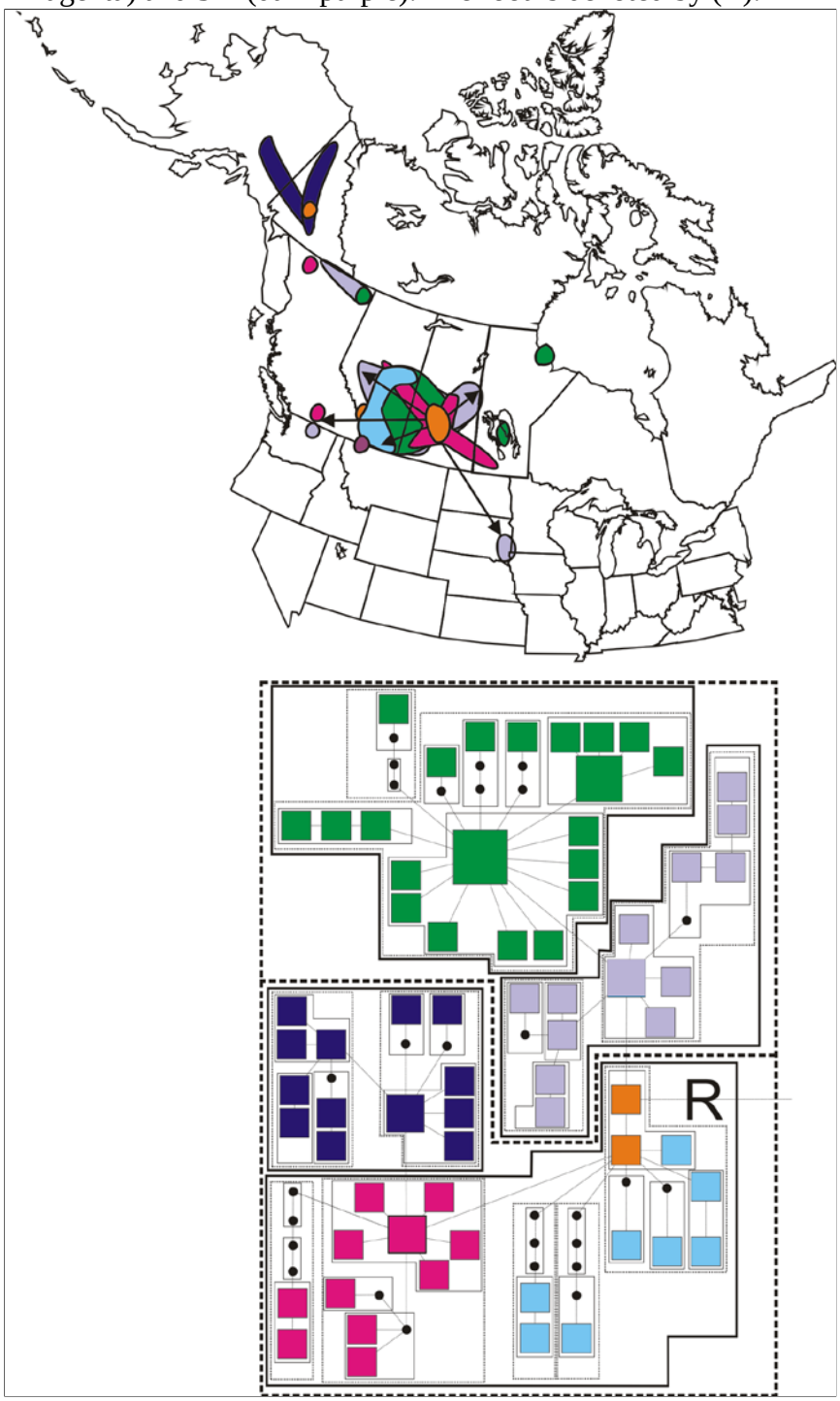
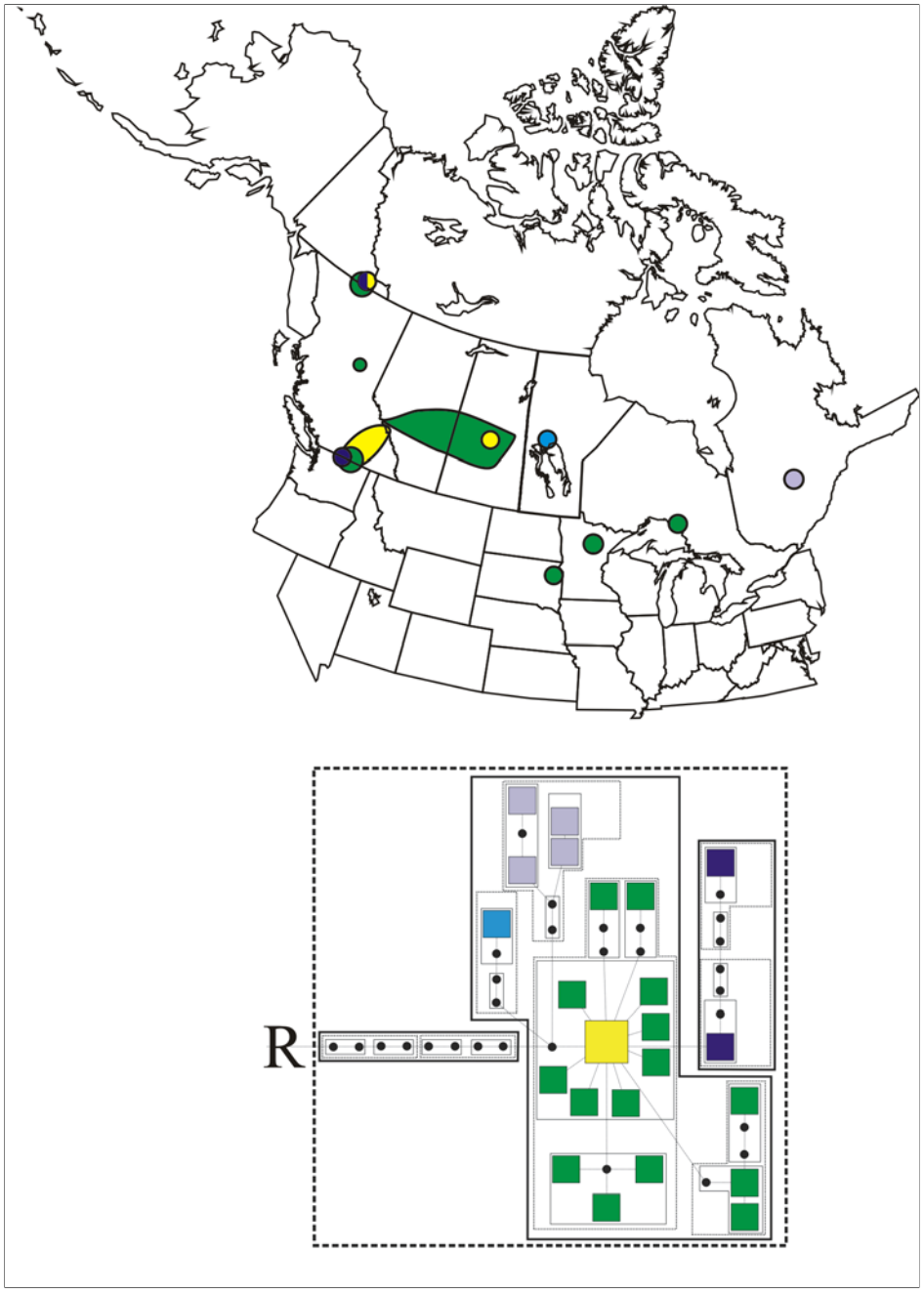


Figure 3.33: Map showing geographical distribution of clade 4-3 in species 7
Clade 4-3 haplotypes possess a geographic distribution represented on the map (top) and their corresponding network positions are indicated on the insert (bottom). The regions occupied by each clade are marked on the map as follows: 2-12/2-15 (green + yellow), 2-13 (blue), 2-14 (light purple) and 3-6 (purple). The root is denoted by (R).

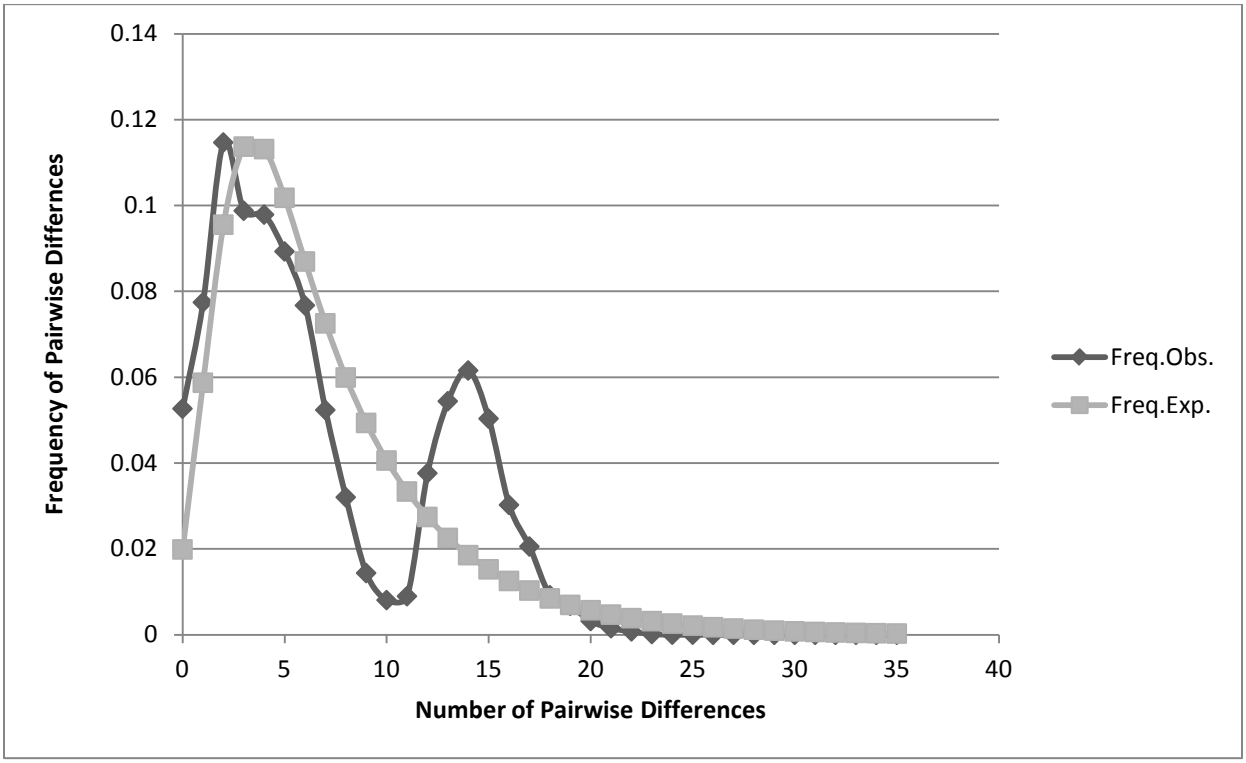


3.5.4 Mismatch distribution and population genetics of species 7

The mismatch distribution was bimodal (fig. 3.34) and the raggedness statistic (table 3.1) rejected the null hypothesis of a demographic expansion (r=0.058, *p=0.002*). The expected frequency curve was estimated from a simulation undergoing constant population expansion with a tau = 2.234.

Tajima's D (T_D=-1.790, *p=0.01*), Ramos-Onsins and Rozas R_2 statistic (R_2=0.0343, *p=0.019*) and Fu's F (Fu_F=-63.07, *p=0.000*) all indicated significant deviations from neutrality (table 3.1). This was consistent with a lineage undergoing demographic expansion.

Figure 3.34: Mismatch distribution of species 7 COI haplotypes
The mismatch distribution of COI haplotypes plotted against an expected frequency curve of a population undergoing constant demographic expansion with $\tau = 2.234$.



3.5.5 Analysis of molecular variance of species 7

The sixty-nine populations were partitioned into six faunal provinces (Cascadia: 5 populations, Great Lakes: 2 populations, Hudson Bay: 34 populations, Mississippi: 4 populations, West Coast: 1 population, Yukon-Mackenzie: 23 populations). The between group component of the AMOVA partitioned into faunal provinces (table 3.20) explained 25.06% (*p<0.001*) of the variance, while both among population and within population variance components were significant ($V_{(AP)}$=43.64, *p<0.001*, $V_{(WP)}$=31.29, *p<0.001*, respectively). The AMOVA partitioned by unglaciated (5 populations) and glaciated (64 populations) regions had a between-group component that only explained 0.23 % of the variance and was not significant.

Table 3.20: AMOVA and SAMOVA of species 7
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * p<0.01, ** p<0.001, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	25.06**	43.64**	31.29**
Unglaciated vs. Glaciated	0.23^{NS}	64.80**	34.97**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	28.14**	25.72**	16.14**
3	**56.11****	**26.19****	**17.70****
4	57.80**	24.84**	17.37**
5	53.80**	27.34**	18.86**
6	61.25**	20.10**	18.64**

The 3-group SAMOVA had a between-group component that explained 56.11% of variance (table 3.20). Neither the 2-group SAMOVA nor the higher level SAMOVA appreciably increased in the between-group variance component. The 3-group SAMOVA resulted in two major partitions of 61 and 7 populations, respectively. The groups were split into east and western groups; the distribution of the second group included habitats located in Manitoba, Quebec and Wisconsin.

3.6 RESULTS: SPECIES 8

3.6.1 Phylogeography of species 8

Species 8 *Hyalella* were isolated from 27 habitats in Florida, in addition to a few disparate sites in Nebraska, Kansas, Tennessee and Mississippi. One hundred and two individuals were sequenced, resulting in 81 COI haplotypes with a nucleotide diversity of π=0.0109 (table 3.1). No other species of *Hyalella 'azteca'* was detected within the Florida Peninsula suggesting this region was lacking much of the species diversity detected in other faunal provinces. Furthermore, no other *Hyalella* species were detected in habitats with species 8, except for sample sites in Tennessee (which included the small morphotype species 5).

The statistical parsimony network was constructed (appendix C, fig. C.11) from 81 COI haplotypes that were nested into seventy-nine 1-step clades (1-1 to 1-79), forty-two 2-step clades (2-1 to 2-42), nineteen 3-step clades (3-1 to 3-19), six 4-step clades (4-1 to 4-6) and the total cladogram consisted of 5 steps.

3.6.2 Outgroup analysis of species 8

The HKY model with a proportion of invariant sites (I=0.501) and a gamma distributed (α=1.096) rate variation was selected on the basis of the Akaike information criterion. The neighbor-joining analysis and parsimony trees were constructed with species 1 haplotype S1H029 as the outgroup. The neighbor-joining and parsimony analyses placed the root within clade 2-11 along the connection to haplotype S8H042 or within clade 2-7, respectively (appendix C, fig. C.11).

3.6.3 Nested clade analysis of species 8

The nested clade analysis of species revealed 9 significant geographical associations, the majority of which were within the 4-step tier (table 3.21).

Table 3.21: Nested clade analysis of species 8 COI haplotypes
Only the clades with significant geographical associations are included. The inference key shows the decision-tree along with the conclusion. IO: Inconclusive Outcome, IBD: Isolation by Distance, AF: Allopatric Fragmentation, LDD: Long Distance Dispersal or PF: Past Fragmentation of a Larger Range. See appendix for Dn, Dc and p-values.

Clade	Inference Key	Conclusion
5-1	1-2-3-5-6-7-YES:	IBD with LDD
4-1	1-2-3-4-NO:	IBD
3-1	No inference	No inference
2-5	1-2-3-4-NO:	IBD
4-3	1-2-3-4-NO:	IBD
3-11	1-2-3-4-NO:	IBD
4-4	1-2-3-4-NO:	IBD
4-5	1-19-20-2-11-17-NO:	IO
4-6	1-2-3-5-6-7-YES:	IBD with LDD

The NCA of the total cladogram indicated isolation by distance with long distance dispersal because the tip clade 4-6 had a significantly large nested clade distance (Dn=1195.26, *p<0.001*) and was partially co-distributed with clade 4-1 (fig. 3.35 and fig. 3.36). If clade 4-6 possessed a geographically range distinct from the other nested clades, the inference key would have inferred long distance colonization. The NCA of clades 4-1, 4-3 and 4-4 indicated isolation by distance because of tip clades with significantly small clade distances ($Dc_{(3-6)}$=54.95, *p<0.001*, $Dc_{(3-10)}$=49.01, *p=0.012* and $Dc_{(3-12)}$=71.89, *p<0.007*, respectively). The significant geographical associations noted within clade 4-5 were insufficient to derive an inference. The

NCA of clade 4-6 returned an outcome of isolation by distance with long distance diseprsal because the I-T (interior minus tip) nested clade distance was significantly large ($I\text{-}T_{Dn}$=646.66, *p<0.*001) in addition to the tip clade (3-18) having significantly small clade and nested clade distances (Dc=94.30, *p<0.001*, Dn=368.66, *p<0.001*, respectively). The interior clade 3-17 was restricted to a single locality in the Apalachicola River system, whereas the tip clades 3-18 and 3-19 were located in Kansas, Mississippi and Nebraska (fig. 3.36). The inference key returned an inconclusive outcome for clade 3-4 along with isolation by distance for clades 2-5 and 3-11.

3.6.4 Mismatch distribution and population genetics of species 8

The mismatch distribution was unimodal (fig. 3.37) and the raggedness statistic (table 3.1) did not reject the null hypothesis of a demographic expansion (r=0.028, *p=0.062*). The expected frequency curve was estimated from a simulation undergoing constant population expansion with a tau = 6.752.

Figure 3.35: Map showing geographical distribution of clades 4-1, 4-2 and 4-3 of species 8
Clades 4-1, 4-2 and 4-3 haplotypes possess a geographic distribution represented on the map (top) and their corresponding network positions are indicated on the insert (bottom). The regions occupied by each clade are marked on the map as follows: Clade 3-1 (green), 3-2 (orange), 3-3 (light purple), 3-4 (yellow), 3-5 (blue), 3-6 (red) and 4-3 (light blue). The roots are denoted by R-NJ and R-MP for the neighbor-joining and maximum parsimony analyses, respectively.

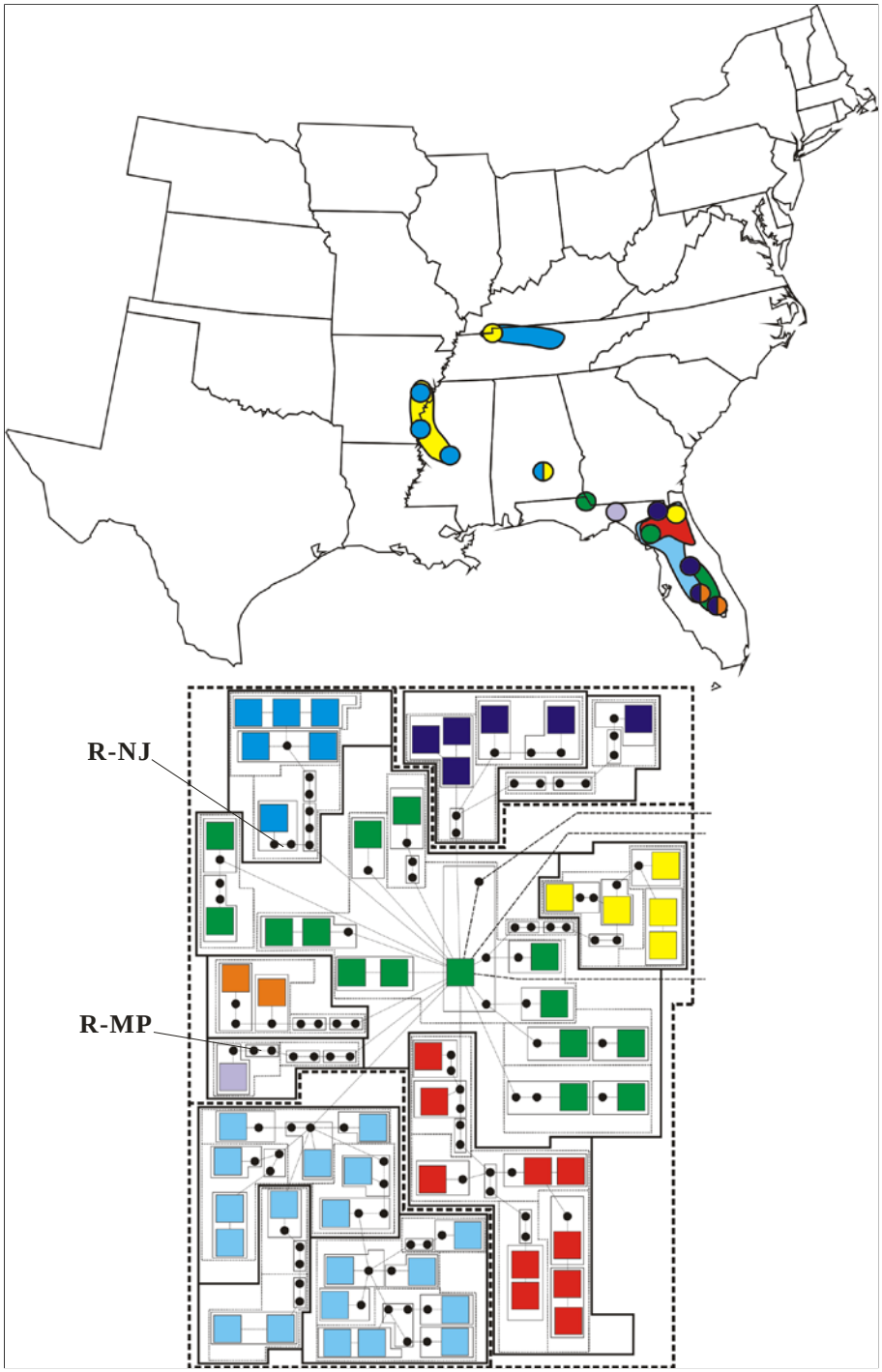


Figure 3.36: Map showing geographical distribution of species 8 clades 4-4, 4-5 and 4-6
Clades 4-4, 4-5 and 4-6 haplotypes possess a geographic distribution represented on the map (top) and their corresponding network positions are indicated on the insert (bottom). The regions occupied by each clade are marked on the map as follows: Clade 4-4 (blue), clade 4-5 (red) and clade 4-6 (yellow + red).

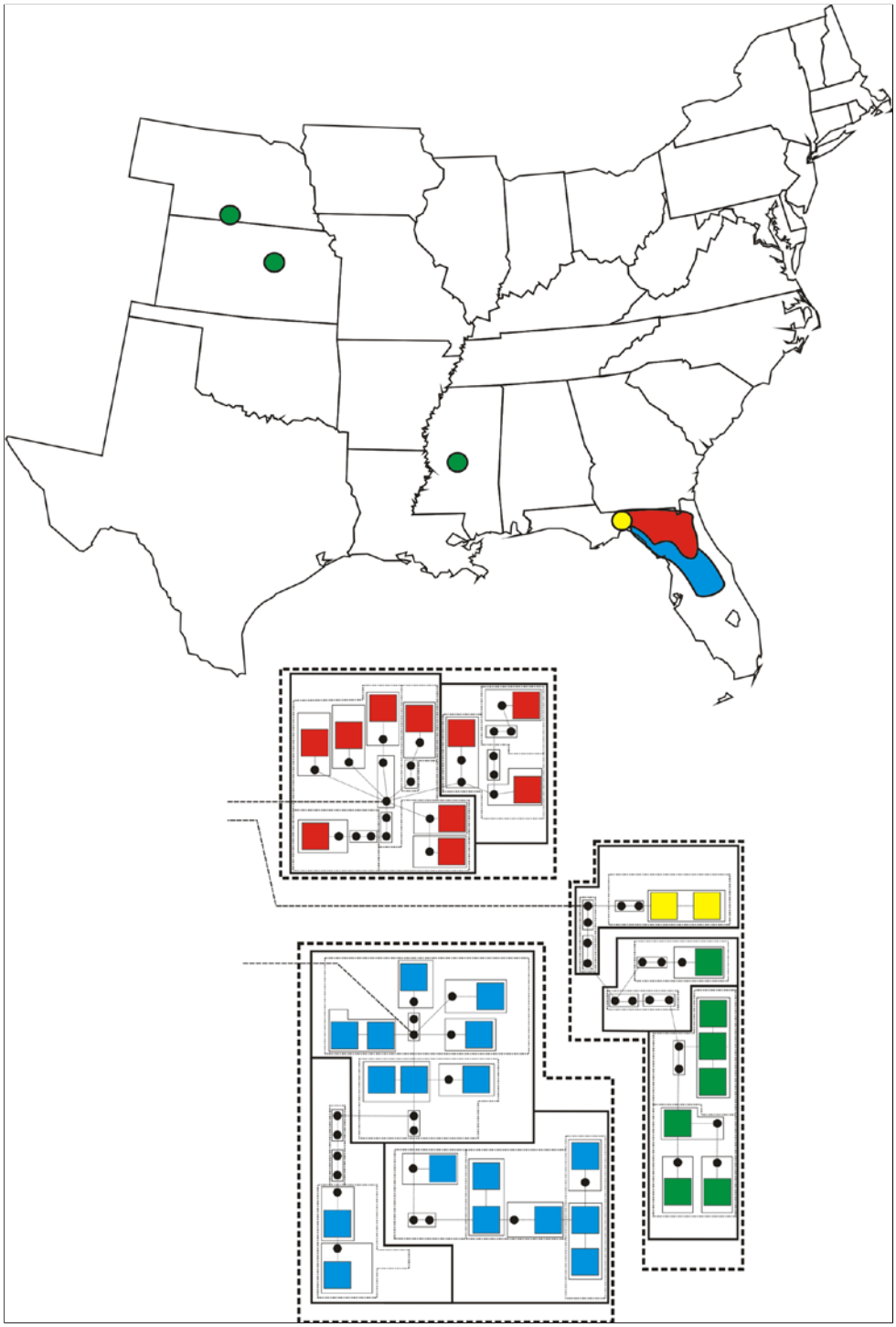
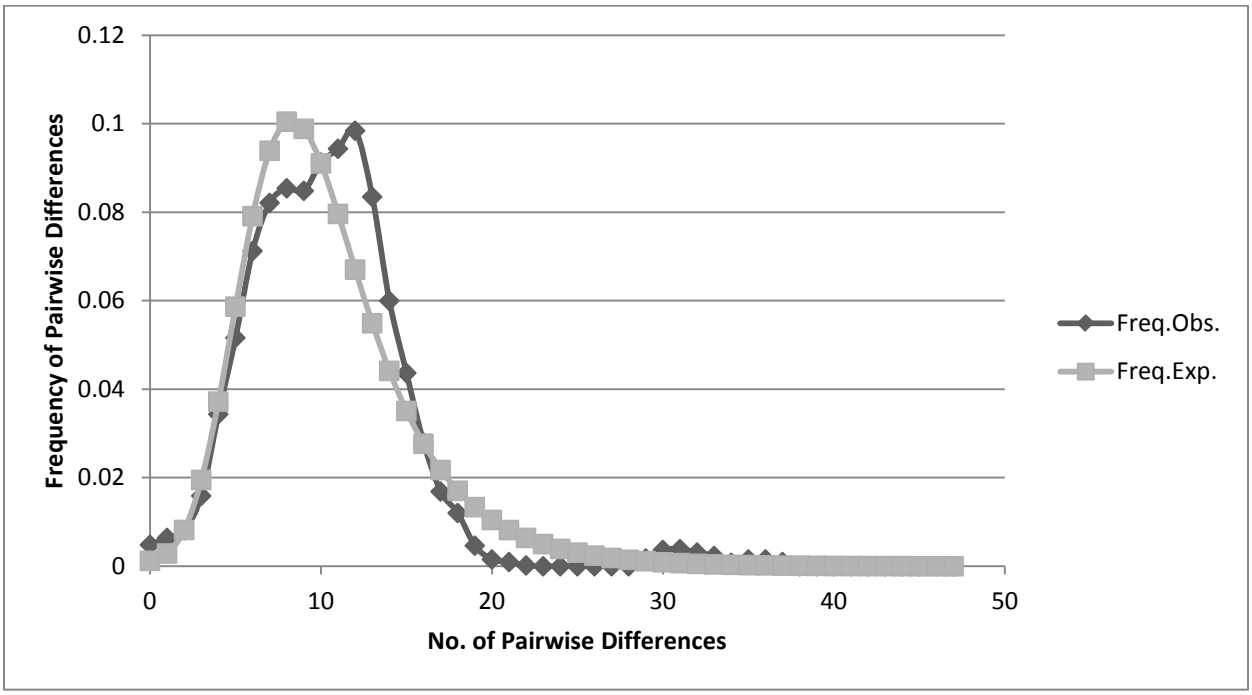


Figure 3.37: Mismatch distribution of species 8 COI haplotypes
The mismatch distribution of COI haplotypes plotted against an expected frequency curve of a population undergoing constant demographic expansion with $\tau = 6.752$.



3.6.4 Analysis of molecular variance of species 8

The twenty-seven populations were partitioned into the faunal provinces (Mississippi: 7 populations and Southeastern: 20 populations). The between group component of the AMOVA partitioned into faunal provinces (table 3.22) explained 19.86% ($p<0.001$) of the variance. No populations possessing clade 8 individuals were detected north of the glacial limit.

Table 3.22: AMOVA and SAMOVA of species 8
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * p<0.01, ** p<0.001, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	19.86**	31.51**	48.63**
Unglaciated vs. Glaciated	-----	-----	----

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	46.09*	18.44**	35.47**
3	44.91**	18.79**	36.31**
4	41.06**	20.15**	40.27**
5	39.37**	20.36**	41.44**
6	36.13**	15.03**	48.84**

All levels of the SAMOVA indicated significant between group variance components (table 3.22); however none of the partitions consisted of two or more groups with at least three populations. Regardless of the number of groups, the populations located in the Florida peninsula consistently formed the largest assemblage.

3.7 DISCUSSION

The large-bodied species of the *Hyalella* '*azteca*' complex represents a broadly distributed group of taxa with present-day distributions that have been shaped by the glacial cycles of the Pleistocene. The patterns of mtDNA diversity within these species were affected by spatial fragmentation of lineages between and likely within glacial refugium along with the effects of range expansions into glaciated regions. While all large-bodied groups (except species 8) were located north of the glacial limit, the spatial structure of these species was highly varied; some species (2 and 7) and lineages (1A, 1D and 1E) possessed large distributions.

Genetic diversity within a range expansion is now receiving more attention in the literature (Excoffier *et al.,* 2009). Initial models were constructed on the basis of short distance dispersal of migrants into neighboring demes resulting in patterns of isolation by distance (Slatkin & Excoffier, 2012). However, this did not fully explain genetic patterns detected in growing populations (Hallatschek & Nelson, 2010). The process of allele surfing is analogous to genetic drift as neutral or slightly deleterious haplotypes can go to fixation in populations along an expansion front. This can exacerbate the repeated founder effects and isolation by distance associated with the serial colonization of neighboring demes. Genetic patterns of populations undergoing range expansions are also affected by long distance dispersal (Pierce *et al.,* 2014). If long distance colonization events are rare they can result in an 'embolism', whereby haplotypes jump ahead of the expansion wave (Excoffier *et al.,* 2009). This can form a region of locally adapted populations that resist colonization of the surfing allele. This pattern is predicted to be more likely when dispersal routes are limited along a corridor (Bialozyt *et al.,* 2006). More recent simulations have indicated that if these events are common, then haplotypes from the

source populations can mitigate the effects of repeated colonization events or allele surfing (Ray & Excoffier, 2010). The genetic consequences of post-Pleistocene range expansions in white-winged doves were countered by reoccurring multi-directional gene flow (Pruett *et al.,* 2011).

The large-bodied *Hyalella 'azteca'* species expanded their ranges into glaciated North America after the retreat of the Pleistocene glaciers. The genetic patterns within these growing populations of species 1 (lineages A, C, D, E, F + G), 2 and 7 are consistent with a myriad of dispersal patterns; short distance colonization (isolation by distance), allele surfing and long distance colonization (table 3.20). Some lineages of species 1 (B and H) were largely restricted to unglaciated regions. Species 8 was the only large-bodied group that was wholly restricted south of the glacial limit.

3.7.1 Spatial structure among large-bodied *Hyalella* species

As is commonly seen among fishes, species assemblages can indicate the boundaries of faunal provinces (Avise, 2000). These freshwater ecoregions in Canada are often defined by current and historical connectedness of drainages and the species' refugial origins (Taylor, 2010). Spatial congruence among multiple fish species is often taken as evidence of common dispersal routes (Bossu *et al.,* 2013). This does not appear to be the case among the large-bodied *Hyalella* species. The distributions among lineages within species 1 and the major clades of species 2 and 7 are not constrained to a consistent faunal province and their distribution patterns are complex.

The AMOVA/SAMOVA of species 1 lineage B, lineage I and species 7 possessed the lowest within-population variance components (11.02, 15.71 and 17.70, respectively) along with

high *Fst* scores (0.890, 0.823 and 0.823, respectively). These reflect lineages that were affected by population structure to a greater extent. In the species 1 lineages, these groups are largely restricted to unglaciated regions and suggests sorting of haplotype variation among older populations. On the other hand, species 7 is distributed throughout glaciated North America. This species delineates into two distinct groups that are separated by 8 mutation steps (clades 4-1/4-2 and clade 4-3) that have undergone demographic expansions (as indicated by the bimodal mismatch distribution). The spatial distributions indicate that the two groups occupy overlapping regions, but there appears to be sorting of haplotype diversity at the population scale. This may suggest that whichever haplotype was among the initial colonizers became the predominant form within that population resisting subsequent migrations from the other clades. These invading individuals may persist at lower frequencies, resulting in higher F_{ST} and lower within population variance component values.

Species 1 lineage A, F, G, H and species 2 had high within-population variance components ranging from 29.08 (species 1 lineage H) to 49.25 (species 1 lineage F), along with Fst scores between 0.507 (species 1 lineage F) and 0.710 (species 1 lineage H).

Five of the species 1 lineages (A, B, E, G and I) had significant between group variance components when partitioning the variation among faunal provinces, but in all cases these values were much lower than the groups identified in the SAMOVA. Faunal provinces are determined on the basis of fish species endemism and are generally consistent with the major drainages in North America. The freshwater cladoceran *Sida crystallina* possesses a similar distribution in glaciated regions to *Hyalella 'azteca'* with patterns being shaped by several refugial lineages (Cox & Hebert, 2001). The between-group COI divergences (4 – 7%) were comparable to the

distances among lineages within *Hyalella* species 1 (2 – 8%). *Sida crystallina* is capable of passive egg dispersal via waterfowl allowing it to cross watersheds. Where *Sida crystallina* differs from *Hyalella* is that the spatial patterns of its phylogroups are coincident with the boundaries of the faunal provinces and there is little admixture of refugial lineages (Cox & Hebert, 2001). As such, the *Hyalella* species within glaciated regions are less restricted to interconnected waterways while dispersing from refugial habitats. Given the broad distribution of some of the *Sida* lineages, both species seem to have been highly successful at exploiting the proglacial lake system during their recolonization of glaciated regions. However, given the lack of geographic patterns coincident with faunal provinces and the prevalence of admixture within suture zones of the Great Lakes and the Midwest, *Hyalella* seems to possess a greater capacity to move among watersheds.

The distribution of *Hyalella 'azteca'* is impressive. Some phylogroups (lineage 1D) have extremely shallow genetic divergences across large distances, a pattern more consistent with the cladoceran *Daphnia pulex* species complex (Crease *et al.,* 2012) or the amphipod *Gammarus lacustris* (Witt not published, 2015), whereby some haplotypes were detected in populations with a continental distribution in glaciated North America. The *Daphnia pulex* species complex distributed across North America and the spatial patterns indicate very little regionalism in glaciated regions (Crease *et al.,* 2012). Similar to *Sida*, *Daphnia* disperses during a diapaused egg stage through directly connected waterways or via animal vector.

While some *Hyalella* groups (species 1 lineage B) had distributions more limited in scale, the general trend of the large-bodied morphotypes possessed broad distributions that spanned multiple faunal provinces. In the spectrum of freshwater invertebrates, these *Hyalella* groups

have a greater degree of regionalism than *Daphnia pulex*, but are capable of passing between watersheds to a greater extent than *Sida crystallina*. The spatial distributions of the large-bodied morphotypes greatly exceed that expected of passively dispersing freshwater species that depend on interconnected proglacial systems, such as fish (Borden & Krebs, 2009; Stepien *et al.*, 2009; Berendzen *et al.*, 2010; April *et al.*, 2013) and invertebrate glacial relicts like *Mysis diluviana* (Dooh *et al*., 2006).

3.7.2 Genetic patterns within large-bodied *Hyalella* species

The dominant pattern at all levels of the nested clade analysis was isolation by distance (table 3.20). This indicates that the most prevalent means by which the large-bodied morphotypes disperse is through short distance colonization events. As new habitats are colonized, derived haplotypes found new populations further along the expansion wave. This results in a pattern of isolation by distance, which can be seen through the network overlays on the geographic maps and by the nested clade analyses. This pattern has been detected into species with patchy distributions, which is often the case with freshwater organisms as a natural consequence of the freshwater system (Elderkin *et al.*, 2008).

Table 3.20: Summary of nested clade analysis inferences
The inferences are indicated on the basis of the nested clade analysis for all species 1 lineages, species 2 five-step clades and species 7. The three major inferences are fragmentation (Fr), restricted gene flow with isolation by distance (IBD) and range expansion (RE). The last column indicates long distance dispersal (LDD) and is marked yes for either long distance colonization or recurrent gene flow with long distance dispersal. A "yes" with a "?" indicates that the only instances with that inference were ones with multiple possible answers.

Species (lineage/clade)	Fr	IBD	RE	LDD
1(A)	No	Yes	Yes	Yes
1(B)	No	Yes	Yes	No
1(D)	No	Yes	No	Yes
1(E)	No	Yes	No	Yes?
1(F)	Yes?	Yes	Yes	No
1(G)	No	Yes	Yes	No
1(H)	Yes?	Yes	Yes	No
1(I)	No	Yes	No	Yes
2 (5-1)	Yes	Yes	Yes	Yes
2 (5-2)	Yes?	Yes	Yes	Yes?
2 (5-3)	Yes?	Yes	Yes	Yes
7	Yes?	Yes	Yes	Yes
8	Yes	Yes	No	Yes

3.7.3 Allele surfing within expanding ranges

What is most evident from the survey of such diverse groups is that large-bodied *Hyalella* species are not beholden to one dispersal strategy, but a combination of short and long distance dispersal along with allele surfing. The haplotype networks of species 1 lineage A, C, and E all possess characteristics that might indicate allele surfing; interior clades with ranges that encompass the combined distributions of their derived counterparts. Within lineage A, the distribution of clade 3-2 is consistent with this pattern, while the other major clade (3-4) possesses a genetic pattern of isolation by distance derived from short distance dispersal among neighboring habitats. The distribution of genetic variation within lineage C and E is shallow in

comparison to other species 1 lineages, with populations that were only detected within glaciated North America. This pattern is more consistent with the *Daphnia pulex* species complex (Crease *et al.*, 2012) or Monarch Butterflies (Pierce *et al*., 2014). Allele surfing in species 2 is far less apparent. In most of the clades, the distributions of derived haplotypes extend the species range (interior clades are located nearer to refugial sources) as mutations occur along the expansion front. This results in patterns of isolation by distance consistent with short distance dispersal among neighboring demes.

Within species 7, the interior haplotype S7H004 (within clade 3-1) possesses a spatial pattern consistent with allele surfing. It is broadly distributed among previously glaciated habitats (occupying 24 sampled sites) and the range overlaps that of the derived haplotypes.

3.7.4 Long distance dispersal among large-bodied *Hyalella* species

Long distance dispersal from source habitats can mitigate the reduction in genetic diversity within populations along the expansion front. In *Hyalella* species, the potential for long distance colonization is high because of its ability to disperse via waterfowl (Swanson, 1984) and the potential to colonize new habitats with a single gravid female (Wellborn & Capps, 2013). All large-bodied species had clade distributions with disjunct populations and the NCA identified long distance dispersal in several cases (table 3.20). Still, many of the lineages within species 1 lacked the sample size necessary to ascertain significant geographical associations at the lower levels of the analysis. This disproportionately affects more recent dispersal events. Older groups will have had more time to establish larger distributions, as such the NCA is well suited to detect

long distance colonization events employed during the initial expansion into glaciated regions. Species 1 lineage B and H both have distributions that included sites north of the glacial limit that suggest a long distance colonization event, although the analysis of these groups did not yield significant geographical associations. Species 1 lineage I was located in disparate regions that the SAMOVA partitioned into Beringian and Cascadian groups. This leads to two potential scenarios, either a larger range was fragmented by the glacial cycles or there was long distance colonization. Given the level of diversity within lineage I, there is no reason to suggest that the divergence of these two groups predates the Pleistocene. The level of divergence among the derived group (Cascadian) is similar to that of expansion groups in lineages A and E.

The species 2 NCA inferred patterns of long distance dispersal at the higher tiers of the network, as a result of the large average distance of populations from the geographic centers of the clades. More so than any other *Hyalella* group, species 2 has demonstrated a profound ability to colonize glaciated regions. Simulations have shown that haplotypes that jump ahead of the expansion front can establish populations that halt the dispersal of surfing allele (Hallatschek & Nelson, 2010). This is most likely to occur when long distance dispersal is rare and the route of dispersal is along a corridor. This distribution pattern apparent in species 2 has been referred to as an ‘embolism’ and is a consequence of long distance colonization. Allelic variation within this region will be more similar to the source populations than the expanding groups. The distribution of species 2 clade 5-2 suggests that individuals with ancestral haplotypes (interior to the network) were able to colonize portions of the Northwest Territories ahead of the expansion wave. The alternative hypothesis is that the interior haplotype expanded throughout glaciated North America and was subsequently fragmented and recolonized by *Hyalella* possessing derived haplotypes. Fragmentation and long distance colonization can produce similar

distribution patterns, however the NCA considers colonization more likely if organism is capable of long distance dispersal. Furthermore, fragmentation inferred at higher tiers of the hierarchical clades is expected to be associated with nested clades possessing distributions restricted to the disjunct regions. This is a consequence of new mutations arising within the fragmented regions. This pattern is not evident within clade 5-2.

3.7.5 *Hyalella 'azteca'* suture zones

Organisms with strong dispersal capabilities can overcome potential barriers to gene flow and possess genetic patterns consistent with isolation by distance (Waples & Gaggiotti, 2006), but this does not exempt these species from the effects of glacial cycles. Latch *et al* (2014) identified refugial lineages of mule deer (*Odocoileus hemionus*) consistent with genetic patterns of allopatric divergence followed by post-Pleistocene dispersal. The resulting effect was broad spatial ranges characterized by patterns of isolation by distance and the formation of suture zones. With the high levels of divergence between *Hyalella 'azteca'* species, it is likely that many cladogenic splits predate the Pleistocene, however many of the lineages and clades therein have been greatly affected by fragmentation from the repeated glacial cycles. The expansion of these previously allopatric groups has resulted in potential suture zones. The most prominent of these is in the habitats surrounding the Great Lakes. Every large-bodied *Hyalella* species, except 8, and many of the species 1 lineages possess a spatial distribution with a border within this region. These habitats have more sympatric populations of *Hyalella* species than any other region in North America. The Great Lakes region has been identified as an important suture zones in fishes (Borden & Krebs, 2009; April *et al.*, 2013) and amphibians (Rissler & Smith,

2010). Even glacial relicts, like *Mysis diluvania,* had greater haplotype diversity in the Great Lakes region (Dooh *et al.,* 2006).The other area within glaciated North America having populations that possess high levels of sympatry and diversity was in central Canada. While many of the clades within species 2 and 7 have distinct distributions, they share a common dispersal route through habitats in the interior of Saskatchewan and Alberta (see below). Lake Agassiz was in important dispersal route for many freshwater species from southern refugia (Rempel & Smith, 1998; Stepien *et al.,* 2009; Aldenhoven *et al.,* 2010). The Great Lakes region represents an important suture zone for small-bodied *Hyalella* species, which will be more fully explored in the next chapter.

3.7.6 Sectoring within species 1 lineage A

The SAMOVA partitioned the species 1 lineage A populations into three major assemblages (fig. 3.9), two of which were consistent with 3-step clades (3-2 and 3-4). Excoffier and Petit (2008) described this pattern as “sectoring”, whereby regions with populations undergoing demographic expansion have uniformly distributed allele frequencies and sharp frequency shifts along the borders. As new habitats are colonized, they can undergo a process of local adaptation and specialization that can resist subsequent migration by competitors who would occupy similar niches (De Meester *et al.,* 2002). However, multiple species of *Hyalella* are known to persist in the same environment in spite of competition for similar niches (Cothran *et al.,* 2013). The spatial distribution of two of the 3-step clades in lineage A was divided by the St. Lawrence River. Several fish species are similarly oriented, with both eastern and western phylogroups (April *et al.,* 2013): Western Blacknose Dace, Johnny Darter, American Pickerel, Golden Shiner

and the Fathead Minnow. This partition was attributed to dispersal from distinct refugia resulting in secondary contact within the Great Lakes region. The distributions of populations east and west of the St Lawrence River were attributed to Atlantic and Mississippian refugia, respectively. However, in lineage 1A, the root of the haplotype network is within a phylogroup (clade 2-6) that is located in northern Ontario and Quebec (fig. 3.8). As such, there is little evidence to suggest that the clades within species 1 lineage A have different refugial sources. In fact, the network suggests that the region surrounding the St. Lawrence River was the terminal end of a range expansion in western Ontario. As such, there is little evidence to suggest that the different lineage A clades (which possess less than 0.5% divergence) are from different refugial sources, which may suggest that its spatial patterns are the result of monopolization and sectoring.

3.7.7 Overview of species 1 refugia

A well-established trend in phylogeography is the tendency for allelic variation within populations to decrease in glaciated regions (Elderkin *et al.,* 2008). As such, glacial refugia should contain older populations with higher levels of diversity, which can be used to trace the post-Pleistocene dispersal of *Hyalella 'azteca'*. Despite of the extent to which this study explores the distribution of species 1, only a few lineages (B, H and I) likely included populations that have persisted through the Pleistocene. Several lineages (A, C, D and E) possess genetic patterns that suggest only expansion populations were detected. There are a few potential reasons for this. Primarily, the data may be missing. Continued sampling in northern United States and a thorough study of the Missouri River system is still required. The network analyses of most

species 1 lineages indicated a northward range expansion (except lineages C, D and I), thus this river system remains a prime candidate as a refuge. Alternatively, the ancestral populations may have gone extinct. The impact of the glaciers extends well beyond the limit of the ice sheets, resulting in lower sea levels that allowed for the existence of coastal refugia that would now be submerged (Bermingham & Avise 1986; Avise, 2000).

Only one population (NY003) was detected south of the glacial limit, thus it is doubtful that the sampling range of this study included the refugial populations of lineage A. The capability of *Hyalella 'azteca'* for long distance colonization makes determinations of refugial origins difficult. As is the case with many species 1 lineages, either the sampling area did not cover the refugial regions or those populations have since gone extinct.

3.7.8 Species 1 lineage B

Lineage B haplotypes were largely restricted to the Great Lakes region with some sites in Quebec and Manitoba. The presence of populations in glaciated regions ensures that recolonization has occurred since the Pleistocene, but this lineage lacks many of the genetic patterns associated with rapid population expansion. While the network and SAMOVA suggests two distinct groups, there is too little data to determine whether they shared a common refuge. The nested clade analysis suggests that both groups colonized the Great Lakes region prior to dispersing to the other previously glaciated habitats. Perhaps the first group (clade 3-5) was located in the southern watersheds of Lake Michigan and Lake Erie, alternatively the second

group (clade 3-6 + clade 4-3) initially colonized the habitats surrounding Lake Superior. The proglacial Lake Duluth has been identified as a potential dispersal route of fresh water fishes from the Mississippi River refugium into western Lake Superior (Stepien *et al.*, 2009). However, none of the sampling in ice-free regions detected lineage B individuals. The genetic patterns of lineage B are in stark contrast with other species 1 groups within the Great Lakes region (lineages 1A and 1E) and commonly share the same habitat. While species 1 lineage A has broadly distributed central haplotypes and mismatch distributions associated with demographic expansions, lineage B is more indicative of stationary populations. The spatial structure does not indicate a stepping stone dispersal, but rather limited long distance dispersal following an initial colonization of the Great Lakes region. Given the high rate of sympatry with species 1 lineages A and E, this may indicate a later migration of lineage B individuals into habitats with established *Hyalella* populations. Monopolization and local adaptation of species 1 lineage A and E populations may limit the signal of demographic expansion within species 1 lineage B.

3.7.9 Species 1 lineage C

The entirety of species 1 lineage C was contained in a single 1-step clade with the interior haplotype S1H180 inhabiting nine of the twelve locations (fig. 3.14). A unimodal mismatch distribution along with the significant deviations from neutrality detected by the Tajima's D, Ramos-Onsins & Rozas R_2 and Fu's F statistics (table 3.1) imply a rapid demographic expansion. This coupled with haplotype S1H180's broad distribution implies allelic surfing on a grand scale. Only three of the sampled localities were located south of the glacial limit, which makes determination of refugial origins difficult. Furthermore, the populations within these

habitats lack the haplotype diversity expected in older populations. Since the central haplotype was detected across such a large region, at least 4 refugia are plausible (Beringian, Missourian, Nahanian and Mississippian).

3.7.10 Species 1 lineage D

The bimodal distribution of lineage D's mismatch distribution coupled with the separation of clades 3-10 and 3-11 by numerous inferred mutations indicates two distinct groups. Although, the lack of sufficient sample size muddles any potential assessment regarding refugial origin or patterns in post-Pleistocene dispersal. Fu's F, Tajima's D as well as Ramos-Onsins & Rozas R_2 were unable to detect any deviations from neutrality and while the nested clade analysis detected long distance dispersal it did not return evidence of a range expansion. Although, the fact remains that lineage D occupies habitats that were once glaciated. This pattern at first may seem to be similar to lineage C, but it lacks a broadly distributed central haplotype coupled with a star phylogeny in the network.

3.7.11 Species 1 lineage E

The spatial range of lineage E was largely congruent with lineage A, and may share similar refugial origins, but the lack of samples from unglaciated regions makes any such determination speculative. The current geographical distribution of lineage E was predominately shaped by the allele surfing of haplotype S1H154. Rapid demographic expansion was supported by the unimodal mismatch distribution, as well as the raggedness, Tajima's D, Ramos-Onsins & Rozas

R_2 and Fu's F statistics (table 3.1). The nested clade analysis of clade 3-12 did not resolve whether the expansion of lineage E into glaciated regions was a consequence of short or long distance movement, although both are likely contributors. Given the overlapping distribution with lineage A, they potentially share a similar history, exploiting similar dispersal paths through pro-glacial lakes formed during the glacial melt.

3.7.12 Species 1 lineage F

Lineage F was predominately distributed in western Canada and included mountainous freshwater sites throughout the Rockies (fig. 3.23). The unimodal mismatch distribution (fig. 3.24) in conjunction with Tajima's D, Fu's and Ramos-Onsins & Rozas R_2 statistics (table 3.1) were consistent with a demographic expansion. Specifically, the nested clade analysis inferred an expansion into habitats north of the Alaskan Range. The tip clades in the network are connected to ancestral haplotypes, which were collected from a site within the watershed of Great Slave Lake. The distribution of lineage F was similar to that of lineages C and D, in that all three groups include vastly disparate localities throughout central Canada with range limits that end within the Great Lakes region. This broad spatial pattern is similar to the fish species lake cisco (*Corogenos artedi*), which was attributed to a Mississippian refugium (Turgeon & Bernatchez, 2001). Fish migrants re-colonized the Mackenzie drainage by dispersing through the Lake Agassiz - Clearwater River channel (Rempel & Smith, 1998). The rationale of Turgeon & Bernatchez (2001) was that this broad spatial pattern was not consistent with the distribution of fish species originating from Berningian refugia. This also makes intuitive sense in that the most broadly distributed species of *Hyalella* in glaciated regions would be the ones first able to exploit

the early proglacial lake system formed from the southern end of the retreating glaciers. Furthermore, if lineages C, D and F share a Mississippian refuge with lineages A, B and E, this might help explain some of the deeper splits in the species 1 phylogeny. There are three major groupings of lineages that are supported by high bootstrap support: group 1 (lineages A to F), group 2 (lineage G) and group 3 (lineages H, I and the population in Cooper Lake, WA). The cladogenic split between group 1 and the other two groups may represent fragmentation of western and eastern lineages into distinct allopatric refugia. With group 1 lineages surviving the Pleistocene glaciations south of the ice sheets, while groups 2 and 3 persisting in Cascadian, Beringian or Nahannian refugia.

3.7.13 Species 1 lineage G

The distribution of lineage G haplotypes suggests multiple potential glacial refugia. The most prevalent inference in the nested clade analysis was a pattern of restrictive gene flow with isolation by distance. This indicates that freshwater habitats in British Columbia were largely colonized by short distance dispersal events. The mismatch distribution, Tajima's *D*, Ramos-Onsins & Rozas R_2 and Fu's *F* all are consistent with rapid demographic expansion. The only pattern of contiguous range expansion was inferred during the NCA at the total cladogram level, with an increase in range within the Yukon and Northwest Territories.

The refugial origins of lineage G are difficult to interpret from its present-day distribution. The central haplotypes of clade 3-15 were associated with individuals collected from the the Liard River system. Molecular evidence in Lake Whitefish distinguished a race

separate from the Beringian clade that was collected from the Liard River and South Nahanni river systems. In conjunction with geological evidence, Foote *et al.,* (1992) identified the Nahanni National Park in the Northwest Territories as a potential ice-free refugium during the last set of glacial advances for white fish. Similar studies have drawn the same conclusion for Lake Trout (Wilson & Hebert, 1998), Arctic Grayling (Stamford & Taylor, 2004) and the Redside Shiner (Houston *et al.,* 2014). This region was located between the Laurentian and Cordilleran ice sheets, and may have been a potential haven for *Hyalella 'azteca'*. Furthermore, Chuchi Lake is situated in this region and contained a divergent set of species 1 haplotypes (S1H350 and S1H349 in fig. 3.3) that didn't group with any of the lineages. However, unlike many freshwater fish species, *Hyalella 'azteca'* is capable of long distance overland dispersal (Swanson, 1984) and the root of the lineage G network was placed within clade 2-37, which is located in southern Alberta (AB030). While the genetic patterns are easier to understand once lineage G individuals began dispersing out of the Liard River system, how these populations were founded initially remains a mystery. On the basis of proximity, either Beringia or Nahanni are possible, although the location of the ancestral root suggests a southern refugium.

3.7.14 Species 1 lineage H

Lineage H was the most diverse of all species 1 lineages (table 3.1). The COI network identified several distinct groups separated by internal edges of inferred haplotypes, the distributions of which were predominately located in ice-free regions south of the Cordilleran glaciers. Lineages undergoing demographic expansions have an excess of singletons and the increasing population sizes prevents the loss of new alleles to genetic drift. However, these clades are replete with

inferred haplotypes. Greater genetic diversity is expected in stationary populations located within or near to the glacial refugia. The nested clade analysis did identify three potential range expansions in conjunction with long distance colonization. This is further supported by the Fu's *F*, Tajima's *D* and R_2 statistics, which are generally very sensitive to demographic expansions (Ramos-Onsins & Rozas, 2002). Clade 5-3 represents the oldest of these, with interior clades distributed in Oregon and tip clades located in Nevada and Washington State. Interestingly, the mismatch distribution was multimodal and the raggedness statistic rejected a null hypothesis of demographic expansion, but the frequency spectrum clearly marks a distinct peak at τ=6.652. This measure is used to infer the age of the expansions on the basis of the peak ($\tau=4N_e\mu$ where N_e is the effective population size and μ is the per locus per generation mutation rate), which is consistent with the range expansion of clade 5-3. Given the diversity of lineage H haplotypes and the mismatch distribution, the estimates of this expansion's date would predate those of the other species 1 lineages, which had much smaller values of tau (table 3.1). However, *Hyalella* populations within this region have been described with indications of accelerated mtDNA evolutionary rates (Witt *et al.*, 2008). Additionally, the NCA identified two other range expansions (clades 3-20 and 2-46). Clade 3-20 demonstrated a pattern of contiguous range expansion with tip clades located on Vancouver Island. Clade 2-46 is probably the most recent of the expansions and included localities in southern British Columbia along with long distance colonization of the Liard River system. Interestingly, the distributions of lineage G and H haplotypes form non-overlapping zones in British Columbia that may suggest possible "sectoring" (Petit and Excoffier, 2009).

3.7.15 Species 1 lineage I

The distribution of lineage I populations were partitioned into Beringian and Cascadian assemblages. The SAMOVA further supported this subdivision, albeit with a slight modification of grouping the populations north of the Alaska Range with the Cascadian assemblage. The mismatch distribution was not uniformly unimodal and the raggedness statistic rejected the null hypothesis of demographic expansion. This was not consistent with Fu's *F*, Tajima's *D* and Ramos-Onsins & Rozas R_2, which all indicated significant departures from neutrality. The nested clade analysis inferred a pattern of isolation by distance at all tiers of analysis along with long distance dispersal within clade 4-9. The root of the network was placed within clade 3-26, which suggests a Beringian refugium. The absence of lineage I individuals in the previously glaciated regions of British Columbia suggests that the colonization of the southern habitats occurred via long distance colonization. Furthermore, this event could predate the retreat of the Wisconsinin glaciations.

3.7.16 Dispersal of species 2 from southern refugia

Broad-scale relationships within species 2 show genetic and spatial patterns consistent with dispersal of allopatric groups from southerly refugia. The nested clade analysis and haplotype networks of the 5-step clades suggest that at least two distinct refugia; clades 5-1 and 5-3 from the Missouri River along with clade 5-2 from the Mississippi River. Their present day distributions suggest sequential short distance dispersal events resulting in genetic isolation by

distance (table 3.20). Given the enormous ranges of species 2, it suggests strong dispersal ability over short distances through the consecutive colonization of neighboring demes.

A consistent pattern in species 2, as well as species 7, is that each of the major 5-step clades that recolonized glaciated North American did so through an initial colonization of central Canada. The retreat of the Wisconsinin glaciations (~13 kya) created a complex lake system in southern Saskatchewan and Manitoba (Klassen, 1994). This freshwater geography of this region was highly turbulent during the retreat of the glaciers, though around 13 kya a proglacial lake system flowed southeast through Lake Agassiz and was connected to the Mississippi River (Stepien *et al.,* 2009). The proglacial Lake Peace flowed into Lake Edmonton, Lake Souris and eventually into Lake Agassiz. This was a tumultuous system of frigid water that experienced heavy sediment load from the turbulent discharge of meltwaters (Rempel & Smith, 1998). As the glaciers retreated northwards, the outflow of Lake Agassiz switched direction, heading into the Hudson Bay and Mackenzie River drainages (~8.5 kya). This system permitted the dispersal of freshwater species into previously glaciated regions of Lake Winnipeg (Stepien *et al.,* 2009). The presence of extant *Hyalella 'azteca'* species 2 in the Mississippi River and Missouri River basins make those regions strong contenders as refugia. The overlapping distributions of 3-step members of clade 5-1 pinpoint central Canada as the ancestral source of the ensuing eastward expansion. The rate heterogeneity of the COI sequences prevents accurate dating of these expansions elements, but it is not an unreasonable expectation that colonization events were largely shaped by the glacial retreats and fluvial events of the late Pleistocene and early Holocene.

3.7.17 Species 2 genetic and spatial patterns

Within species 2, clade 3-1 (appendix C, C.10) is by far the largest of the 3-step clades and the most broadly distributed (fig. 3.26). Nested within this group is the tip clade 2-1 marked by isolation by distance and the interior clade 2-2 that the NCA indicated a pattern consistent with a recent range expansion. The central haplotype S2H162 is broadly distributed throughout the range of clade 2-1 (13 sample sites) and *de novo* mutations have begun appearing in geographically restricted tip clades. The potential for long distance dispersal was inferred because clades within the network may be genetic neighbors yet geographically separated. The NCA infers that an initial range expansion occurred in Alberta, perhaps into the proglacial Lake Bassono system. The differential timings of the 5-1 clade range expansions reflect the dynamic nature of the changing freshwater landscape in response to the retreat of the glaciers. Of the twelve significant effects, four of them indicated range expansions with at least one at every tier going from the 2-step to the 5-step level. Rate heterogeneity remains a potential confounder and could affect any estimations regarding the ages of events. Albeit, it is unlikely that differential mutation rates will alter the order of events among nested groups. Of principal interest are the contradictory directions of the clade 5-1 range expansions. The 4-1 group expands eastward towards the Great Lakes region, whereas clade 4-2 extends in direct opposition. If we take the internal placement of clade 3-2 as the ancestral group that precipitated both expansions, the formation of clade 3-3 must predate its expansion into clade 3-4's habitats, but this does not resolve the absence of clade 3-2 haplotypes in the intermediary region. Either the 3-2 group skipped this area during their expansion eastward via waterfowl vectors or the intervening population went extinct prior to colonization by the 3-4 group. Such an extinction event would have to affect the entire intervening region between clades 3-2 and 3-3. This area directly

coincides with the size of proglacial Lake Agassiz. It is conceivable that clade 5-1 individuals dispersed through the Peace-Agassiz channel into the proglacial lake system of the Great Lakes region (i.e. Lake Duluth). When ice dams of Lake Agassiz broke at the start of the Younger Dryas (Murton *et al.*, 2010), the massive outburst of flood waters could have displaced large numbers of populations and habitats.

Clade 5-3 contains five 3-step clades (3-12 to 3-16) that were obtained from habitats in an area stretching from Iowa to Alaska. The central and ancestral clade was 3-12 and was collected from sites in central Alberta. This continues the trend seen with the previous 5-step clades, by which the expansions emanate from central Canada. The nested clade analysis of clade 3-15 indicated a contiguous range expansion north into Alaska and British Columbia.

3.7.18 Repeated expansions within clade 5-2 (species 2) from a Mississippian refugium

Clade 5-2 has four 3-step clades (3-8 to 3-11) that were broadly distributed through the western Great Lakes region, the American Midwest, British Columbia and Northwest Territories (fig. 3.28). The NCA indicated multiple range expansions at varying tiers of the nested species network expanding into previously glaciated habitats. The interior clade 3-11 was distributed throughout Missouri, Iowa and South Dakota, which indicates that the unglaciated regions of the Missouri River basin were the likely refugial source of clade 5-2. As seen in clade 5-1, two distinct range expansions are inferred at different tiers (the 2-step and in this case the 4 and 5-step), implying that recolonization of glaciated regions was a dynamic process as more regions became unglaciated.

Clades 3-11 and 2-21 possess patterns of contiguous range expansions that likely originated in the Mississippi basin (fig. 3.28). Clade 3-11 expanded its range along both its southern and northern fronts, and the range of clade 2-21 indicated colonization of habitats in southern Saskatchewan. The central haplotype of clade 3-9 (S2H002) was sampled in 9 sites that split into two broad groups; a Wisconsin/Iowa and northern Canada group. Either this haplotype occupied a much larger range and was subsequently fragmented, or a long distance dispersal event resulted in a colonization of a habitat ahead of the expansion wave. This phenomenon has been referred to as an 'embolism', a subset of sectoring (Excoffier *et al.,* 2009). If a species capable of long distance dispersal jumps ahead of the short-distance colonization events normally associated with a surfing allele, this haplotype can disperse back towards the expansion front. These locally adapted populations can resist invasion by the surfing allele.

3.7.19 Species 7 genetic and spatial patterns

Species 7 successfully recolonized glaciated North America through a combination of long distance colonization, short distance dispersal and allele surfing. The mismatch distribution indicates two distinct demographic expansion signals that coincide with the pattern indicated by the network. The first expansion consists of two 4-step clades (4-1 and 4-2) and the second coincides with the remaining 4-step clade (4-3). The relationship among the 3-step clades is more ambiguous and the spatial patterns indicate a complicated history. Similar to species 2, this species may have exploited the Peace-Agassiz water connection (Rempel & Smith, 1998). All the sample sites located in unglaciated regions are tip clades and this coupled with the lack of contiguous range expansions inferred at higher tiers, is a strong indication that the sampling area

did not include habitats retaining refugial populations. The localities in glaciated regions were predominately centered in the proglacial lake systems of Bassano and Agassiz and may indicate a Missouri River refugium, but given the potential for long distance dispersal, other refugia cannot be completely discounted.

3.7.20 Species 8 genetic and spatial patterns

Wellborn & Broughton's (2008) study of *Hyalella* communities in Oklahoma strongly suggests that the species 8 range extends west of the Mississippi River. The analysis within this thesis considered populations distributed throughout the Gulf-Atlantic Coastal Plain, but also included sites within the Mississippi River drainages. The AMOVA indicated that partitioning the covariance according to faunal provinces was significant. This pattern places groups to the east and west of the Apalachicola River (partitioned according to Atlantic and Gulf drainages) and is consistent across multiple taxa; including the gopher tortoise (Ennen *et al.*, 2012), Gulf Coast box turtle (Butler *et al.*, 2011), green anole lizard (Campbell-Staton *et al.*, 2012) and southern leopard frog (Newman & Rissler, 2011). This is one of the more prominent biogeographic signals that formed the basis of Avise (2000) study into phylogeography and is further explored in Soltis *et al.* (2006). While many of these species have reciprocally monophyletic groups, the species 8 assemblages possess COI haplotypes that are in different sections of the haplotype network. Specifically, the distribution of clades 3-4, 3-5 and 3-17 are located in the Mississippi drainages, but their COI haplotypes in the network are connected to each other through clades that are located on the Florida Peninsula. Furthermore, these clades are separated by edges of inferred haplotypes within the network that were either not sampled or have gone extinct. This

suggests that phylogroups within the central continent have begun sorting according to allopatric lineages and that the Florida Peninsula was colonized by an interior haplotype; likely recently given that the the mismatch distribution and population genetics statistics still bear the signal of rapid demographic expansion. The value of tau ($\tau = 6.75$) would place this expansion prior to those seen in *Hyalella* species colonizing glaciated regions (species 2, $\tau = 5.44$), however heterogeneous rates of COI evolution (chapter 2 and Witt, 2006) makes these estimations difficult. Similar genetic patterns were detected in populations of least killifish (*Heterandria formosa*) in which mismatch distributions and coalescent simulations indicated rapid demograghic expansion within the Florida Peninsula (Bagley *et al.*, 2013). The distribution of species 8 indicates a greater degree of diversity within populations at the northern part of the Peninsula. The interior haplotypes of the network were collected from these northern sites and the derived individuals were detected to the south. This is in contrast to the expansion of killifish, which was traced to a southern refugium that likely resulted from a previous contraction in species range (Bagley *et al.*, 2013). While a southwardly expansion in species 8 may not be the only possibility, the genetic and spatial patterns suggest this conclusion.

3.8 CONCLUSIONS

The large bodied *Hyalella* species represent a diverse and broadly distributed group of species. While species 8 was not detected in glaciated regions, all of these species bear genetic and spatial patterns consistent with range expansions during or after the Pleistocene. These species have dynamic ranges and their distributions have been shaped by a combination of both short and long distance dispersal. However, population structure with genetic patterns of isolation by

distance indicates that long distance dispersal events, though significant, are likely rarer than either active dispersal via swimming or short distance jumps via animal vectors to neighboring habitats. That being said, several species 1 lineages along with species 2 and 7 possess a profound capacity for colonization. These species possess spatial distributions that span thousands of kilometers, which could only have been colonized since the last glacial maximum.

Many of the species 1 lineages along with species 2 and 7 possessed spatial and genetic patterns consistent with groups that persisted through the glaciations in either a Missouri River or Mississippi River refugium. This suggests that the more broadly distributed groups in glaciated North America were able to exploit the proglacial water system formed from the meltwaters of the retreating ice sheets. The Lake Peace to Lake Agassiz waterways would have provided a westward dispersal corridor for early migrants and the dynamic shifts of this water system can help explain the genetic and spatial patterns of large-bodied *Hyalella* species.

What is evident from this study is that the distributions of some of the large-bodied species have not been fully characterized. The Missouri River watersheds of the American Midwest need to be studied to better characterize the refugial origins of species 1, 2 and 7. Alternatively, the limits of species 8 range remain unexplored in the American south and within the Gulf Coast drainages.

Chapter 4

Phylogeography of Small-Bodied *Hyalella* Species

4.1 INTRODUCTION

This chapter continues the biogeographical analysis (see chapter 3) of the *Hyalella 'azteca'* species complex with a focus on the small-bodied members (species 3, 4, 5, 6 and 9). The aim is to compare the spatial and genetic patterns of small-bodied species to their larger counterparts. While not universal, Cope's Rule has been applied to a large number of organisms (Heim *et al.,* 2015) and suggests that the size of species will increase over generational time. However, the study of the ecological or evolutionary mechanisms that cause the persistence of smaller body sizes remains poorly characterized (Blanckenhorn, 2000). With respect to the *Hyalella 'azteca'* complex, the distribution of species is determined in part by ecological mechanisms resulting in the discrete partitioning of populations. As a consequence of their size, the large ecomorphs possess a competitive advantage with respect to the small-bodied species (Wellborn, 2002). In order to meet the higher resource requirement of size, these species can engage in aggressive foraging strategies (Blanckenhorn, 2000). In habitats containing both morphotypes, the smaller amphipods experience a greater degree of wounding either from larger morphs or predation by *Gammarus lacustris* (Cothran *et al.,* 2013). The advantages of larger size come at greater risk of predation in localities with centrarchid fish. Stomach content analysis of *Lepomis* fish species indicated that large-bodied *Hyalella* were experiencing predation at five-times the rate of their smaller counterparts (Wellborn, 1994a; Wellborn & Cothran, 2007). The complex interaction between competitive ability and predation avoidance may cause the persistence (or at least delay extinction) of populations containing both morphotypes (Cothran *et al.,* 2013).

To what extent does size variation in morphotypes relate to the spatial patterns of *Hyalella* species? In large-bodied species pairing success (mating behavior) was greater in individuals with larger body and posterior gnathopod size (Wellborn and Bartholf, 2005; Cothran & Jeyasingh, 2010). If large-bodied species experience selective pressures (both natural and sexual) towards an increased gnathopod size and *Hyalella* is known to disperse via animal vectors by grasping, it is not an unreasonable expectation that larger individuals may possess an inherent dispersal advantage. This chapter will explore the genetic and geographic patterns of the small-bodied species of *Hyalella 'azteca'* in comparison to the large-bodied species examined in chapter 3.

4.2 MATERIALS AND METHODS

See chapter 3 for an overview of sampling and sequencing procedures, in addition to the methodology for the statistical parsimony network, nested clade analysis (NCA), mismatch distributions, raggedness, Tajima's D, Fu's F, Ramos-Onsins and Rozas's R_2 statistic, analysis of molecular variance (AMOVA) and structural analysis of molecular variance (SAMOVA).

4.3 RESULTS: SPECIES 3

4.3.1 Overview small-bodied *Hyalella 'azteca'* species

Three hundred of the five hundred and fifteen sample sites were inhabited by small-bodied *Hyalella 'azteca'* species, which were predominately detected in habitats within the eastern half of North America. At least 60% of the populations were an admixture of two or more *Hyalella* species (appendix 1). The small morphotypes were frequently detected in habitats with other small-bodied species (48%) as opposed to their large-bodied counterparts (28%) and a smaller fraction of sites (12%) contained at least three distinct members of the cryptic complex. The nucleotide diversities ranged from 0.9% in lineage A of species 4 to 1.8% within species 6 (table 4.1). The G+C content (table 4.1) was fairly consistent across all members of the complex (42.8% - 46.0%).

4.3.2 Phylogeographic analysis of species 3

Three hundred and ninety species 3 individuals were collected and sequenced from 81 localities broadly distributed throughout the eastern half of North America (fig. 4.1). Seventy-four percent of these habitats were inhabited by at least one additional *Hyalella* species (35% with a small-bodied, 25% with a large-bodied morphotype and 14% with both). Species 3 was distributed in the eastern half of North America and as far north as Lake Winnipeg, MB (fig. 4.1).
Three-hundred and ninety COI sequences (196 unique haplotypes) were unambiguously connected into a single statistical parsimony network (fig. 4.2) with a maximum connection limit of 10 mutational steps as determined by the 95% parsimony criterion. The haplotypes were nested into one-hundred and ten 1-step clades (1-1 to 1-110), forty-nine 2-step clades (2-1 to 2-

49), seventeen 3-step clades (3-1 to 3-17), four 4-step clades (4-1 to 4-4), two 5-step clades (5-1 and 5-2) and the total cladogram was a 6-step clade.

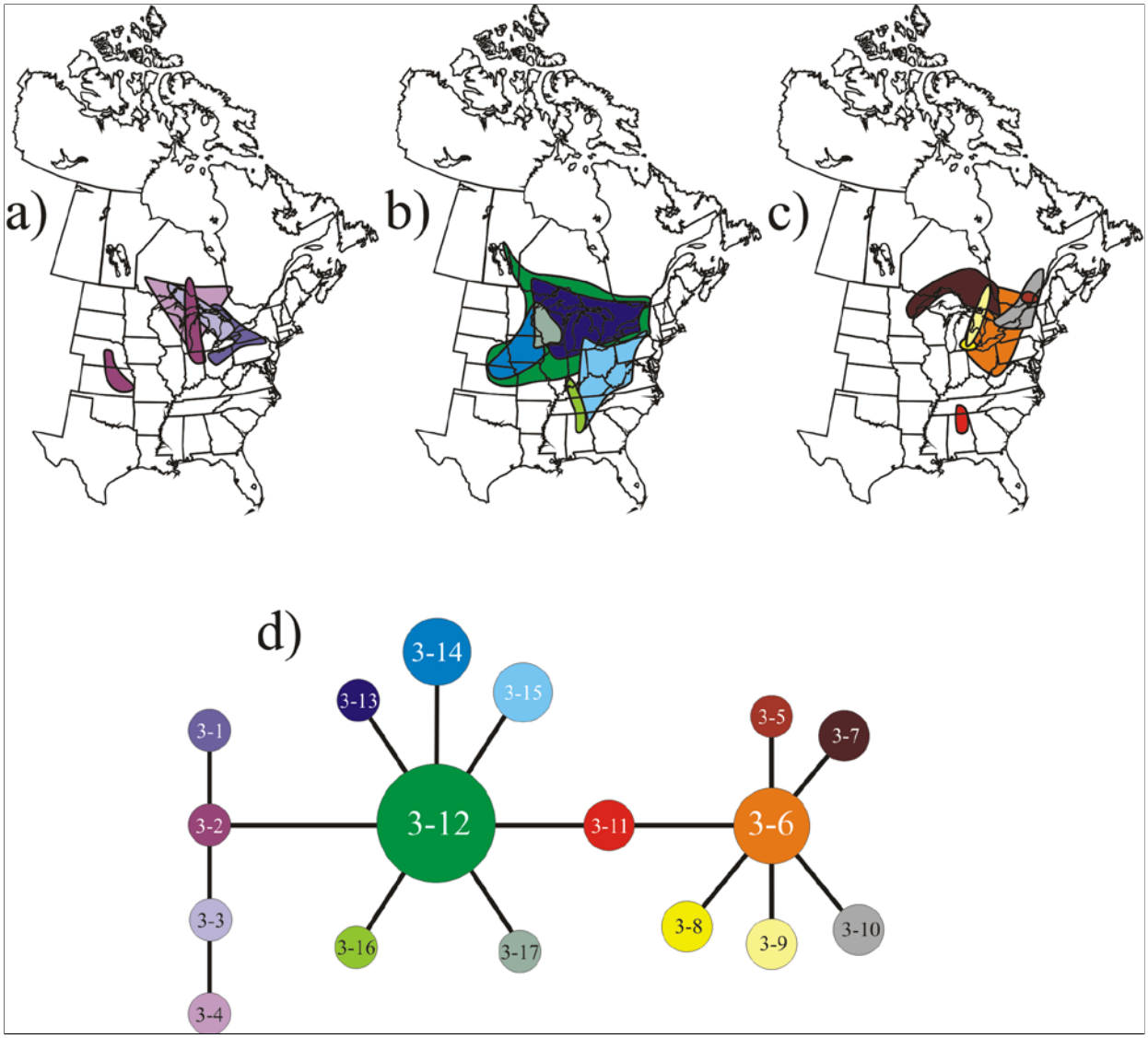
4.3.3 Outgroup analysis of species 3

The root of the haplotype network was determined by comparing phylogenetic trees determined by neighbor-joining and parsimony analyses (see chapter 2). The model of molecular evolution selected by the Akaike information criterion was a TPM2uf model with a proportion of invariant sites ($I = 0.590$) and gamma distributed rate variation ($\alpha = 1.381$). The NJ-tree placed the root along the edge connecting clades 2-43 and 2-44, but the parsimony analysis inferred a root along the edge between clades 1-95 and 1-99 within clade 2-43 (fig. 4.2). As such, both of these were located within clade 3-15, which will be considered an interior clade for the nested clade analysis. Species 1 lineage G haplotype (S1H209) was used as the outgroup.

Table 4.1: Population, sample and genetic statistics for each small-bodied species (lineage) of *Hyalella 'azteca'*

Species (Lineage)	No. of populations	No. of specimens	No. of Haplotypes	Nucleotide Diversity	Raggedness	Tau	Tajima's D	R_2	Fu's F
3	81	390	196	0.01390	0.004 (*p=0.001*)	4.550	-2.092 (*p=0.001*)	0.026 (*p=0.005*)	-260.098 (*p<0.001*)
4 (A)	12	47	29	0.00916	0.005 (*p<0.001*)	3.361	-2.013 (*p=0.006*)	0.043 (*p=0.002*)	-15.691 (*p<0.001*)
4 (B)	2	3	3						
4 (C)	1	6	3						
4 (D)	1	5	1						
4 (E)	1	1	1						
5 (A)	55	187	77	0.01006	0.005 (*p=0.002*)	3.238	-2.069 (*p=0.002*)	0.030 (*p=0.005*)	-34.179 (*p<0.001*)
5 (B)	8	21	26	0.00879	0.200(*p=0.0239*)	4.625	-1.220 (*p=0.088*)	0.0646 (*p=0.113*)	-8.811 (*p=0.006*)
5 (C)	3	15	9						
5 (D)	4	22	4						
5 (E)	3	12	4						
6	116	387	181	0.01875	0.003 (p<*0.001*)	7.162	-1.678 (*p=0.016*)	0.034 (*p=0.017*)	-187.40 (*p<0.001*)
9	18	53	40	0.01202	0.017 (*p=0.156*)	4.850	-1.774 (*p=0.018*)	0.0565 (*p=0.034*)	-13.00 (*p<0.001*)

Figure 4.1: Map showing the geographical distribution of species 3
The colored regions on the maps above refer to distributions of the respective clades in the network: a) clades 3-1 to 3-4, b) 3-12 to 3-17, c) 3-6 to 3-10 and d) condensed network of 3-step clades. The network (below) represents the associations among the COI 3-step clades of species 3 (figure 4.3) and the size of the nodes are proportional to their sample sizes.



4.3.4 Nested clade analysis of species 3

The nested clade analysis of species 3 COI haplotypes returned significant geographical associations in 22 clades (table 4.2): 14 were best modeled by restricted gene flow with isolation by distance, 1 allopatric fragmentation, 1 contiguous range expansion, 3 inconclusive outcomes, 1 inadequate geographic sampling and 2 associations with multiple potential outcomes.

Clade 5-1 was detected in a set of habitats broadly distributed throughout the species range (fig. 4.3), however only a few individuals were detected. Within the network, these haplotypes were separated by long edges possessing multiple inferred mutations (appendix C, fig. C.12). In order to resolve a symmetrical clade formed during the nesting of the 3-step clades, clade 2-5 was included within clade 3-1 on the basis of sample size. Within clade 5-1, the nested clade analysis of clades 3-1 and 3-2 returned significant geographical associations, but the inference key indicated inconclusive outcomes. The difference in interior and tip nested clade distances for both 3-step clades was significantly small (I-T $Dn_{(3\text{-}1)}$ = -421.34, $p = 0.017$, I-T $Dn_{(3\text{-}2)}$ = -129.46, $p = 0.001$), which provides insufficient information for the inference key to determine an outcome. Clade 5-1 is a poorly understood group and its populations possess no clear geographical pattern. The haplotypes detected in previously glaciated regions were those most closely linked to clade 5-2 (haplotype S3H114 was one mutational step from haplotype S3H070) and their inclusion in clade 5-1 was likely an artifact of the nesting procedure. The remaining haplotypes were restricted to regions south of the glacial limit and could represent older lineages that have persisted in ice-free regions. However, the placement of the root in clade 3-15 would indicate that these haplotypes were derived. As such, they could represent a group of COI sequences with an accelerated rate of evolution.

Clade 5-2 represents the majority of the COI haplotypes within species 3 (179 of the 196 unique haplotypes) and includes 20 of the 22 significant geographical associations. Clades 4-3 and 4-4 were nested in clade 5-2 and were separated by 6 inferred mutational steps (fig. 4.2a). The clade 5-2 NCA indicated isolation by distance with restricted gene flow because the tip clade (4-3) had a significantly small clade distance (Dc=547.10, $p<0.001$). The derived clade 4-3 was detected in populations within previously glaciated regions, but the distribution of its direct ancestral clade (4-4) was predominately distributed south of the glacial limit. This implies that while there was expansion of clade 5-2 into glaciated regions, the spatial patterns did not indicate either allele surfing or long distance colonization. The genetic pattern of isolation by distance was more consistent with stepping-stone migration resulting from short distance dispersal among neighboring demes. However, some of the clades at lower tiers possess disjunct distributions that may suggest long distance movement (see below).

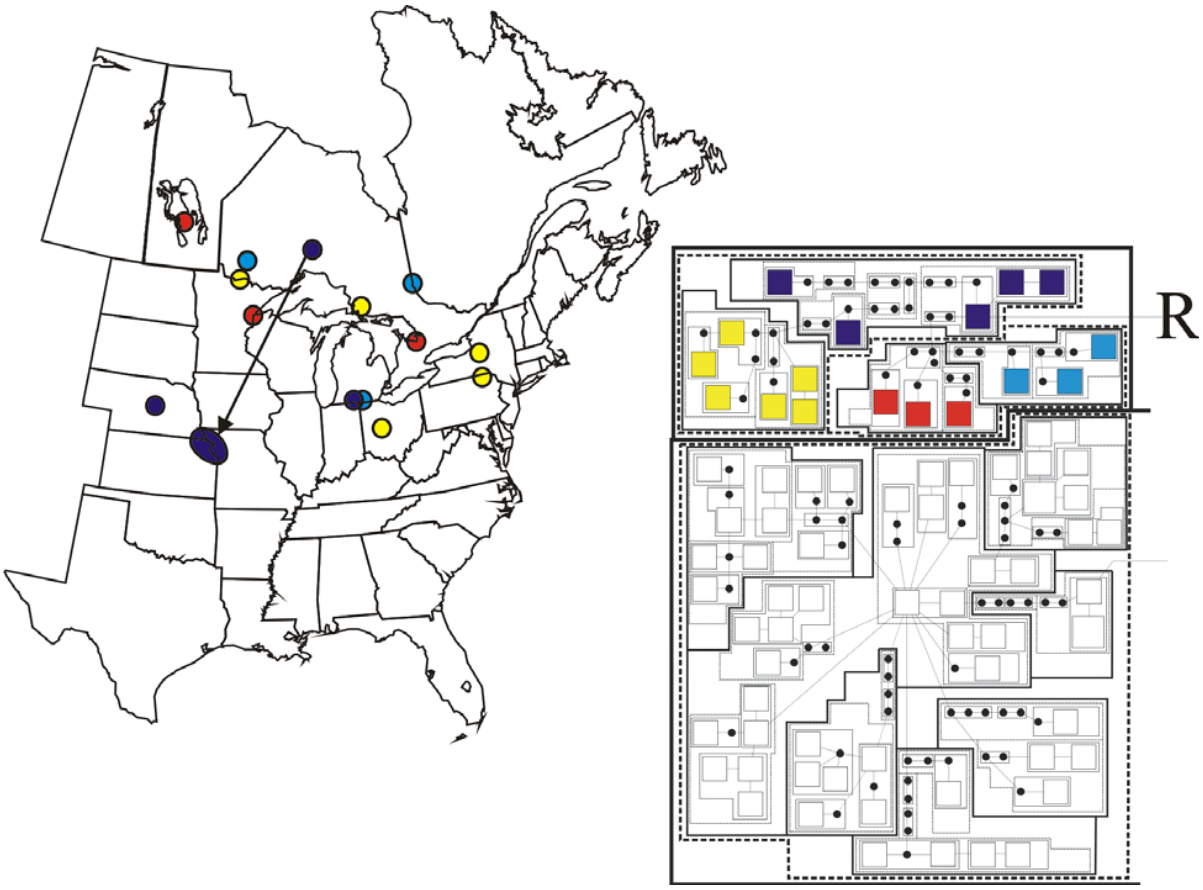
Clade 4-3 consists of seven 3-step clades (3-5 to 3-11) that were predominately distributed throughout the Great Lakes/St. Lawrence watershed (fig. 4.3). The major exceptions were clades 3-7 and 3-11, which were located in northern Quebec and southern Tennessee, respectively. Many of the 3-step clades (3-7 to 3-11) were separated from the interior of the network by edges of inferred haplotypes (appendix C, fig. C.12a). Clade 3-7 was collected from localities in northern Quebec, whereas clades 3-8 and 3-9 were restricted to habitats located in southern Michigan. Clade 3-10 was distributed throughout New York and Quebec (fig. 4.3). Interestingly, clade (3-11) intervening between clades 4-3 and 4-4 possessed a distribution isolated in the Interior Plains of Tennessee and Georgia (fig. 4.3). Clade 4-3 NCA returned multiple possible inferences: long distance colonization, past fragmentation of a larger range, past range expansion followed by subsequent fragmentation or past fragmentation followed by

range expansion. This outcome was derived from several tip clades with significantly small clades distances ($Dc_{(3\text{-}7)}$=16.07, $p<0.001$, $Dc_{(3\text{-}8)}$=37.56, $p<0.001$, $Dc_{(3\text{-}10)}$=176.54, $p<0.001$) and an interior clade (3-11) possessing significantly small clade and large nested clade distances (Dc=80.65, $p<0.001$, Dn=1210.45, $p<0.001$, respectively). The dominant spatial pattern the NCA identified was the large geographic distance between the distribution of clade 3-11 and the other 3-step clades. This likely represents a long distance colonization event within clade 5-2, which traces back its origin through a haplotype (S3H164) located in northern Quebec (QC020). According to the NCA, at some point since the retreat of the Pleistocene glaciations, species 3 individuals colonized the watersheds of Hudson Bay in an area consistent with the glacial lake Ojibway (Pielou, 1991) prior to expanding into the Great Lakes and St. Lawrence River regions.

Table 4.2: Nested clade analysis with inference key for species 3
The inference key shows the decision-tree along with the conclusion. AF = Allopatric Fragmentation, CRE = Contiguous Range Expansion, LDC = Long Distance Colonization, LDD, IBD = Isolation by Distance, IO = Inconclusive Outcome, IGS = Inadequate Geographical Sampling, PRE = Past Range Expansion followed by subsequent fragmentation, PFR= Past Fragmentation followed by Range Expansion and PF = Past Fragmentation of a Larger Range. See appendix for full table of Dc, Dn and p-values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key
5-1	No Inference
4-1	No Inference
3-1	1-2-11-17-NO: IO
3-2	1-2-11-17-NO: IO
5-2	1-2-3-4-NO: IBD
4-3	1-2-3-5-6-13-YES: LDC, PF, PRE or PFR
3-5	1-2-11-17-4-NO: IBD
3-6	1-2-3-4-NO: IBD
2-14	1-2-3-4-NO: IBD
3-10	No Inference
2-21	No Inference
1-47	1-19-20-2-11-12-13-14-YES: CRE or LDC or PF
3-11	No Inference
2-23	No Inference
1-50	1-19-20-NO: IGS
4-4	1-2-3-4-NO: IBD
3-12	1-2-3-4-NO: IBD
2-24	1-2-3-5-6-7-8-NO: IBD or LDD
2-30	No Inference
1-68	1-19-20-2-3-4-NO: IBD
2-37	1-2-3-4-NO: IBD
3-13	1-19-20-2-11-17-NO: IO
2-39	No Inference
1-90	1-2-3-4-NO: IBD
3-14	No Inference
2-42	1-2-3-4-NO: IBD
3-15	1-2-3-4-NO: IBD
2-43	1-2-3-4-NO: IBD
1-95	1-2-11-12-NO: CRE
2-44	1-19-20-2-3-4-9-NO: AF
3-17	1-19-20-2-3-4-NO: IBD

Figure 4.2: Map showing the geographical distribution of species 3 clade 5-1
Clade 5-1 haplotypes possess a geographic distribution represented on the map (left) and their corresponding network positions are indicated on the insert (right). Nested within clade 5-1 are clades 3-1 (yellow), 3-2 (purple), 3-3 (red) and 3-4 (blue). The arrows on the map indicate potential dispersal routes. The root is denoted by (R).



If clade 3-11 was excluded from the analysis of clade 4-3, because its inclusion was likely an artifact of the nesting procedure (the haplotypes were one mutational step from clade 4-4), then the genetic patterns of the remaining 3-step clades can be assessed. The 3-step clades with the small clade distances (3-7, 3-8 and 3-10) had distributions that were mostly non-overlapping with interior clades and their positions in the network were separated from the ancestral groups (appendix C, fig. C.12) by a high number of inferred mutational steps (3 to 5 mutations). The inference key interprets this pattern as allopatric fragmentation, splitting groups within the Great Lakes region.

Nested within clade 4-3 were several clades with significant geographical associations (clades 3-6, 2-14, 1-47 and 1-50). Clades 3-6 and 2-14 were hierarchically related to each other and the inference key indicated isolation by distance for both (table 4.2) as a result of tip clades with significantly small clade and nested distances (clade 3-6: $Dc_{(2\text{-}13)}$=215.25, *p=0.013,* $Dn_{(2\text{-}13)}$=224.19, *p<0.001*, $Dc_{(2\text{-}14)}$=204.76, *p<0.001* and 2-14: $Dc_{(1\text{-}36)}$=0, *p<0.001*). The inference for clade 2-14 was further supported by the interior clade (1-34) that was collected from two geographically disparate habitats, one in Indiana and the other in West Virginia. This resulted in a significantly large distance between the interior and tip clade distances ($I\text{-}T_{Dc}$ = 211.52, *p<0.001*), which was consistent with isolation by distance. Clade 1-47 had a tip haplotype (S3H178) with a significantly small clade and large nested clade distance (Dc=38.87, *p<0.001*, Dn=306.89, *p<0.001*, respectively) and an interior haplotype (S3H180) with significantly small clade and nested clade distances (Dc=74.01, *p<0.001,* Dn=130.27, *p<0.001*, respectively). Essentially, the interior haplotype (S3H180) was restricted to populations within Quebec, whereas haplotype S3H178 was located in the New York Finger Lakes region. The inference key interprets this as a contiguous range expansion, long distance colonization or past fragmentation

of a larger range. The NCA of clade 1-50 may indicate allopatric fragmentation, but additional sampling of populations in Alabama and Tennessee are required.

Clade 4-4 (appendix C, fig. C.12b and fig. C.12c) was composed of a broadly distributed interior clade 3-12 (fig. 4.4) and numerous tip clades (3-13 to 3-17). These tip clades possessed more restricted ranges (fig. 4.5 and 4.6), but when conjoined, they approximate the region occupied by the interior clade 3-12. The nested clade analysis of clade 4-4 indicated isolation by distance because the interior clade (3-12) had a significantly large nested clade distance (Dn=732.97, $p=0.008$) and several tip clades possessed significantly small clade and nested clade distances ($Dc_{(3\text{-}14)}$=245.10, $p<0.001$, $Dc_{(3\text{-}15)}$=289.11, $p<0.001$, $Dc_{(3\text{-}17)}$=346.80, $p=0.001$, $Dn_{(3\text{-}17)}$=344.67, $p<0.001$). Additionally, the I-T clade and I-T nested clade distance were significantly large ($I\text{-}T_{Dc}$=401.79, $p<0.001$, $I\text{-}T_{Dn}$=129.53, $p=0.003$, respectively), which would infer isolation by distance even if none of the tip clades had significantly small clade distances.

Figure 4.3: Map showing the geographical distribution of species 3 clade 4-3
Clade 4-3 haplotypes possess a geographic distribution represented on the map (left) and their corresponding network positions are indicated on the insert (right). Nested within clade 4-3 are clades 3-5 (orange), 3-6 (green + yellow + blue), 3-7 (dark purple), 3-8/3-9 (dark green), 3-10 (light purple) and 3-11 (red). The root is denoted by (R).

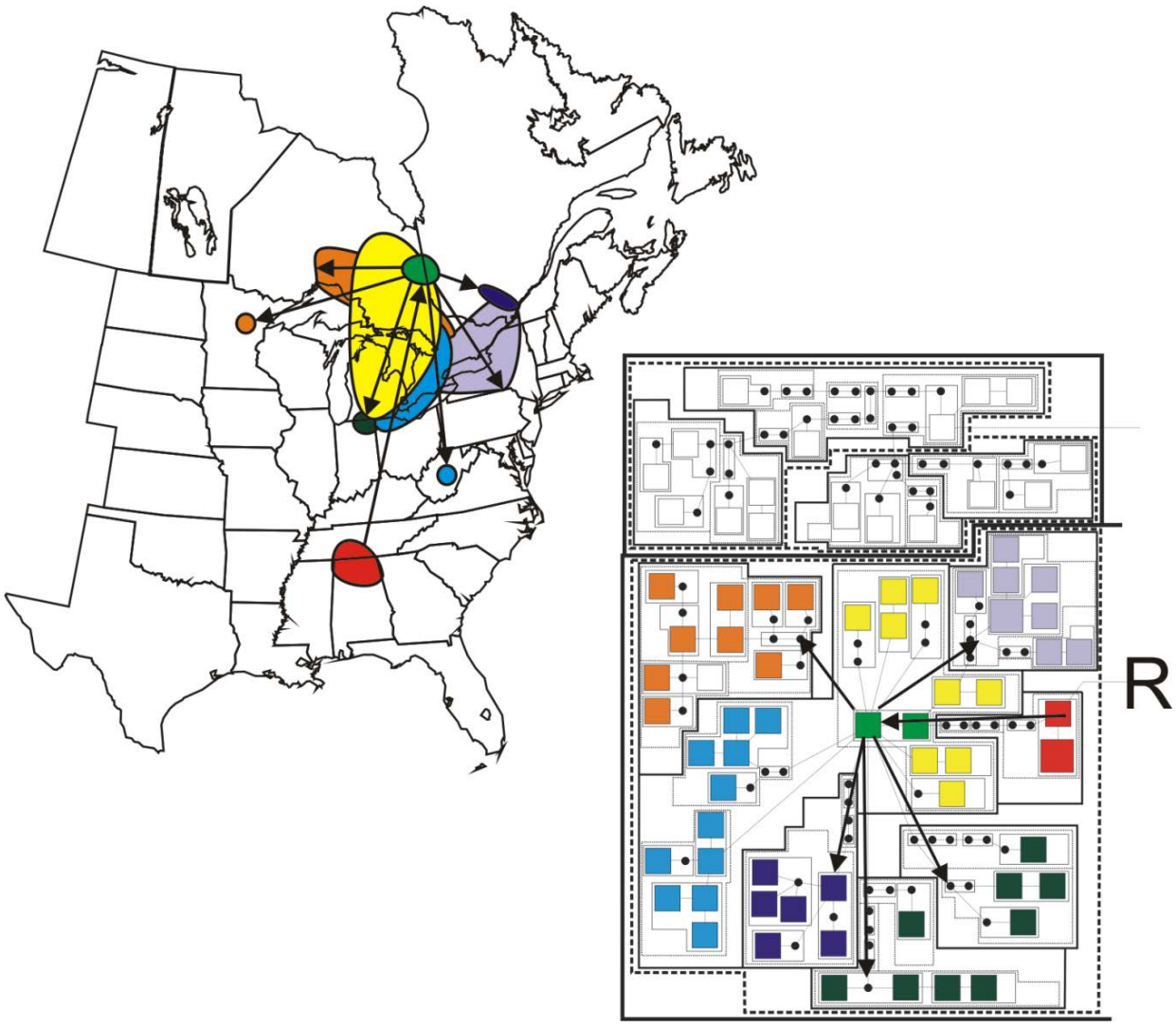


Figure 4.4: Map showing the distribution of species 3 clade 3-12
Clade 3-12 haplotypes possess a geographic distribution represented on the map (left) and their corresponding network positions are indicated on the insert (right). Nested within clade 3-12 are haplotype S3H070 (green), clade 1-49 (green + red) and clade 2-24 (green + red + blue). The distribution of tip clades nested within clade 3-12 is indicated by the yellow regions. The root is denoted by (R).

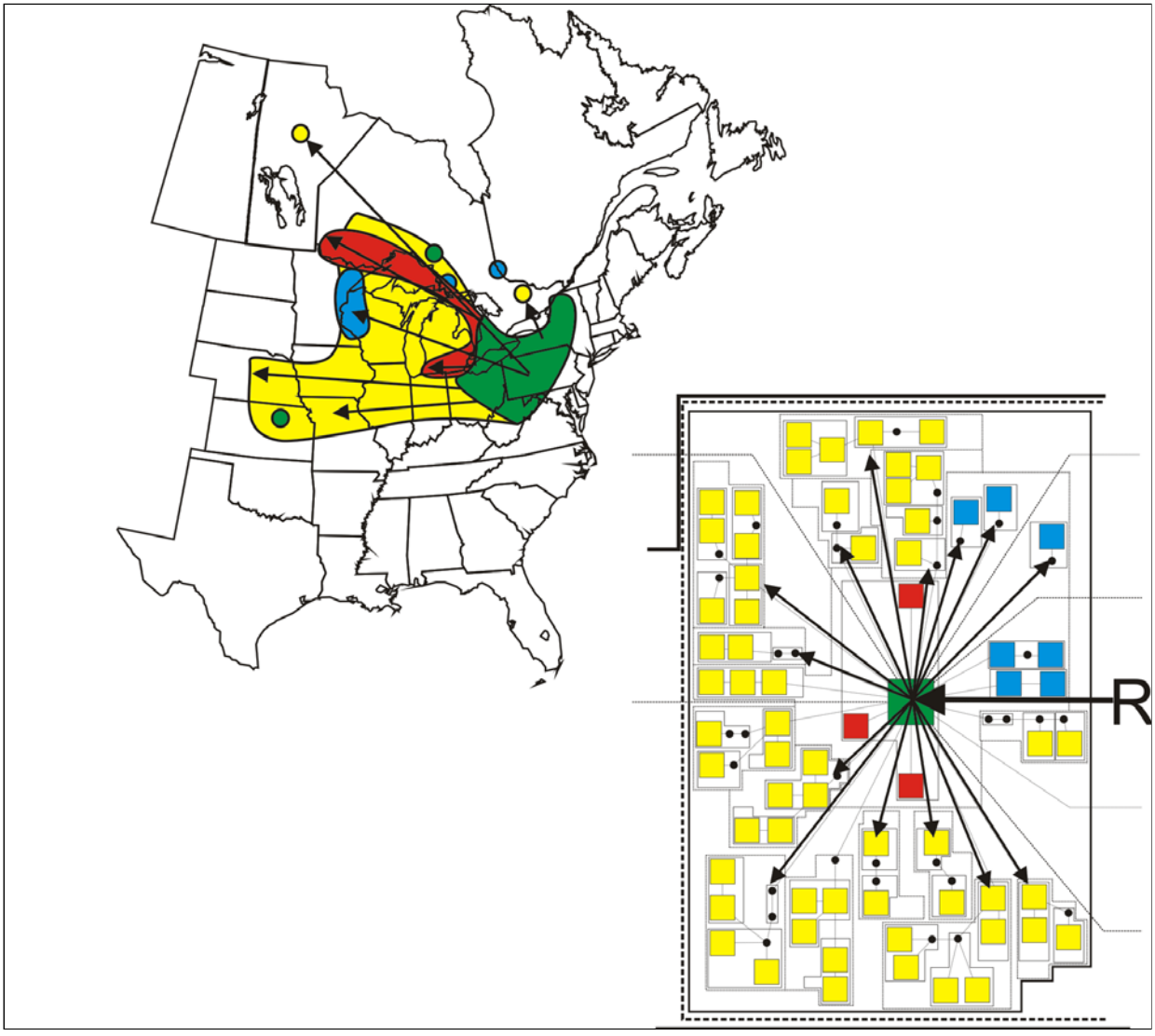


Figure 4.5: Map showing the distribution of species 3 clades 3-13, 3-14, 3-16 and 3-17
The haplotypes of clades 3-13, 3-14, 3-16 and 3-17 have geographic distributions represented on the map (left) and their corresponding network positions are indicated on the insert (right). The distributions of the 3-step clades are color coded as follows: clade 3-13 (light purple), clade 3-14 (red + yellow), clade 3-16 (dark green + dark purple) and clade 3-17 (blue). These clades connect directly to haplotype S3H070 (fig. 4.5) whose distribution is marked by green on the map.

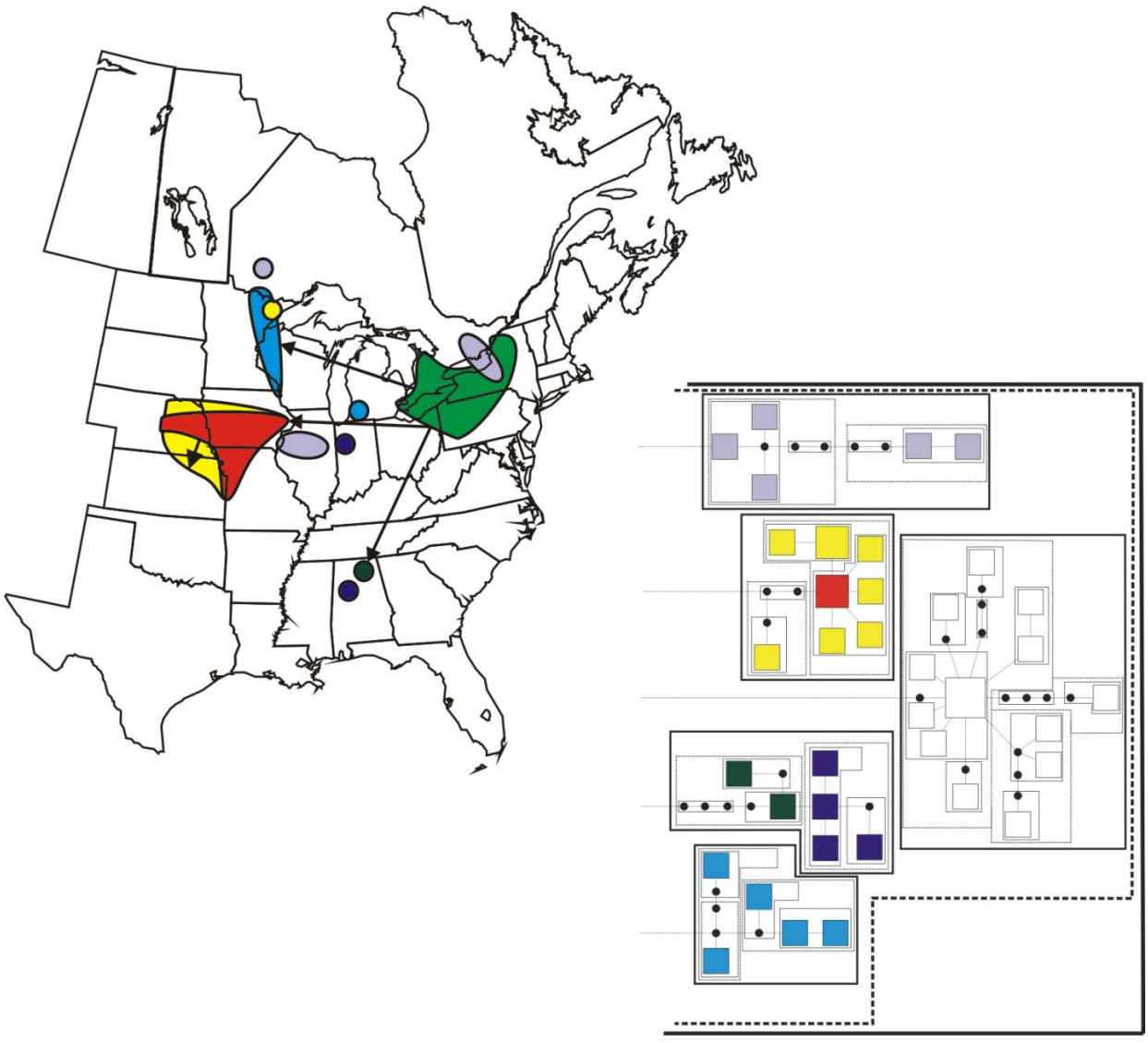
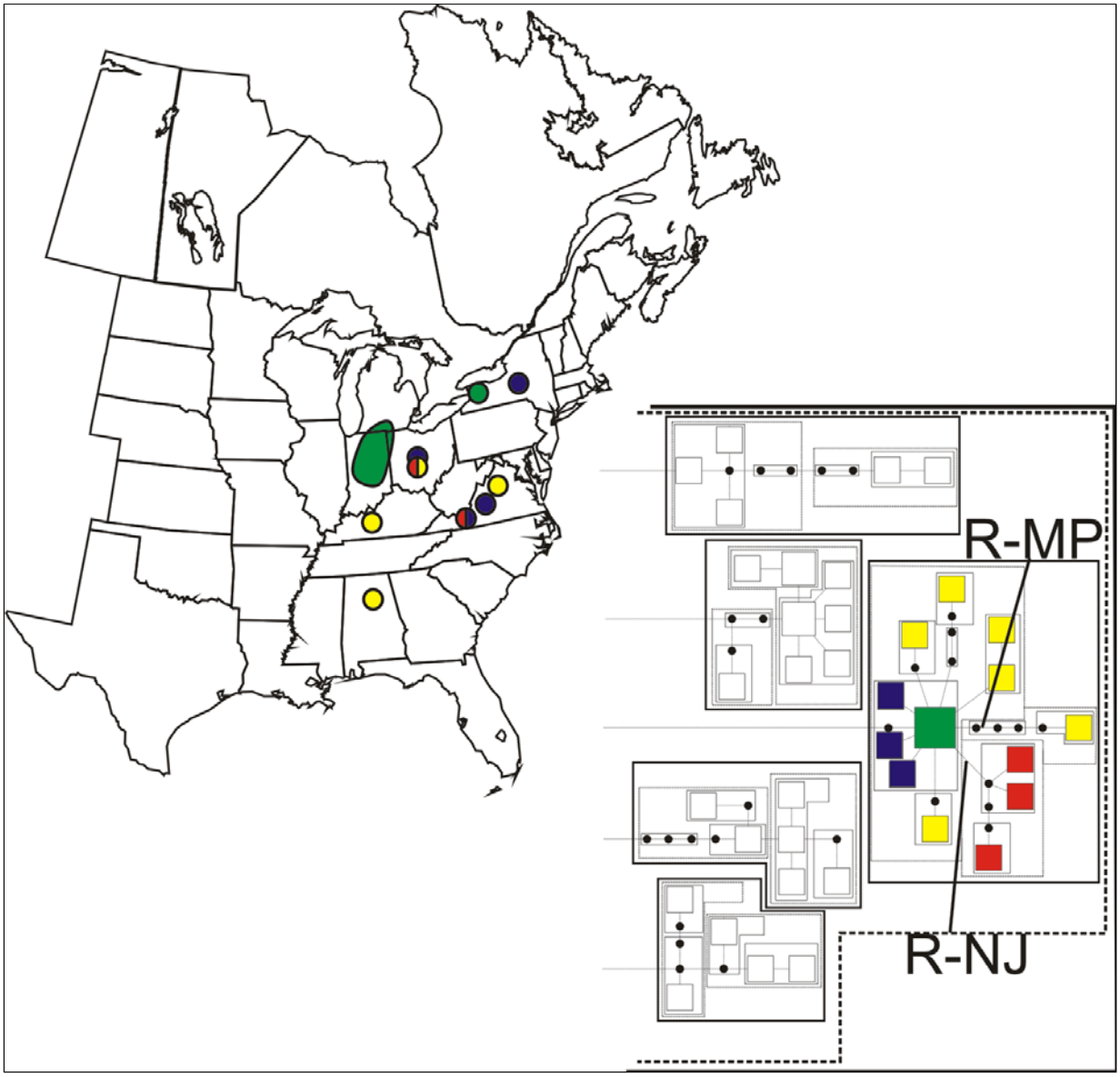


Figure 4.6: Map showing the distribution of species 3 clade 3-15
Clade 3-15 haplotypes have geographic distributions represented on the map (left) and their corresponding network positions are indicated on the insert (right). The distributions are color coded as follows: haplotype S3H002 (green), clade 1-92 (green + purple), clade 2-43 (green + purple + yellow) and clade 2-45 (red). R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.



The remaining clades nested within clade 4-4 with 12 significant geographical associations (clades 3-12, 3-13, 3-15, 3-17, 2-24, 2-37, 2-42, 2-43, 2-44, 1-67, 1-90 and 1-97). Clade 3-12 possesses an interior haplotype (S3H070) that was detected in 12 habitats with 37 connecting edges in the haplotype network (appendix C, fig. C.12b). This pattern approaches a star phylogeny and is consistent with a range expansion. The spatial pattern of clade 3-12 is interesting. The interior clade is distributed throughout the Great Lakes and St. Lawrence River watersheds along with a disjunct population in Kansas (fig. 4.4). This clade (2-24) bears evidence of long distance colonization because its tip clade (1-53) had a significantly small clade distance and large nested clade distance (Dc=0, $p=0.006$, Dn=1704.94, $p<0.001$). However, at the 3-step level, the dominant pattern is isolation by distance. The NCA of clade 3-12 detected an interior clade (2-24) with a significantly large nested clade distance (Dn=823.13, $p=0.002$) and several tip clades with significantly small clade and nested clade distances ($Dc_{(2\text{-}25)}$=41.93, $p<0.001$, $Dc_{(2\text{-}26)}$=86.96, $p<0.001$, $Dc_{(2\text{-}29)}$=132.99, $p<0.001$, $Dc_{(2\text{-}32)}$=0, $p<0.001$, $Dn_{(2\text{-}32)}$=204.19, $p<0.001$, $Dc_{(2\text{-}34)}$=31.65, $p<0.001$ and $Dc_{(2\text{-}37)}$=332.72, $p<0.001$). As such, the colonization of both glaciated and unglaciated regions by clade 3-12 individuals appears to have originated from Atlantic drainages. The colonization of habitats progressed westward into the Great Lakes and Mississippi River watersheds through a combination of long distance colonization and short distance dispersal. Clades 2-37 and 1-68 were also nested within clade 3-12 and the inference key indicated isolation by distance for both, as a consequence of tip clades possessing significantly small clade distances (clade 2-37: $Dc_{(1\text{-}83)}$=56.32, $p=0.012$; clade 1-68: $Dc_{(S3H029)}$=0, $p=0.012$).

Three of the 3-step clades nested within clade 4-4 had significant geographical associations (3-13, 3-15, and 3-17), but the inference key returned an inconclusive outcome for clade 3-13.

Clade 3-15 has a central haplotype (S3H002) with a large sample size (10 haplotypes within 4 localities) and two tip clades (2-44 and 2-45) separated by multiple inferred mutations (appendix C, fig. C.12c). The interior haplotype S3H002 is located in populations in Indiana, Michigan and New York, whereas the tip clades further expand the range of clade 3-15 into habitats throughout the New York Finger Lakes region as well as the Shenandoah Valley in Virginia (fig. 4.6). The clade 3-15 NCA indicated isolation by distance because the tip clade (2-44) had significantly small clade and nested clade distances (Dc=149.91, *p=0.003*, Dn=174.33, *p<0.001*, respectively) as well as an interior clade (2-43) possessed significantly large clade and nested clade distances (Dc=317.34, *p=0.002*, Dn=318.62, *p=0.002*, respectively). Two clades nested within clade 3-15 had significant geographical associations: clades 2-43 and 1-95. The NCA of clade 2-43 indicated isolation by distance because the tip clades (1-93 and 1-95) had significantly small clade distances (Dc=0, *p<0.001*, Dc=0, *p<0.001*, respectively). Alternatively, the NCA of clade 1-95 indicated a contiguous range expansion because the interior haplotype (S3H002) had a significantly small clade distance (Dc=147.16, *p<0.001*).

Clade 3-17 is distributed within southern Michigan and the western Great Lakes region (fig. 4.5). The NCA indicated a genetic pattern of isolation by distance because the tip clade (2-49) has significantly small clade and nested clade distances (Dc=165.65, *p=0.001*, Dn=322.80, *p=0.007*, respectively)

At the lower tiers, there were five additional geographic associations nested within clade 4-4. The NCA of clades 1-90, 2-42 and 2-43 indicated isolation by distance because the tip

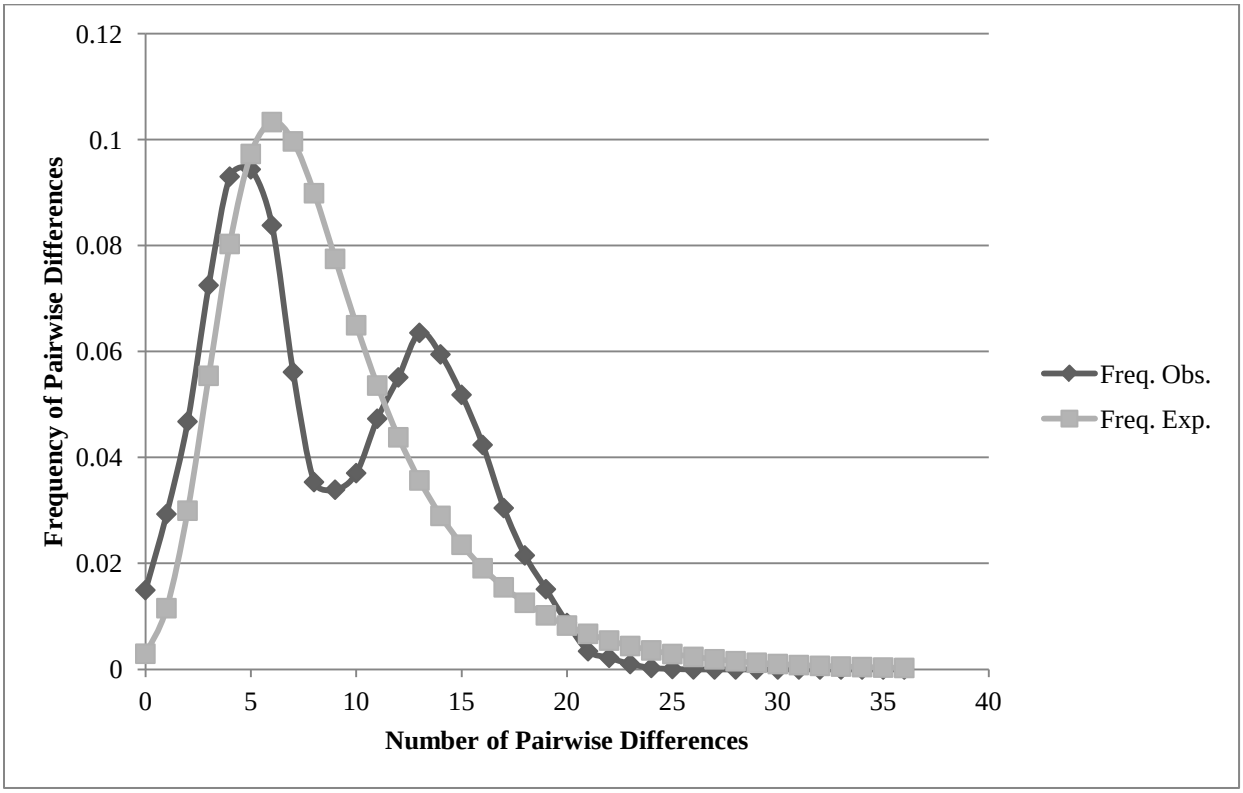
clades were associated with significantly small clade and nested clade distances (clade 1-90: $Dc_{(S3H016)}$=0, *p=0.012*. Clade 2-42: $Dc_{(1-90)}$=190.50, *p=0.001*, $Dn_{(1-90)}$=226.37, *p=0.004*. Clade 2-43: $Dc_{(1-93)}$=0, *p=0.001*, $Dc_{(1-95)}$=0, *p<0.001*). The NCA of clade 1-95 indicated contiguous range expansion because the interior haplotype (S3H002) had a significantly small clade distance (Dc=147.16, *p<0.001*). This haplotype is restricted to habitats in Indiana, New York and Michigan, whereas the tip haplotypes were located in the Shenandoah Valley, Ohio and New York (fig. 4.6). The inference key returned a pattern of allopatric fragmentation for the analysis of clade 2-44 because the interior clade (1-97) possessed a significantly large nested clade distance (Dn=176.58, *p=0.018*), but is allopatric with respect to its tip clade (1-98). The interior clade occured in Ohio, the tip clade was detected in a non-overlapping region within the Shenandoah Valley (fig. 4.6).

4.3.5 Mismatch distribution and population genetics of species 3

The clade 3 mismatch distribution was bimodal (fig. 4.8) with a significant raggedness statistic (*r* =0.004, *p < 0.001*). The smooth bimodal distribution suggests that two expansions may have occurred. While this may indicate expansions of different ages, two co-occurring expansions involving haplotypes in distinct sections of a network will produce a bimodal distribution.

Tajima's *D*, Fu's *F* and R_2 all indicated significant deviations from neutrality (*D*=-2.092, *p=0.001*, *F*=-260.098, *p<0.001* and R_2=0.0262, *p=0.005*, respectively). These results were consistent with a lineage undergoing demographic expansion.

Figure 4.7: Mismatch distribution of species 3
The mismatch distribution of species 3 COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with tau = 4.550.



4.3.6 Analysis of molecular variance of species 3

The seventy-nine populations were partitioned into five faunal provinces (Central Appalachian: 4 populations, Great Lakes: 30 populations, Hudson Bay: 11 populations, Mississippi: 33 populations, Southeastern: 1 population). The between group component of the AMOVA partitioned into faunal provinces (table 4.3) explained 10.21% ($p<0.001$) of the variance. The AMOVA partitioned by unglaciated (45 populations) and glaciated regions (34 populations) had a between-group component that explained 14.13 % of the variance.

Table 4.3: AMOVA and SAMOVA of species 3
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * $p<0.01$, ** $p<0.001$, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	10.21**	43.13**	46.16**
Unglaciated and Glaciated	14.13**	40.97**	44.90**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	22.98*	38.37**	38.65**
3	43.02**	27.48**	29.51**
4	41.96**	26.37**	31.67**
5	40.41**	25.98**	33.60**
6	45.32**	19.98**	34.70**

The 3-group SAMOVA had a between-group component that explained 43.02% of variance (table 4.4). Neither the 2-group SAMOVA nor the higher level SAMOVA appreciably increased the between-group variance component. The 3-group SAMOVA resulted in two major partitions of 74 and 4 populations, respectively. The smaller group consisted of 4 populations from Quebec (QC002, QC024, QC028 and QC046), and included haplotypes in clades 3-7 and 3-10. Both of these clades were separated from the interior clades by multiple inferred haplotypes and the nested clade analysis of clade 4-3 (in which both 3-step clades were nested) indicates the potential for fragmentation of a past larger range or long distance colonization. The

SAMOVA suggests that the populations in Quebec were distinct from the rest of species 3. A single habitat in Iowa (IA005) was the only member of the third partition.

4.4 RESULTS: SPECIES 4

Sixty-two species 4 individuals were collected from 13 sites that were located in the southeastern United States along with an additional locality in New Brunswick. Species 4 co-occurred with other small bodied species in four habitats, and large bodied species within two (appendix A).

4.4.1 Nested COI haplotype network of species 4

The statistical parsimony network consisted of three distinct lineages A, B and C with a maximum 95% limit of 10 mutational steps, and two non-grouped haplotypes (S4H033 and S4H037). The majority of haplotypes (32) were nested within lineage A (appendix C, fig. C.13). Lineage A consisted of a 5-step clade composed of twenty 1-step clades (1-1 to 1-20), twelve 2-step clades (2-1 to 2-12), four 3-step clades (3-1 to 3-4) and two 4-step clades (4-1 and 4-2). Lineage B consists of three haplotypes that form a network composed of two 1-step clades (1-21 and 1-22) and a single 2-step clade (2-14). Lineage C has three haplotypes, but forms a single 1-step clade (1-23). Two remaining haplotypes (S4H033 and S4H037) were at least 10 mutational steps from any other haplotype.

Lineage A haplotypes were geographically divided into a group within the southeastern Gulf and Atlantic drainages, as well as a western assemblage near the Mississippi River (fig. 4.9). Lineage B was only detected in Louisiana (LA001) and lineage C in a roadside pond near the Arkansas River (fig. 4.9). The two haplotypes that didn't group with the other lineages were

discovered in Juniper's Lake, FL, (S4H033) and the Digdoguash River, New Brunswick (S4H037).

4.4.2 Outgroup analysis of species 4

The Tamura-Nei model of molecular evolution was selected by the Akaike information criterion with gamma distributed rate variation (α=0.159). The parsimony and neighbor-joining analyses placed the roots within clades 2-9 and 2-10, respectively. The species 1 lineage G haplotype (S1H209) was used as the outgroup.

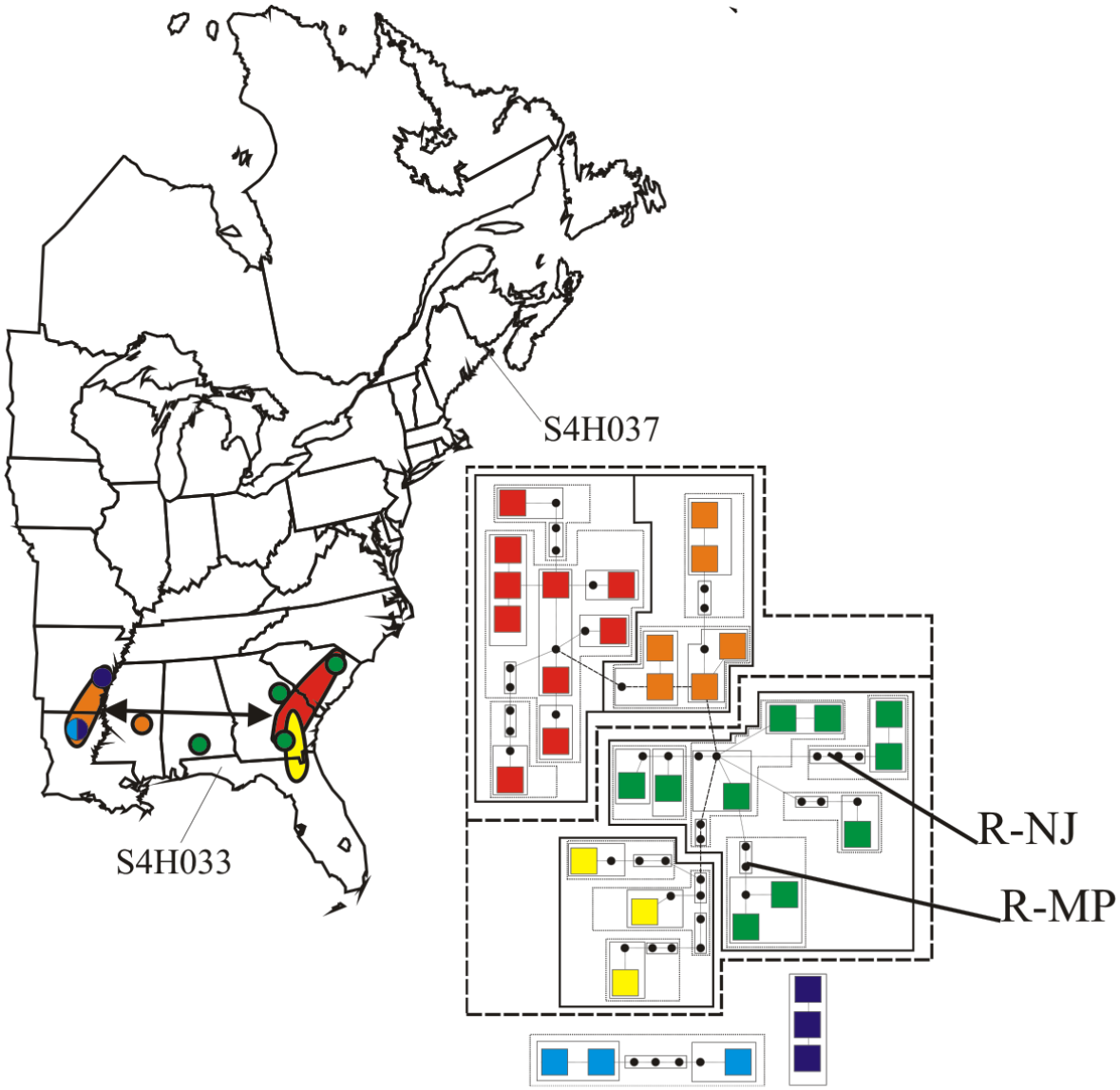
4.4.3 Nested clade analysis of species 4

The NCA of lineage A detected two significant geographical associations (table 4.4), but the haplotypes within the remaining networks had insufficient sample size or localities. The NCA of clade 4-1 returned multiple possible outcomes, including fragmentation of a past larger range, fragmentation followed by a range expansion or long distance colonization. These resulted from the interior clade (3-2) having a significantly small clade distance (Dc=146.07, *p=0.017*), and the tip clade (3-1) possessing significantly small clade and large nested clade distances (Dc=126.11, *p<0.001*, Dn=480.68, *p<0.001*, respectively). The separation of clade 4-1 into disjunct regions (fig. 4.10) may indicate the fragmentation of a historically larger range or long distance colonization. Interestingly, the position of clade 3-2 in the network is between clades 3-1 and 3-3, both of which have distributions within the Atlantic drainages. However, the habitats containing clade 3-2 haplotypes were located along the Gulf. On the basis of the roots inferred by neighbor-joining and maximum parsimony, clade 3-3 is the ancestral group, which implies repeated migration between the Gulf and Atlantic regions. The NCA of clade 3-3 indicated isolation by distance because the I-T clade and nested clade distances were significantly large (I-T_{DC}=292.24, *p<0.001*, I-T_{Dn}=228.15, *p=0.005*, respectively). Root ambiguity had no effect on the NCA inferences.

Table 4.4: Nested clade analysis with inference key of species 4
The #-step clades refer the nested network (fig. 4.8). The inference key numbers refer to the progression of steps outlined in Templeton (2011). LDC = Long Distance Colonization, PLR = Past Larger Range with subsequent extinction, PF = Past Fragmentation followed by Range Expansion and IBD = Isolation by Distance. Refer to appendix for full table of Dc, Dn and p-values.

Clade	Inference Key
4-1	1-2-11-12-13-YES: PLR, LDC or PF
3-3	1-2-11-17-4-NO: IBD

Figure: 4.9: Map showing distribution of species 4
The distributions of the clades are denoted on the map (left) and refer to the clades color coded on the network (right). The clades are denoted as follows: Clade 3-1 (red), clade 3-2 (orange), clade 3-3 (green), clade 3-4 (yellow), clade 2-14 (blue) and clade 1-23 (purple). The roots are denoted by R-NJ and R-MP and refer to the neighbor-joining and maximum parsimony analyses, respectively.

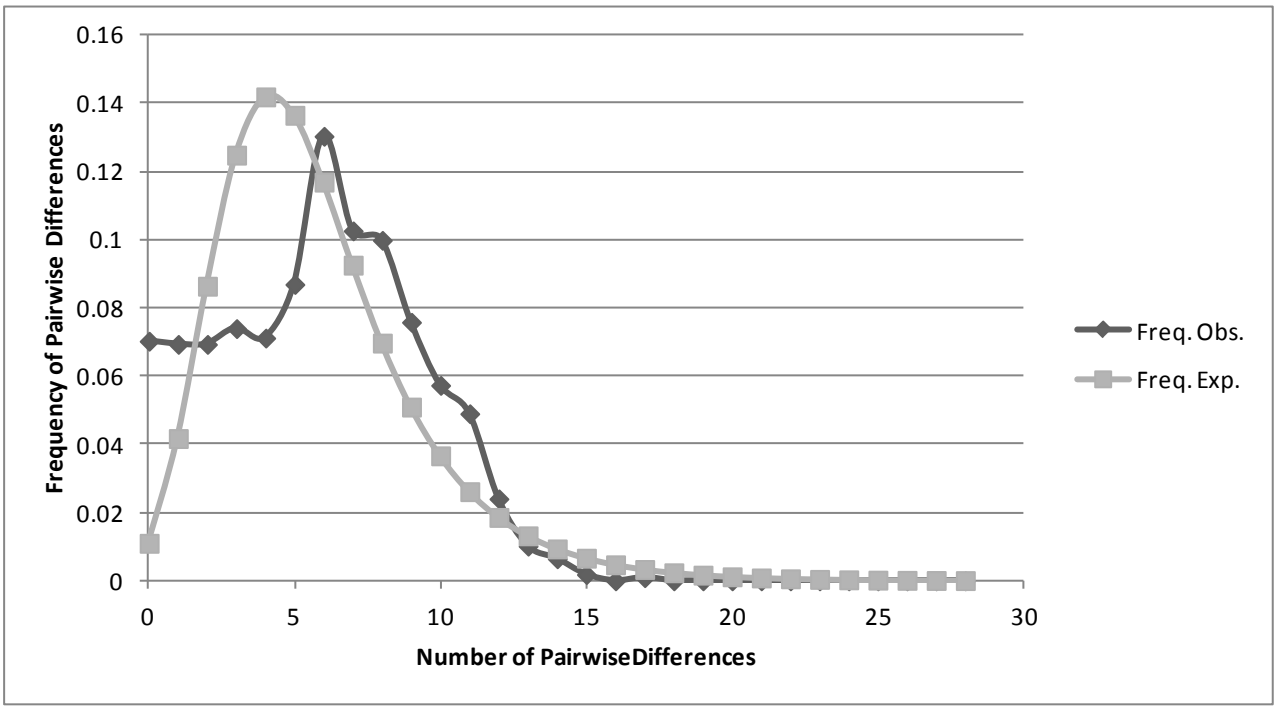


4.4.4 Mismatch distribution and population genetics of species 4

The mismatch distribution of lineage A (fig. 4.10) resulted in a ragged multimodal line and the raggedness statistic rejected the null hypothesis of population expansion (r=0.005, $p<0.001$) when compared against a simulated population undergoing constant demographic expansion (tau=3.361). The other clade 4 lineages (B through E) had sample sizes too small to conduct a mismatch analysis, and the inclusion of those haplotypes with lineage A did not significantly alter the distribution (not shown).

Tajima's D, Fu's F and R_2 statistics indicated significant departures from neutrality consistent with a demographic expansion (D=-2.01, $p<0.001$, F=-15.69, $p<0.001$ and R_2=0.0433, $p=0.002$).

Figure 4.9: Species 4 lineage A mismatch distribution
The mismatch distribution of species 4 lineage A COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with tau = 3.361.



4.4.5 Analysis of molecular variance of species 4

The thirteen populations were partitioned into four faunal provinces (Central Appalachian: 1 population, Great Lakes: 1 population, Mississippi: 2 populations, Southeastern: 9 populations). The between group component of the AMOVA partitioned into faunal provinces (table 4.5) explained 20.92% ($p<0.001$) of the variance. Species 4 was not detected in previously glaciated regions, with the exception of the single location in New Brunswick.

Table 4.5: AMOVA and SAMOVA of species 4
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * p<0.01, ** p<0.001, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	22.95^{NS}	31.64**	45.41**
Unglaciated and Glaciated	----	----	----

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	56.67^{NS}	19.84**	23.49**
3	34.05*	30.56**	35.40**
4	51.90**	9.46**	38.64**
5	46.04**	10.53**	43.43**
6	53.29**	2.90**	43.81**

The 5-group SAMOVA had a between-group component that explained 46.04% of variance (table 4.4) and was the first setup to produce two or more partitions of at least 3 populations. The first assemblage consisted of three disjunct populations from Arkansas, Louisiana and New Brunswick (AR001, CH001 and NB002). The second group had 7 populations located in the southeastern section of the range (AL002, FL017, GA003, GA005, GA007, GA008 and NC006), which is the most consistent partition at all levels of the SAMOVA.

4.5 RESULTS: SPECIES 5

Three-hundred and three species 5 individuals (130 haplotypes) were collected from 72 localities (appendix A). These small-bodied amphipods were rarely discovered with other large-bodied morphotypes (only 11% of sites) and more commonly co-occurred with other small-bodied *Hyalella* species (42% of sites).

4.5.1 Nested COI haplotype network of species 5

The statistical parsimony network returned five distinct lineages (A to E) and an ungrouped haplotype (S5H102) on the basis of the parsimony criterion (maximum connection limit at 95% was 10 mutational steps). Lineage A comprised the largest of the five networks (187 sequences) and was detected in 55 locations (appendix C, fig. C.13). The 73 haplotypes were nested into forty-two 1-step clades (1-1 to 1-42), twenty 2-step clades (2-1 to 2-20), nine 3-step clades (3-1 to 3-9), three 4-step clades (4-1 to 4-3) and the total cladogram consisted of 5-steps. The nested network of species 5 lineage B (appendix C, fig. C.14) consisted of twenty-eight haplotypes (S5H103 to S5H130) nested into fifteen 1-step clades (1-43 to 1-57), seven 2-step clades (2-21 to 2-28), three 3-step clades (3-10 to 3-12) and a single 4-step clade (4-4). Species 5 lineage C (fig. 4.14) possessed nine haplotypes nested into six 1-step clades (1-58 to 1-63), three 2-step clades (2-29 to 2-31), two 3-step clades (3-13 and 3-14) and the total cladogram consisted of 4-steps (4-5). Species 5 lineage D (appendix C, fig. C.15) had four haplotypes nested in a single 1-step clade (1-64). Similarly, species 5 lineage E (appendix C, fig. C.15) had four haplotypes, though these were nested into three 1-step clades (1-65 to 1-67), three 2-step clades (2-32 to 2-34) and a single 3-step clade (3-15).

4.5.2 Outgroup analysis of species 5

The HKY model of molecular evolution was selected by the Akaike information criterion with a proportion of invariant sites (I=0.408) and gamma distributed rate variation (α=0.875). The parsimony and neighbor-joining analyses placed the roots within clades 4-2 and 4-3, respectively. The root inferred by maximum parsimony was along the edge between clade 2-15 and 2-17, and the root inferred by neighbor-joining was within clade 2-20 between haplotypes S3H072 and S3H073. Species 1 lineage G haplotype (S1H209) was used as the outgroup.

.

4.5.3 Nested clade analysis of species 5

The nested clade analyses of lineages A and B returned significant geographical associations (table 4.6), while the remaining lineages did not have sufficient sample sizes. Of the ten clades in lineage A with significant permutation tests, five were best modeled by a pattern of restricted gene flow with isolation by distance, four by contiguous range expansions and one with multiple potential inferences (contiguous range expansion, long distance colonization or past fragmentation).

Table 4.6: Nested clade analysis with inference key of species 5 lineage A
The #-step clades refer to the nested network specific for species 5 (fig. 4.11). The inference key numbers refer to the progression of steps outlined in Templeton (2011). CRE = Contiguous Range Expansion, LDC = Long Distance Colonization, PF = Past Fragmentation, LDD = Long Distance Dispersal and IBD = Isolation by Distance with Restricted Gene Flow. Only lineage A had significant geographical associations. See the appendix for Dc, Dn and p-values. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key
5-1	1-2-3-5-6-7-YES: IBD with LDD
4-1	1-2-11-12-13-14-YES: CRE, LDC or PF
3-1	No Inference
2-1	1-2-3-4-NO: IBD
3-4	1-2-11-12-NO: CRE
3-5	1-2-11-12-NO: CRE
2-13	1-2-3-4-NO: IBD
1-27	1-2-3-4-NO: IBD
4-2	No Inference
3-6	1-2-3-4-NO: IBD
4-3	1-2-11-12-NO: CRE
3-8	1-2-11-12-NO: CRE

The NCA of the total lineage A cladogram indicated a pattern of isolation by distance with long distance dispersal (regardless of root ambiguity). This resulted from the tip clade 4-1 having a significantly small clade distance (Dc=928.66, *p=0.006*) and the tip clade 4-3 having significantly small clade and large nested clade distances (Dc=44.14, *p<0.001*, Dn=1492.57, *p=0.005*, respectively). The interior clade 4-2 has a very broad range, consisting of habitats in at least three of the disjunct regions occupied by lineage A: New Brunswick, South Carolina and the Gulf drainages in the southeastern USA (these sites are denoted by the orange and dark purple regions in fig. 4.11). Comparatively, the tip clades 4-1 and 4-3 possess far more limited geographical distributions. The roots inferred by the neighbor-joining and maximum parsimony analysis indicate that habitats along the Gulf drainages may contain the ancestral haplotypes to lineage A.

The NCA of clade 4-1 inferred a range expansion because the interior clade 3-1 had significantly small clade and nested clade distances (Dc=409.62, *p<0.001*, Dn=651.03, *p<0.001*, respectively). However, the tip clade 3-5 had significantly small clade and large nested clade distances (Dc= Dc=687.774, *p<0.001*, Dn=1129.09, *p<0.001*, respectively), which indicates the potential for long distance colonization or past fragmentation of a larger range. The interior clade (3-1) has a distribution that encompasses habitats in New Brunswick, Maine, Quebec and Ontario (fig. 4.11). The tip clades were located in ice-free regions throughout the eastern United States along the western side of the Appalachian Mountains (fig. 4.11).

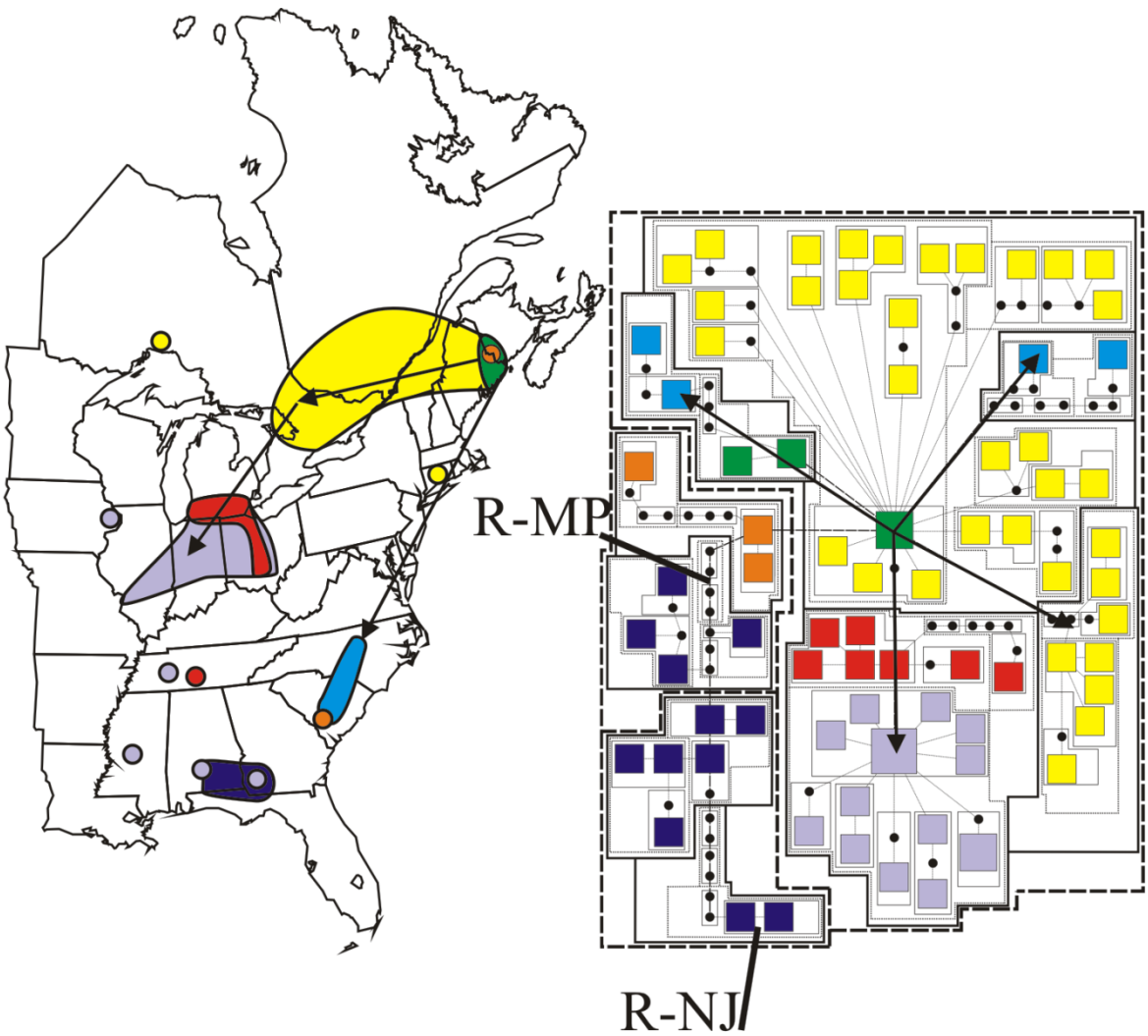
Five lower tier clades were nested within clade 4-1 with significant geographical associations (1-27, 2-1, 2-13, 3-4 and 3-5). The NCA of three of these clades indicated patterns of isolation by distance as a result of possessing tip clades/haplotypes with significantly small clade distances (clade 1-27: $Dc_{(S5H009)}$=0, *p<0.001*. clade 2-1: $Dc_{(1-3)}$=0, *p<0.001*. clade 2-13:

$Dc_{(1\text{-}29)}$=0, *p<0.001*, $Dc_{(1\text{-}32)}$=162.95, *p<0.001*). The NCA for clades 3-4 and 3-5 both indicated contiguous range expansions because they possess interior clades (2-9 and 2-11, respectively) with significantly small clade distances ($Dc_{(2\text{-}9)}$= 29.74, *p<0.001*, $Dc_{(2\text{-}11)}$=84.24, *p<0.001*, respectively). Both of these expansions appear to have progressed westward. The interior clade 2-9 haplotypes were located in New Brunswick, while the tip clades of clade 3-4 were distributed throughout Quebec and Ontario. Within clade 3-5, the interior clade 2-11 occupies a region in southern Michigan and Ohio with the derived clades located in the Mississippi River Basin (fig. 4.11).

The NCA of clade 4-3 inferred contiguous range expansion because the interior clade 3-8 had significantly small clade and nested clade distances (Dc=31.92, *p=0.010*, Dn=26.19, *p=0.010*, respectively) and the tip clade 3-9 had a significantly large nested clade distance (Dn=172.43, *p=0.010*). This analysis assumes the root inferred by maximum parsimony, otherwise all nested clades are interiors and no inference was possible.

Two 3-step clades (3-6 along with 3-8) with significant geographical association were nested within clades 4-2 and 4-3, respectively. The NCA of clade 3-6 indicated isolation by distance because the tip clade (2-14) had significantly small clade and nested clade distances (Dc=0, *p=0.015*, Dn=820.42, *p=0.015*, respectively). In the case of clade 3-8 the inference key returned an outcome of contiguous range expansion because the interior clade (2-18) was restricted to a single habitat (Flint River, GA) resulting in a significantly small clade distance (Dc=0, *p=0.023*). A range expansion was further supported by a tip clade (2-19) with a significantly large clade distance (Dc=23.45, *p=0.021*) and the I-T clade distance was significantly small ($I\text{-}T_{Dc}$=-23.45, *p=0.021*). The tip clade (2-19) was distributed among habitats in southern Georgia and Mississippi (fig. 4.11).

Figure: 4.10: Map showing distribution of species 5 lineage A
The distributions of the clades are denoted on the map (left) and refer to the clades colored on on the network (right). The regions occupied by each clade are color coded on the map as follows: 3-1 (green +yellow), 3-2 (blue + green), 3-3 (blue), 3-4 (yellow), 3-5 (red + light purple), 3-6 (orange), 3-7 (dark purple) and 3-8 (light purple). R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively. The arrows represent the direction of range expansion inferred from the nested clade analysis.



The nested clade analysis of lineage B identified 4 clades (2-21, 3-10, 3-11 and 4-4) with significant geographical associations (table 4.7). The NCA of the total cladogram (clade 4-4) indicated isolation by distance because the tip clade had significantly small clade and nested clade distances (Dc=104.20, *p<0.001*, Dn=156.21, *p<0.001*, respectively). The inference key returned multiple outcomes for each of the 3-step clades. The possible outcomes for clade 3-10 include contiguous range expansion, long distance colonization and past fragmentation of a previously larger range. The tip clade 2-22 was detected in a single habitat in northern Virginia and possessed significantly small clade and large nested clade distances (Dc=70.56, *p<0.001*, Dn=83.54, *p=0.003*, respectively). The interior clade (2-21) had significantly small clade and nested clade distances (Dc=0, *p<0.001*, Dn=157.44, *p=0.006*, respectively) and was distributed throughout the Shenandoah Valley and North Carolina (fig. 4.12). Given the vagility of *Hyalella* species, long distance colonization remains the default explanation (rather than fragmentation) according to the inference key (Templeton, 2011). A larger than average number of mutational steps between the tip clade with reversed significance (small clade distance and large nested clade distance) may indicate allopatric fragmentation, but clade 2-21 and 2-22 were only separated by two inferred haplotypes.

The NCA of clade 3-11 indicated either long distance colonization, past range expansion or past fragmentation of a larger range. The interior clade (2-23) had significantly small clade and large nested clade distances (Dc=37.39, *p<0.001*, Dn=432.12, *p<0.001*, respectively) along with a spatial distribution located in North Carolina and Virginia. The tip clades (2-24 and 2-25) were located in a distinct region in New York and Pennsylvania. This indicates that the northward expansion of this clade either occurred through long distance colonization or short contiguous jumps followed by local extinction in the intervening habitats. The NCA of clade 2-

21 indicated isolation by distance because the I-T clade distance was significantly large (I-T_{Dc}=42.22, *p=0.006*).

Table 4.7: Nested clade analysis with inference key of species 5 lineage B
The #-step clades refer to the nested network specific for species 5 (appendix C, fig. C.14). The inference key numbers refer to the progression of steps outlined in Templeton (2011). CRE = Contiguous Range Expansion, PRE: Past Range Expansion, LDC = Long Distance Colonization, PF = Past Fragmentation, LDD = Long Distance Dispersal and IBD = Isolation by Distance with Restricted Gene Flow. Only lineage A had significant geographical associations. See the appendix for Dc, Dn and p-values.

Clade	Inference Key
2-21	1-2-3-4-NO: IBD
3-10	1-2-5-6-13-14-YES: CRE, LDC or PF
3-11	1-2-11-12-13-YES: LDC, PRE or PF
4-4	1-2-3-4-NO: IBD

The nested clade analyses of lineages C, D and E revealed no significant geographical associations. Lineage C individuals were detected in the drainages of the Apalachicola River and Mississippi River (fig. 4.13). Similarly, lineage E was detected in Arkansas and Louisiana (fig. 4.13). Lineage D occurred within Georgia and North Carolina (fig. 4.13).

Figure: 4.11: Map showing distribution of species 5 lineage B
The distributions of the clades are denoted on the map (left) and refer to the clades colored in the network (right). The regions occupied by each clade are marked on the map as follows: 3-10 (green), 3-11 (blue + purple + red), and 3-12 (yellow). The root is denoted by (R).

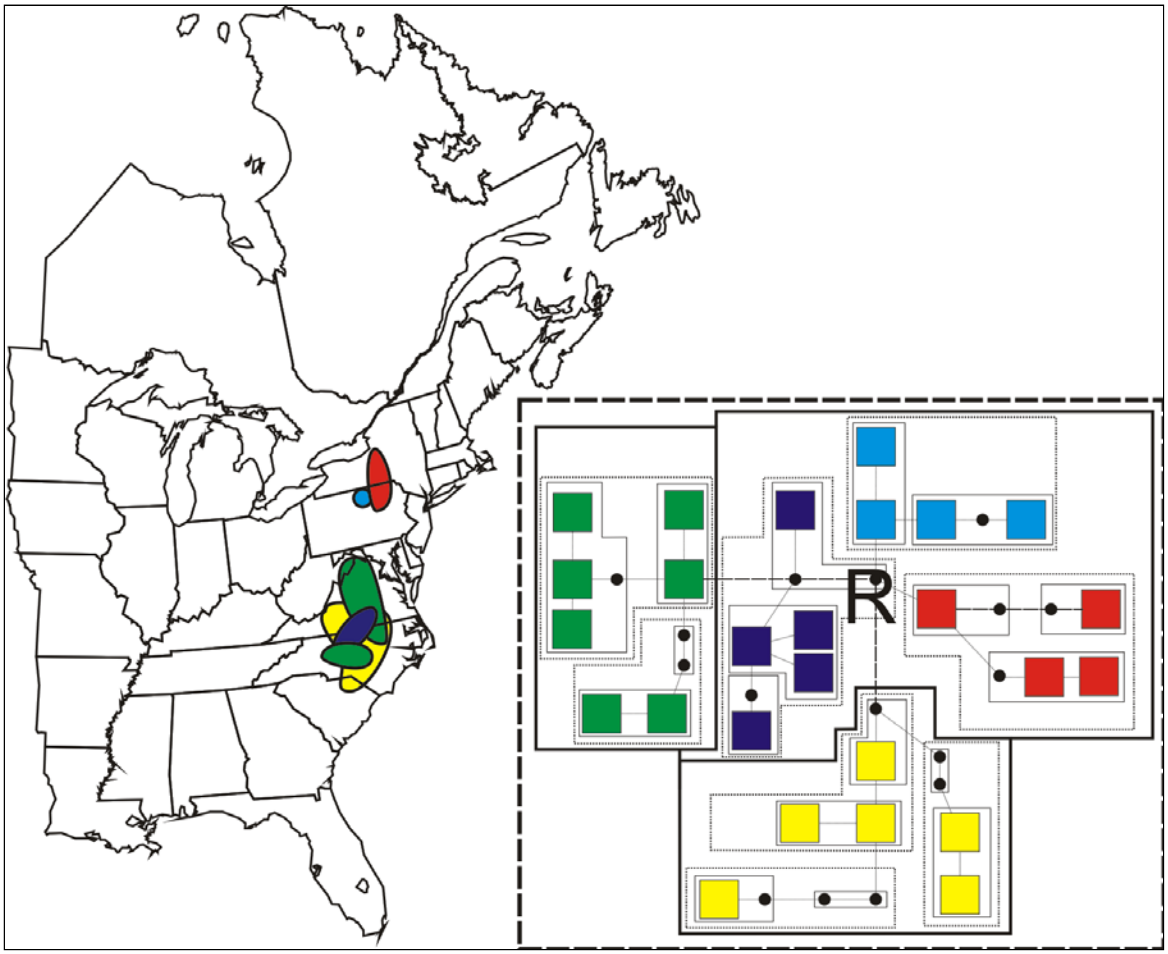
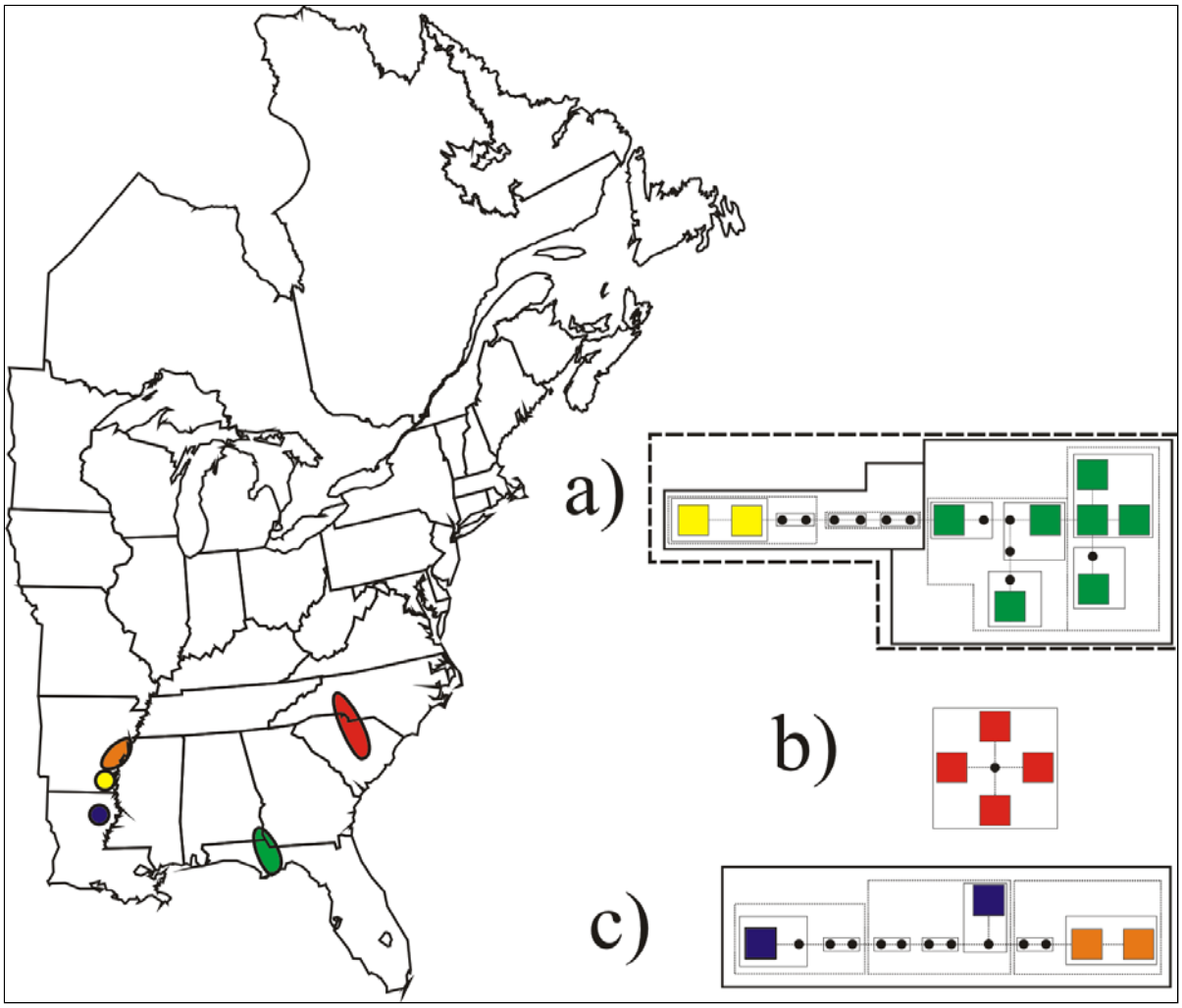


Figure: 4.12: Map showing distribution of species 5 lineages C, D and E
The distributions of the clades are denoted on the map (left) and refer to the clades marked on the network (right). The regions occupied by each clade are marked on the map as follows: a) lineage C (green + yellow), b) lineage D (red) and c) lineage E (purple + orange).



4.5.4 Mismatch distributions and population genetics of species 5

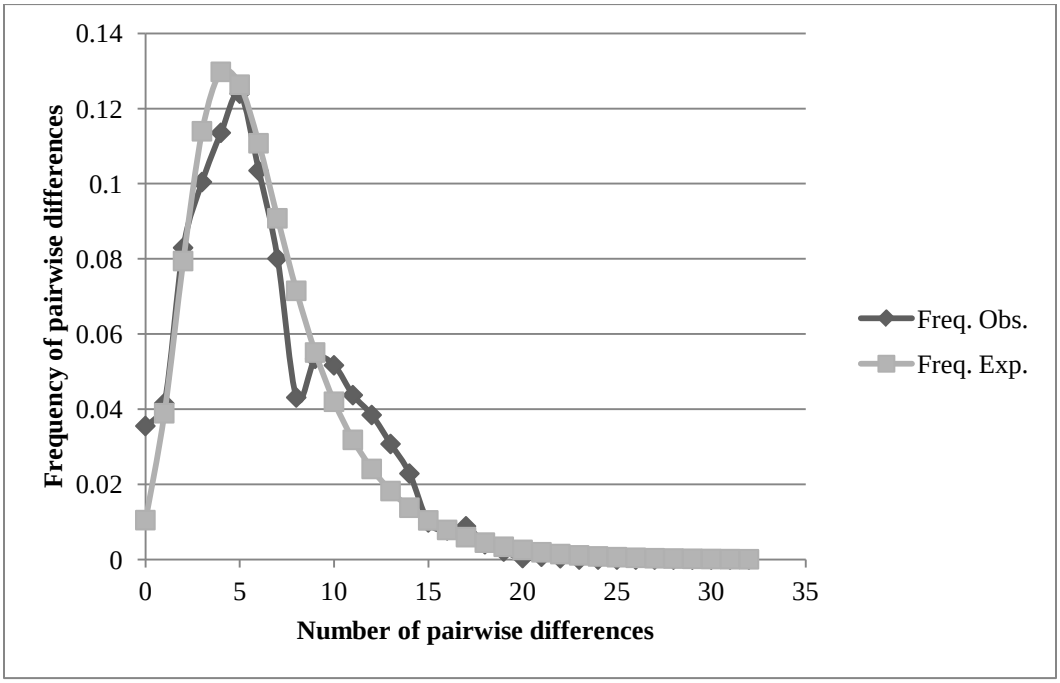
The mismatch distribution of species 5 lineage A was bimodal (fig. 4.14a) and the raggedness statistic rejected (r=0.005, $p=0.002$) the null hypothesis of a demographic expansion when compared against a simulated population using tau=3.328.

The mismatch distribution of species 5 lineage B was unimodal (fig. 4.14b) and the raggedness statistic (r=0.200, $p=0.024$) did not reject the null hypothesis of a demographic expansion when compared against a simulated population using tau=4.625.

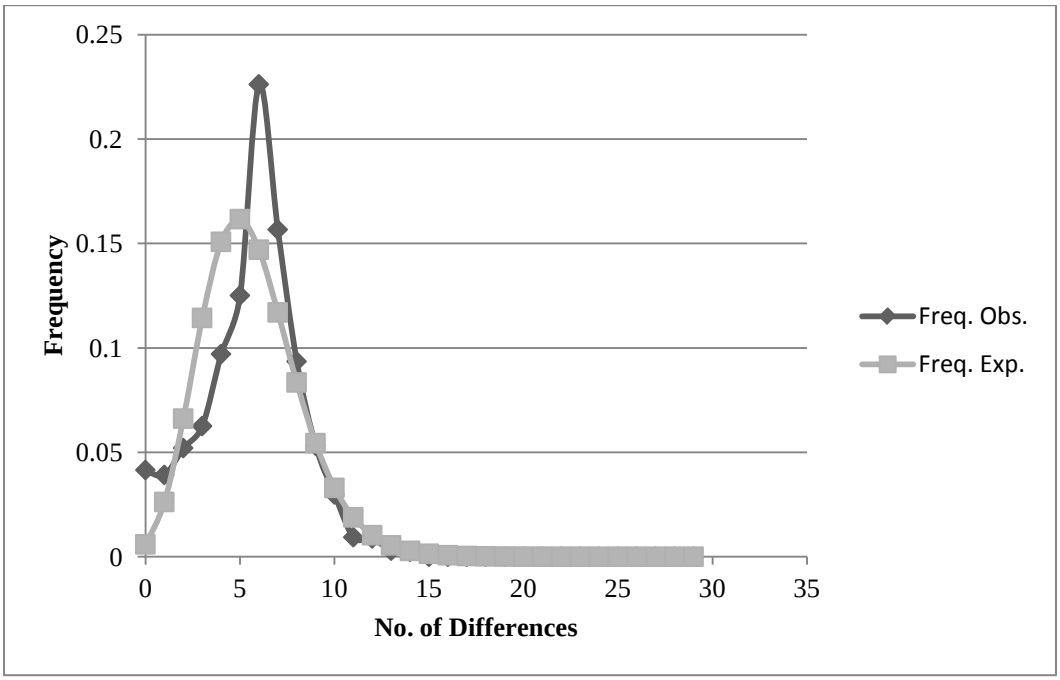
The Tajima's *D*, Fu's *F* and R_2 statistics of lineage A (table 4.1) all detected significant departures from neutrality consistent with a demographic expansion (D_T=-2.069, $p=0.002$, F_F=-34.179, $p<0.001$, R_2=0.030, $p<0.001$). Whereas, only Fu's F was significant with respect to lineage B (D_T=-1.220, $p=0.088$, F_F=-8.811, $p<0.001$, R_2=0.0646, $p<0.113$).

Figure 4.13: Mismatch distributions of species 5
The mismatch distribution of a) species 5 lineage A COI haplotypes was plotted against an expected curve of a population undergoing constant demographic expansion with tau = 3.328 and b)species 5 lineage B was plotted against an expected curve of a population undergoing constant demographic expansion with tau = 4.625

A)



B)



4.5.5 Analysis of molecular variance of species 5 lineage A

The AMOVA partitioned on the basis of faunal provinces had six groups of populations (Central Appalachian: 18 populations, Great Lakes: 18 populations, Hudson Bay: 1 populations, Northern Appalachian: 14 populations, Mississippi: 16 populations and Southeastern: 6 populations) and the between-group variance component was 27.56% (table 4.8). The AMOVA grouped according to glaciated and unglaciated regions had a between-group component of 9.50% (glaciated: 33 populations and unglaciated: 40 populations). This was not too surprising given the spatial structure of lineage A which possesses a pattern of isolation by distance across multiple faunal provinces.

None of the partitions considered in the SAMOVA consisted of two or more groups with at least 3+ populations each.

Table 4.8: AMOVA and SAMOVA of species 5 lineage A
V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * $p<0.01$, ** $p<0.001$, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	27.56**	54.29**	18.05**
Unglaciated and Glaciated	9.50*	72.03**	18.47**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	18.13*	62.76**	19.12**
3	52.80**	35.60**	11.60*
4	46.24**	39.77**	13.99**
5	51.98**	34.61**	13.41**
6	45.15**	41.14**	13.71**

4.6 RESULTS: SPECIES 6

Three-hundred and eighty-seven species 6 individuals (181 haplotypes) were collected from 112 localities that were distributed through the eastern half of North America. Seventy-one percent of these locations were cohabited by another *Hyalella* species, predominately species 1, 3 and 5. The majority of sites were located in the Mississippi River and Great Lakes watersheds, although the species range includes habitats in the Hudson Bay drainages within Quebec as well as sites along the east coast of the USA.

4.6.1 Nested haplotype network of species 6

The 181 COI haplotypes of species 6 nested into a single 6-step clade (appendix C, fig. C.16) with ninety-nine 1-step clades (1-1 to 1-99), forty-three 2-step clades (2-1 to 2-43), twenty 3-step clades (3-1 to 3-20), seven 4-step clades (4-1 to 4-7) and three 5-step clades (5-1 to 5-3).

4.6.2 Outgroup analysis of species 6

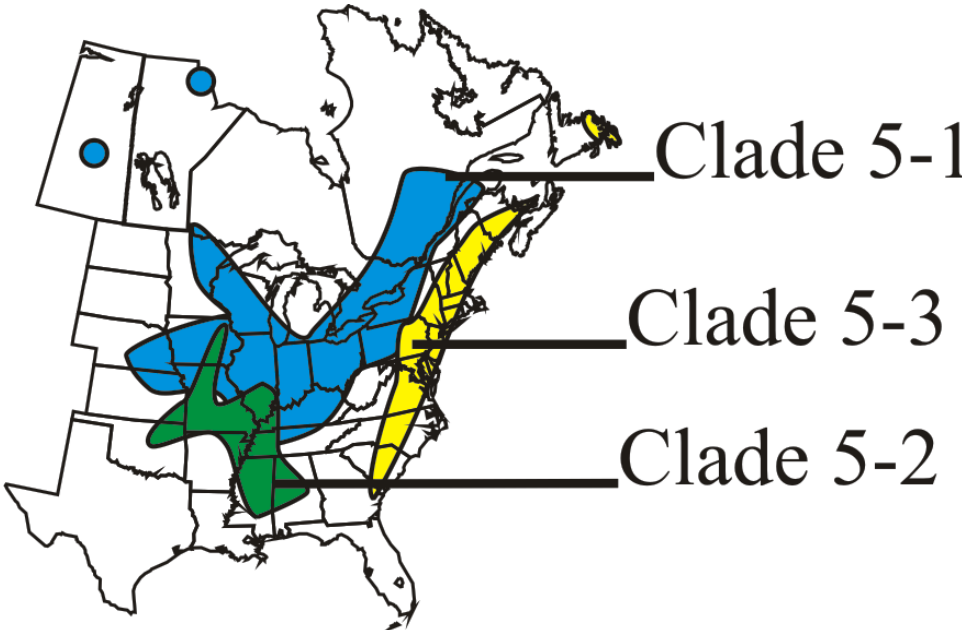
The TIM2 model of molecular evolution was selected by the Akaike information criterion with a proportion of invariant sites (I=0.598) and gamma distributed rate variation (α=1.466). The analyses were conducted using *Hyalella* species 1 haplotype S1H209 as the outgroup. The roots were inferred by neighbor-joining and maximum parsimony within clades 2-7 and 2-10, respectively. The root of the MP tree was along the edge between clade 1-24 and 1-25 and the root of the NJ analysis was between clades 1-28 and 1-29. The species 1 lineage G haplotype (S1H209) was used as the outgroup.

4.6.3 Nested clade analysis of species 6 clade 6-1

The species 6 NCA detected 21 significant geographical associations (table 4.7) with multiple genetic patterns inferred at all levels of the analysis. Nine associations indicated isolation by distance, four contiguous range expansions and the remaining eight had multiple potential inferences. Eight of the inferences indicated the potential for long distance dispersal.

The inference key indicated multiple outcomes for the total cladogram: long distance colonization, fragmentation of a past larger range or past range expansion. The interior clade 5-2 had significantly small clade and large nested clade distances (Dc=423.63, $p<0.001$, Dn=1063.76, $p<0.001$, respectively) along with the tip clade (5-3) having a significantly large nested clade distance (Dn=1228.21, $p<0.001$). Both clades possessed distinct distributions (fig. 4.15). Clade 5-2 occupied a region adjacent to the Mississippi River in the southern USA and clade 5-3 was located in drainages along the eastern coast (fig. 4.15). As such, the spatial pattern indicates either a long distance colonization event from across the Appalachian Mountains or a previously larger range that has since been fragmented and the intervening space colonized by clade 5-1 individuals.

Figure: 4.14: Map showing distribution of species 6
The distributions of the 5-step clades are displayed on the map as follows: clade 5-1 (blue), clade 5-2 (green) and clade 5-3 (yellow).



4.6.4 Nested clade analysis of species 6 clade 5-1

The clade 5-1 NCA indicated restricted gene flow with isolation by distance because the tip clade (4-1) had significantly small clade and nested clade distances (Dc=473.71, *p<0.001*, Dn=557.62, *p<0.001*, respectively). The interior clade (4-2) possessed a distribution that was much broader than clade 4-1, resulting in a significantly large nested clade distance (Dn=848.39, *p<0.001*). This analysis assumes the root as inferred by neighbor-joining. The NCA using the root inferred by parsimony considers clade 4-1 an interior clade. The inference key still returns isolation by distance as the dominant pattern but with a different decision path (1-2-11-17-4-NO as opposed to 1-2-3-4-NO). Nested within clade 5-1 were several clades with significant geographical associations (clades 1-8, 1-13, 2-2, 2-4, 2-6, 3-1, 3-2 and 4-1) that need to be considered in order to unravel this group's demographic history. The NCA of clade 4-1 indicated isolation by distance with long distance dispersal regardless of the root location. This was because the tip clade 3-3 had significantly small clade and large nested clade distances (Dc=245.09, *p=0.006*, Dn=683.26, *p=0.001*, respectively). The NCA of clade 3-1 indicated isolation by distance because both the tip clades (2-1 and 2-4) had significantly small clade distances (Dc=173.33, *p<0.001*, Dc=240.93, *p<0.001*, respectively) along with an interior with a significantly large nested clade distance (Dn=410.00, *p=0.016*). The localities possessing these tip clades formed a distribution that was overlapping with that of the interior clade (fig. 4.16). The inference key indicated contiguous range expansion with regard to clade 2-2 because the interior clade (1-8) had a significantly small clade distance (Dc=232.06, *p<0.001*). This clade's distribution was adjacent to the Mississippi River and the tip clades were located throughout the Great Lakes region. The NCA of clade 1-8 indicated isolation by distance because haplotype S6H009 had 4 sequences restricted to a single habitat resulting in a significantly small clade

distance (Dc=0, *p<0.001*). Although clade 2-3 did not have a significant geographical association, the NCA of its interior clade (1-13) indicated either past fragmentation or long distance colonization. Haplotype S6H031 had significantly small clade and large nested clade distances (Dc=0, *p=0.011*, Dn=376.61, *p<0.001*, respectively) and was detected in a single disjunct habitat (fig. 4.16).

Table 4.9: Nested clade analysis with inference key of species 6

The #-step clades refer to the nested network specific for species 6 (appendix C, fig. C.16). The inference key numbers refer to the progression of steps outlined in Templeton (2011). CRE = Contiguous Range Expansion, PRE = Past Range Expansion, LDC = Long Distance Colonization, IBD = Isolation by Distance with Restricted Gene Flow, LDD = Long Distance Dispersal, PLR = Past Larger Range with subsequent extinction, and PF = Past Fragmentation followed by Range Expansion. The "Clade" column is staggered to show the nested clade structure and all clades in a given hierarchy are included regardless of the NCA outcome.

Clade	Inference Key
6-1	1-2-11-12-13-YES: LDC, PLR or PRE
5-1	1-2-3-4-NO: IBD
4-1	1-2-3-5-6-7-8-YES: IBD with LDD
3-1	1-2-3-4-NO: IBD
2-2	1-2-11-12-NO: CRE
1-8	1-2-3-4-NO: IBD
2-3	No Inference
1-13	1-2-3-5-15-NO: PF or LDC
3-2	1-2-11-12-NO: CRE
2-4	1-2-3-5-15-NO: PF or LDC
3-3	No Inference
2-6	1-19-20-2-3-5-15-NO: PF or LDC
4-2	1-2-3-4-NO: IBD
3-5	1-2-11-12-NO: CRE
2-14	1-2-3-4-NO: IBD
3-6	1-2-3-4-NO: IBD
2-16	1-19-20-2-11-12-13-YES: LDC or PLR or PRE
4-3	1-2-3-4-NO: IBD
3-9	1-2-3-5-6-TOO FEW: CRE or LDC or IBD with LDD
5-2	No Inference
4-4	No Inference
3-11	1-19-20-2-3-4-NO: IBD
4-5	1-2-3-4-NO: IBD
3-13	No Inference
2-31	1-2-3-4-NO: IBD
3-14	1-2-11-12-NO: CRE
5-3	1-2-11-12-NO: CRE
4-6	1-2-3-5-1-5-NO: PF or LDC
4-7	1-2-3-4-NO: IBD
3-19	1-2-3-4-NO: IBD

The NCA of clade 3-2 indicated contiguous range expansion because the interior clade (2-5) had significantly small clade and nested clade distances (Dc=338.63, *p=0.011*, Dn=347.76, *p=0.008*, respectively). Furthermore, the tip clade (2-6) had significantly large clade and nested clade distances (Dc=722.46, *p<0.001*, Dn=776.76, *p<0.001*, respectively). The interior clade was predominately distributed within the Great Lakes region, whereas the region occupied by clade 2-6 encompassed disparate sites in Nebraska and West Virginia (fig. 4.16). The NCA of clade 2-4 (nested within clade 3-2) indicated either past fragmentation or long distance colonization because the tip clade (1-18) had significantly small clade and large nested clade distance (Dc=0, *p=0.002*, Dn=649.47, *p=0.003*, respectively). This tip clade was restricted to a single locality in South Dakota. According to the inference key, long distance colonization was favored over fragmentation because the clade was located at a lower tier in the nested analysis and *Hyalella* is known to possess the ability to disperse via waterfowl (Swanson, 1984).

Multiple clades nested within clade 4-1 have demonstrated clear patterns of contiguous range expansions in conjunction with both short and long distance dispersal. The overall trend of dispersal seemed to progress along an east to west orientation, with clade 6 individuals migrating throughout the glaciated regions of the southern Great Lakes region.

Clade 4-2 NCA inferred a pattern of isolation by distance because the tip clade (3-6) had significantly small clade and nested clade distances (Dc=296.14, *p<0.001*, Dn=551.55, *p<0.001*, respectively). The NCA of clade 3-5 (nested within clade 4-2) indicated a pattern of contiguous range expansion because its interior clade (2-14) had a significantly small clade distance (Dc=438.63, *p<0.001*). Clade 2-14 was distributed throughout Quebec and New York State and the clade has a genetic pattern consistent with isolation by distance because a tip clade (1-36) had a significantly small clade distance (Dc=0, *p=0.001*). The NCA of clade 3-6 indicated isolation

by distance because a tip clade (2-17) had a significantly small clade distance (Dc=0, *p=0.002*) and contained a 2-step clade (2-16) for which the inference key inferred a pattern of long distance colonization. The interior clade (1-41) had significantly small clade and nested clade distances (Dc=149.46, *p=0.019,* Dn=223.95, *p=0.019,* respectively) along with a tip clade (1-42) with significantly small clade and large nested clade distances (Dc=0, *p=0.019*, Dn=655.72, *p=0.019,* respectively). This tip clade was restricted to a single habitat in Pennsylvania (fig. 4.17), which the NCA infers as a long distance dispersal event from the watersheds of the St. Lawrence River. Clade 4-2 represents a broadly distributed group of species 6 populations. The interior clades possess distributions that have expanded throughout the Great Lakes region and reached the Churchill River watershed. The clade 4-2 haplotype network indicates that the ancestral haplotypes were located along the eastern edge of the range. Similar to clade 4-1, the genetic patterns suggest an east to west gradient going from ancestral to derived haplotypes.

Nested within clade 4-3 was a clade (3-9) with a significant geographical association. Clade 3-9 had an interior clade with a significantly small nested clade distance (Dn=272.07, *p=0.008*), but too few nested clades with concordant geographical patterns to distinguish patterns of contiguous range expansion, long distance colonization or isolation by distance with long distance dispersal.

4.6.5 Nested clade analysis of species 6 clade 5-2

Nested with clade 5-2 were three clades with significant geographical associations (2-31, 3-11 and 3-14). The NCA of clade 3-11 indicated a pattern of isolation by distance among populations adjacent to the Mississippi River because the tip clades (2-27 and 2-29) had a significantly small

clade distances (Dc=0, *p<0.*001, Dc=0, *p<0.001*, respectively). Clade 3-14 was associated with a range expansion because its interior clade had a significantly small clade distance (Dc(2-31)=0, *p<0.001*). The ancestral haplotypes were detected in populations in Iowa, Kansas and Missouri, while the derived groups were detected in two habitats in Oklahoma (OK011 and OK002). Although clade 3-13 did not possess a significant geographical association, its interior clade (2-31) had a significantly large clade distance (Dn=305.98, *p<0.001*) and tip clades (1-71 and 1-72) with significantly small clade distances (Dc=5.73, *p=0.001*, Dc=0, *p=0.001*, respectively). This was consistent with multiple potential outcomes: long distance colonization, past larger range or past range expansion, which all can be coupled with subsequent fragmentation. The two clades were separated by multiple inferred haplotypes, which supports fragmentation of the distribution into a northern group in Missouri and Iowa along with a southern group adjacent to the Mississippi River (fig. 4.18).

4.6.6 Nested Clade Analysis of species 6 clade 5-3

The clade 5-2 NCA indicated a contiguous range expansion. The interior clade 4-6 possessed a significantly small clade distance (Dc = 366.52, *p<0.001*) and distribution throughout the Southeastern and Central Appalachian faunal provinces (fig. 4.18). The clade 4-6 NCA indicated either past fragmentation or long distance colonization as a consequence of the significantly small clade distance among tip clades 3-16 and 3-18 (Dc = 0, *p<0.001* and Dc = 62.90, *p<0.001*, respectively) as well as the significantly large nested clade distance within tip clade 3-17 (Dn = 506.59, *p<0.001*). The nested clade analysis was unable to resolve the two competing explanations; either species 6 dispersed among faunal provinces via avian vectors or a previously larger range had become fragmented. Clade 4-7 NCA indicated a pattern of isolation by distance

because the tip clade 4-1 had significantly small clade distances (Dc = 17.62, *p=0.003*). The patterns of species 6 within these groups indicate several distinct spatial groups along the eastern coast. The high tier clades indicate range expansion and long distance colonization along with lower tier clades with patterns of isolation by distance. As such, the northward colonization of species 6 within clade 5-3 potentially predates the colonization of the Great Lakes region by species 6 clade 5-1.

Figure: 4.15: Map showing distribution of species 6 clade 4-1
The distributions of the clades are denoted on the map (left) and refer to the clades marked on the network (right). The regions occupied by each clade are marked on the map as follows: 2-1 (green), 2-2 (red), 2-3 (dark purple), 3-2 (yellow) and 3-3 (light purple). R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.

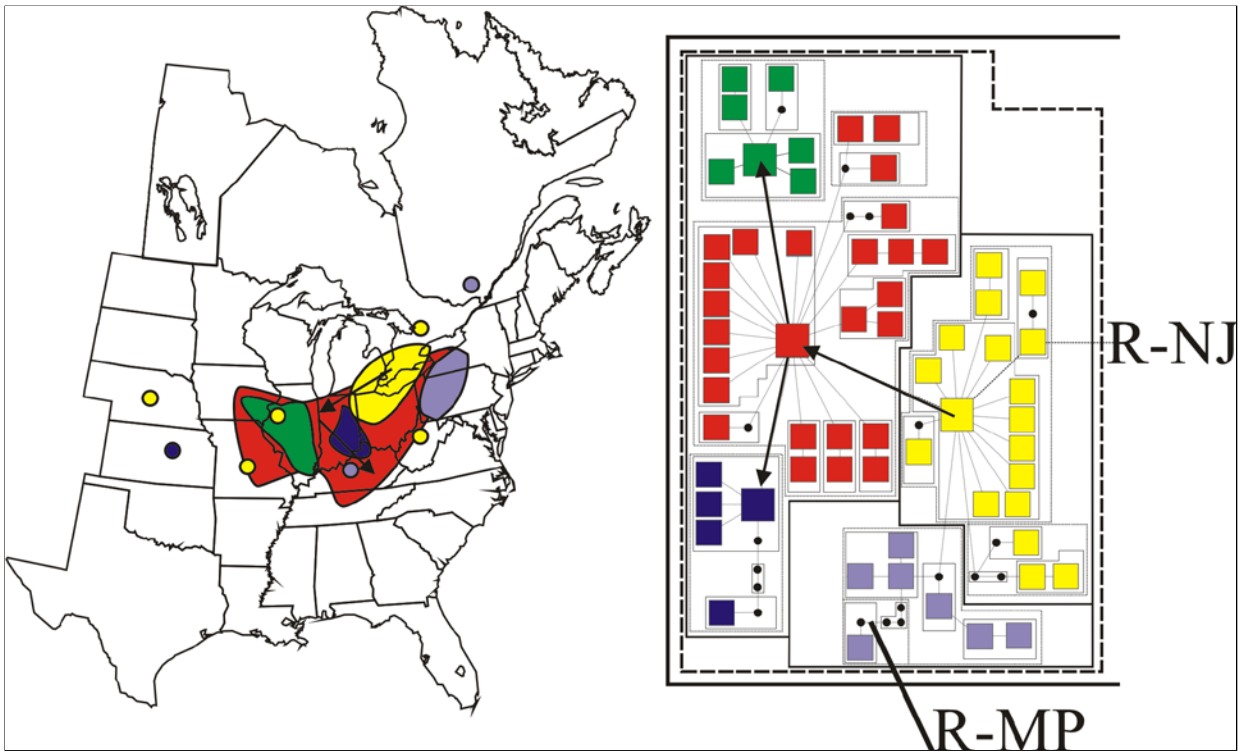


Figure: 4.16: Map showing distribution of species 6 clades 4-2 and 4-3
The distributions of the clades are denoted on the map (left) and refer to the clades marked on the network (right). The regions occupied by each clade are marked on the map as follows: clade 3-4 (green), clade 3-5 (yellow), clade 3-6 (red), clade 3-7 (blue) and clade 4-3 (dark purple). R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.

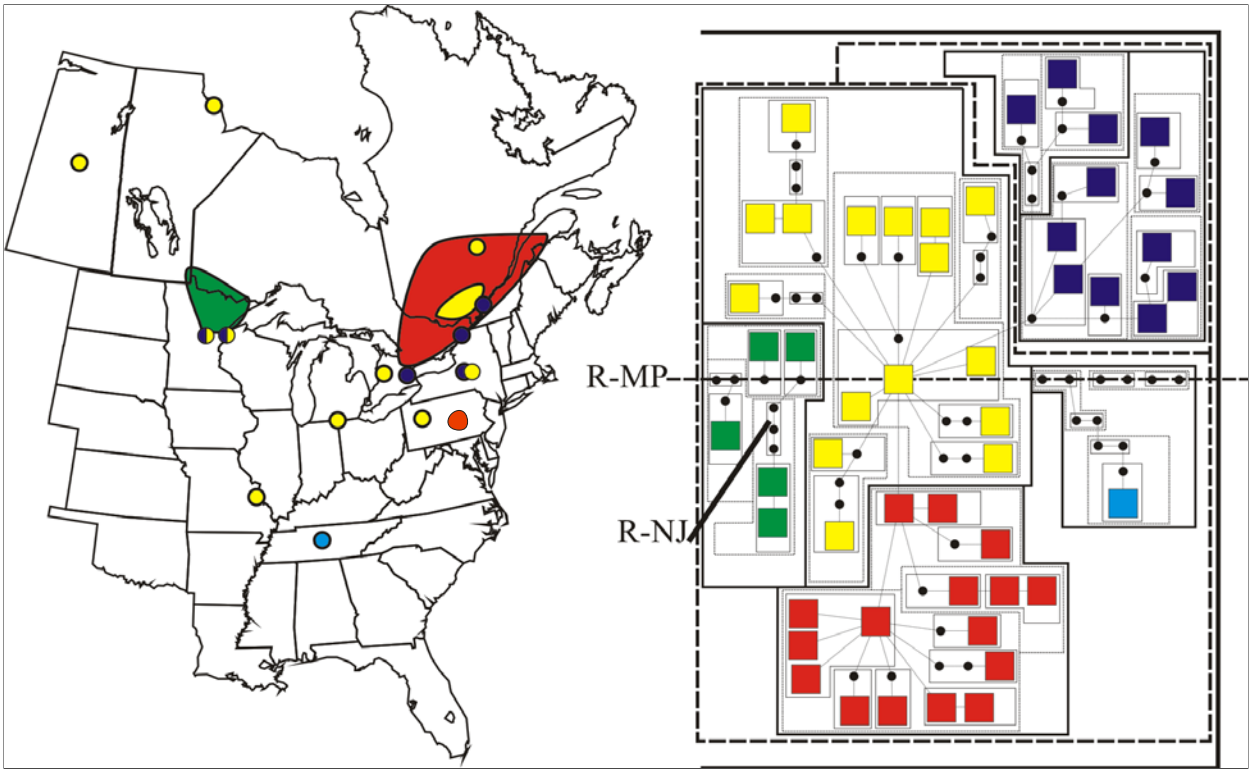
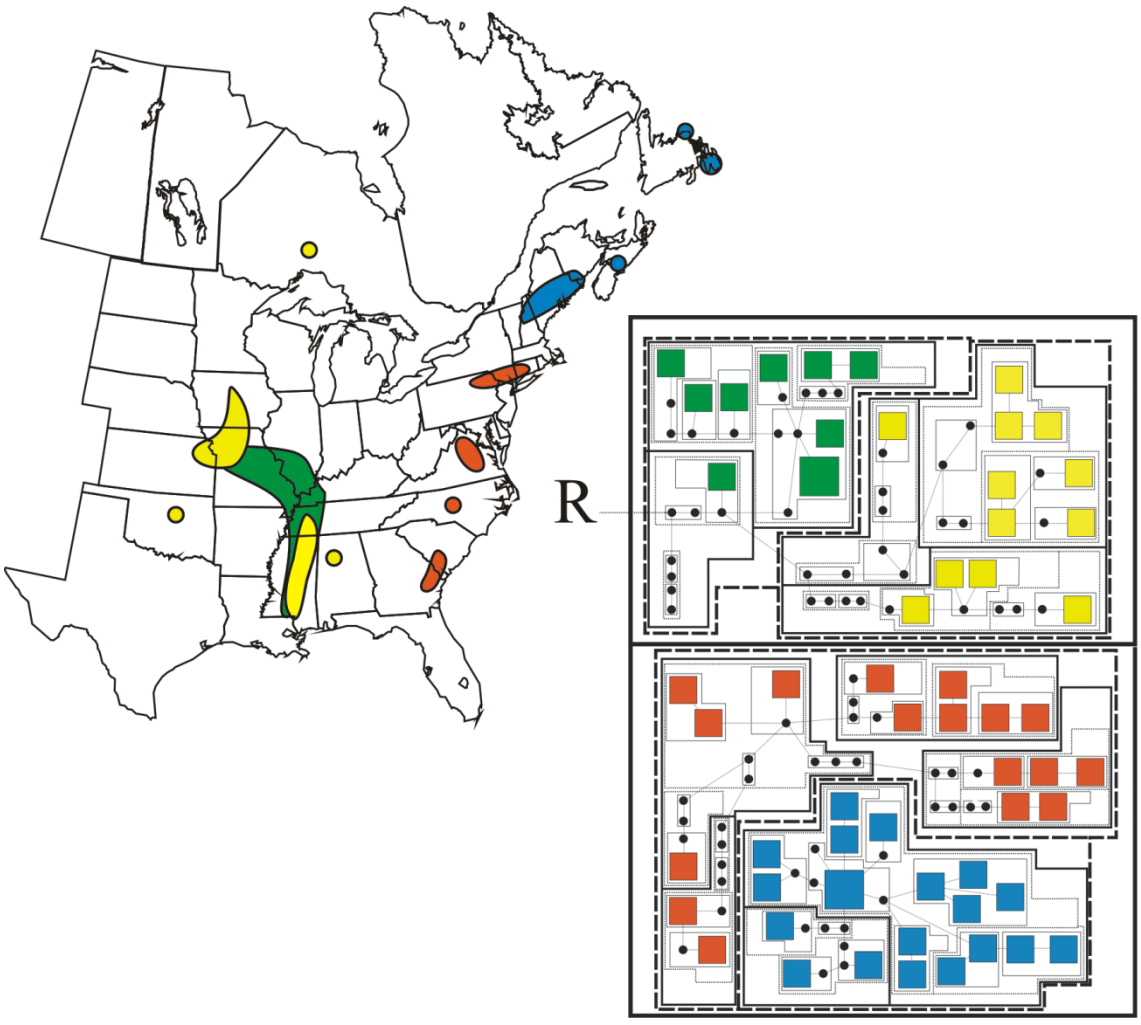


Figure: 4.17: Map showing distribution of species 6 clades 5-2 and 5-3
The distributions of the clades are denoted on the map (left) and refer to the clades marked on the network (right). The regions occupied by each clade are marked on the map as follows: clade 3-10 (green), clade 3-11 (green), clade 3-12 (yellow), clade 3-13 (yellow), clade 3-14 (orange), clade 4-6 (orange) and clade 4-7 (blue). The root is denoted by (R).



4.6.7 Mismatch distribution and population genetics of species 6

The mismatch distribution of species 6 formed a multimodal curve (fig. 4.19) in conjunction with a raggedness statistic (r=0.003, *p<0.001*) that rejected the null hypothesis of a population expansion when compared against a simulated population using tau = 7.162.

The Tajima's D and Ramos-Onsins & Rozas' R_2 statistics did not detect any deviations from neutrality (Dt=-1.678, *p=0.016*, R_2=0.034, *p=0.017*). Although, Fu's F statistic was more consistent with a lineage undergoing demographic expansion (F_f=-187.40, *p<0.001*).

4.6.8 Analysis of molecular variance of species 6

The species 6 AMOVA partitioned on the basis of faunal provinces had six groups of populations (Central Appalachian: 6 populations, Great Lakes: 41 populations, Hudson Bay: 7 populations, Northern Appalachian: 15 populations, Mississippi: 37 populations and Southeastern: 6 populations) and the between-group variance component was 26.13% (table 4.9). The AMOVA grouped according to glaciated and unglaciated regions had a between-group component of 6.87% (glaciated: 63 populations and unglaciated: 49 populations).

The 5-group SAMOVA of species 6 maximized the between group variance component (53.43%) and partitioned the populations into three major groups (group 1: 39 populations, group 2: 29 populations, group 3: 41 populations, group 4: 1 population and group 5: 1 population). The other SAMOVAs either had lower between-group variance components (table 4.9) or had fewer notable partitions of 3 or more populations. The only exception was the 6-group SAMOVA, which had a higher between-group variance component, but the sixth partition only contained a single population. The three major groups identified in the 5-group SAMOVA were

consistent with the 5-step clades identified in the haplotype network (Group 1 = clade 5-3, group 2 = clade 5-2 and group 3 = clade 5-1). The F_{ST} value was determined using the 5-group partition (F_{ST} = 0.821, *p<0.001*).

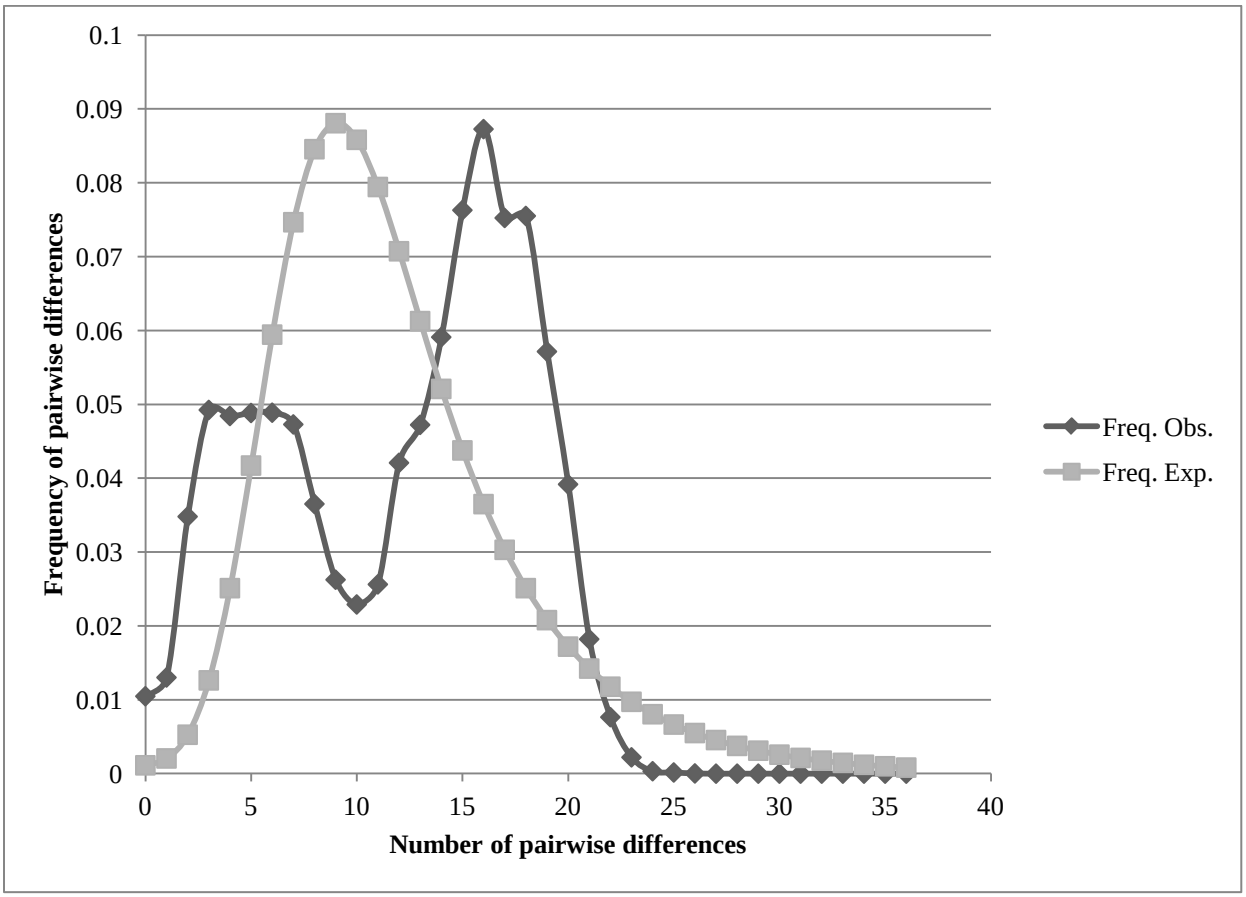
Table 4.10: AMOVA and SAMOVA of species 6

V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * p<0.01, ** p<0.001, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	26.13**	54.22**	19.65**
Unglaciated and Glaciated	6.87**	72.69**	20.44**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	19.04**	62.68**	18.28**
3	39.22**	47.86**	12.92**
4	36.72**	48.22**	15.06**
5	43.03**	39.10**	17.87**
6	49.61**	34.47**	15.92**

Figure 4.18: Mismatch distribution of species 6
The mismatch distribution of COI haplotypes plotted against an expected curve of a population undergoing constant demographic expansion with a tau = 7.162



4.7 RESULTS: SPECIES 9

Fifty-three species 9 individuals (31 haplotypes) were collected from 18 habitats that were located along the western side of North America (British Columbia and Washington State). The COI phylogeny of *Hyalella* lineages indicated that species 9 was most closely allied with species 6 (chapter 2), although the spatial distributions of the two species were located on different sides of the Rocky Mountains. Of the 18 habitats containing species 9, four were cohabitated by large-bodied *Hyalella*: species 1 (3 sites) or species 1 and 2 (1 site).

4.7.1 Nested haplotype network of species 9

Thirty of the species 9 COI haplotypes nested into a single 5-step clade (appendix C, fig. C.17) with twenty-three 1-step clades (1-1 to 1-23), fifteen 2-step clades (2-1 to 2-15), six 3-step clades (3-1 to 3-6) and three 4-step clades (4-1 to 4-3). Haplotype S9H002 did not group with the other species 9 sequences because the number of mutational steps exceeded the parsimony criterion. This haplotype was detected in a single habitat within British Columbia (BC024).

4.7.2 Outgroup analysis of species 9

The HKY model of molecular evolution was selected by the Akaike information criterion with a proportion of invariant sites (I=0.641) and gamma distributed rate variation (α=0.897). The parsimony and neighbor-joining analyses were conducted using species 1 haplotype S1H209 as the outgroup. The neighbor-joining analysis placed the root between haplotypes S9H008 and S9H009 in clade 1-6. The root inferred by maximum-parsimony was along the edge connecting

clade 3-3 to clade 3-5. However, this potentially places the ancestral haplotype in either clade, given the number of inferred haplotypes along this edge.

4.7.3 The nested clade analysis of species 9

The nested clade analysis identified 3 clades with significant geographical associations (table 4.10), but none lower than the 4-step tier (clades 4-2, 4-3 and 5-1). The NCA of total cladogram (clade 5-1) indicated a pattern of contiguous range expansion because the interior clade 4-2 had a significantly small clade distance (Dc=99.62, $p<0.001$) and the I-T clade distance was significantly small (I-T Dc = -60.14, $p=0.009$). The inference key returned an outcome of contiguous range expansion with respect to clade 4-2 because the tip clade 3-4 had a significantly large clade distance (Dc=283.53, $p<0.001$) and the I-T clade distance was significantly small (I-T Dc=-249.50, $p=0.002$). Similarly, the NCA of clade 4-3 indicated contiguous range expansion because a tip clade (3-6) had a significantly large clade distance and the I-T clade distance was significantly small (I-T Dc=-136.06, $p=.017$). Clade 4-2 populations (specifically its interior clade 3-3) were located within Oregon State (fig. 4.20) and the geographical distributions of the tip clades (3-4, 4-1 and 4-3) were detected at the southern limit of the Cordilleran Ice Sheet (fig. 4.20). All three of these tip clades were linked to an inferred haplotype within clade 3-3 that was not detected in this study, although the area between the interior and tip clade distributions remains largely unexplored. The root of the network indicates that clade 3-3 was ancestral, which suggests a northward migration of species 9 out of refugial populations located south of the Cordilleran Ice Sheet.

Figure: 4.19: Map showing distribution of species 9
The distributions of the clades are denoted on the map (left) and refer to the clades marked on the network (right). The regions occupied by each clade are marked on the map as follows: 4-1 (purple), 4-2 (yellow) and 4-3 (blue + red). The root is denoted by (R).

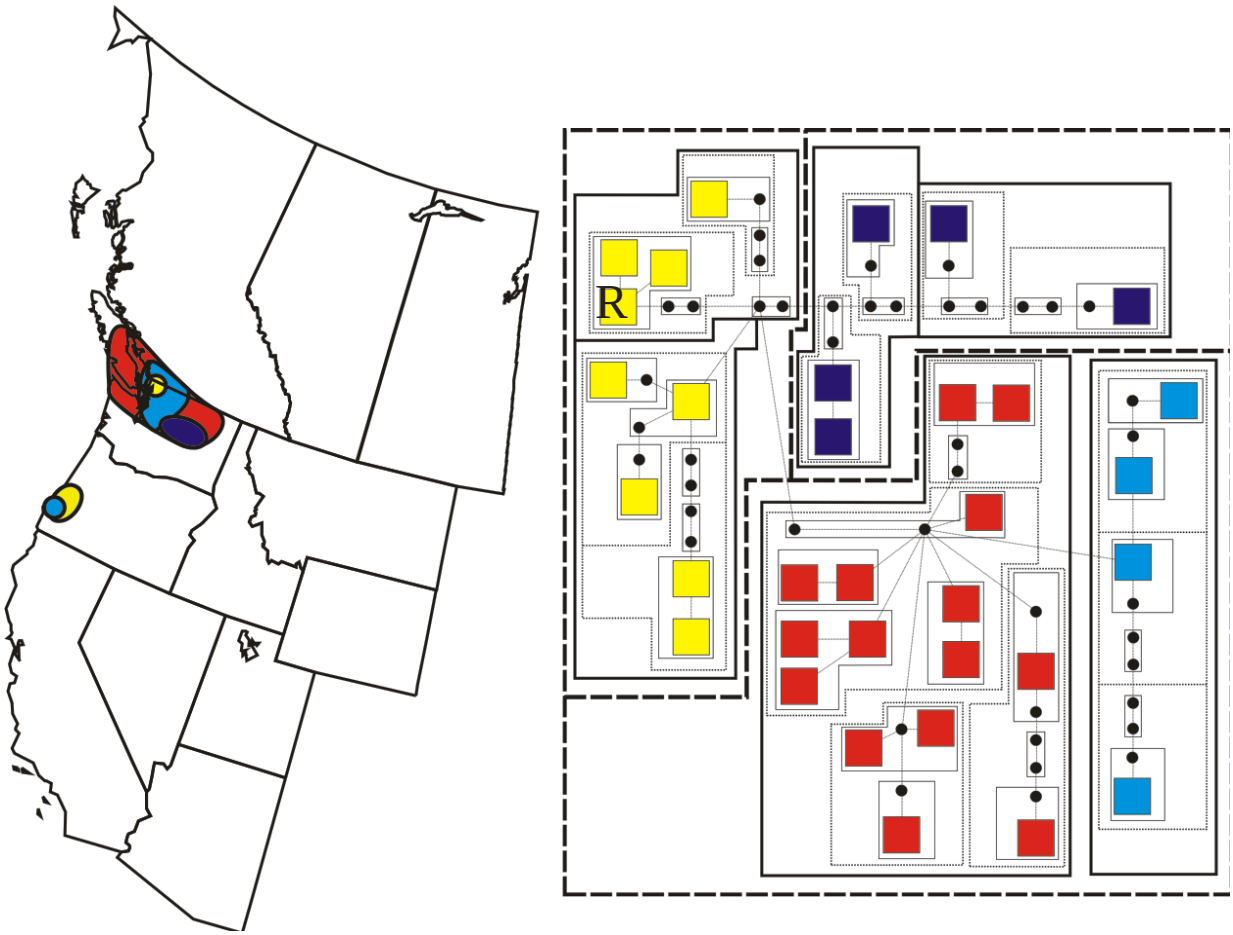


Table 4.11: Nested clade analysis with inference key of species 9
The #-step clades refer to the nested network specific for species 9 (appendix C, fig. C.17). The inference key numbers refer to the progression of steps outlined in Templeton (2011). CRE = Contiguous Range Expansion. Dc, Dn and p-values are included in appendix B.

Tip/Node	Inference Key
4-2	1-2-11-12-NO: CRE
4-3	1-2-11-12-NO: CRE
5-1	1-2-11-12-NO: CRE

4.7.4 The mismatch distribution and population genetics of species 9

The mismatch distribution of species 9 formed a multimodal curve (fig. 4.21). However, the raggedness statistic (r=0.017, *p=0.156*) did not reject the null hypothesis of a population expansion when compared against a simulated population using tau = 4.850.

The Tajima's *D* and Ramos-Onsins & Rozas' R_2 statistics did not detect any deviations from neutrality (*D*=-1.774, *p=0.018*, R_2=0.0565, *p=0.034*). Although, Fu's *F* statistic was more consistent with a lineage undergoing demographic expansion (*F*=-13.00, *p<0.001*).

4.7.5 Analysis of molecular variance of species 9

The species 9 AMOVA partitioned on the basis of faunal provinces had two groups of populations (Cascadian: 11 populations and West Coast: 9 populations) and the between-group variance component was 26.13% (table 4.10). The AMOVA partitioned according to glaciated and unglaciated regions had identical population partitions as the AMOVA grouped on the basis of faunal provinces.

Figure 4.20: Mismatch distribution of species 9 haplotypes
The mismatch distribution of COI haplotypes plotted against the expected curve of a population undergoing constant demographic expansion with a tau = 4.850.

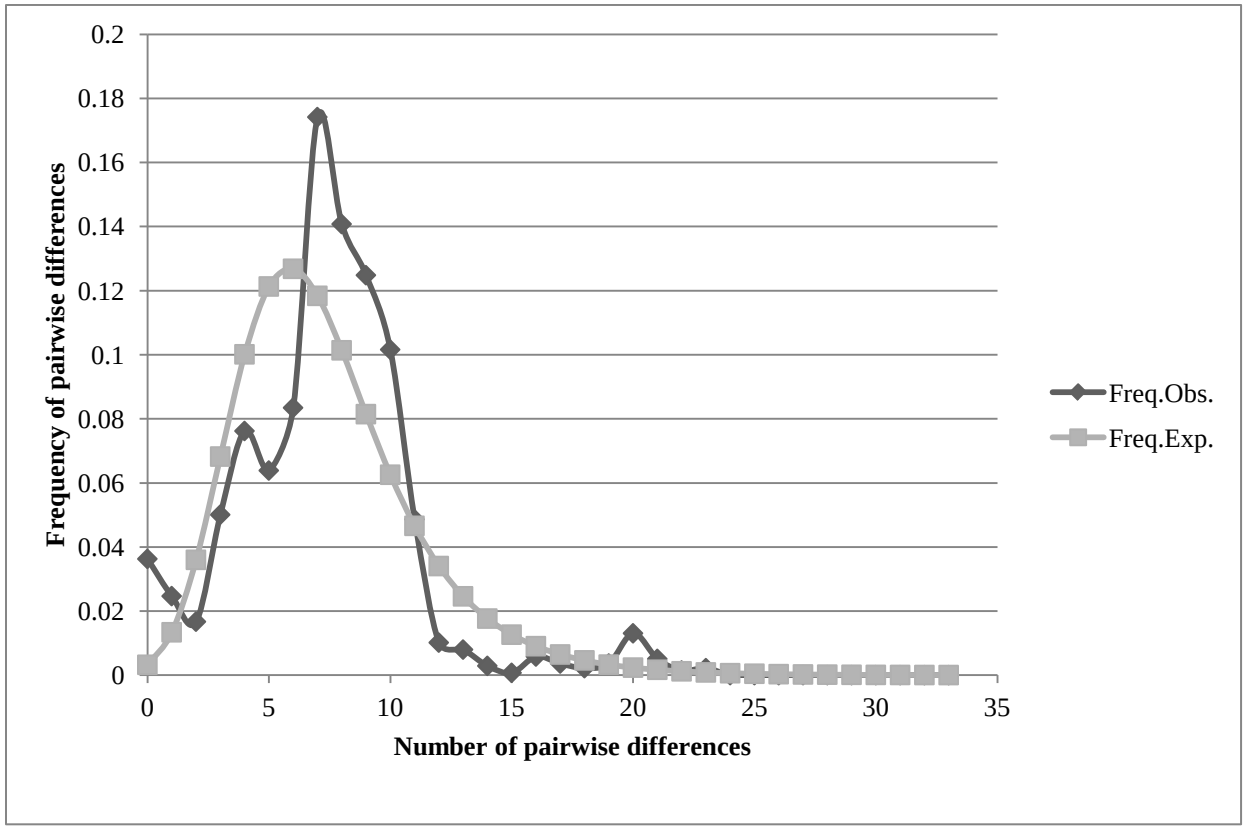


Table 4.12: AMOVA and SAMOVA of species 9

V(BG) = Variation Between Groups, V(AP) = Variation Among Populations within groups and V(WP) = Variation Within Populations. * $p<0.01$, ** $p<0.001$, NS – Not Significant

AMOVA	V(BG)	V(AP)	V(WP)
Faunal Provinces	26.13**	54.22**	19.65**
Unglaciated and Glaciated	26.13**	54.22**	19.65**

SAMOVA Groups	V(BG)	V(AP)	V(WP)
2	33.18**	35.15**	31.67**
3	36.71**	30.79**	32.50**
4	39.88**	27.22**	32.90**
5	40.93**	24.90**	34.17**
6	41.14**	23.45**	35.40**

The 2-group SAMOVA of species 9 (table 4.10) had a between group variance component (33.18%) and partitioned the populations into two major groups (group 1: 6 populations and group 2: 14 populations). The first partition coincides with clade 3-3. None of the other SAMOVA tests had appreciably higher between-group variance components, nor did they detect any further major subdivisions.

4.8 DISCUSSION

A principal outcome of this study is the characterization of the ranges and geographic distribution of the small-bodied *Hyalella 'azteca'*. Previously published studies of these species have largely focused on the Great Lakes region (Wellborn, 1994; Witt & Hebert, 2000; Wellborn & Broughton, 2008), which is now known to consist of a small fraction of their spatial ranges. Species 3, 5 and 6 were the most broadly distributed small-bodied members of the *Hyalella* complex and their distributions were sympatric over large areas throughout the eastern United States and Canada. Alternatively, species 4 and 9 possess far more restricted geographical ranges. The demographic histories of species south of the glacial limit can be difficult to assess because they persisted in more stable regions during the Pleistocene glaciations allowing for greater spatial heterogeneity (Hewitt, 2001).

4.9.1 Range expansions of the widely distributed small-bodied species into the Great Lakes region

All three of the widely distributed small-bodied species (3, 5 and 6) possess genetic patterns consistent with demographic expansion. However, unlike their large-bodied counterparts, the extent of the range expansion is more limited. The nested clade analyses of species 3 (clades 4-3 and 4-4), species 5 (lineage A) and species 6 (clade 5-1) suggest that these expansions originated from either Atlantic refugia (species 5) or the eastern United States. Fascinatingly, the networks and NCA indicate that these three species colonized the Great Lakes region in a pattern that went from east to west. In comparison, the large-bodied species likely originated from Missouri River, Mississippi River or Beringian refugia and they colonized the Great Lakes region from west to

east. This has resulted in a very complex suture zone, with some habitats possessing up to four distinct *Hyalella* species.

4.9.2 Genetic patterns within the small-bodied species

The most readily identifiable genetic patterns were isolation by distance with recurrent gene flow and range expansions among the small bodied *Hyalella* species (table 4.9). Species 3, 5 and 6 possessed phylogroups or lineages that displayed evidence for range expansions into the Great Lakes region. Additionally, the NCA detected a range expansion within species 9 along the southern limit of the Cordilleran ice sheet. Spatial structure was detected in species 4 (all lineages), species 5 (all lineages) and species 6 (clades 5-2 and 5-3), which the inference key interpreted in two ways: either fragmentation of a past larger range or long distance dispersal (table 4.9). These regions were often separated by large distances and the lower tier clades often possessed patterns of isolation by distance within the nested clades.

Table 4.13: Summary of nested clade analysis inferences for small-bodied lineages
The inferences are indicated on the basis of the nested clade analysis for all small-bodied species lineages with significant geographical associations. The three major inferences are fragmentation (Fr), restricted gene flow with isolation by distance (IBD) and range expansion (RE). The last column indicates long distance dispersal (LDD) and is marked yes for either long distance colonization or recurrent gene flow with long distance dispersal. A "yes" with a "?" indicates that the only instances with that inference were ones with other possible scenarios.

Species (lineage/clade)	Fr	IBD	RE	LDD
3	Yes?	Yes	Yes	Yes
4a	Yes?	Yes	No	Yes?
5a	Yes?	Yes	Yes	Yes
5b	Yes?	Yes	Yes?	Yes?
6	Yes?	Yes	Yes	Yes
9	No	No	Yes	No

4.9.3 Species 3 genetic and spatial patterns

Species 3 is one of the more broadly distributed small-bodied species, occupying habitats in both previously glaciated and ice-free regions. The network indentified three genetically and spatially distinct groups (clades 5-1, 4-3 and 4-4). The genetic patterns within species 3 indicated two distinct expansions into the Great Lakes region. The Tajima's D, Fu's F and Ramos-Onsins & Rozas' R_2 all indicated departures from neutrality consistent with a demographic expansion with respect to species 3. Additionally, the mismatch distribution was bimodal, which indicates structure among the mitochondrial lineages with each potentially undergoing demographic expansion. This pattern is consistent with the structure of clades 4-3 and 4-4, which are separated by 6 inferred haplotypes through clade 3-11. Co-dispersal of distinct phylogroups (with a bimodal mismatch distribution) into the Great Lakes regions has been previously detected in the freshwater mussel species *Actinonaias ligamentina* (Elderkin *et al.*, 2008).

Clade 5-1 was a poorly resolved group within the haplotype network and the sampling area included geographically disparate populations with small sample sizes. Interestingly, the haplotypes within clade 5-1 are separated by long edges of inferred haplotypes and their distribution lacks any coherent spatial signal. This branch of the COI network may represent haplotypes from older clades that have diminished in distribution. Several freshwater species south of the glacial limit have genetic and spatial patterns consistent with range contraction during the Pleistocene (Bagley *et al.*, 2013). Species 3 clades undergoing range expansion (see below) may have swept through clade 5-1 range and replaced its populations.

Clades 4-3 and 4-4 are separated by six inferred mutations and the mismatch distribution has a bimodal curve that suggests that both groups are undergoing demographic expansions. The clades nested within clade 4-3 are dominated by patterns of isolation by distance as a result of broadly distributed interior clades and geographically restricted tip clades. At the level of clade 4-3, some of 3-step tip clades (3-8, 3-9, 3-10 & 3-11) are both genetically (4-5 inferred mutations) and spatially distinct (fig. 4.4). The NCA suggests either fragmentation or long distance dispersal. Taken in conjunction with the demographic expansion indicated by the mismatch distributions and deviations from neutrality suggested by the population genetics statistics (Fu's F, Tajima's D and Ramos-Onsins & Rozas R_2), this pattern suggests a range expansion of species 3 clade 4-3 into the Great Lakes region. Ultimately, clade 4-3 traces back to clade 3-11 in Tennessee, however, the COI haplotypes linking these two groups were either not sampled or have gone extinct. Unlike the other small-bodied clades that recolonized the Great Lakes region (see below), the interior haplotypes were only detected in populations within glaciated regions.

Within clade 4-4, the central haplotype was detected in twelve populations predominately located in the St. Lawrence River drainage along with ice-free regions in Pennsylvania and West Virginia. This suggests that this clade persisted south of the Laurentide Ice Sheet during the Pleistocene glaciations within a Northern Appalachian refugium. The interior clade 2-24 indicated the potential for long distance dispersal to the southern Great Lakes region. Long distance colonization can accelerate the process of expansion and lead to patchiness in the spatial distribution of haplotype frequencies (Excoffier *et al.,* 2009). Even with the removal of clade 4-3 haplotypes, the mismatch distribution and population genetics statistics (not shown) were consistent with demographic expansion. Clade 4-4's genetic and spatial patterns suggest post-Pleistocene range expansion out of a Northern Appalachian refugium into the Great Lakes and northern Mississippi River watersheds.

4.9.4 Species 4 genetic and spatial patterns

Species 4 possesses the most spatially restricted distribution of any of the *Hyalella* species considered in this study and all the haplotypes (aside from S4H037) were detected in ice-free regions. Many *Hyalella* species in the North American Great Basin have only been detected in single habitats (Witt *et al.,* 2008). The statistical parsimony networks split the haplotypes into three distinct lineages. Lineage A was detected in two regions (Gulf and Atlantic drainages) that could have been the result of past fragmentation or long distance migration between demes. The Gulf clade 3-2 intervenes between the Atlantic clades (3-1 & 3-3) within the haplotype network. The NCA suggests either migration between groups or allopatric fragmentation coupled with incomplete lineage sorting. Despite the large COI divergences between lineages, the haplotypes

within all three were co-distributed among habitats along the Mississippi River. The multimodal mismatch distribution was inconsistent with a demographic expansion, however the population genetics statistics indicated significant departures from neutrality. This suggests a more dynamic process of colonization and extinction within this region that likely stretches back before the Pleistocene.

4.9.5 Species 5 genetic and spatial patterns

Species 5 has a deep phylogenetic split with high boostrap support that separates the lineages into two distinct groups (figure 4.1). The first assemblage consists of lineages A, C and E with distributions that are predominately north, west and south of the Appalachian Mountains. Alternatively, Lineages B and D are located east of the Appalachian Mountains within habitats in North Carolina, South Carolina and New York. Although, a few species 5 lineage A haplotypes were detected in the habitats of the second group, these are potentially the result of recent introgression (see chapter 5). The AMOVA strongly indicated that the differentiation within regions was greatly surpassed by the differentiation between groups because many of the lineages (aside from A) are restricted to separate faunal provinces. Species 5 possesses a strong signal of spatial structure, indicative of population subdivision with weak dispersal among provinces. Although, lineage B was separated into two disjunct regions, which may have been the result of long distance colonization or past fragmentation of a larger range.

Lineage A was the most diverse group in species 5 and unlike the other lineages, had a distribution that indicated a strong capability for long distance dispersal. The central haplotype of clade 4-1 was located in habitats in New Brunswick and Maine with the distribution of tip clades

indicating dispersal throughout the Great Lakes region as well as more distant localities in the drainages of the Mississippi River. The geographical distributions of the derived clades form distinct regions that imply a combination of short and long distance colonization events. The nested clade analysis identified multiple clades possessing patterns consistent with long distance colonization and contiguous range expansions. The mismatch distribution was multimodal, but possessed a large peak consistent with a simulated population undergoing demographic expansion along with several smaller jagged peaks. This pattern implies that lineage A possesses some elements undergoing demographic expansion in conjunction with more stationary populations. This is also supported by the departures from neutrality indicated by the Tajima's D, Fu's F and Ramos-Onsins & Rozas R_2. On the basis of the nested clade analysis, the Great Lakes region was colonized prior to the Mississippi River. This is similar to the spatial and genetic patterns possessed by species 3 clade 3-12 and species 6 clade 4-1 (see below).

4.9.6 Species 6 genetic and spatial patterns

The genetic patterns within species 6 indicate a complex history with the potential for both long distance colonization and allopatric fragmentation. Some degree of range expansion has invariably occurred, which is supported by the detection of species 6 haplotypes in glaciated regions and indicated by the nested clade analysis at several tiers. However, on the basis of the mismatch distribution and measures of departures from neutrality, there does not appear to be an overriding pattern of demographic expansion. The expansive elements of species 6 are restricted to clade 5-1. This group represents the northern wing of the species 6 distribution, but even those sections of the network that encompass haplotypes detected in glaciated regions do not possess

the characteristics commonly associated with demographic expansions; namely a pattern of a star phylogeny in the network centered around a broadly distributed interior haplotype with an excess of rare alleles (Excoffier *et al.,* 2008).

An interesting element of the spatial distribution of species 6 haplotypes is the recurring pattern of disjunct distributions. Both the AMOVA and SAMOVA indicated that differentiation between populations within regions had a far greater variance component than the within population variance component. Additionally, seven of twenty significant geographical associations identified by the NCA were the result of fragmentation or long distance colonization. Allopatric fragmentation is often coupled with a greater number of inferred haplotypes between areas in the network (Posada *et al.,* 2000). The longer that two groups are isolated from each other, the greater the divergence. Alternatively, if there are few mutational steps, especially at lower tiers in the nested haplotype network, then long distance dispersal is a valid explanation of the genetic patterns. Within species 6, both of these outcomes likely exist to some degree. Clade 5-1 possesses a greater potential for long distance colonization and contiguous range expansion, especially in previously glaciated regions near the Great Lakes and the St. Lawrence River. Alternatively, clades 5-2 and 5-3 possess long edges of inferred haplotypes that indicate past fragmentation of larger ranges. The distribution of clade 5-2 encompasses distinct regions within the Mississippi River drainage along with a few disparate habitats in Oklahoma and Ontario. Unlike its large-bodied counterparts that likely persisted in Mississippian or Missourian refugia, clade 5-2 was not detected in glaciated regions. Clade 5-3 is located in discrete areas along the eastern coast of North America from Georgia to Newfoundland. Similar to the spatial patterns identified in species 5, the Appalachian Mountains continue to represent a potential barrier to gene flow that splits distinct genetic groups. The

haplotype network of this clade indicates that the haplotypes in the southern regions are ancestral, indicating a northward progression either through long distance colonization or fragmentation of a larger range.

4.9.7 Species 9 genetic and spatial patterns

Species 9 is restricted to a region along the southern limit of the Cordilleran Ice Sheet. The nested clade analysis indicated a contiguous range expansion throughout this area at the 4-step and 5-step tiers that likely coincides with the retreat of the cordilleran ice sheet. However, Tajima's *D* and Ramos-Onsins & Rozas' R_2 statistics indicated no significant departures from neutrality that would be consistent with a demographic expansion.

4.9: CONCLUSIONS

The small-bodied species 4, 5 and 6 (clades 5-2 and 5-3) along with the large-bodied species 8 consist of clades that are spatially partitioned into eastern and western groups along the southeastern Coastal Plain. As discussed in chapter 3, this pattern has been detected in multiple species (Soltis *et al,* 2006, Newman & Rissler, 2011, Butler *et al.,* 2011, Campbell-Staton *et al.,* 2012 Ennen *et al.,* 2012), and partially formed the basis of Avise's (2000) conceptualization of the field of phylogeography. Species 3, 5 and 6 possessed spatial and genetic patterns with range expansions that progressed out of eastern refugia into the Great Lakes region. In species 4, 5 and 6, the geographic distributions of these groups extend along the eastern coast, indicating that the partitions are potentially shaped by limited dispersal across the Appalachian Mountains and north into the maritime provinces; notably recent introgression among species 5 populations.

Chapter 5

Comparison between large and small-bodied *Hyalella* species

5.1: INTRODUCTION

Species with a greater ability to disperse will erode the isolation by distance signal and result in a weaker slope when plotting pairwise population F_{ST} values against pairwise population geographic distances. Traditionally, F_{ST} is calculated on the basis of heterozygosity, however an F_{ST} analogue (ϕ_{ST}) can be estimated with mitochondrial sequences, which converts sequence divergences and haplotype frequencies into Euclidean vectors (Excoffier & Lischer, 2010). The dispersal ability of the *Hyalella* species will be assessed using comparisons of isolation by distance curves (Diniz-Filho *et al.,* 2013). Correlation between ϕ_{ST} and geographic distance matrices will be assessed with a Mantel test (Mantel, 1967), which operates on the basis that if two matrices are correlated then the sum of the products, known as the Mantel correlation coefficient (*r*), will be greater than those derived from random permutations of the same data. This process is repeated and the percentage of scores above the observed *r*-value is considered the *p*-value. The slope of the isolation by distance curve can be estimated usng reduced major axis regression (Bohonak, 2002), which is considered appropriate for ecological analyses where the independent variable is measured with error (Hellberg, 1994). This process has been used to disentangle spatial patterns among ecotypes in the anadromous prickly sculpin (Dennenmoser *et al.,* 2014) and compare genetic diversity among populations of freshwater mussels (*Cumberlandia monodonta*) in glaciated/unglaciated regions (Inoue *et al.,* 2014).

The objective of this chapter is to determine if differences in patterns of isolation by distance can differentiate small-bodied *Hyalella* from their large-bodied counterparts. High rates

of dispersal or migration among demes can prevent local adaptation and homogenize genetic variability across regions. As seen in chapter 3, the large-bodied *Hyalella* species were detected throughout glaciated North America; species 1, 2 and 7 possessed broad distributions that included several faunal provinces. This was especially notable in species 1 lineage C and species 7, whereby ancestral haplotypes (S1H139 and S7H004, respectively) were located in nearly every population for that species. In stationary populations (as might be expected south of the glacial limit), ancestral haplotypes will eventually be detected throughout a species range as a consequence of migration. However, species 1 lineage C and species 7 have mismatch distributions that indicate recent as well as rapid demographic expansion. Although not as pronounced, species 1 lineage A and species 2 clade 5-1 also had broadly distributed ancestral haplotypes. Large-bodied *Hyalella* possess larger gnathopods, which may aid in dispersal by grasping to waterfowl. This is supported by anecdotal evidence in the Canadian prairies, in which *Hyalella* have successfully colonized temporary freshwater potholes without direct access to drainage systems. In this chapter the following hypothesis will be tested; large-bodied *Hyalella* species are better dispersers than their small-bodied counterparts and will have shallower regression slopes with positive Mantel correlations between pairwise ϕ_{ST} and geographic distances. Alternatively, if no patterns among isolation by distance curves are apparent, then additional factors are restricting the dispersal of *Hyalella* lineages; limited access to proglacial dispersal routes and ecological interactions among *Hyalella* species as well as centrarchids are potential explanations. Furthermore, a Mantel test only indicates the presence of spatial structure with regard to the genetic variability and interpretations must be stated with caution as isolation by distance may not be the sole explanation for observed patterns (Miermans, 2012). Specifically, recolonization of previously glaciated habitats may result in

geographical clustering of populations and Mantel tests may not be sufficient to differentiate this from patterns of isolation by distance (Miermans *et al.*, 2011).

5.2: MATERIALS AND METHODS

Pairwise ϕ_{ST} values were determined using arlequin 3.5 (Excoffier & Lischer, 2010) using a Tamura-Nei model to assess genetic distance. Only populations with 5 or more sequences were considered to limit the number of pairwise ϕ_{ST} scores of 1 and the analysis used *Hyalella* lineages with at least 5 or more such populations. The matrices of pairwise ϕ_{ST} and geographic distances were compared using a mantel test (10000 repetitions) in Arlequin 3.5 Excoffier & Lischer, 2010). The coordinates were plotted on a scatterplot in Excel 2010 and the reduced major axis regression line was estimated by RMA v 1.17 (Bohonak, 2002), with 95% confidence intervals derived from 10,000 bootstrap replicates.

5.3 RESULTS

The major lineages within six members of the *Hyalella* complex (species 1 (A, H, I), 2, 3, 5 (A), 6, 7) were analyzed for patterns of isolation by distance. These groups represent the major lineages within the species complex that re-colonized glaciated regions. The other species (or lineages) were either completely restricted to unglaciated areas of North America or had only a few populations with five or more samples. The mantel tests indicated a significant positive association between pairwise ϕ_{ST} and geographic distances in all groups analyzed except species 1 lineages H and I (large-bodied morphotype) along with species 3, as well as species 5 lineage A (small-bodied morphotypes).

5.3.1 Large-bodied *Hyalella* isolation by distance curves

Of the large-bodied *Hyalella* analyzed, species 1 (lineages A and I), 2 and 7 indicated positive correlations between genetic ϕ_{ST} and geographic distance (table 5.1). The isolation by distance analysis of species 1 lineage A (fig. 5.1) included 38 populations with pairwise ϕ_{ST} (ranging from –0.1803 and 1) and geographic distances spanning 9.4 to 1214.6 km for which the Mantel test indicated a positive correlation (r=0.253, $p < 0.001$). The isolation by distance curve of species 1 lineage H (6 populations) had a negative slope (fig. 5.2) and the mantel test was not significant (r=0.320, $p = 0.116$, respectively). Similar to lineage A, the isolation by distance curve of species 1 lineage I had a positive slope (fig. 5.3) and the mantel test indicated a significant positive correlation between pairwise ϕ_{ST} and geographic distances (r=0.260, $p = 0.0429$). Sixteen lineage I populations were considered that encompassed a range between 1.8 and 2326 km.

Table 5.1: Summary of *Hyalella* Isolation by distance curves
This table includes the statistics for the isolation by distance curves including the species (lineage), number of populations analyzed, Mantel Test (with significance) and the Reduced Major Axis (RMA) regression with 95% confidence intervals determined by 10,000 boostraps.

Species (Lineage)	Morphotype	Number of populations	Mantel Test	RMA Slope	RMA 95% Confidence Intervals (lower, upper)
1 (A)	Large	38	0.253 ($p<0.001$)	1.11×10^{-3}	1.05×10^{-3}, 1.18×10^{-3}
1 (H)	Large	6	0.320 ($p=0.116$)	-5.20×10^{-4}	-6.00×10^{-4}, -4.50×10^{-4}
1 (I)	Large	16	0.261 ($p=0.043$)	1.81×10^{-4}	1.46×10^{-3}, 2.16×10^{-3}
2	Large	51	0.498 ($p<0.001$)	2.71×10^{-4}	2.59×10^{-4}, 2.82×10^{-4}
3	Small	41	-0.03 ($p<0.621$)	-4.71×10^{-4}	-5.03×10^{-4}, -4.95×10^{-4}
5 (A)	Small	18	0.140 ($p=0.174$)	2.20×10^{-4}	1.88×10^{-4}, 2.69×10^{-4}
6	Small	26	0.229 ($p=0.009$)	3.32×10^{-4}	3.06×10^{-4}, 3.66×10^{-4}
7	Large	69	0.730 ($p<0.001$)	4.08×10^{-4}	3.77×10^{-4}, 4.39×10^{-4}

Figure 5.1: Species 1 lineage A isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The mantel test was significant for a positive correlation among ϕ_{ST} and geographic distance ($p < 0.001$).

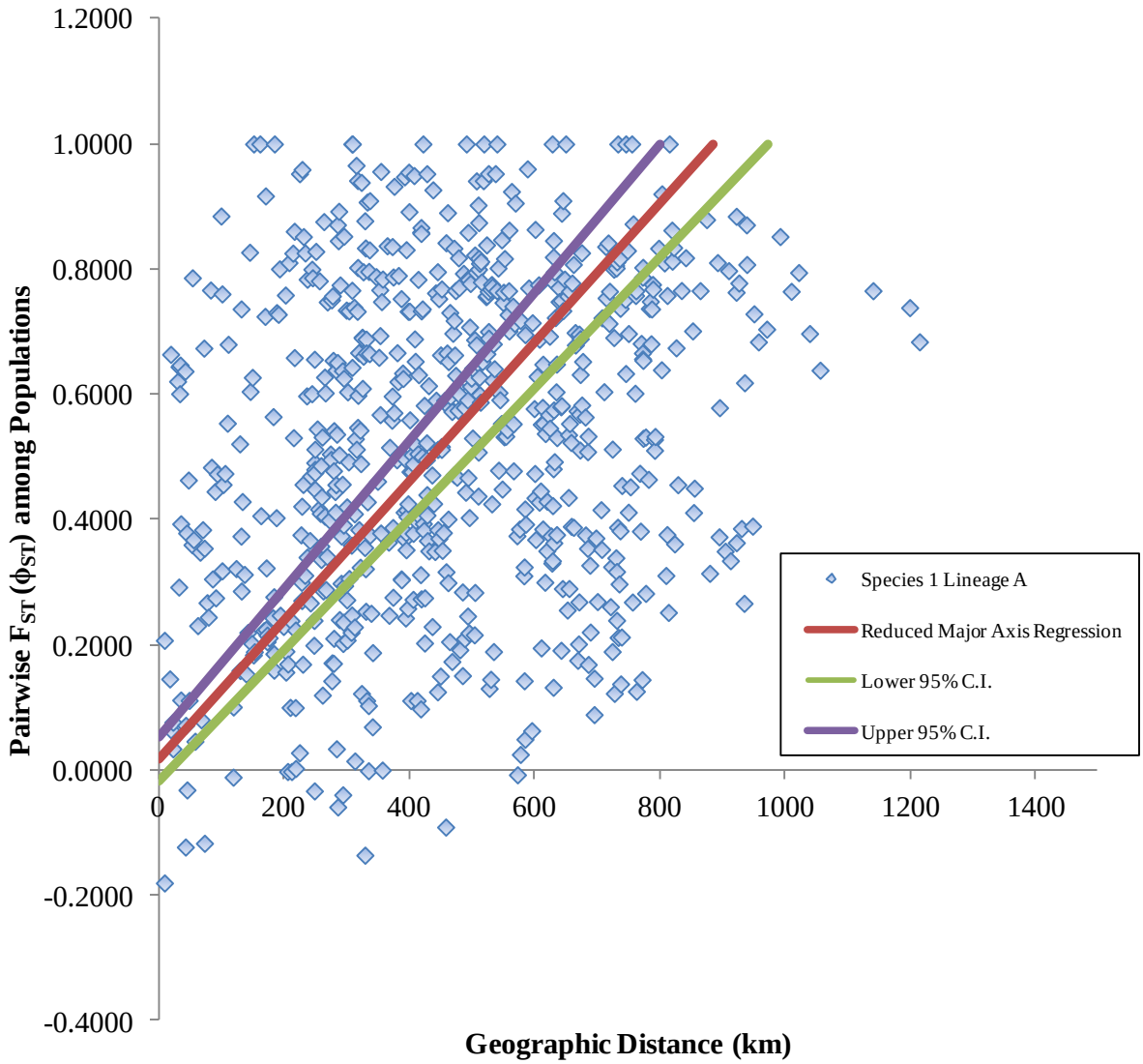


Figure 5.2: Species 1 lineage H isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was not significant for a positive correlation among ϕ_{ST} and geographic distance ($p = 0.116$).

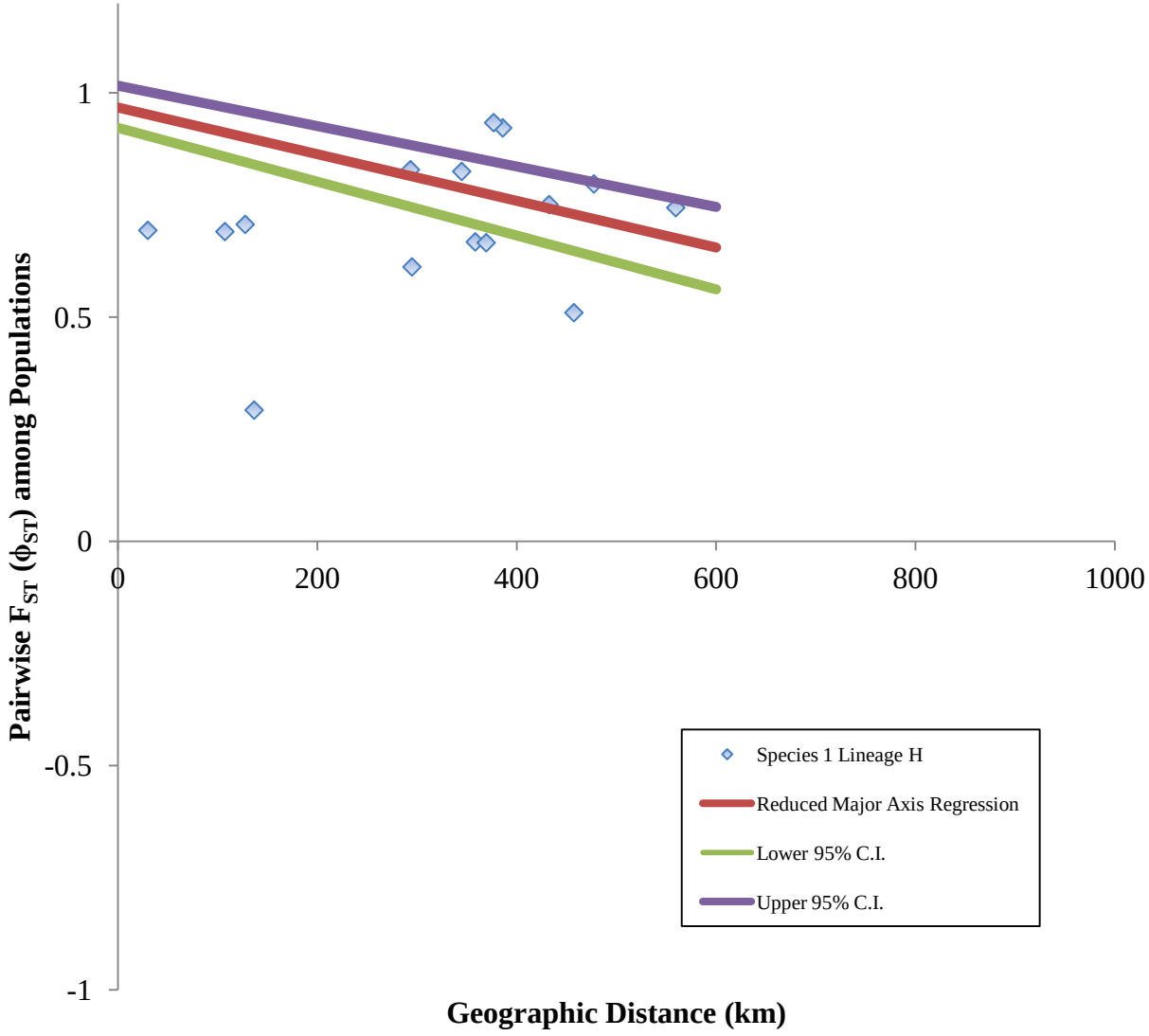
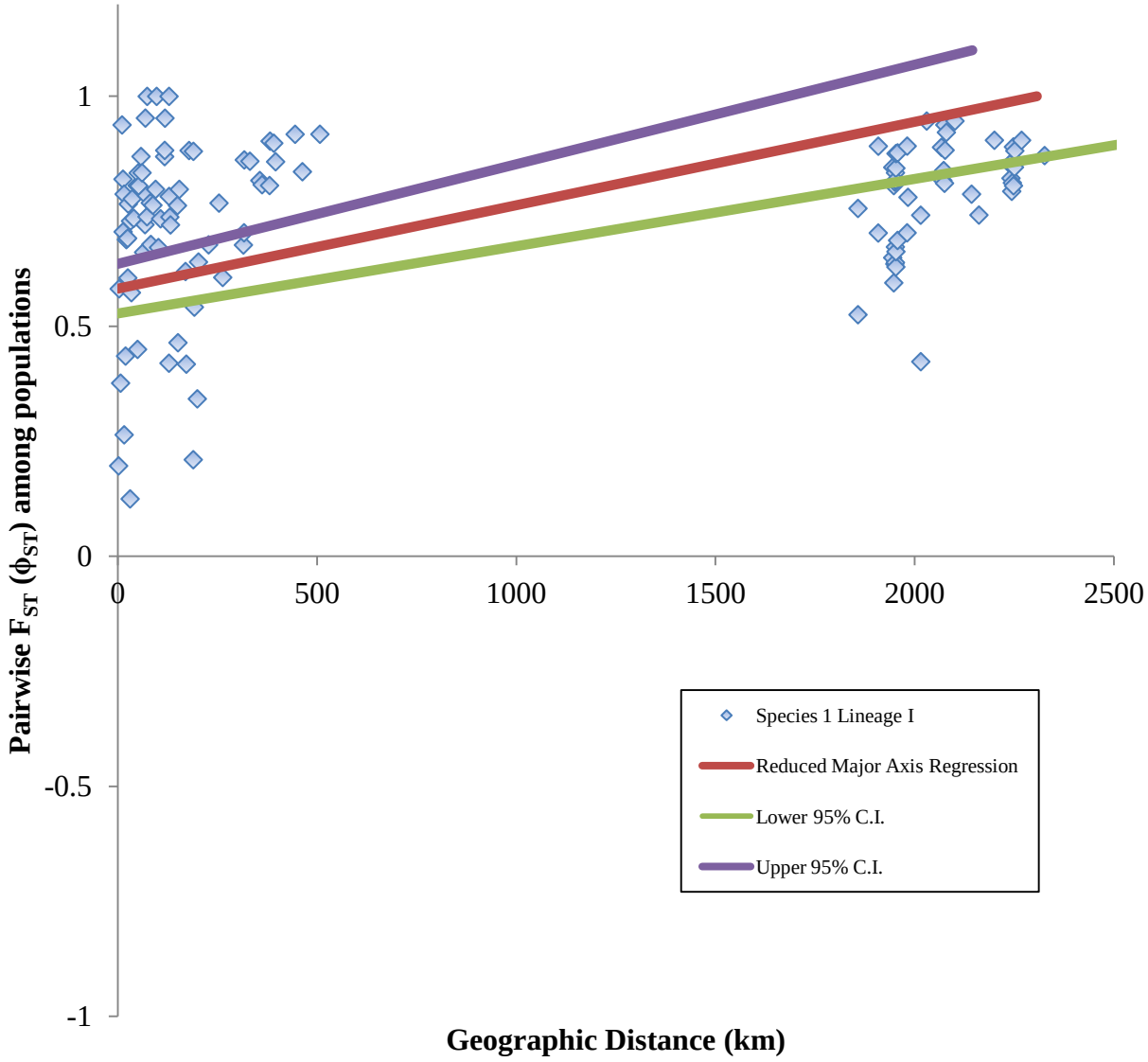


Figure 5.3: Species 1 lineage I isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was significant for a positive correlation among ϕ_{ST} and geographic distance ($p = 0.0429$).



The species 2 analysis of isolation by distance included 51 populations (table 5.1) that were separated by distances of 4.7 to 4520.7 km. The reduced major axis regression had a positive slope (fig. 5.4) and the Mantel test indicated a positive correlation (r=0.498, $p < 0.001$).

The species 7 isolation by distance curve had a positive slope (fig. 5.5) with a significant correlation between ϕ_{ST} and geographic distance (r=0.73, $p < 0.001$) with 27 populations spanning a range from 8.4 to 4185 km.

5.3.2 Small-bodied *Hyalella* isolation by distance curves

The isolation by distance analyses of species 5 lineage A (fig. 5.7) and species 6 (fig. 5.8) had positive correlations between pairwise ϕ_{ST} and pairwise geographic distance (r=0.14, $p=0.174$ and r=0.229, $p=0.009$, respectively). The pairwise geographic distances within species 5 lineage A (18 populations) and species 6 (36 populations) ranged from 4.5 – 4234 km and 12.7 – 3124 km, respectively. The species 3 analysis included 41 populations that spanned distances from 28.6 to 2979 km and the Mantel test did not indicate a significant correlation (r= -0.03, $p = 0.621$) resulting in a negative reduced major axis regression slope (fig. 5.6).

Figure 5.4: Species 2 isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was significant for a positive correlation among ϕ_{ST} and geographic distance ($p < 0.001$).

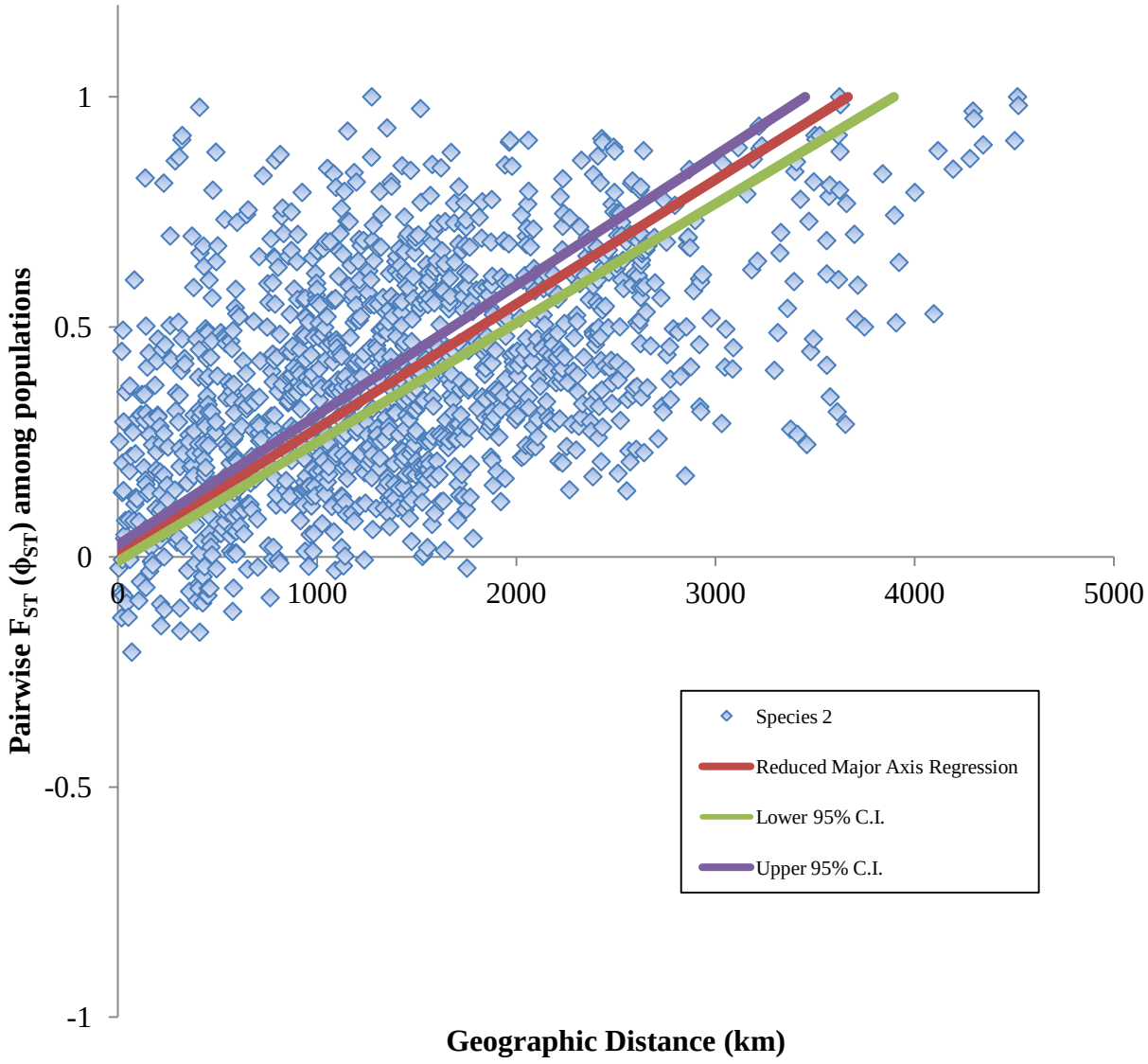


Figure 5.5: Species 7 isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was significant for a positive correlation among ϕ_{ST} and geographic distance ($p < 0.001$).

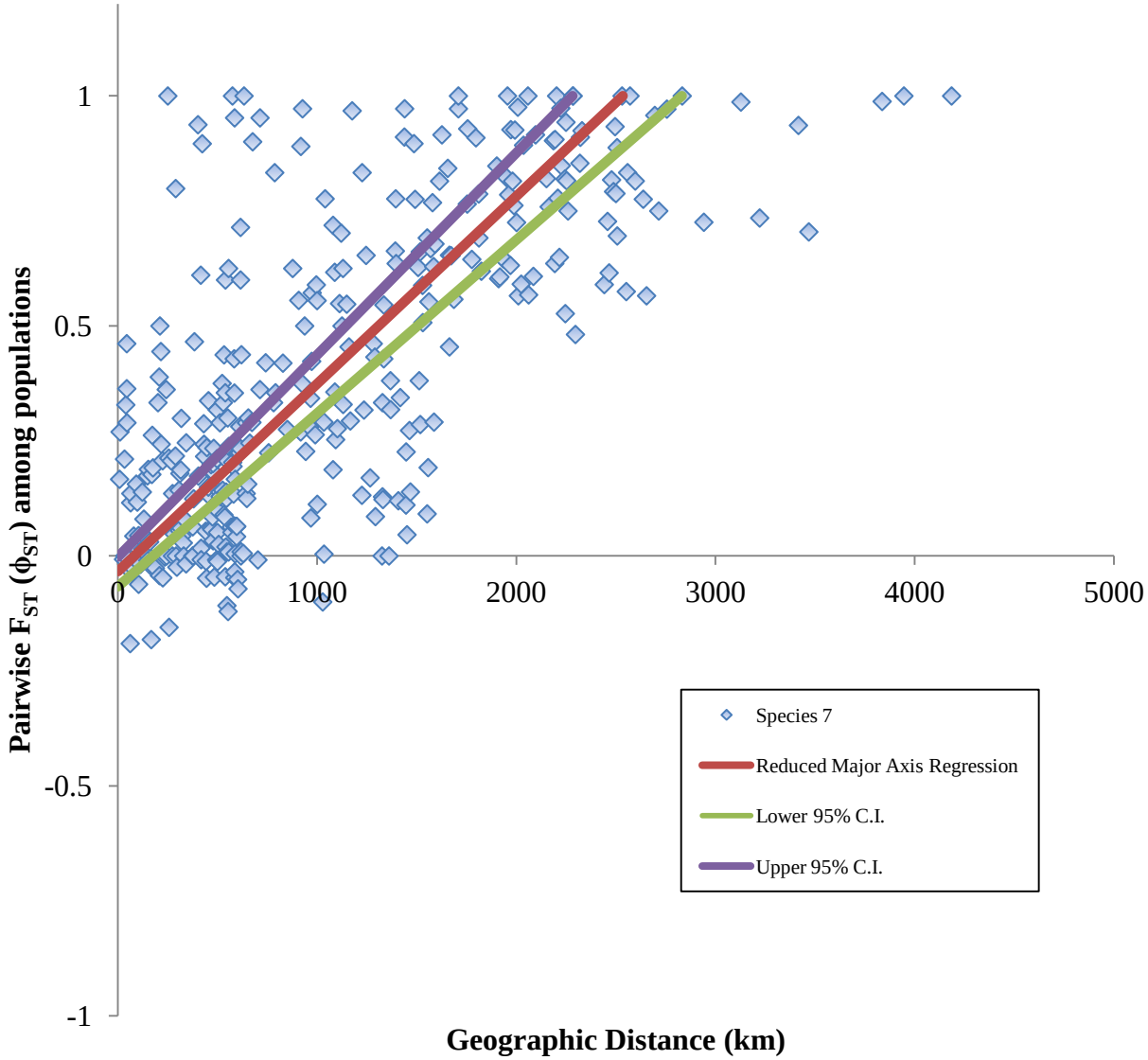


Figure 5.6: Species 3 isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was not significant for a positive correlation among ϕ_{ST} and geographic distance ($p < 0.621$).

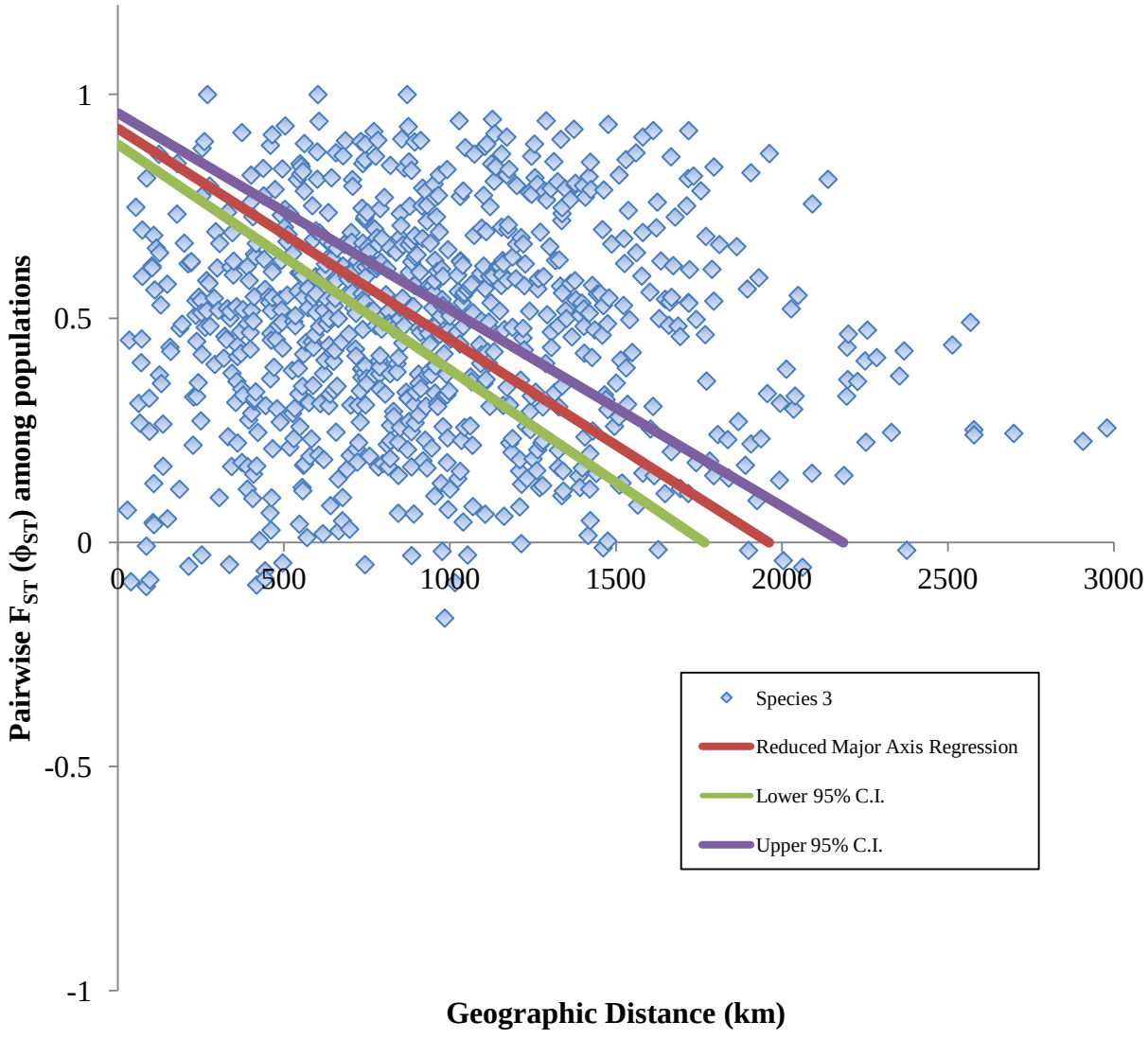


Figure 5.7: Species 5 lineage A isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was not significant for a positive correlation among ϕ_{ST} and geographic distance ($p = 0.174$).

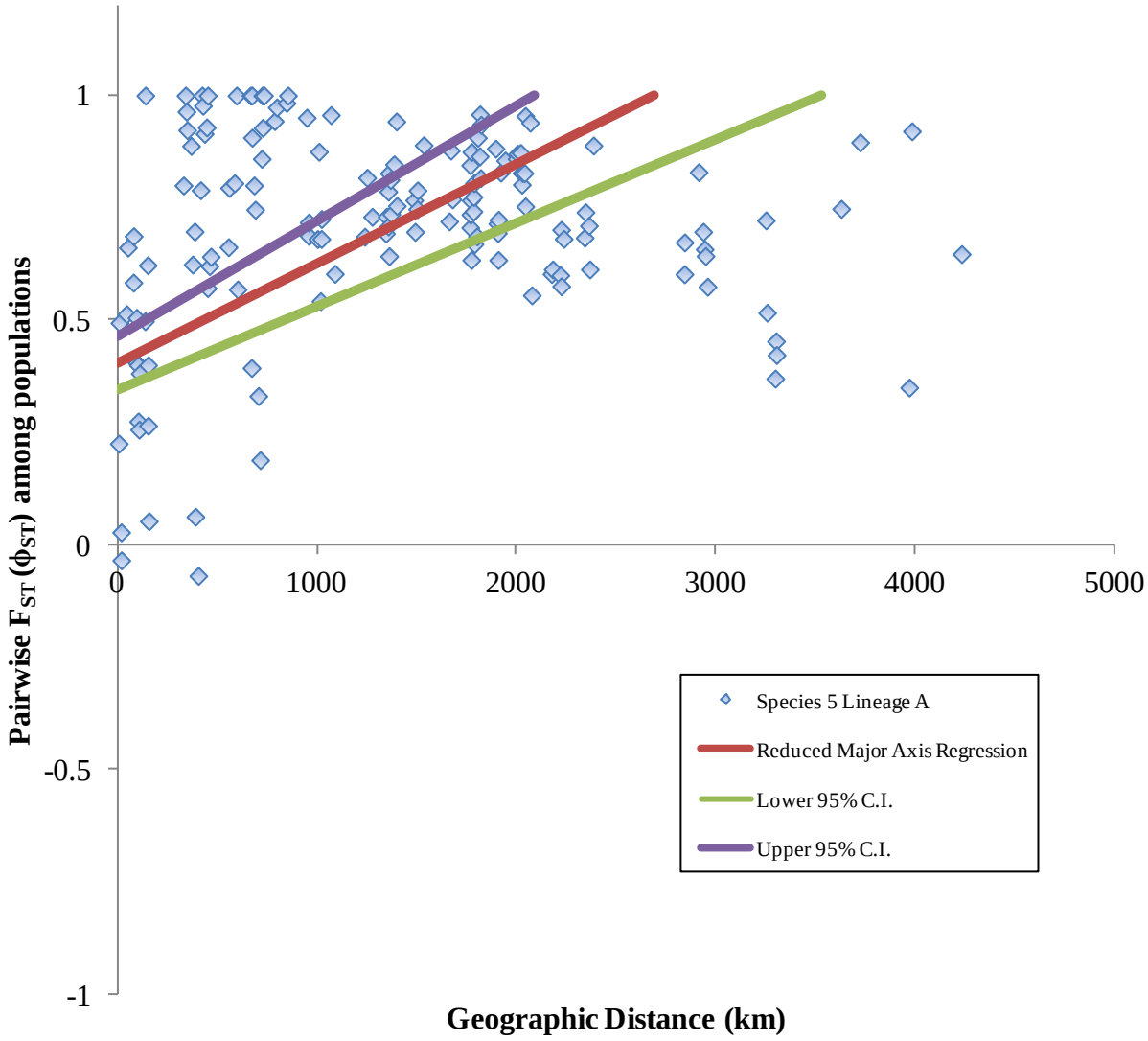
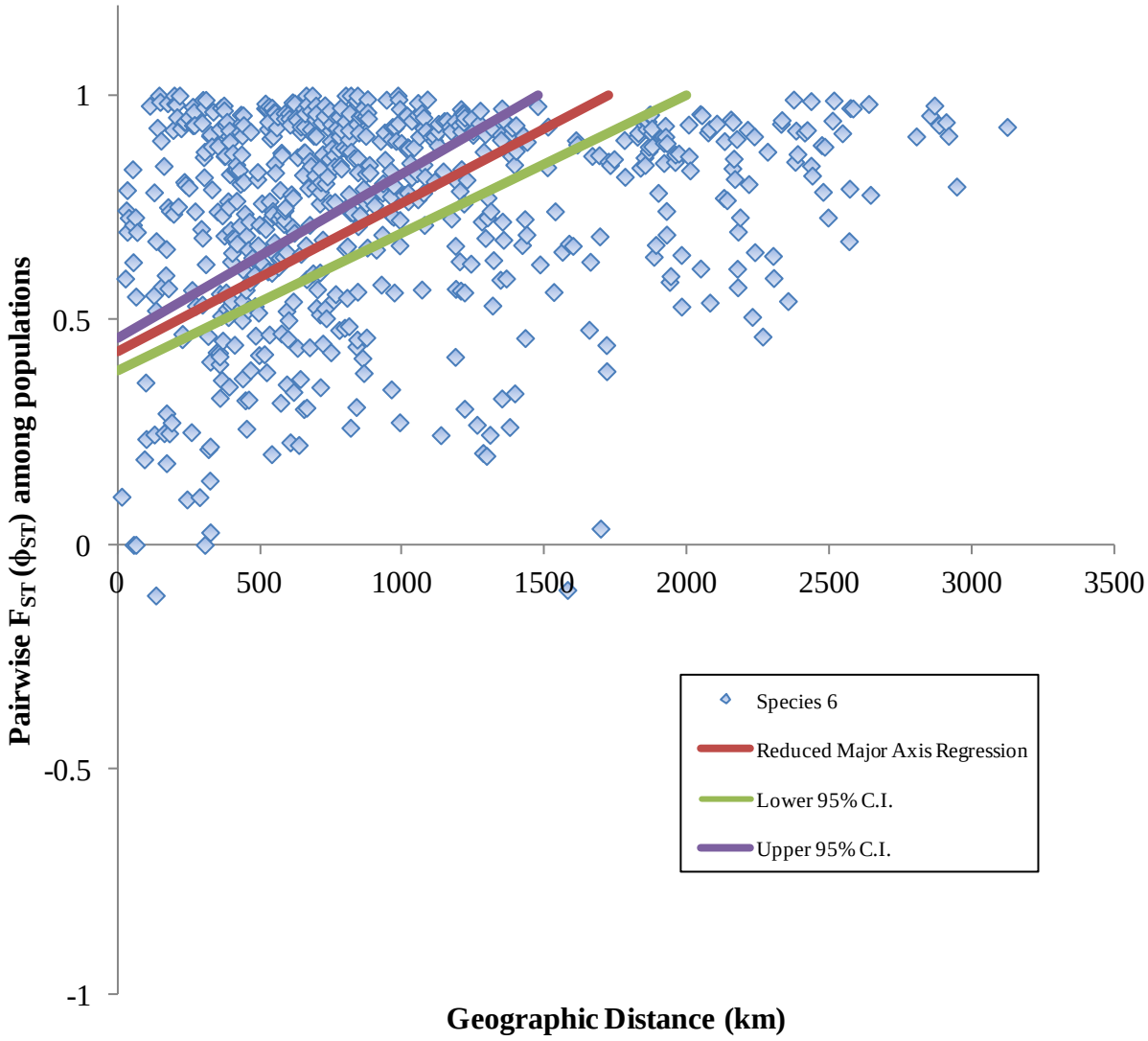


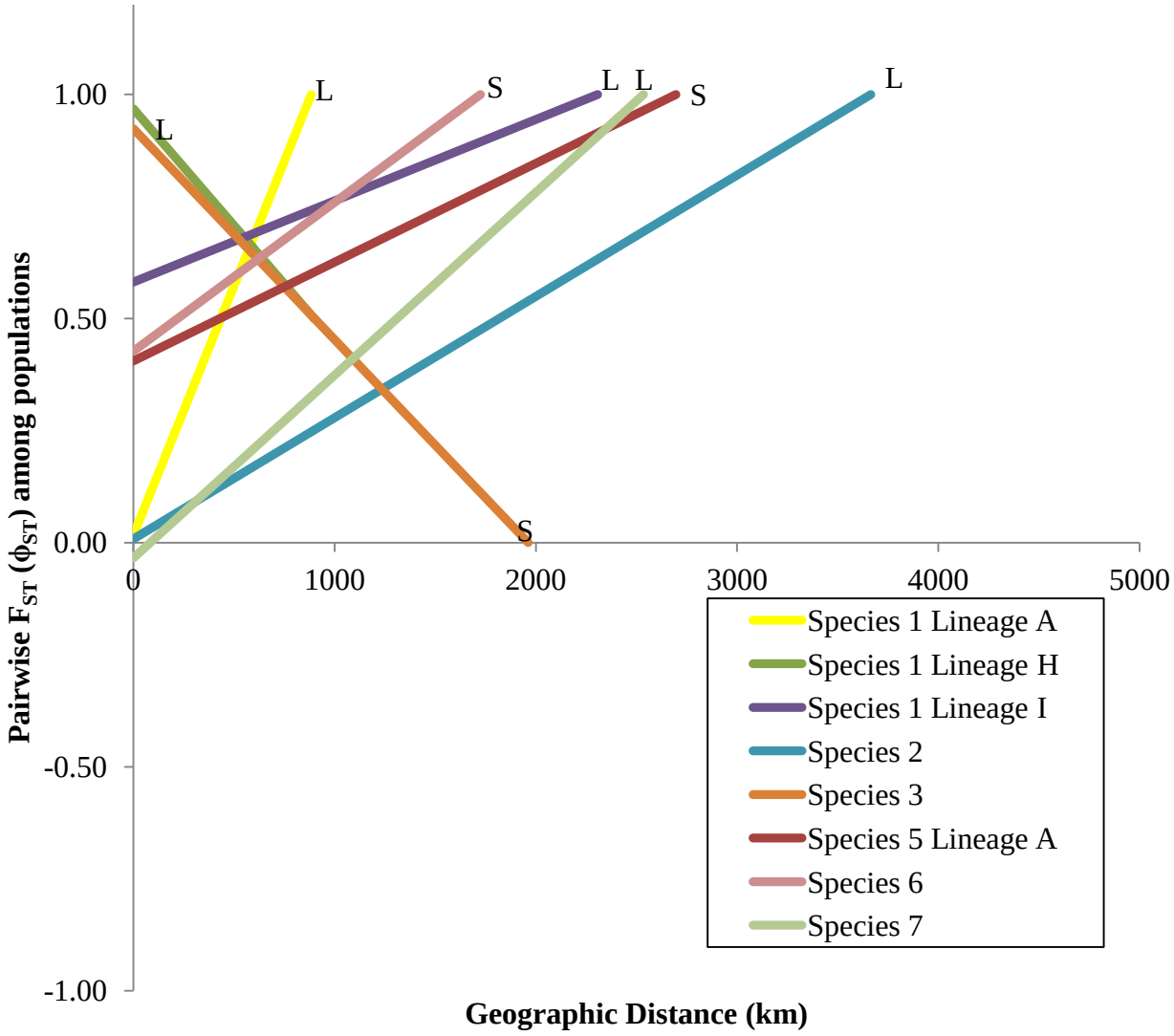
Figure 5.8: Species 6 isolation by distance curve
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (confidence limits determined through bootstrapping). The Mantel test was significant for a positive correlation among ϕ_{ST} and geographic distance ($p = 0.009$).



5.4: DISCUSSION

The reduced major axis regressions of the isolation by distance curves were plotted on the same graph to highlight possible differences in patterns among small and large-bodied *Hyalella* (fig. 5.9). Species 1 lineage A and species 7 had the steepest slopes, while species 1 lineage H (large-bodied) and species 3 (small-bodied) had negative regression slopes that indicate no association among population ϕ_{ST} and population geographic distances. The species of *Hyalella* with positive correlations had reduced major axis regressions with slopes that ranged from 1.11×10^{-3} (species 1 lineage A) to 1.81×10^{-4} (species 1 lineage I) with slopes of both small and large-bodied morphotypes falling in between. This pattern was not consistent with the hypothesis that large-bodied *Hyalella* are better dispersers than their small-bodied counterparts. The analysis of species 3 includes a clade of deeply divergent COI haplotypes (clade 5-1) that was detected throughout the species range. The distribution of species 3 was potentially the result of two distinct COI clades (5-2 and 5-3) sweeping through the Great Lakes and Mississippi River basin replacing the potentially older clade 5-1 haplotypes. As such, the negative isolation by distance slope may be the result of these more divergent haplotypes influencing the pairwise ϕ_{ST}. What was evident from the comparison of regression slopes was that there was no clear pattern that distinguishes small and large-bodied *Hyalella* lineages.

Figure 5.9: Reduced major axis regressions of the *Hyalella* isolation by distance curves
Pairwise ϕ_{ST} values between populations were plotted against population geographic distances with a reduced major axis regression (see fig 5.1 – fig 5.8). Species 1 lineage (A, H, I), species 2 and 7 are large-bodied *Hyalella*. Species 3, 5 lineage A and 6 are small-bodied *Hyalella*. The large and small-bodied *Hyalella* regressions are denoted as L and S, respectively.



5.4.1: Comparison of large and small-bodied *Hyalella* distributions

The distribution of the *Hyalella* species complex is the result of Pleistocene vicariance in conjunction with post-glacial dispersal. The Great Lakes suture zone represents an admixture of both large and small morphotypes. The larger *Hyalella* species in glaciated North America (species 1, 2 & 7) are likely derived from western refugia and those lineages that persisted south of the glacial limit (aside from the Cascadian species 1 lineage I) were able to exploit the proglacial lake system of the Peace-Agassiz corridor. This resulted in the broad distribution of larger *Hyalella* species throughout the Great Lakes, Hudson Bay and Yukon-Mackenzie drainages. The NCA of species 1 lineage A, species 2 and species 7 (chapter 3) provide clear evidence that they invaded the Great Lakes regions from the west. However, the true origins of many of the large-bodied *Hyalella* cannot be known for certain without sampling the refugial habitats with a focus on the Missouri River basin in the northern United States.

The small-bodied *Hyalella* species (3, 5 and 6) possess spatial patterns more consistent with re-colonization of eastern previously glaciated regions from populations originating in Appalachian or Atlantic refugia. Freshwater fish species that dispersed out of Atlantic refugia (April *et al.,* 2013*)* have exploited the pro-glacial lake system that connected proglacial Lake Hall (now the Finger Lakes) through the Lake Ontario basin and into the Lake Champlain basin. With respect to the small-bodied *Hyalella*, this pattern is most apparent in species 6 clade 4-2. Alternatively, species 3 clade 5-2, species 5 lineage A and species 6 clade 4-1 re-colonized the Great Lakes region followed by the Mississippi River basin. In the case of the species 3, its clades may have swept over existing or declining populations that had undergone range contraction during the Pleisotcene. *Diporeia hoyi* (Usjak, 2010) has a similar split in regional groups within the Great Lakes; however, both groups were attributed to western refugia

(Mississippian and Missourian). As a glacial relict that is restricted to deep cold lake waters, this invertebrate disperses through connected waterways only. The nested clade analyses of the small bodied *Hyalella* clearly suggest an eastern origin and their dispersal must have allowed for overland dispersal via animal vectors (chapter 4). Interestingly, the longitudinal orientation of the dispersal vector is not what would be expected by avian transport whose flight paths are along latitudinal lines (Figuerola *et al.,* 2002). Other small bodied groups (species 4, 5 lineages B/C and 9) were not located north of the glacial limit and possess far more limited geographical distributions than their northern counterparts.

To what extent is the great range encompassed by the large-bodied *Hyalella* in glaciated North America a product of greater dispersal ability or being in the right place at the right time to exploit the Lake Agassiz dispersal corridor? While species 5 lineage A and species 6 have ranges that are almost as great as their large-bodied counterparts, they include both glaciated and unglaciated regions. Many of the species 1 lineages along with species 2 and 7 occupy expansive regions in glaciated North America that had to have been re-colonized since retreat of the glaciers. However, significant patterns of isolation by distance have been detected in the large-bodied species. Genetic variability within species 1 lineage A, species 2 and 7 was strongly correlated with genetic distance. This suggests that dispersal in large-bodied *Hyalella* was largely unidirectional, a pattern that is more consistent with a stepping-stone model rather than an island model (Slatkin & Excoffier, 2012).

As explored in chapter 3, the spatial patterns in large-bodied *Hyalella* resemble more closely *Sida crystallina* (Cox & Hebert, 2001) and freshwater fish species (Borden & Krebs, 2009; April *et al.,* 2013) than the broad dispersal patterns of *Daphnia pulex* (Crease *et al.,* 2012). The NCA clearly demonstrates the potential for long distance dispersal and the ranges of these

species encompass multiple watersheds, indicating potential dispersal via avian vectors. However, the extent of this means of dispersal is insufficient to overcome the patterns of isolation by distance. In comparison to small-bodied *Hyalella* in glaciated regions, the differences in spatial distributions seem to be shaped by shared dispersal pathway rather than differences in dispersal abilities. While the large-bodied *Hyalella* re-colonized glaciated regions from the west, the small-bodied species dispersed from the east into the Great Lakes and the Mississippi River basin.

Chapter 6

Phylogeographic Analysis of the *Hyalella* First Internal Transcribed Spacer

6.1 INTRODUCTION

Over-reliance on a single molecular marker can reduce the confidence in inferences derived from spatial and genetic patterns. Congruence among molecular markers can provide a greater degree of confidence, especially when the morphological characters are cryptic or complex (Cunha *et al.,* 2015). Furthermore, the differences between the process of inheritance between mitochondrial and nuclear markers (haploid vs. diploid, uniparental vs. biparental) can provide different resolutions regarding the demographic history of a species. Mitochondrial markers evolve more rapidly as a consequence of lower effective population sizes and only provide the demographic history of the maternal line. To counter these potential biases, nuclear markers provide another line of analysis to study the effects of the Pleistocene glaciations and may supply insights into the broad spatial and genetic patterns of the *Hyalella* species complex. The first internal transcribed spacer (ITS1) of the small ribosomal subunit gene is receiving greater attention in the literature as the search for a barcoding gene that can delineate species boundaries within algae, fungi, animals and protists continues (Wang *et al.,* 2015). While second internal transcribed spacer (ITS2) has received greater attention in eukaryotes, with an online database consisting of sequences and secondary structures (Coleman, 2013), ITS1 may be a useful marker if its structure can be determined. Furthermore, very few studies that analyze ITS1 include multiple samples from the same individual (Harris & Crandall, 2000), which may overlook intragenomic variation (Ruiz-Estévez *et al.,* 2015).

6.1.1 Concerted evolution of multigene families

The small ribosomal subunit is part of a multigene family with repeated units that are potentially located within several different loci, either on the same or different chromosomes. This ribosomal DNA (rDNA) gene family consists of a tandem repeat of several genes (Sloan *et al.*, 2013); reading from 5' to 3' are the first external transcribed spacer, 18S rDNA gene, the first internal transcribed spacer, 5.8S rDNA gene, the second internal transcribed spacer, 28S rDNA gene, the second external transcribed spacer and finally the intergenomic spacer (IGS). These paralogous subunits are derived from genomic duplication events and were once thought to evolve independently of each other (Brown *et al.,* 1972). As such, it was expected that each copy of IGS would possess a distinct evolutionary history and the magnitude of molecular divergences within a species should be as great as between species. However, Brown *et al.,* (1972) discovered that the IGS divergences were much greater between *Xenopus* congenerics than the corresponding intraspecies variation. Distinct copies of the IGS were not evolving independent of each other. To account for this, a new model of concerted evolution was proposed, whereby repeated intraspecific genomic elements are homogenized by recombination among subunits (Ganley & Kobayashi, 2007). Several possible mechanisms have been proposed, including unequal crossing-over, gene conversion, replication slippage, duplicative transposition and retrotransposition (Dover, 1982; Eickbush & Eickbush, 2007).

Unequal crossing-over and gene conversion involve unreciprocal transfer of genetic information from one locus to another. With unequal crossing-over, large amounts of DNA can be inserted into the receiving strand, resulting in the complete duplication of genes.

Homogenization of the repeats will occur over multiple cross-over events as the genes are gained and lost, since the genome can only accommodate a limited number of functional duplicates before there is a cost to fitness (Kobayashi, 2011). On the other hand, gene conversion inserts a much smaller band of DNA and will not increase the number of genes (Nei and Rooney, 2005). Gene conversion is initiated by a double stranded break followed by degradation of the 5' ends. The 3' overhang forms heteroduplexes with other paralogs, resulting in recombination and homogenization (Eickbush & Eickbush, 2007). Replication slippage is an error in the DNA replication mechanism whereby tandem repeats are gained or lost. Transposition, or retrostransposition if there is an RNA intermediate, occurs if the small ribosomal subunit is bordered by transposable elements and inserts itself in other places throughout the genome

The rate of concerted evolution is indirectly proportional to the mutation rate; the larger the rate, the greater amount of variation that will persist. Similarly, the number and location of the loci can have an effect on the rate of homogenization. The centromere and telomere are replete with repeated elements that could impair chromosome segregation if the size of the region varies (Charlesworth, 2012). Thus, any unequal crossing-over events would be selected against and this could prevent homogenization among separate loci. Lower rates of chromosomal interchange, as seen with X-Y chromosomal pairings in *Drosophila melanogaster,* can also prevent recombination and lead to independent lineages (Charlesworth, 2012).

6.1.2 Three-dimensional structure of rRNA sequences and analysis of covariation

The functionality of rRNA molecules depends on the folding of the transcribed sequence into a three-dimensional structure (Westhof *et al.*, 2011). Alignment of rRNA sequences must take into

account the structure of the molecule to ensure the comparison of homologous regions. A robust element of rRNA structures is that there are numerous possible sequences that can fold into the same three-dimensional configuration (Gutell *et al*, 2002). However, determination of the structure can be difficult. Historically, a minimum energy optimality criterion has been used to identify regions of the rRNA sequence capable of forming helical structures (Coleman, 2013). A variety of programs exist (MFOLD, RNAfold) that attempt to reconstruct the structure by minimizing the amount of necessary energy needed to maintain the base pair bonds (Washietl *et al.*, 2012). This approach can be problematic for numerous reasons. While some of the energy rules used by the minimum energy optimality criterion are determined experimentally, others are derived or are estimates. Furthermore, these programs must contend with the distinct possibility that temperature-based folding variation will override any attempt to ascertain a common structure (Armbruster, 2001). The difficulty in ascertaining folding structures is exacerbated within freshwater poikilotherms, as their internal temperatures will vary according to the environment. *Hyalella 'azteca'* occupies a wide range of latitudes (see chapters 2 – 4). Additionally, the minimum energy optimality criterion cannot account for external processing. The ultimate shape of the rDNA structure can be influenced by third party molecules that can enforce non-optimal base pair bonds (Fujita *et al.* 2009).

In order to determine the structure of the first internal transcribed spacer, an analysis of covarying sites is necessary. Helices in these molecules are maintained by the base pairings and mutations along one side of its stem can disrupt the structure. As such, in order to maintain the functionality of the molecule, either there must be a compensatory base pair mutation on the complementary strand or the structure must be enforced by a third-party molecule (Wolf *et al.*, 2013). For example, if a C – G pairing had a mutation resulting in A – G, then a mutation of the

G to U would maintain the bond. Single compensatory mutations are also possible with the nitrogenous base T. In rRNA molecules, Uracil can bind with either Guanine or Adenine. Therefore, an A – U mutation to G – U would not require a second compensatory mutation. The best sequences for this type of analysis are phylogenetically similar ones, whereby conserved elements provide anchor points for the alignment (Gutell *et al,* 2002). Distantly related sequences can be problematic, because too much variation will erroneously lead to the comparison of non-homologous regions. Two completely dissimilar sequences will appear as if all the sites had undergone compensatory mutations. Compensatory mutations in structurally conserved regions of the internal transcribed spacers are often correlated (~93%) with species boundaries (Wolf *et al.,* 2013). Covariation among sites can be ascertained through a chi squared test that compares the observed occurrences of a base pair at a given site (i.e A-T) against the expected number of the 16 possible combinations as determined from the number of each nucleotide at the sites.

The phylogenetic utility of ribosomal RNA multigene families derives from the within-species similarity of the intergenic components and the deep divergences between species (Weider *et al.,* 2005). While the genic regions have little variation as a consequence of strong purifying selection, ITS1, ITS2 and IGS will freely homogenize within a species and may provide a strong indication of genetic barriers among phylogroups. The first internal transcribed spacer has been successfully used to characterize phylogentic relationships in snails (Uit de Weerd *et al.,* 2004; Moore *et al.,* 2015), ostracods (Gandolfi *et al.,* 2001), copepods (Zagoskin *et al.,* 2014), insects (Yeh *et al.,* 2015) and parasitic invertebrates (Stentiford *et al,* 2014).

6.1.3 Objectives

Prior to this study, the first internal transcribed spacer of *Hyalella 'azteca'* had not been characterized and similar attempts in other freshwater amphipods have yielded disjointed sequence alignments (Kornobis & Palsson, 2011). This study will use an analysis of covariation among complementary base pairs to determine common structural elements in *Hyalella* species. The comparison of homologous structures will facilitate the ITS1 sequence alignment. Phylogenetic network analysis of ITS1 may provide a strong independent measure of evolutionary relationships within the *Hyalella 'azteca'* species complex. Concerted evolution should be reinforced by species boundaries as the ribosomal subunits within interbreeding groups become homogenized. Consequently, phylogroups that are reproductively isolated should rapidly diverge. Using the homologous regions, a network of ITS1 sequences can provide an independent estimation of the phylogeographic patterns in COI variation explored in the previous chapters.

6.2 MATERIAL AND METHODS

6.2.1 PCR amplification

At least two *Hyalella 'azteca'* individuals were haphazardly selected from each of the major COI phylogroups previously characterized (see chapters 2 through 4). To reduce the possibility of including heteroduplexes or chimaeric artifacts from the PCR process, DNA from each individual was PCR amplified and cloned twice. The 50 µl PCR mixture contained 5 µl of reaction buffer (10x dilution), 0.1 unit of Taq polymerase (Fisher Scientific Canada Co., Ottawa, ON, Canada), 0.4 uM of the ITS1-F/ITS1-R primers (Sigma-Aldrich Canada Ltd., Oakville, ON,

Canada), 0.2 uM of each dNTP (Fisher Scientific Canada Co., Ottawa, ON, Canada) and 2 µl of the DNA template. The polymerase reaction was conducted on an Eppendorf Mastercycler (Eppendorf Canada Ltd., Mississauga, ON, Canada). The conditions for ITS1 amplification were 94°C for 1 min, 5 cycles of 94°C at 1 min, 45°C at 1 min 15s, 72°C at 1 min, then 24 cycles of 94°C at 1 min, 50°C at 1min 05s, 72°C at 1 min. The reaction finished with a 5 min 15s extension step at 72°C. The amplicons were gel extracted and purified with Qiagen QX1 and PE Wash (Qiagen Inc., Mississauga, ON, Canada).

6.2.2 Cloning of ITS1

The two independent PCR products from each individual were inserted into JM109 competent cells using the pGEM-T Easy Vector System 1 (Promega Corporation, Madison, WI, USA). The ITS1 amplicons were ligated using 5 µl of 2x rapid ligation buffer, 1 µl of pGEM-T Easy Vector, 1 µl of PCR product, 1 µl of T4 DNA ligase and 2 µl of deionized water. The reactions were incubated on ice for 1 hour. Fifty µl of JM109 competent cells were thawed from the -80°C freezer on ice and transferred to the tubes with the ligation cocktail. The cells incubated on ice for 20 minutes were then heat shocked at 42°C in a water bath before being returned to the ice for 2 minutes. Super Optimal broth with Catabolite repression (SOC) solution (950 µl at room temperature) was added to each tube and incubated with shaking at 37°C overnight. Approximately 100 µl of solution was plated onto LB/ampicillin/IPTG/X-galactose plates and incubated overnight at 37°C. Isolated colonies that tested positive for the insert were selected and suspended in solution 1. Six clones were selected and sequenced with the ABI 3130XL capillary sequencer (Applied Biosystems Canada, Streetsville, ON, Canada). The sequences were aligned

against *Gammarus* 18s and 5.8s to ensure the entire ITS1 gene was recovered. These genes are highly conserved; as such they represent clear indicators of the 5' and 3' ends of ITS1 and allow for proper orientation of the sequence. Only those sequences that were long enough to reach 5.8s were included in the study.

6.2.3 Structural alignment and domain analysis

The ITS1 sequences were initially aligned by MUSCLE (Edgar, 2004) with the default settings within each species to identify regions of high similarity among *Hyalella* species, although the alignments were then edited manually on the basis of the folding structure and covariation analysis. The major size variants were identified and those corroborated from independent analyses were included in the study. As described above, each *Hyalella* sample selected for analysis of its ITS1 was independently amplified, cloned and sequenced twice. Sequences that could not be corroborated were excluded from the analysis as potential chimaeras or artifacts of PCR amplification. A 50% majority rule consensus sequence for each ITS1 size variant was constructed. The rationale being that the PCR process introduces an error factor in the amplification of sequences. Since the cloning procedure inserts a single PCR amplicon there is a chance an artifact will be transformed. Even single nucleotide errors can drastically alter the folded rRNA structure and since these artifacts are likely unique, a 50% majority rule consensus sequence will remove them.

The 50% majority rule consensus ITS1 sequences constructed from the alignments of each domain. Regions with conserved elements may persist due to selection or homogenization. Given that some ITS1 variants possess sequence lengths nearly twice that of others, structures

present in one species may be wholly absent in another. Once regions are identified the local structure can be determined. The overriding strategy was that common regions share common structures. The ITS1 sequence in *Hyalella* is composed of a series of regions that can be absent or present in different orders. This process allows the identification of domains in which their base pairs are inherited together. For example, domain III may contain 120 base pairs that are present in its entirety or completely absent. These regions may be a strong indicator of structures, since both sides of the stem are required to form a helix. Once a domain is detected, the most conserved elements were used to query common structures in the other ITS1 sequences. The final alignments for each domain underwent further iteration during the analysis of covariation.

6.2.4 Folding of rRNA structures and manual sequence alignment

Crease & Colbourne (1998) used a folding temperature of 20°C in the analysis of ribosomal structures in *Daphnia pulex*. However, further analysis of the 16S molecule detected little variation in structure with a larger range of temperatures (Ambruster, 2001). The initial ITS1 structures in *Hyalella* were generated according to a free-energy criterion using a 16°C folding temperature (Dr. T. Crease, personal communication, 2015). Beginning with the smaller ITS1 variants, potential helices identified by MFOLD v2.3 (Zuker, 2003) were then enforced in the other *Hyalella* sequences. As a consequence of the highly variable 5' end of *Hyalella* ITS1, the MFOLD program produced inconsistent structures, even within the same species. As such, by forcing known structures, other potential helices can be detected. Species 1 and 2 were folded first because they possessed little intraspecific variation. The sequences of each potential helix were aligned to identify covarying base pair changes. These domains were then enforced in the

MFOLD binding protocols in order to generate an amended structure. These steps were reiterated until all major structures were identified.

Once the major domains were identified, the folding of the helices was enforced at the base of the stems and the remainder of the ITS1 sequence was allowed to bind according to their free energy. This was accomplished in MFOLD 2.3 with its native syntax. For example, the command to force a binding between three base pair combinations would be “F 72 146 3”. In this example, base pair 72 would be bound to 146 followed by 73 to 145 and 74 to 144. Even with this command, MFOLD won’t pair incorrect matches (i.e. A binding C). Long distance binding of domains was discouraged to promote local structures. The purpose of this was to create a starting point to proceed with manual iteration using covariation of base pair changes.

The alignments were modified in MEGA 6.1 (Tamura *et al.*, 2013) and the structures were enforced in MFOLD program’s folding syntax. The conserved regions of the helices were used as anchor points and the intervening regions were aligned and inspected for sequence mutations. Once the final alignments were constructed, the analysis of covariation was conducted using a chi-square test in Excel 2010. The number of observed base pair combinations (16 possible types, i.e. AA, GC) are tabulated for each base pairings and then compared against the expected number of each combination calculated from the observed number of each nucleotide per site. If the observed and expected values are the same, than the changes in each site are independent. Sites that are highly conserved will also produce insignificant *p*-values. If the ITS1 sequence alignments are representative of the structure, then multiple sites should have significant chi-square results.

6.2.5 ITS1 rRNA gene network analysis

Networks were constructed using the TCS 1.21 statistical parsimony program (Clement *et al.*, 2000) using the homologous regions of the unique sequences (table 6.1). In *Hyalella* species with multiple structural variants, only those regions that were common to all sequences were used. *Hyalella* species data sets were merged into the same network analysis if their inclusion did not violate the statistical parsimony limit. The sequences were then denoted according to structural variant, COI lineage or faunal province to highlight genetic and spatial patterns.

6.2.6 ITS1 rRNA gene analysis of molecular variance

An analysis of molecular variance (AMOVA) was conducted within each species using the homologous ITS1 regions (Excoffier *et al.,* 1992). In case with multiple structural variants, only the common regions were used. To test for intraindividual variation, sequences which would normally be grouped into populations were instead grouped according to the individual. These individuals were then assigned to Faunal Provinces to assess the broader spatial patterns of ITS1 variation. *Hyalella* species were included in the same analysis on the basis of the parsimony criterion used in the networks (see above). The AMOVA was conducted using the program Areliquin v3.0 (Excoffier *et al.,* 2005).

6.3 RESULTS

The first internal transcribed spacer of the nine *Hyalella 'azteca'* species analyzed in this study (table 6.1) delineates into eighteen distinct structural variants. Nine major structural domains were detected, eight of which formed helical structures supported by the analysis of covarying mutations (fig. 6.1). Each of these structures was identified in at least two species of *Hyallela*, although no domain was present in all ITS1 variants. Of the 1054 clones generated in this study, 35 were not included in the analysis (32 in species 5 and 3 in species 6) because of noise in the sequence read, or its structure was significantly different from any other independent sequence. Species 5 possessed the shortest ITS1 sequence (348 base pairs), whereas species 6 contained the longest sequence (734 base pairs) and the most sequence length variation (401 – 734 base pairs).

Table 6.1: Summary of ITS1 sequence variation in *Hyalella 'azteca'* species

Species	Shortest Sequence	Longest Sequence	Number of Clones	Unique Sequences	Nucleotide diversity	G+C Content
1	521	551	214	125	0.00932	0.483
2	518	539	87	51	0.00978	0.481
3	432	449	111	69	0.00952	0.462
4	480	507	54	38	0.01902	0.479
5	348	566	325	190	0.03564	0.499
6	401	734	116	65	0.07597	0.483
7	462	492	51	31	0.01168	0.472
8	381	527	74	49	0.02908	0.473
9	445	545	22	15	0.00487	0.480

6.3.1 Structure of domain 1

Helix 1 was present in all structural variants of ITS1 in *Hyalella 'azteca'* except for species 8, and many of the structural variants possessed multiple copies within the same sequence (fig. 6.2). An alignment consisting of all copies of helix I indicated several nucleotide pairings that have potentially undergone compensatory mutations. Double compensatory mutations in the 26 - 40 and 23 – 43 site pairings were the best support for this helical structure because they were located in the most conserved region adjacent to the tip of the helix. The portion of the stem nearer to the base of the helix possesses a great deal of variation. This region can be absent entirely, while the upper region was present in all copies of helix I. Structural variants 5-1 and 5-2 were nearly identical except for a double compensatory mutation at sites 7 and 62 along with a single complementary mutation at site 11. Sites 10 – 59 and 11 – 58 were present in nearly all incarnations of helix I and represent additional double mutations, whereas site 18 represents a single complementary mutation.

Figure 6.1: Alignment of *Hyalella* ITS1 according to rRNA structure
The domains indicated in the legend (right) correspond to those determined by the analysis of covarying base pair mutations. The alignment (left) includes structural variants of *Hyalella azteca* analyzed in this study.

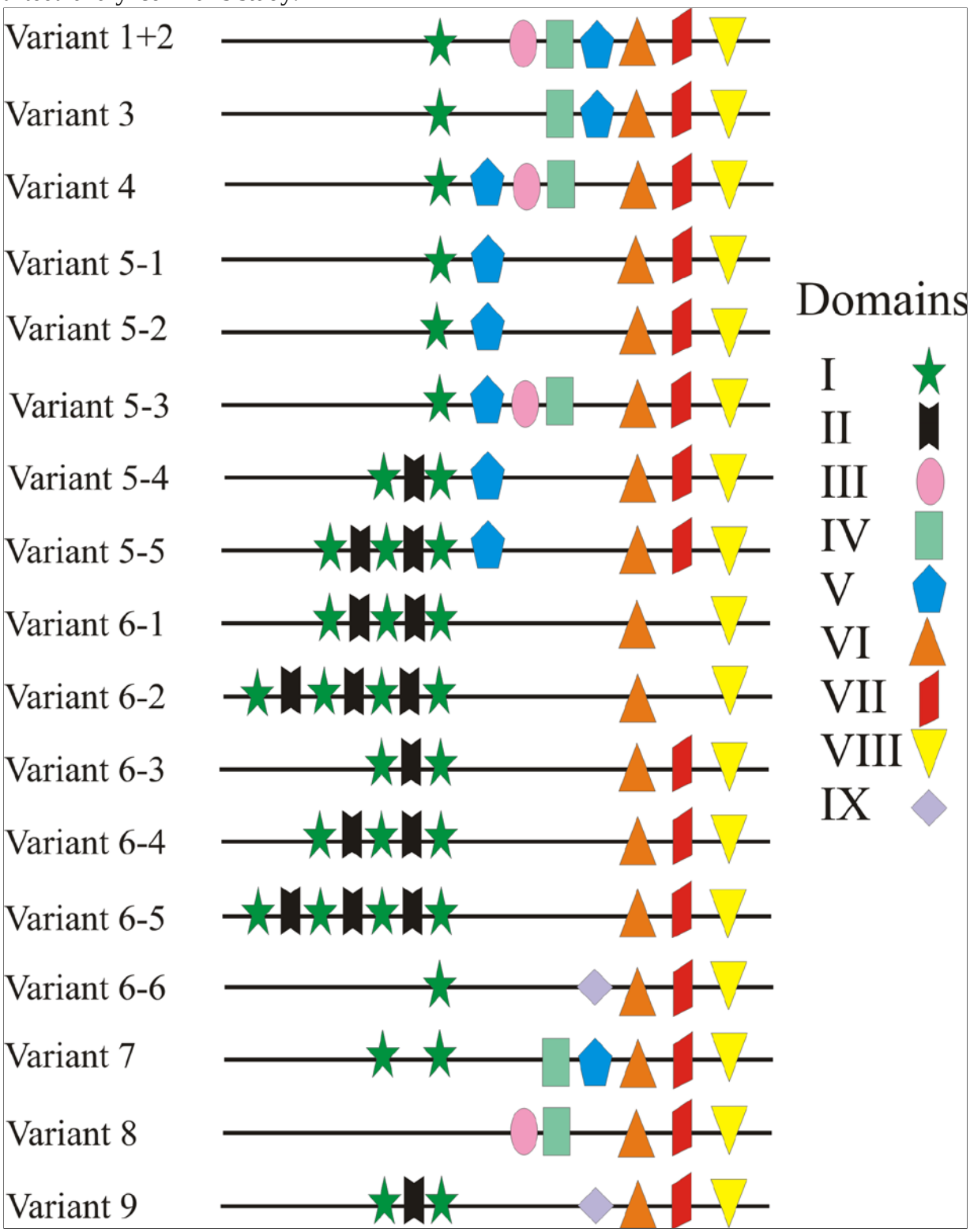
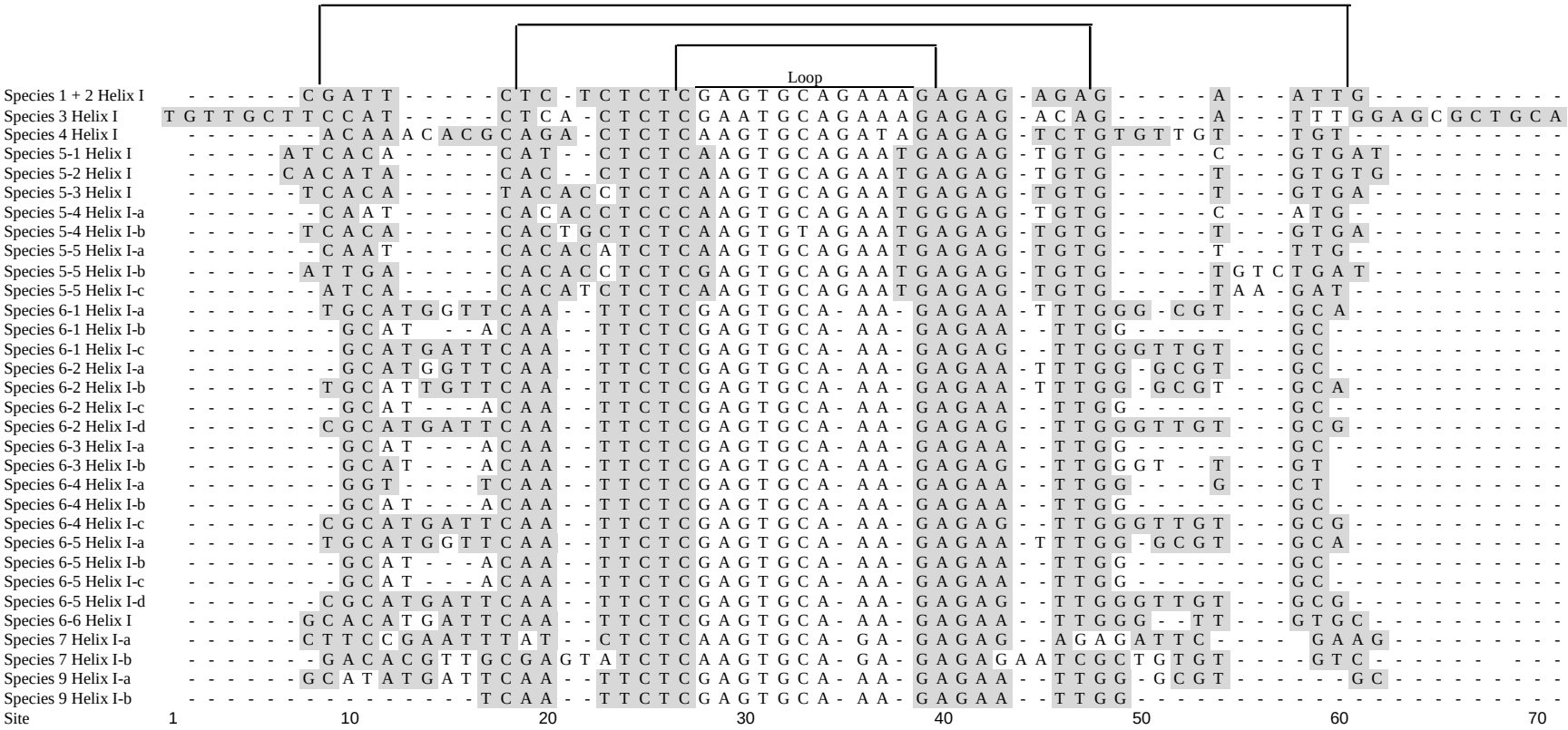


Figure 6.2: Structural alignment of domain I in the ITS1 rRNA gene

The alignment of domain I sequences was determined by structural analysis of the helix. The bases involved in the pairing of the helical structure are shaded, andthe unbound elements are unshaded.



6.3.2 Structure of domain II

Domain II was only detected in structural ITS1 variants that possess multiple copies of helix I (variants 5-4, 5-5, all of species 6 and 9) except for species 7. This segment of ITS1 was inserted between helix I copies. Interestingly, in variant 5-5, the two copies of domain II can bind with each other and extend helix I-b (figure 6.23). This can be further elongated by the binding of helix I-a to I-c (figure 6.24). These regions were not only direct repeats, but when inverted they bind complementarily. In variant 5-4, domain II can form a helix (figure 6.25), but the support for this structure was limited because the amount of intraspecfic or intercopy variation was insufficient to detect compensatory mutations. Species 6 also possesses a repeated domain between helix I copies, except this sequence was not homologous with the same region in species 5 variants. Aside from its position, species 6 domain II shares little in common with variants 5-4 and 5-5; it doesn't form a single consistent helix, nor does it readily bind to other copies of itself to extend helix I. This region seems to function as a structural spacer between copies of helix I.

6.3.3 Structure of domain III

The domain III helical structure was present in five of the ITS1 structural variants (figure 6.3). The alignment constructed from the 100 unique sequences of the initial 618 sequences. The chi-square analysis identified 5 linked covarying sites with significant *p*-values (table 6.2). These sites were located at the base of the helix, while the loop and linked sites surrounding it were highly conserved (fig. 6.3).

Figure 6.3: Structural alignment of domain III in the ITS1 rRNA gene

The alignment of domain III sequences was determined through structural analysis of the helix. Members from each species with domain III were chosen to represent the ITS1 variation. The bases involved in the pairing of the helical structure are shaded and connected by lines. Gaps in the alignment are denoted by a dash (-) symbol. The base pairings with significant chi-square statistics are denoted by the red text.

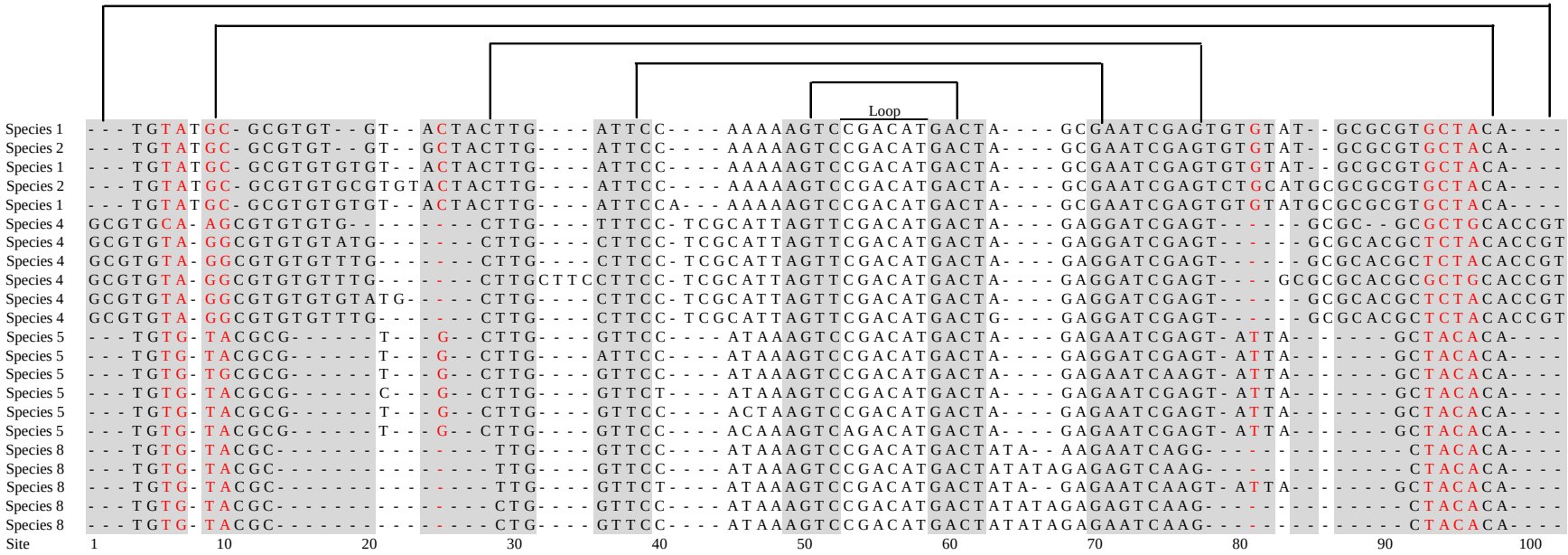


Table 6.2: Comparative chi-squared statistics of domain III

The 5’ and 3’ sites are base pairs within the helix structure and refer to sites in figure 6.3. The significant *p*-values are bolded. Chi-square *p-values* of 1 indicate that the observed values were identical to expected values.

5' Site	3' Site	*p*- value
1	102	1.000
2	101	1.000
3	99	1.000
4	98	1.000
5	97	1.000
6	**96**	**4.136E-05**
7	**95**	**1.154E-22**
9	**94**	**1.453E-17**
10	**93**	**1.578E-19**
12	92	1.000
13	91	1.000
14	90	1.000
15	89	0.163
16	88	1.000
17	87	1.000
18	86	0.879
19	85	1.000
20	84	1.000
21	83	1.000
24	82	1.000
25	**81**	**3.284E-11**
26	80	1.000
27	79	1.000
28	77	1.000
29	76	0.771
30	75	0.903
31	74	1.000
36	73	1.000
37	72	0.767
38	71	0.594
39	70	1.000
49	62	1.000
50	61	1.000
51	60	1.000
52	59	0.340

6.3.4 Structure of domain IV

Similar to helix III, this domain was present in *Hyalella* species 1, 2, 3, 4 and 7. The alignment was constructed from 530 sequences and the chi-square test was conducted on the 74 unique sequences (fig. 6.4). The loop of the helix was highly conserved among *Hyalella* species (sites 45 – 53). Six paired sties, majority of which located near the loop of the helix, had evidence for covariation with significant chi-square *p*-values (table 6.3).

6.3.5 Structure of domain V

Domain V was present in *Hyalella* species 1, 2, 3, 4 and all species 5 structural variants (fig. 5.5). The helix within species 7 possessed a structure sequence consistent with ITS1 domain V, however the alignment of homologous regions did not have high confidence. As such, it was not included in the chi-square analysis. An alignment of 145 unique sequences was constructed from the 702 total sequences (fig. 6.5); species 7 possessed 8 unique version of domain V. The chi-square analysis identified 8 covarying sites with significant *p*-values (table 6.4). Within species 5, some of the structural variants lacked the 3' basal sequence (GCG) at sites 87 to 89 in figure 6.5.

Figure 6.4: Structural alignment of domain IV in the ITS1 rRNA gene

The alignment of domain IV sequences was determined through structural analysis of the helix. Members from each species with domain IV were chosen to represent the ITS1 variation. The bases involved in the pairing of the helical structure are shaded and the unbound elements are unshaded. Gaps in the alignment are denoted by a dash (-) symbol. The base pairings with significant chi-square statistics are denoted by the red text.

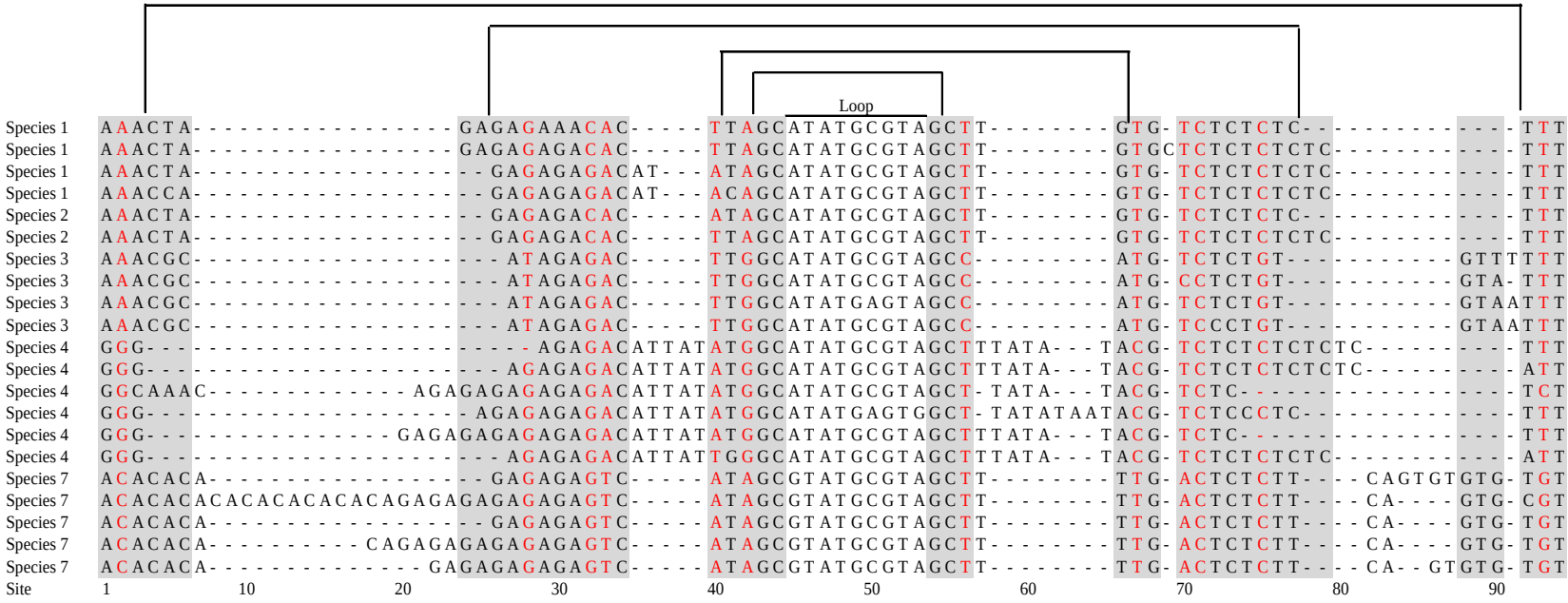


Table 6.3: Comparative chi-squared statistics of domain IV
The 5’ and 3’ sites are base pairs within the helix structure and refer to sites in figure 6.4. The significant *p*-values are bolded. Chi-square *p-values* of 1 indicate that the observed values were identical to expected values.

5' Site	3' Site	*p* -value
1	94	0.668
2	93	**1.051E-13**
3	92	1.000
4	90	1.000
5	89	1.000
6	88	1.000
24	79	1.000
25	78	1.000
26	77	1.000
27	76	1.000
28	75	**1.635E-16**
29	74	0.837
30	73	1.000
31	72	1.000
32	71	**9.648E-11**
33	70	**9.614E-13**
34	68	1.000
40	67	**0.007**
41	66	0.480
42	56	**0.003**
43	55	1.000
44	54	1.000

Figure 6.5: Structural alignment of domain V in the ITS1 rRNA gene

The alignment of domain V sequences was determined through structural analysis of the helix. Members from each species with domain V were chosen to represent the ITS1 variation. The bases involved in the pairing of the helical structure are shaded and the unbound elements are unshaded. Gaps in the alignment are denoted by a dash (-) symbol. The base pairings with significant chi-square statistics are denoted by the red text.

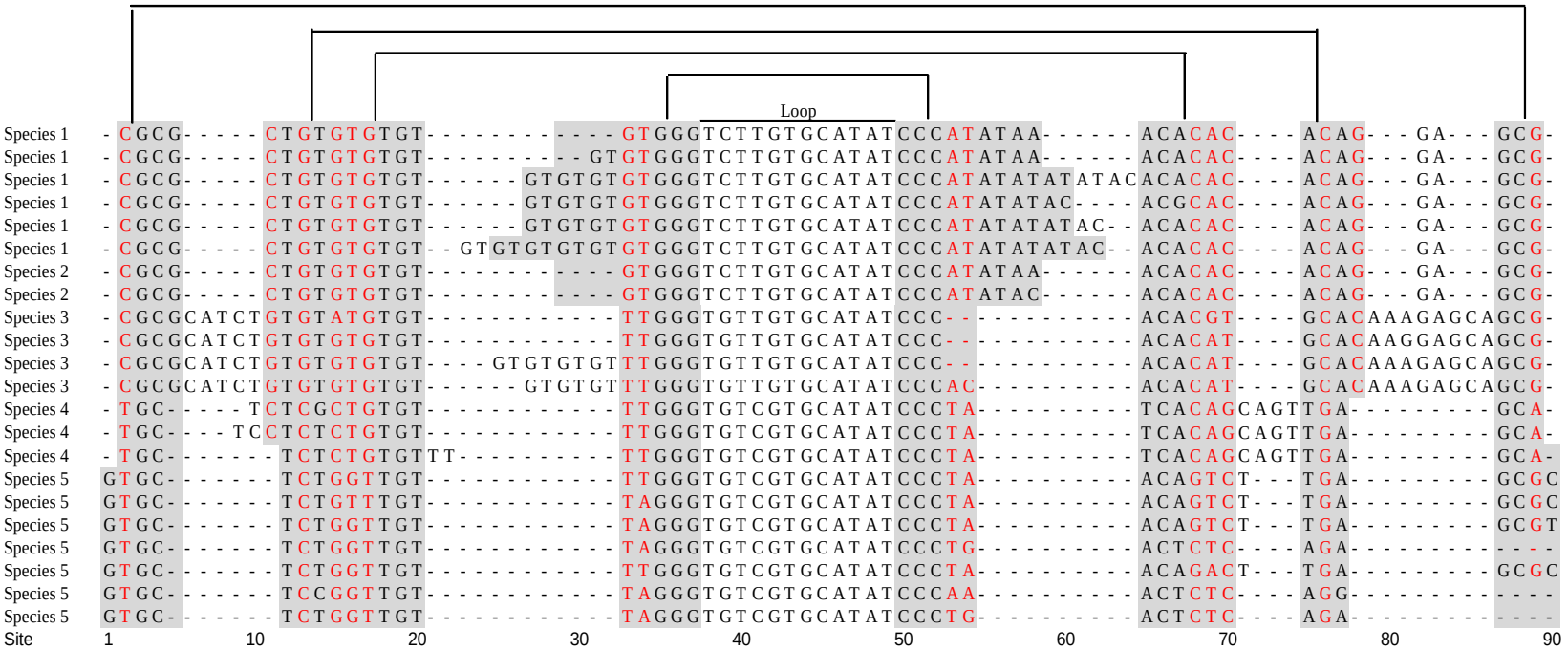


Table 6.4: Comparative chi-squared statistics of domain V
The 5’ and 3’ sites are base pairs within the helix structure and refer to sites in figure 6.5. The significant *p*-values are bolded. Chi-square *p-values* of 1 indicate that the observed values were identical to expected values.

5' Site	3' Site	*p* -value
1	90	1.000
2	**89**	**2.467E-08**
3	88	1.000
4	87	1.000
11	**78**	**5.140E-16**
12	77	0.999
13	**76**	**6.584E-15**
14	75	0.263
15	**70**	**8.680E-32**
16	**69**	**4.910E-27**
17	**68**	**5.195E-10**
18	67	0.998
19	66	0.998
20	65	0.971
33	**54**	**3.680E-14**
34	**53**	**2.439E-08**
35	52	1.000
36	51	1.000
37	50	1.000

6.3.6 Structure of domain VI

The region upstream of domain VI was highly variable in structural composition, and the downstream alignment was far more consistent among *Hyalella* species. Almost all structural variants possess a copy of helix VI (fig. 6.6) except structural variant 6-4. One thousand and sixty-four sequences possessed this domain, resulting in sixty-four unique sequences. The most variable structural region detected was between sites 14 and 18, which formed a structural bubble along the 5' stem. The ITS1 sequence was highly conserved among *Hyalella* species and the chi-squared analysis did not detect any significant tests consistent with covariation (table 6.5).

6.3.7 Structure of domain VII

Domain VII forms a highly conserved helix that was present in all ITS1 structural variants except 6-1 and 6-2. One thousand and fifty-five sequences possessed this domain, resulting in seventy unique sequences. The high degree of similarity had a limited number of identifiable base pair changes (fig. 6.7) and the chi-squared test did not detect any significant covariation (table 6.6).

6.3.8 Structure of domain VIII

Domain VIII folded into the longest helix within the *Hyalella* ITS1 rDNA gene. One-thousand and fifty-five sequences were aligned on the basis of structure and sequence homology (fig. 6.8). The stem regions at the base of the helix were highly conserved in addition to the loop. Two-hundred and fifty-eight unique sequences were identified and used in the chi-squared analysis (table 6.7), which resulted in 8 paired sites with evidence of covariation.

Figure 6.6: Structural alignment of domain VI in the ITS1 rRNA gene
The alignment of domain VI sequences was determined through structural analysis of the helix. Members from each species with domain V were chosen to represent the ITS1 variation. The bases involved in the pairing of the helical structure are shaded and the unbound elements are unshaded. Gaps in the alignment are denoted by a dash (-) symbol.

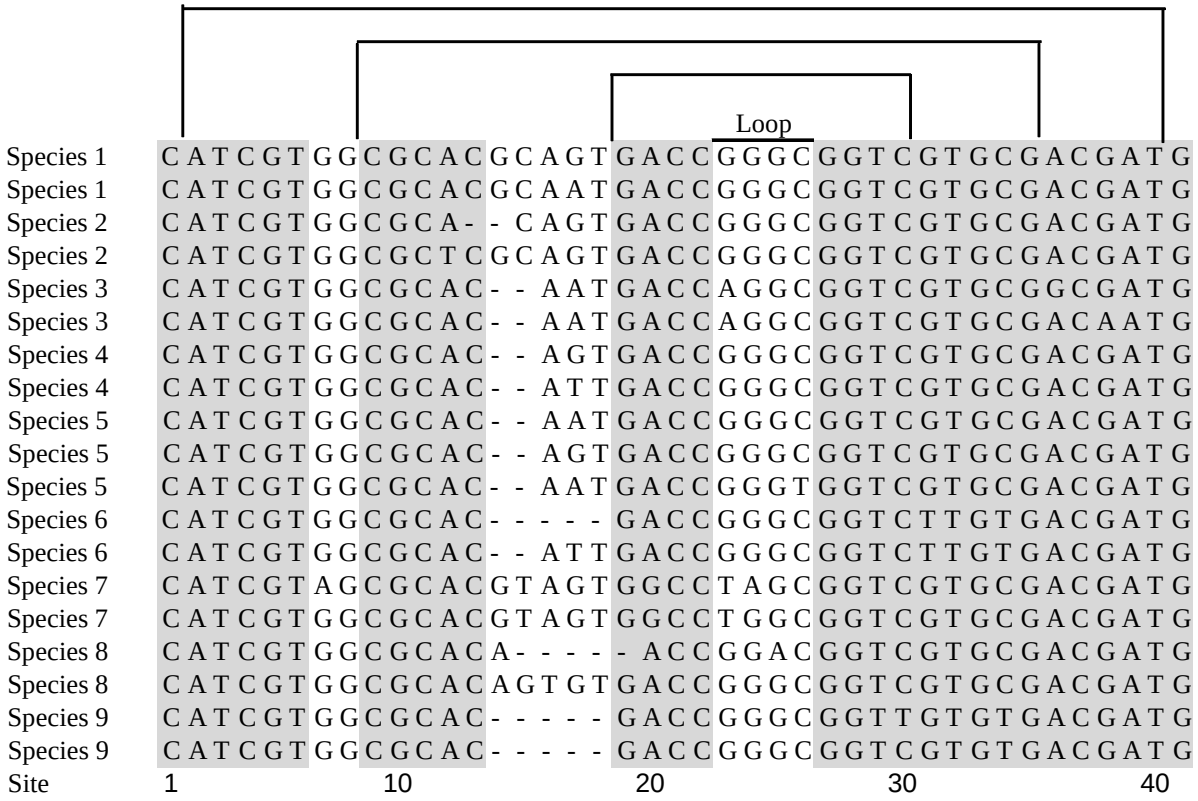


Figure 6.7: Structural alignment of domain VII in the ITS1 rRNA gene
The alignment of domain VII sequences was determined through structural analysis of the helix. Members from each species with domain V were chosen to represent the ITS1 variation. The bases involved in the pairing of the helical structure are shaded and the unbound elements are unshaded. Gaps in the alignment are denoted by a dash (-) symbol.

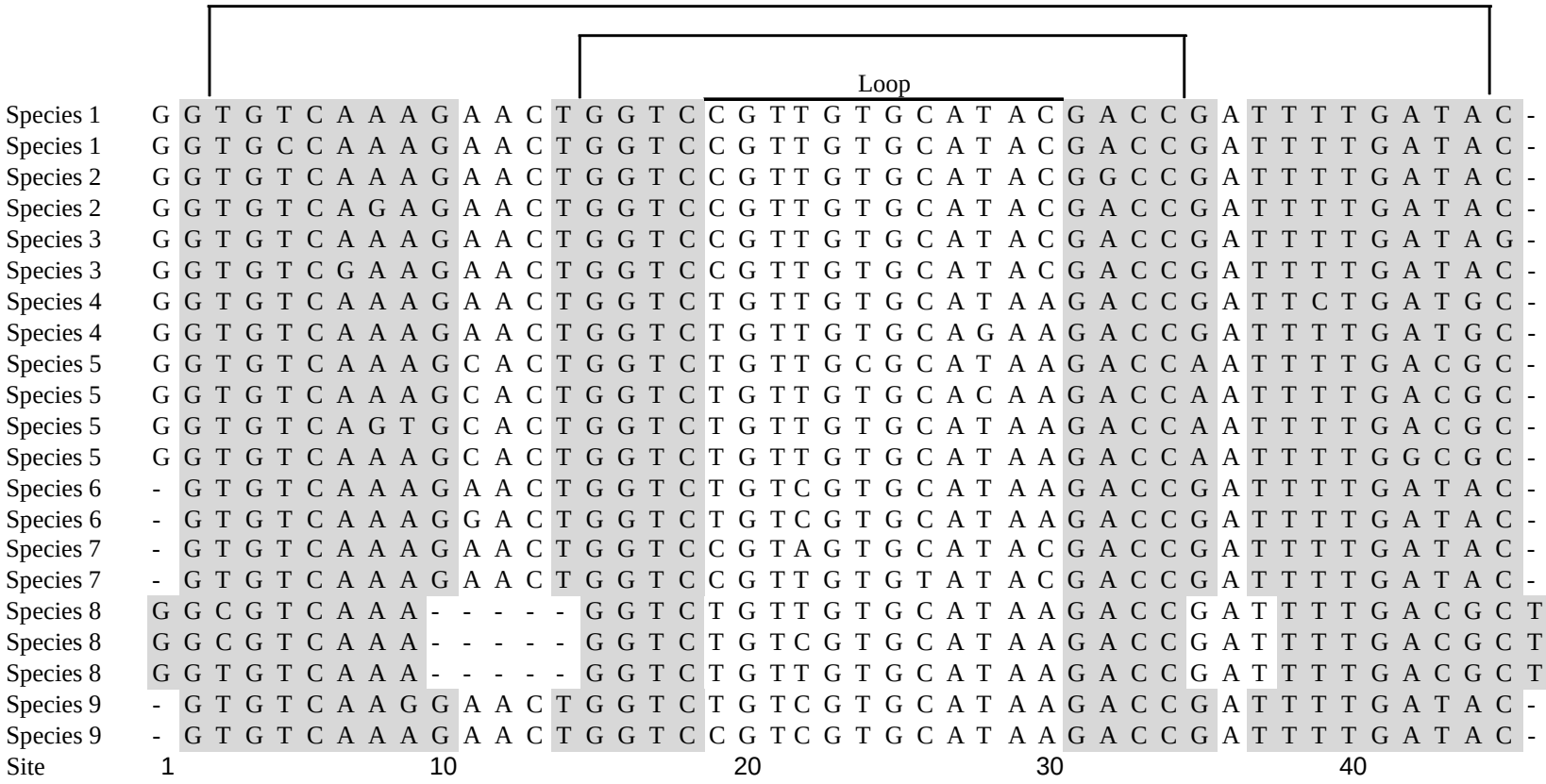


Table 6.5: Comparative chi-squared statistics of domain VI
The 5’ and 3’ sites are base pairs within the helix structure and refer to sites in figure 6.6. No significant *p*-values were detected.

5' Site	3' Site	*p* -value
1	41	0.982
2	40	0.997
3	39	1.000
4	38	1.000
5	37	0.979
6	36	0.993
9	35	1.000
10	34	0.943
11	33	1.000
12	32	1.000
13	31	0.996
19	30	1.000
20	29	1.000
21	28	1.000
22	27	1.000

Table 6.6: Comparative chi-squared statistics of domain VII
The 5’ and 3’ sites are base pairs within the helix structure and refer to sites in figure 6.7. No significant *p*-values were detected.

5' Site	3' Site	*p* -value
1	46	1.000
2	45	1.000
3	44	0.168
4	43	0.582
5	42	0.998
6	41	1.000
7	40	0.960
8	39	0.866
9	38	0.998
10	37	0.999
14	35	0.173
15	34	0.999
16	33	1.000
17	32	1.000
18	31	0.984

Figure 6.8: Structural alignment of domain VIII in the ITS1 rRNA gene

The alignment of domain VIII sequences was determined through structural analysis of the helix. Members from each species with domain VIII were selected to represent the ITS1 variation. The bases involved in the pairing of the helical structure are shaded and the unbound elements are unshaded. Gaps in the alignment are denoted by a dash (-) symbol. The base pairings with significant chi-square statistics are denoted by the red text.

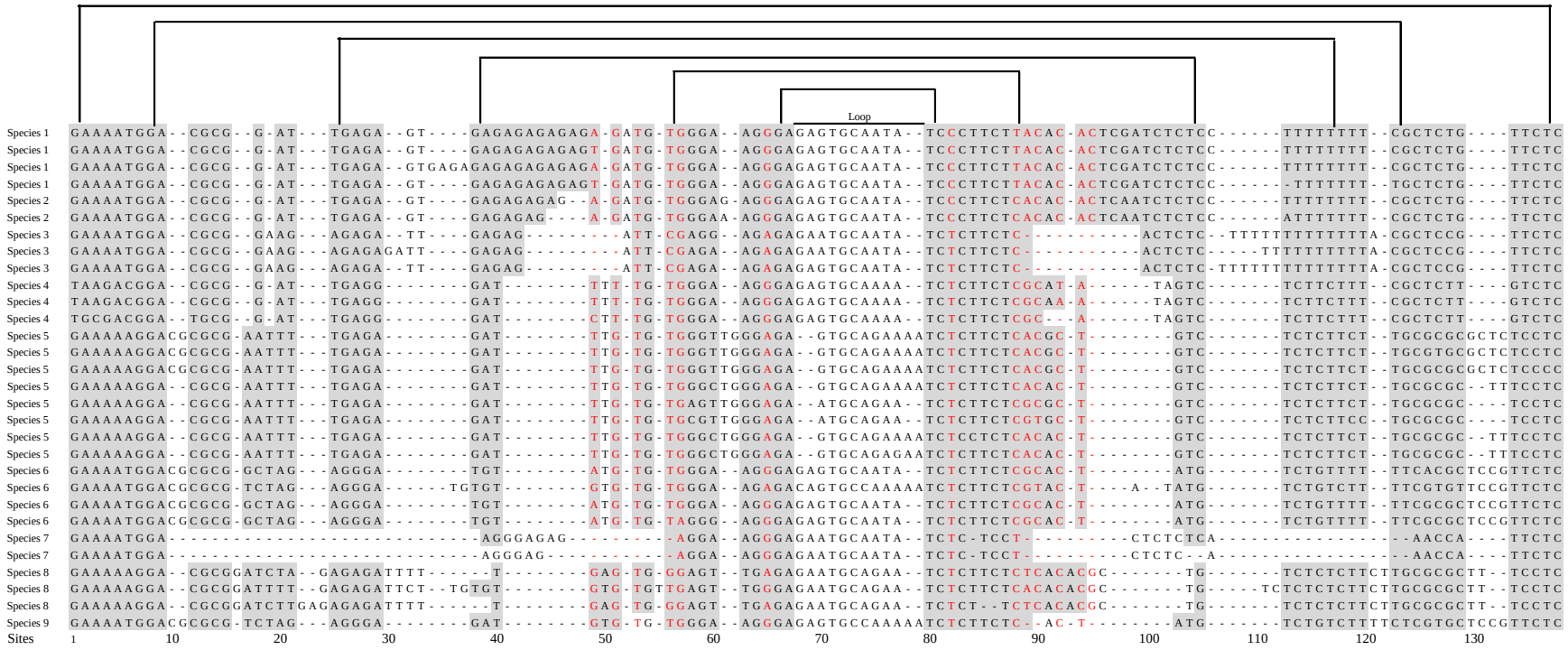


Table 6.7: Comparative chi-squared statistics of domain VIII

The 5’ and 3’ sites are base pairs within the helix structure and refer to sites in figure 6.8. The significant *p*-values are bolded. Chi-square *p-values* of 1 indicate that the observed values were identical to expected values

5' Site	3' Site	*p* -value
67	80	0.977
66	81	1.000
65	**82**	**1.266E-15**
64	83	0.992
63	84	0.107
60	85	0.927
59	86	1.000
58	87	0.105
57	**88**	**7.375E-08**
56	**89**	**8.358E-42**
54	**90**	**0.012**
53	91	0.903
51	**92**	**6.814E-46**
50	93	1.000
49	**94**	**1.150E-25**
48	**95**	**5.371E-40**
47	96	0.679
46	**97**	**2.155E-49**
45	98	0.950
44	99	1.000
43	100	1.000
42	101	1.000
41	102	1.000
40	103	1.000
39	104	1.000
38	105	1.000

6.3.9 Genetic patterns of the ITS1 gene within species 1 and 2

Eighty-six species 1 and 2 *Hyalella* individuals (60 and 26 samples, respectively) were selected on the basis of COI lineage and spatial diversity. Three hundred and ten clones were included in the analysis, resulting in 176 unique sequences. Species 1 and 2 possessed a limited degree of ITS1 variation ($\pi = 0.0093$ and $\pi = 0.0097$, respectively) with a sequence length ranging from 518 to 551 bases (table 6.1). The structure of the rRNA molecule (fig. 6.9) included seven discernible domains (I, III, IV, V, VI, VII and VIII). The structure can form a closed molecule, with binding of the 5' end with the region between domains VI and VII (fig. 6.10). However, the 5' region has too little intraspecific sequence variation to make a determination on the basis of covariation analysis and the other *Hyalella* species have different folding structures.

The species 1 and 2 network of ITS1 sequences (including all domains) revealed very little structuring among species (fig. 6.11), which indicated a high degree of homogenization within the ribosomal subunit. Assignment of ITS1 sequences according to their location in faunal provinces revealed some degree of spatial structuring (fig. 6.12). The AMOVA using faunal provinces as partitions (Cascadia: 21 individuals, Central Appalachian: 1 individual, Great Lakes: 12 individuals, Hudson Bay: 21 individuals, Mississippi: 12 individuals and Yukon-Mackenzie: 19 individuals) indicated a between group variance component of 16.42% (table 6.8). Notably, the groups located within the Atlantic and Cascadian drainages grouped together more readily than sequences sampled from the Mississippi or Yukon-Mackenzie faunal provinces (fig. 6.12). This suggests that the genetic patterns of ITS1 were influenced by spatial structure (like isolation by distance) and that any potential boundaries between species 1 and 2 were insufficient to prevent total homogenization. Both the intra-individual and inter-individual within faunal provinces variance components were significant (48.7% and 34.8%, respectively).

Table 6.8: AMOVA of the ITS1 gene in *Hyalella 'azteca'*
V(AP) = Variation Among Provinces, V(AI) = Variation Among Individuals within provinces and V(WI) = Variation Within Individuals. * $p<0.01$, ** $p<0.001$. Species 1 +2 and species 6 + 9 were combined on the basis of ITS1 sequence networks. Negative values are possible within an AMOVA and should be taken as 0.

Species	V(AP)	V(AI)	V(WI)
1 + 2	16.42**	48.75**	34.83**
3	13.16*	41.10**	45.74**
4	16.11*	29.01**	54.88**
5	6.61	43.57**	49.82**
6 + 9	31.01**	45.97**	23.02**
7	-8.84	56.69**	52.15**
8	54.79*	29.00**	16.21**

Figure 6.9: Structure (open configuration) of the ITS1 consensus sequence in species 1 and 2

The structure is derived from species the 1 and 2 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, III, IV, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites.

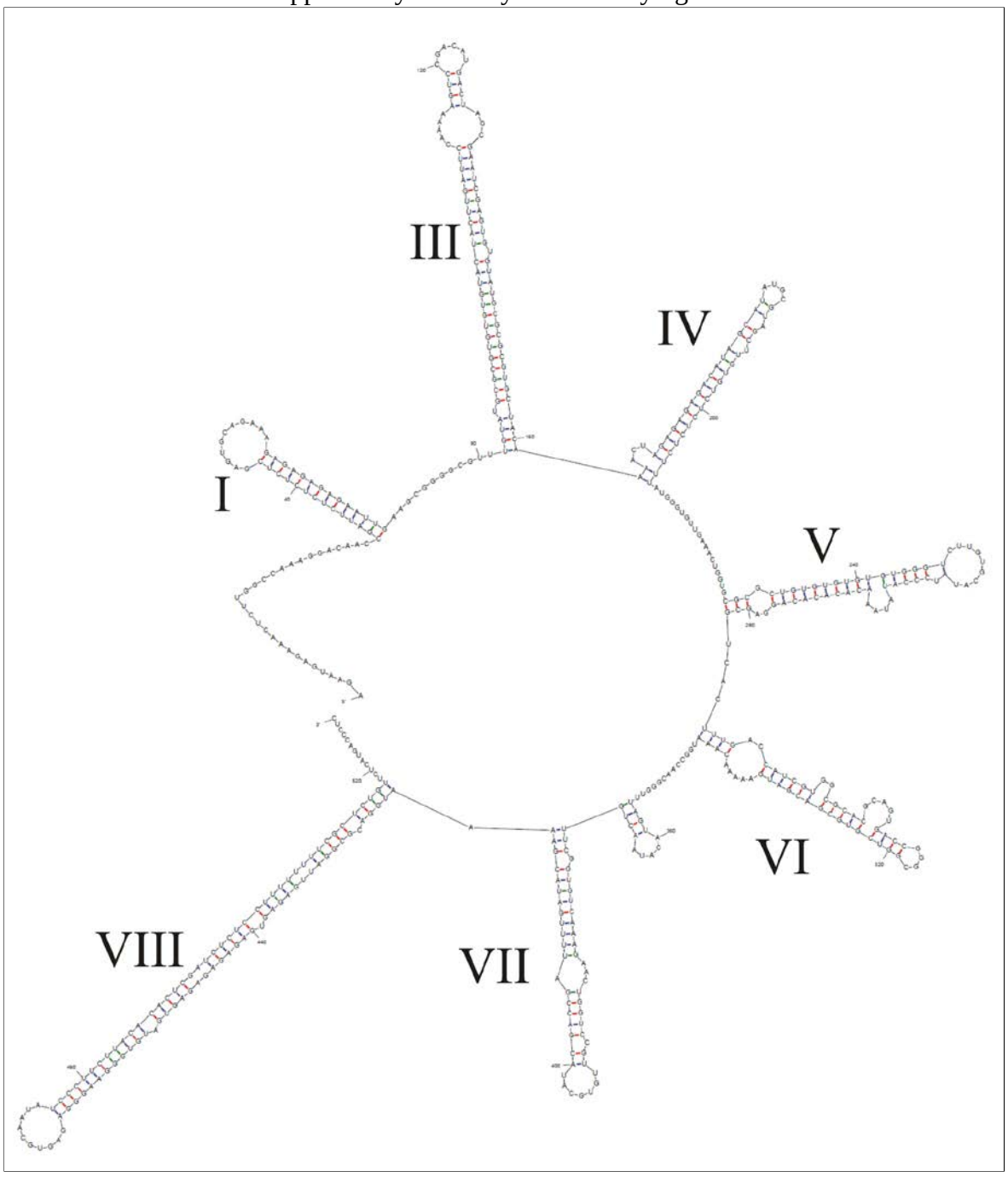


Figure 6.10: Structure (closed configuration) of the ITS1 consensus sequence in species 1 and 2

The structure is derived from the species 1 and 2 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, III, IV, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. The region upstream of domain I and downstream of domain VI were allowed to bind according on the basis of free energy in MFOLD.

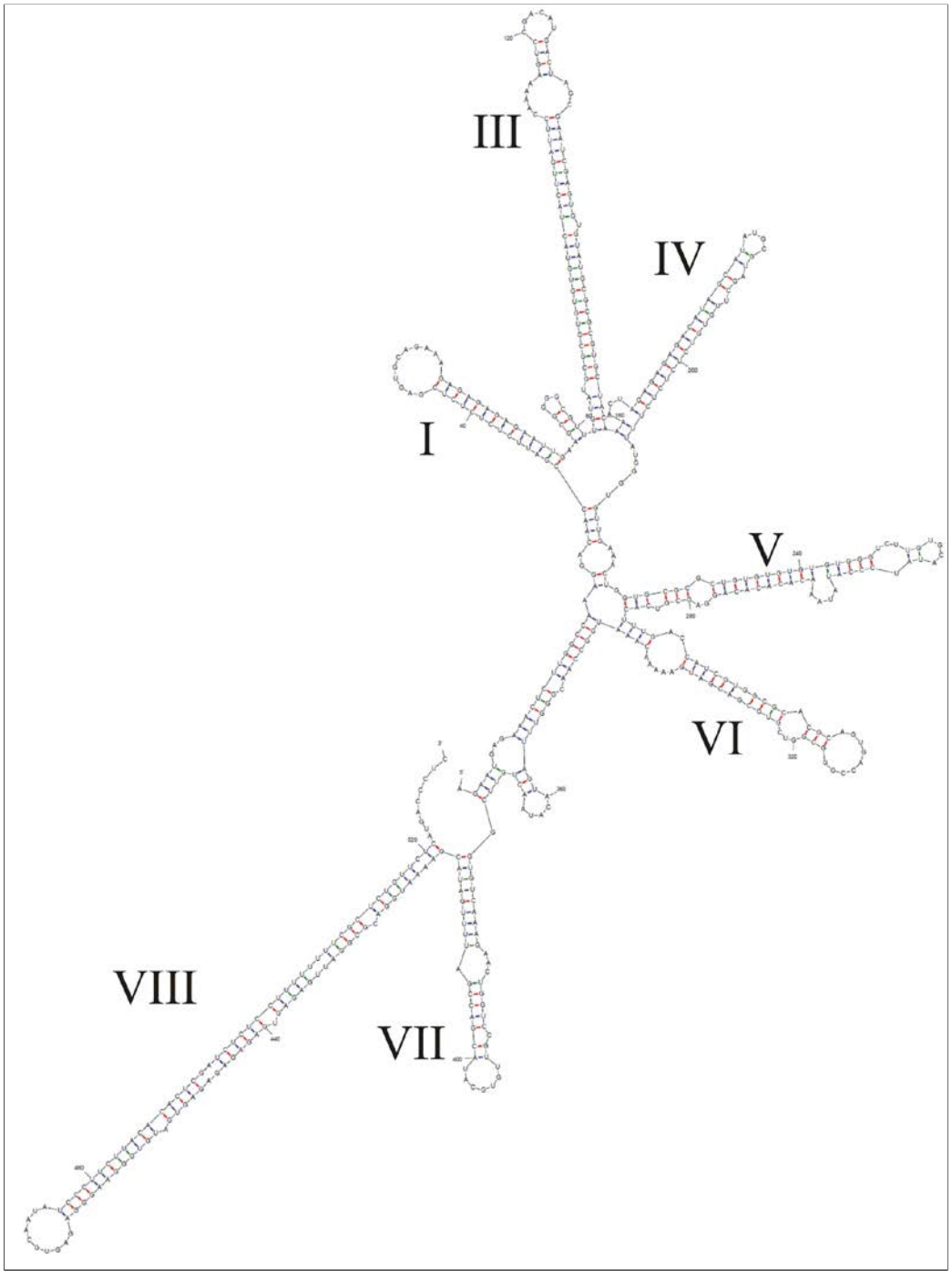


Figure 6.11: ITS1 rRNA gene network of species 1 and 2

The ITS1 parsimony network is colored according to species: species 1 (blue) and species 2 (red). This network was constructed using domains I, III, IV, V, VI, VII and VIII. The sizes of the circles are proportional to the sample size and those nodes representing sequences possessed by both species are colored according to the percentages of each species.

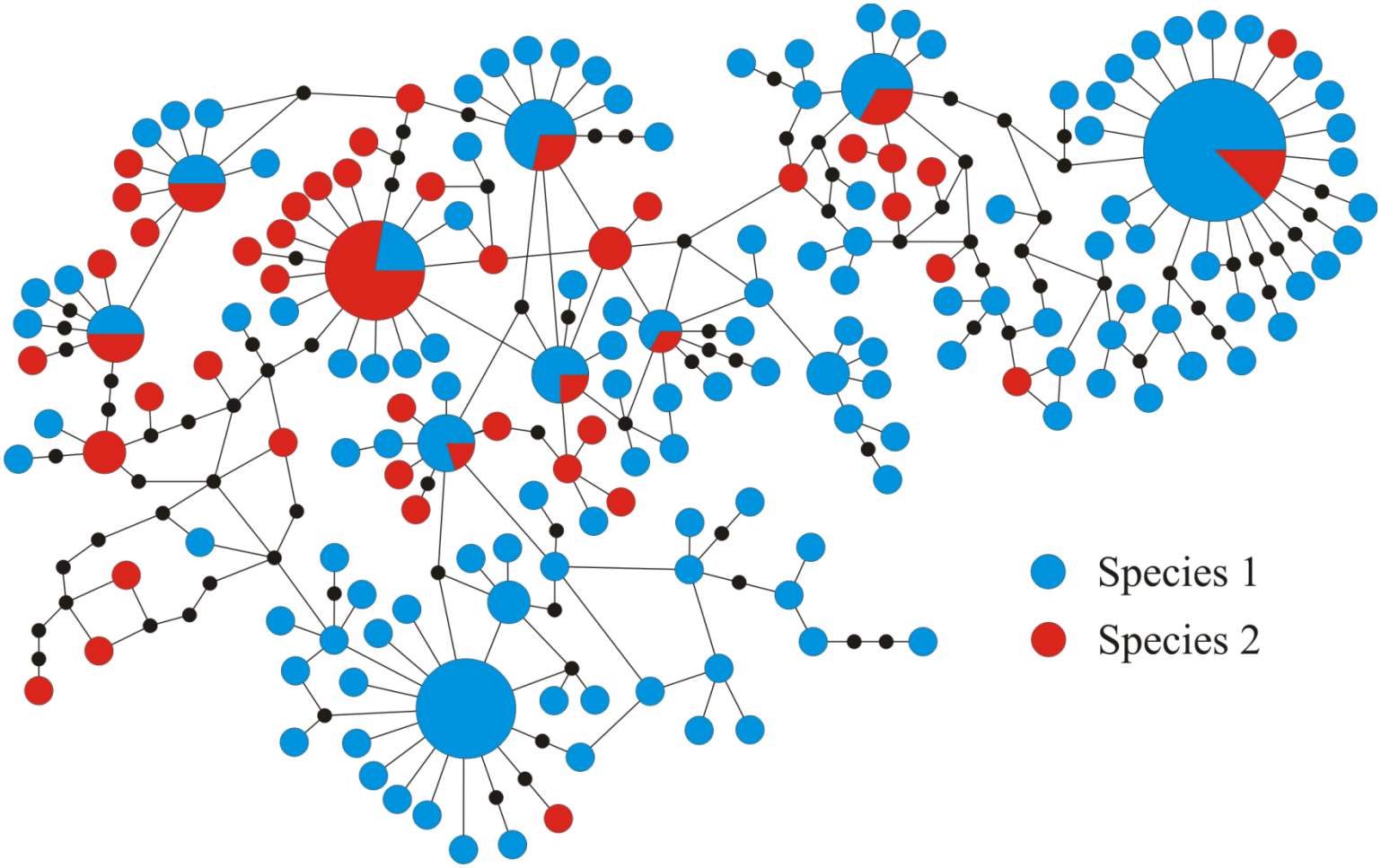
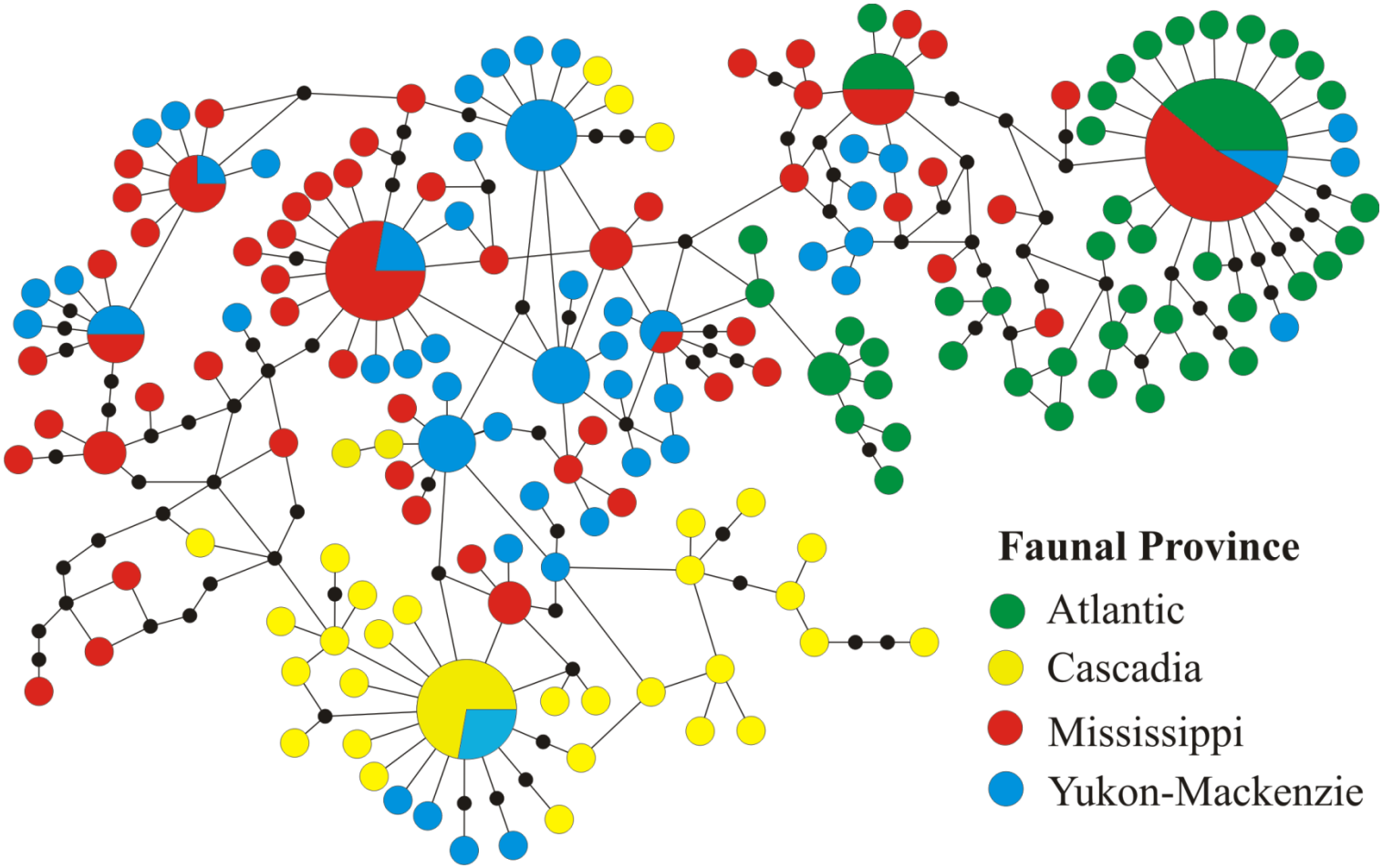


Figure 6.12: ITS1 network of species 1 and 2 with respect to faunal provinces

The ITS1 parsimony network is colored according to geographic regions: Atlantic (green), Cascadia (yellow), Mississippi (red) and Yukon-Mackenzie (blue). This network was constructed using domains I, III, IV, V, VI, VII and VIII. The sizes of the circles are proportional to the sample size and those nodes representing ITS1 sequences from the network in fig. 6.12, and are colored according to faunal province.



6.3.10 Genetic patterns of the ITS1 gene within species 3

Thirty-one species 3 individuals were selected on the basis of the COI phylogeny and spatial distribution (table 6.1). One-hundred and eleven clones were generated resulting in 69 unique sequences. The sequence length ranged from 432 to 449 bases and the nucleotide diversity (π = 0.0095) was similar to species 1 and 2 (π = 0.0093 and π = 0.0097, respectively). The rRNA structure of the ITS1 gene within species 3 possessed domains I, IV – VIII (fig. 6.13). The lack of domains II, III and domain I repeats resulted in a shorter ITS1 fragment. In terms of structure, species 3 was most similar to species 1, 2 and 7. Sequence alignment of these ITS1 groups results in the pairing of more homologous regions than in comparisons to small-bodied *Hyalella* ITS1 sequences (species 4, 5 and 6).

The species 3 network (fig. 6.14) was constructed using the entire ITS1 sequence and indicated some degree of spatial structure. The network was denoted according to the Great Lakes, Hudson Bay and Mississippi River faunal provinces with the latter being split into eastern and western groups (fig. 6.14). Similar to species 1 and 2, the species 3 network indicated a longitudinal gradient that was consistent with isolation by distance. The AMOVA, using faunal provinces (Great Lakes: 12 individuals, Hudson Bay: 3 individuals and Mississippi: 16 individuals), indicated a between group variance component of 13.16% (table 6.8). The intra-individual and inter-individual within provinces variance components represented a far greater portion of the covariation (41.1% and 45.7%, respectively). The pattern most readily apparent was a general shift in ITS1 gene diversity from west to east. ITS1 sequences were more closely linked to individuals that were in closer proximities. This was consistent with the analysis of COI variation that indicated range expansions of clades 4-2 and 4-3 (see chapter 4), which occurred along a longitudinal cline. Furthermore, the disparate COI clade 5-1 *Hyalella* did not possess a

distinct ITS1 sequence that separates it from the general body of ITS1 variation. Either, the nuclear genome has been homogenized through gene flow or the level of divergence between clades was insufficient to derive a unique ITS1 variant.

6.3.11 Genetic patterns of the ITS1 gene within species 4

Fifteen COI individuals were selected for ITS1 sequencing on the basis of lineage and spatial diversity within species 4. Fifty-four clones were generated, resulting in 28 sequences (π = 0.019) with sequence length varying from 480 to 507 bases (table 6.1). The rRNA folding structure of ITS1 in species 4 possessed all the same domains as species 1 + 2. However, the position of helix V was between domains I and III instead of helices IV and VI (fig. 6.15). The ITS1 network consisted of long edges with inferred sequences (fig. 5.16). The association between the COI sequences and ITS1 sequences indicated that there wasn't very much congruence between the markers. While the COI phylogeny indicated distinct monophyletic lineages, the ITS1 network was admixed (fig. 6.16). The AMOVA (Central Appalachian: 1 individual, Mississippi River: 5 individuals and Southeastern: 9 individuals) detected some degree of spatial structure among the faunal provinces (figure 6.17) with a between region variance component of 16.17% (table 6.8). The intra-individual variance component was the highest (54.88%) among all the *Hyalella* species, which was not surprising given the long edges inferred in the network. The between individual within region variance component (29%) was lower than other species (table 6.8), but still significant.

Figure 6.13: Structure of the species 3 ITS1 consensus sequence
The structure is derived from species the 3 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, IV, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites.

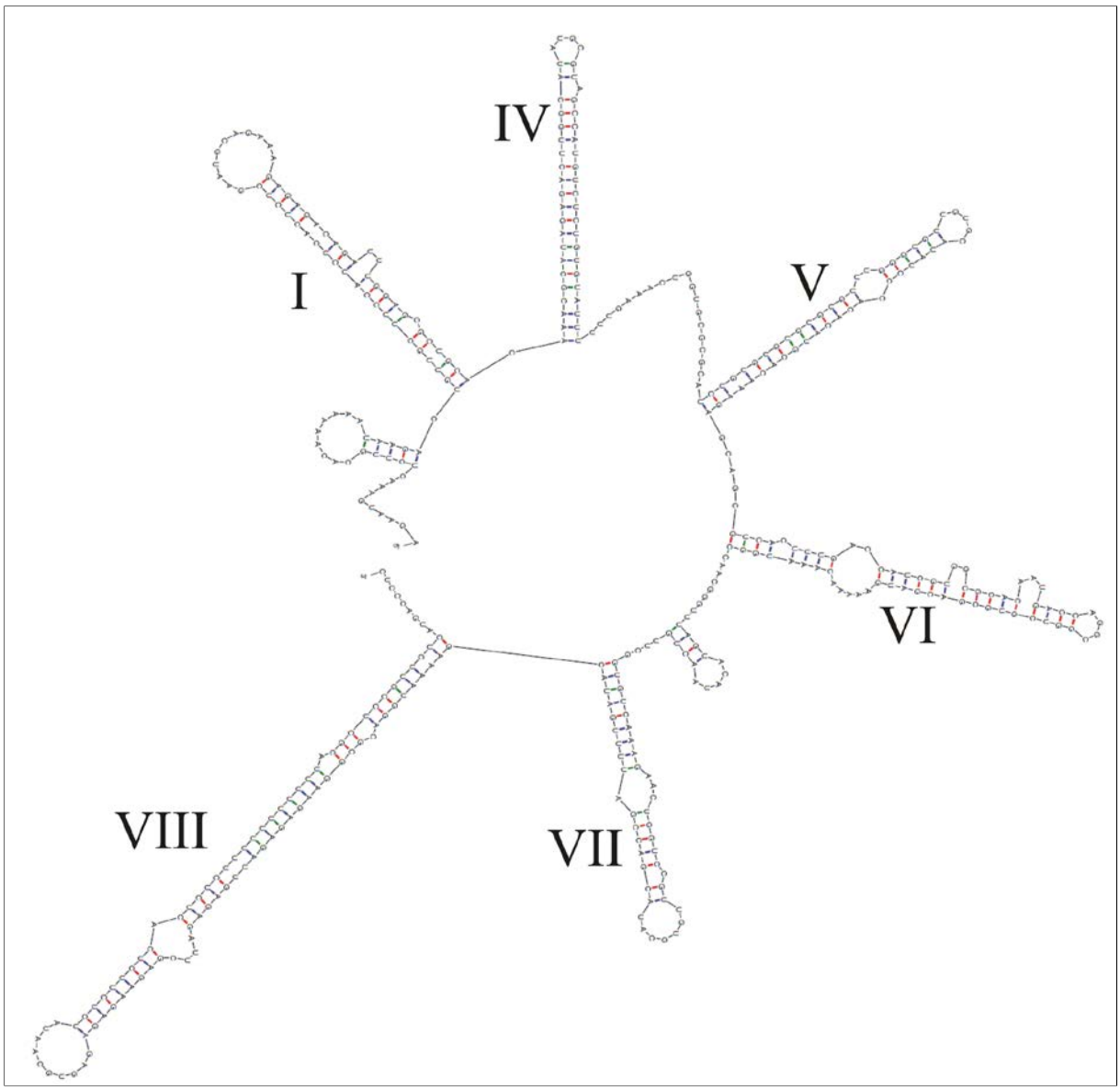


Figure 6.14: Species 3 ITS1 network according to its spatial distribution
The ITS1 parsimony network is colored according to geographic regions listed on the key (right). The Western Mississippi group (red) includes the states of Nebraska, Iowa and Kansas. The Eastern Mississippi group (blue) includes the states of Alabama, Indiana, Kentucky, Michigan and Ohio. Yellow and green represent the Great Lakes and Hudson Bay faunal provinces, respectively. This network was constructed using domains I, IV, V, VI, VII and VIII. The sizes of the circles are proportional to the sample size.

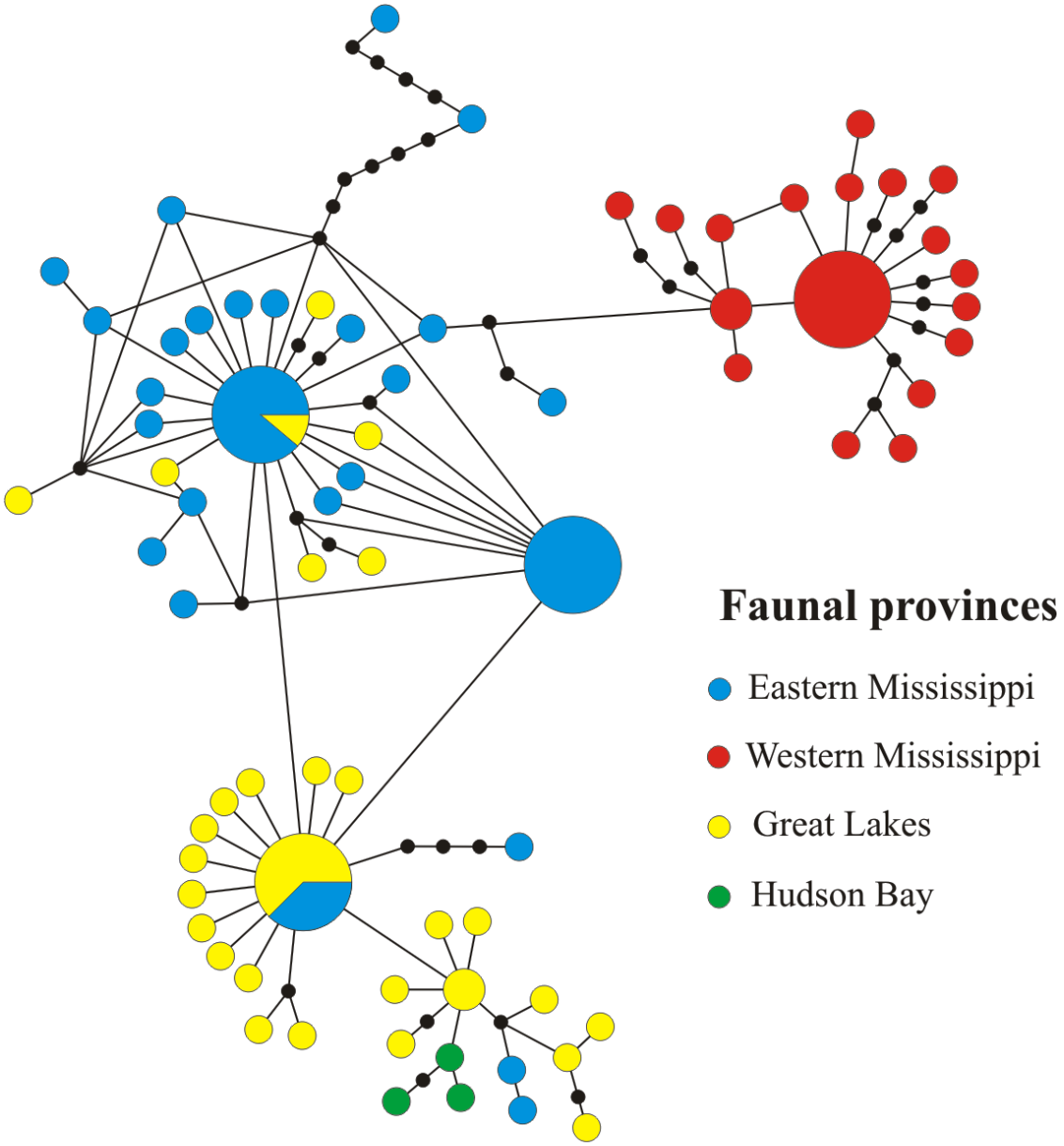


Figure 6.15: Structure of the species 4 ITS1 consensus sequence
The structure is derived from the species 4 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, V, III, IV, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites.

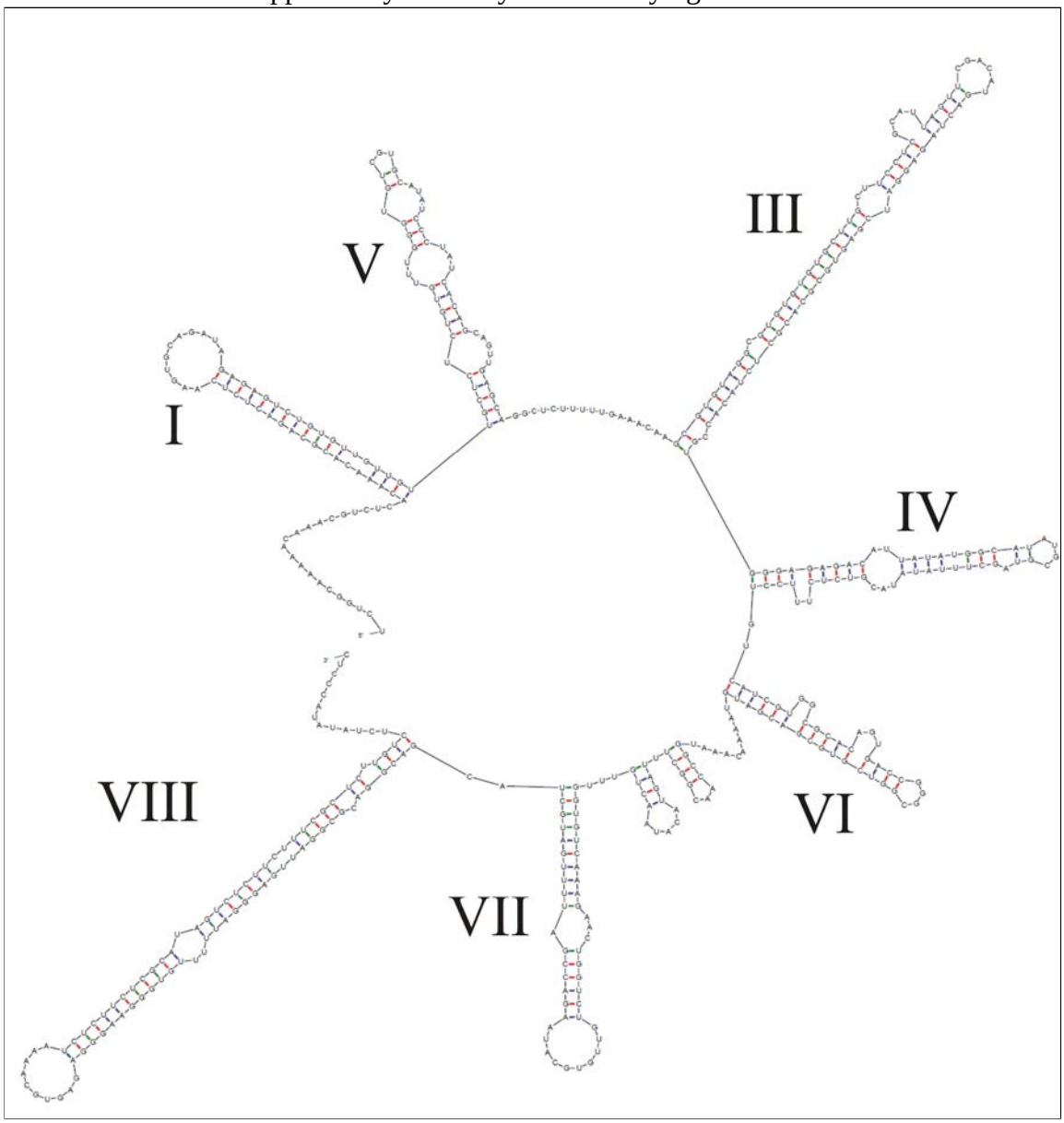


Figure 6.16: Species 4 ITS1 rRNA gene network with respect to COI lineages
The ITS1 network (54 clones) is colored according to COI lineages (right insert): blue (4A), yellow (4B) and red (4C). This network was constructed using domains I, IV-VIII. The sizes of the circles are proportional to the sample size.

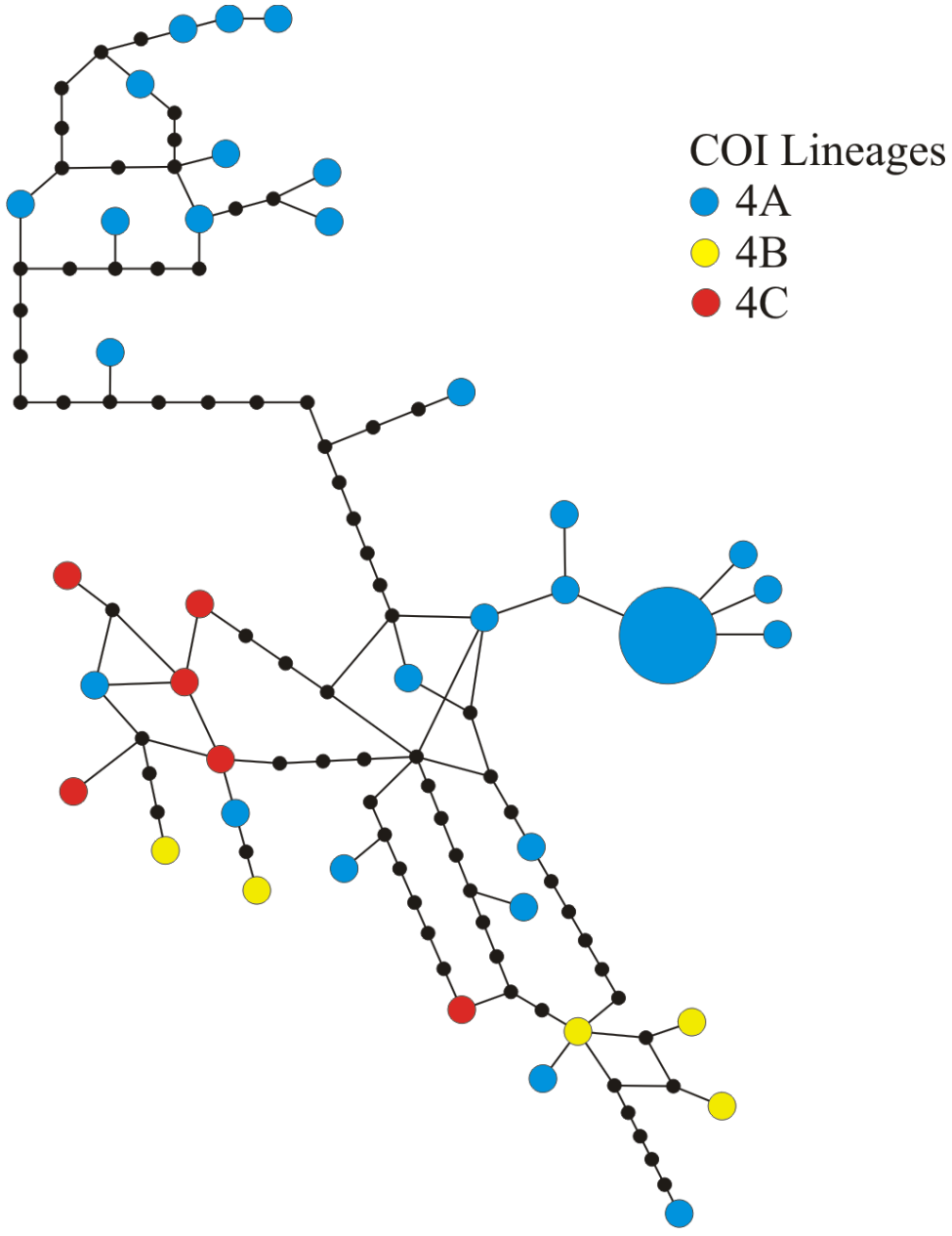
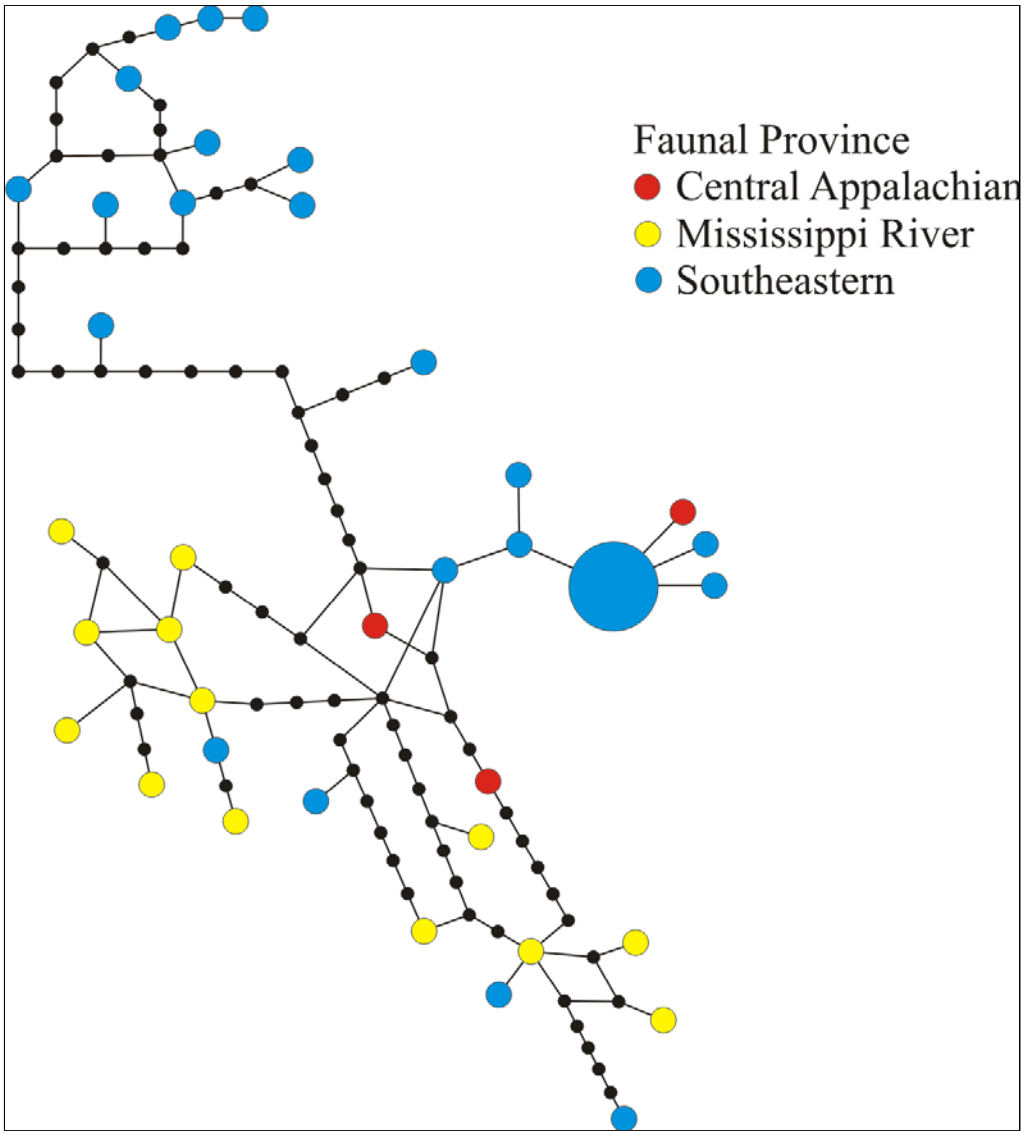


Figure 6.17: Species 4 ITS1 rRNA gene network with respect to faunal province
The ITS1 network (54 clones) is colored according to faunal province (right insert): Central Appalachian (red), Mississippi River (yellow) and Southeastern (blue). This network was constructed using domains I, IV-VIII. The sizes of the circles are proportional to the sample size.



6.3.12 Genetic patterns of the ITS1 gene within species 5

Forty-two *Hyalella* species 5 individuals were haphazardly selected on the basis of COI lineage and spatial diversity. Three-hundred and twenty-five clones were generated resulting in 190 unique sequences. The nucleotide diversity was greater (π = 0.036) than most other *Hyalella* species (except for species 6) with a sequence length variation ranging from 348 to 566 bases (table 6.1). Species 5 possessed five structural variants (figure 6.2). Structural variants 5-1 (fig. 6.18) and 5-2 (fig. 6.19) were nearly identical in structure. Both variants consist of domains I, V, VI, VII and VII, except variant 5-2 contains an additional helix at the 5' end that was not homologous with the other *Hyalella* sequences. Structural variant 5-3 (fig. 6.20) was the only species 5 structure that possesses domains III and IV. Structural variants 5-4 (fig. 6.21) and 5-5 (fig. 6.22) both had multiple copies of helix I, but differed in the number of repeats (fig. 6.2). Multiple configurations of the domain I-II repeats were possible in these two structural variants. On its own, helix I can fold into an independent helix. However, in variant 5-5 it was possible to extend the helix of Ib, by allowing helix Ia to bind to Ic (fig. 6.23). This created an extended helix consisting of five domains (Ia, IIa, Ib, IIb and Ic). There was insufficient variation among structural variant 5-5 ITS1 sequences within species 5 to distinguish these different conformations on the basis of covariation alone.

The ITS1 network constructed from domains VI, VII and VIII (fig. 6.24) did not result in five clear partitions, as would be expected if the structural variants were part of distinct nucleolus organizer regions (NOR). Structural variant 5-3 was the most common and broadly distributed. It was present in all major COI lineages (see chapter 4) and in populations within all the faunal provinces occupied by species 5. Similarly, structural variant 5-5 had a broad distribution and was detected within populations in Michigan, Alabama and South Carolina. The AMOVA

reflects this pattern (table 6.8) with little covariation explained by the between faunal province variance component (6.61%). Structural variants 5-1, 5-2 and 5-4 were present in individuals collected from the southern Mississippi River watersheds and populations in the Southeastern faunal province in North and South Carolina. Structural variant 5-4 splits into two distinct geographic groups: an isolated section derived from populations in Louisiana and a group from South Carolina. The 3' domains within these groups were located in distinct regions within the network. The structural variant 5-4 ITS1 sequences in the South Carolina group shared a close similarity to structural variant 5-2, in both geography and sequence similarity. This may indicate the potential for hybridization between ITS1 structural variants 5-2 and 5-4. Similarly, a splinter group of variant 5-1 alleles (as seen on the upper-left side of the network in fig. 6.25) formed a phylogroup whose 3' domains (VI, VII and VIII) were similar to variant 5-3 sequences. These potential hybrids were also located in the southeastern faunal province, which may indicate that this region had undergone secondary contact among distinct species 5 lineages.

The ITS1 structural variants did not partition in patterns congruent with the COI phylogroups. The most broadly distributed COI lineages (5A and 5B) had individuals containing all five structural variants. The remaining lineages had a more limited selection of structural variants. COI lineages 5C and 5D (fig. 6.25) were largely represented by the most common structural variant 5-3 (fig. 6.26), while COI lineage 5E (fig. 6.25) had individuals with both structural variants 5-1 and 5-3 (fig. 6.25). This lack of congruence among markers suggests that species 5 was maintaining multiple ITS1 structural variants within its populations that may have persisted prior to the divergence of the COI lineages. This suggests multiple independent NORs within the species 5 genome that potentially have had little recombination between loci.

Figure 6.18: Structure of the ITS1 variant 5-1 consensus sequence in species 5
The structure is derived from the structural variant 5-1 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of five structural variants detected in species 5.

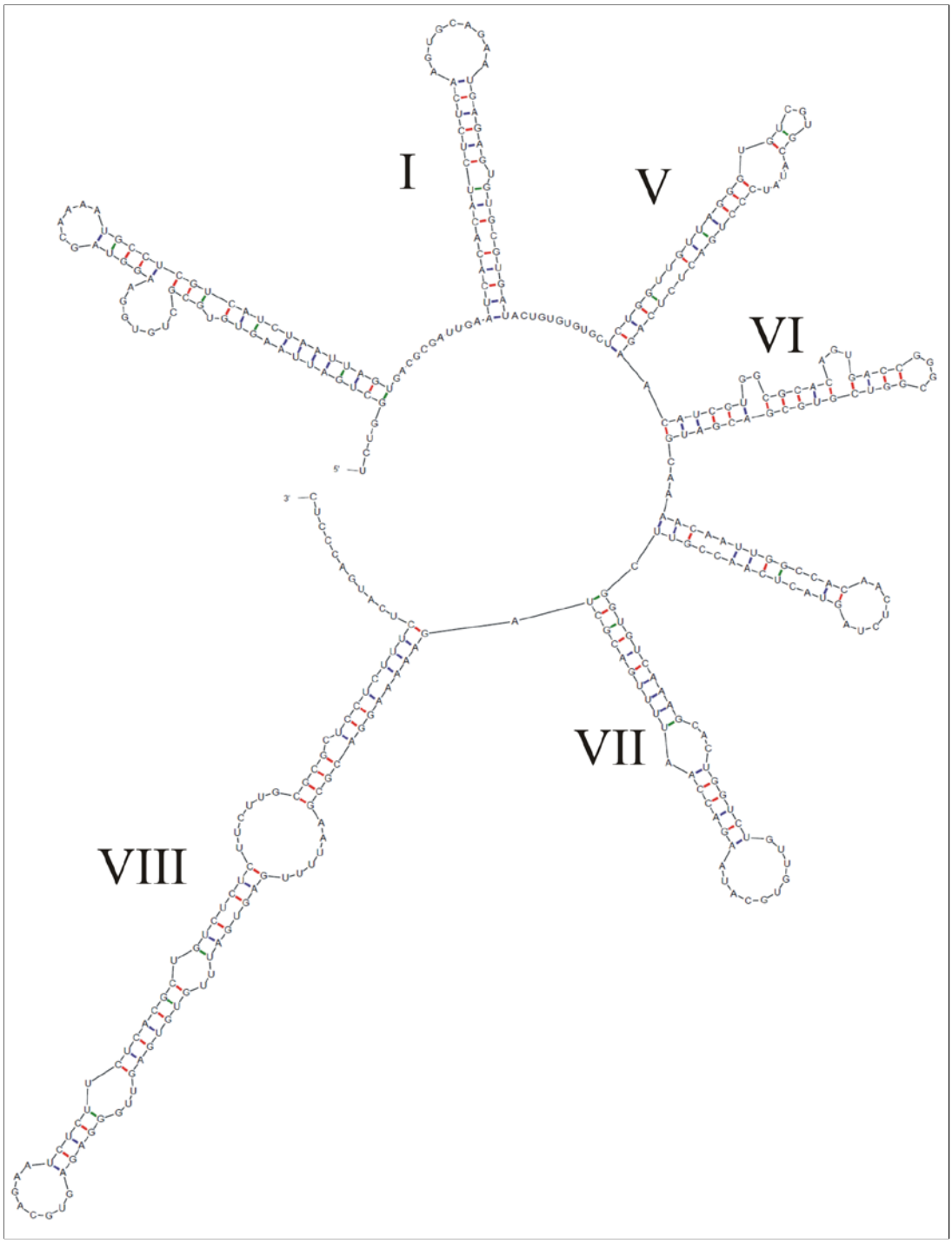


Figure 6.19: Structure of the ITS1 variant 5-2 consensus sequence in species 5
The structure is derived from the structural variant 5-2 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of five structural variants detected in species 5.

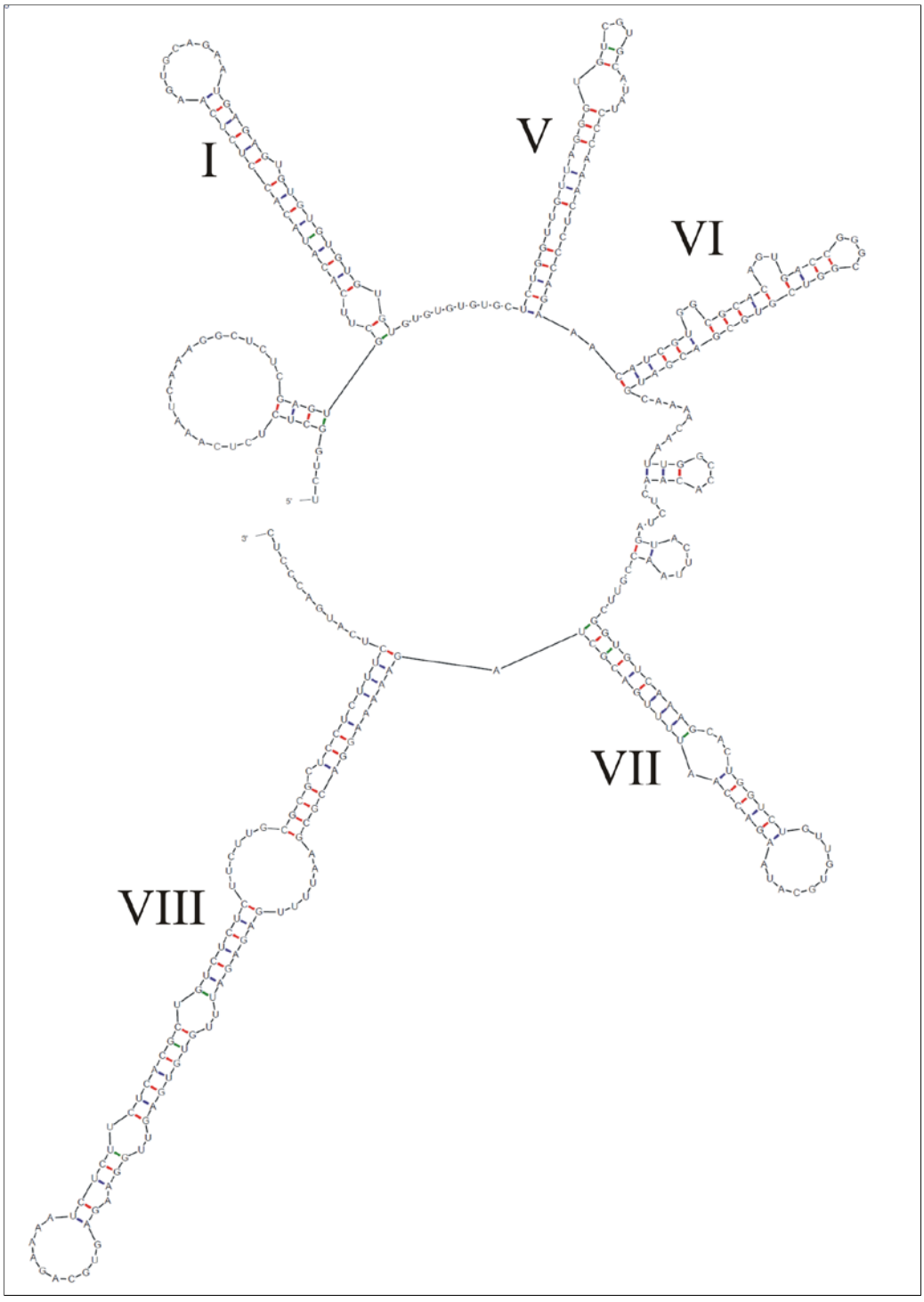


Figure 6.20: Structure of the ITS1 variant 5-3 consensus sequence in species 5
The structure is derived from the structural variant 5-3 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, V, III, IV, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of five structural variants detected in species 5.

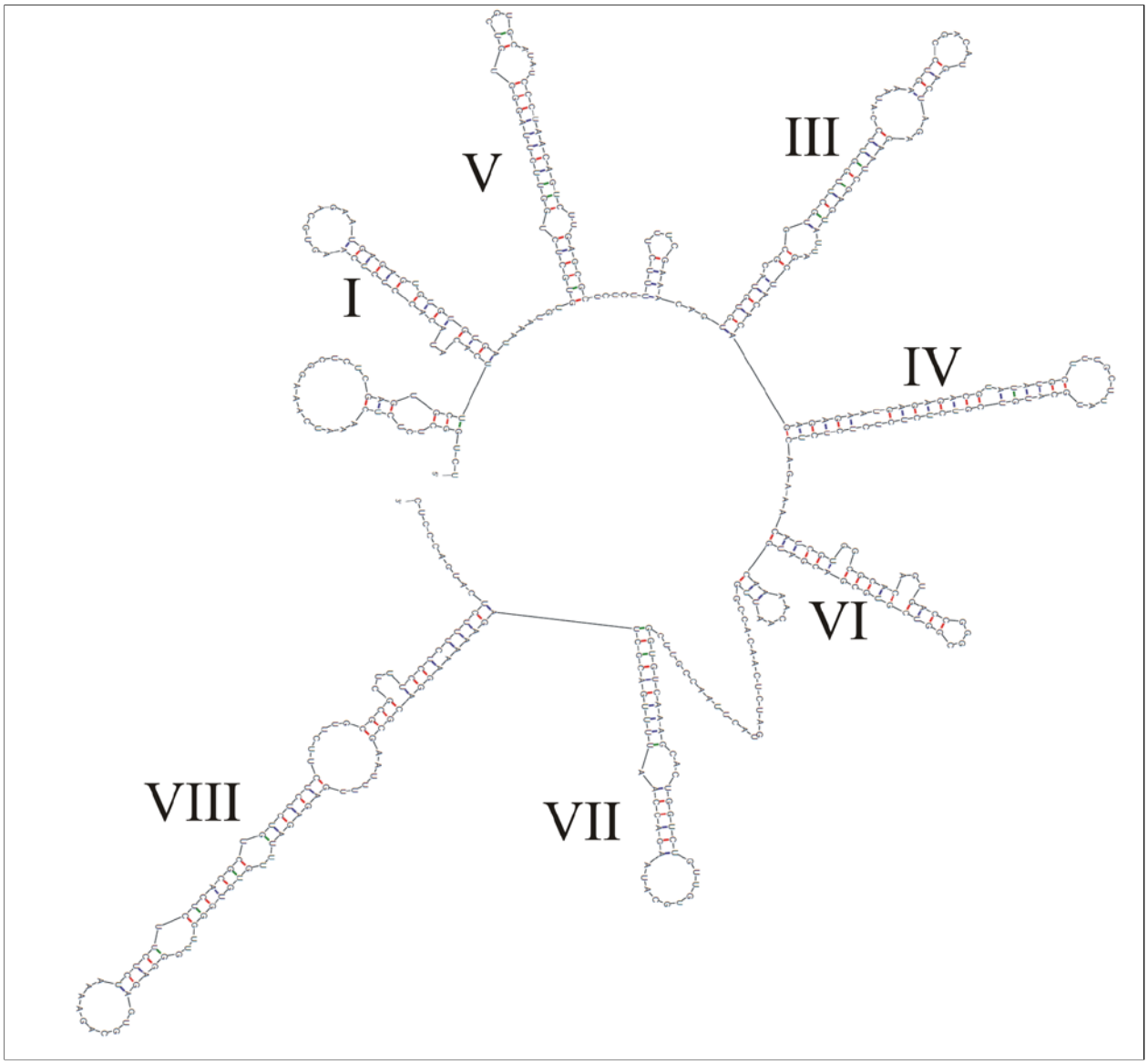


Figure 6.21: Structure of the ITS1 variant 5-4 consensus sequence in species 5
The structure is derived from the structural variant 5-4 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, II, Ib, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of five structural variants detected in species 5.

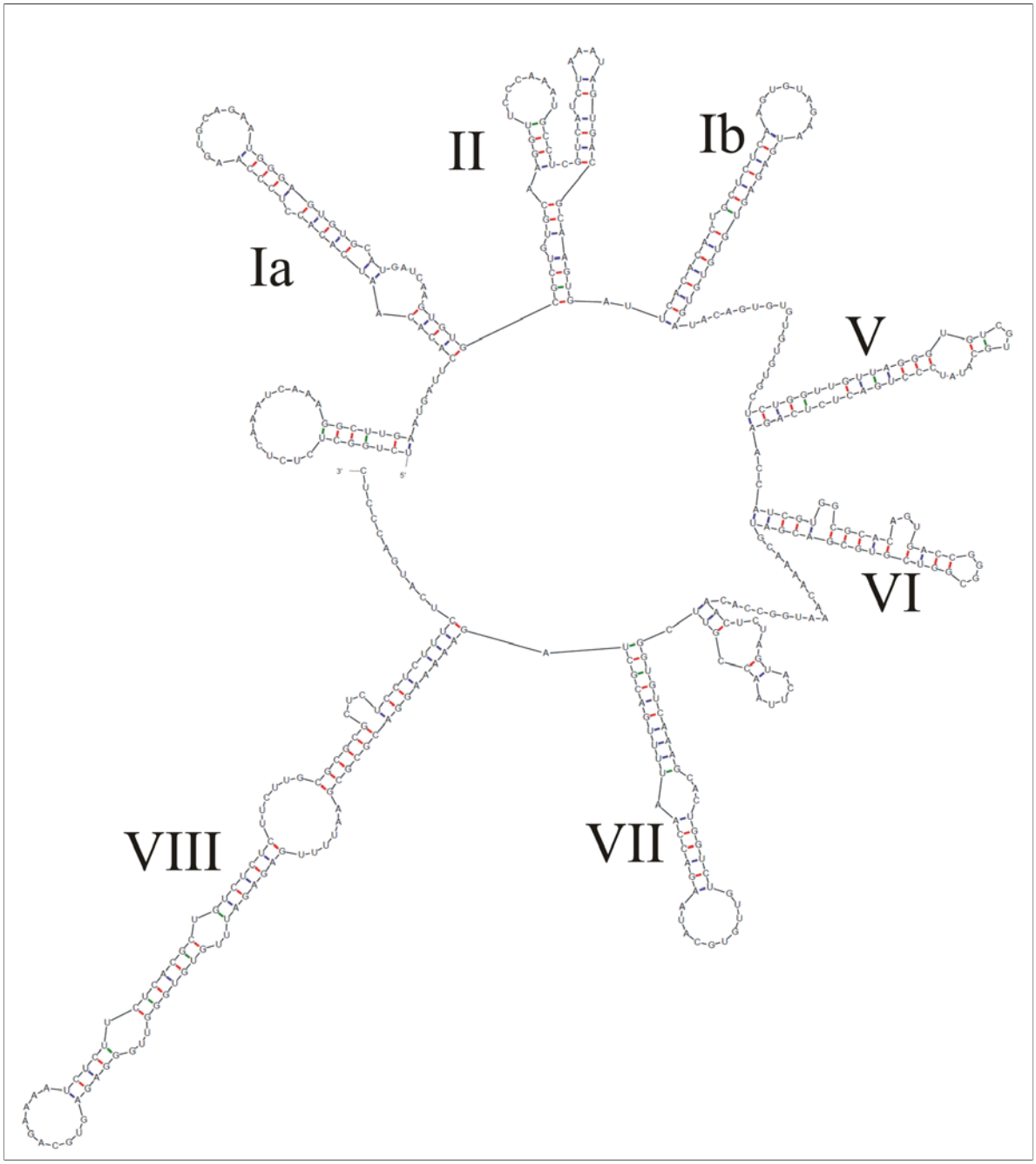


Figure 6.22: Structure of the ITS1 variant 5-5 consensus sequence in species 5
The structure is derived from the structural variant 5-5 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, IIa, Ib, IIb, Ic, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of five structural variants detected in species 5.

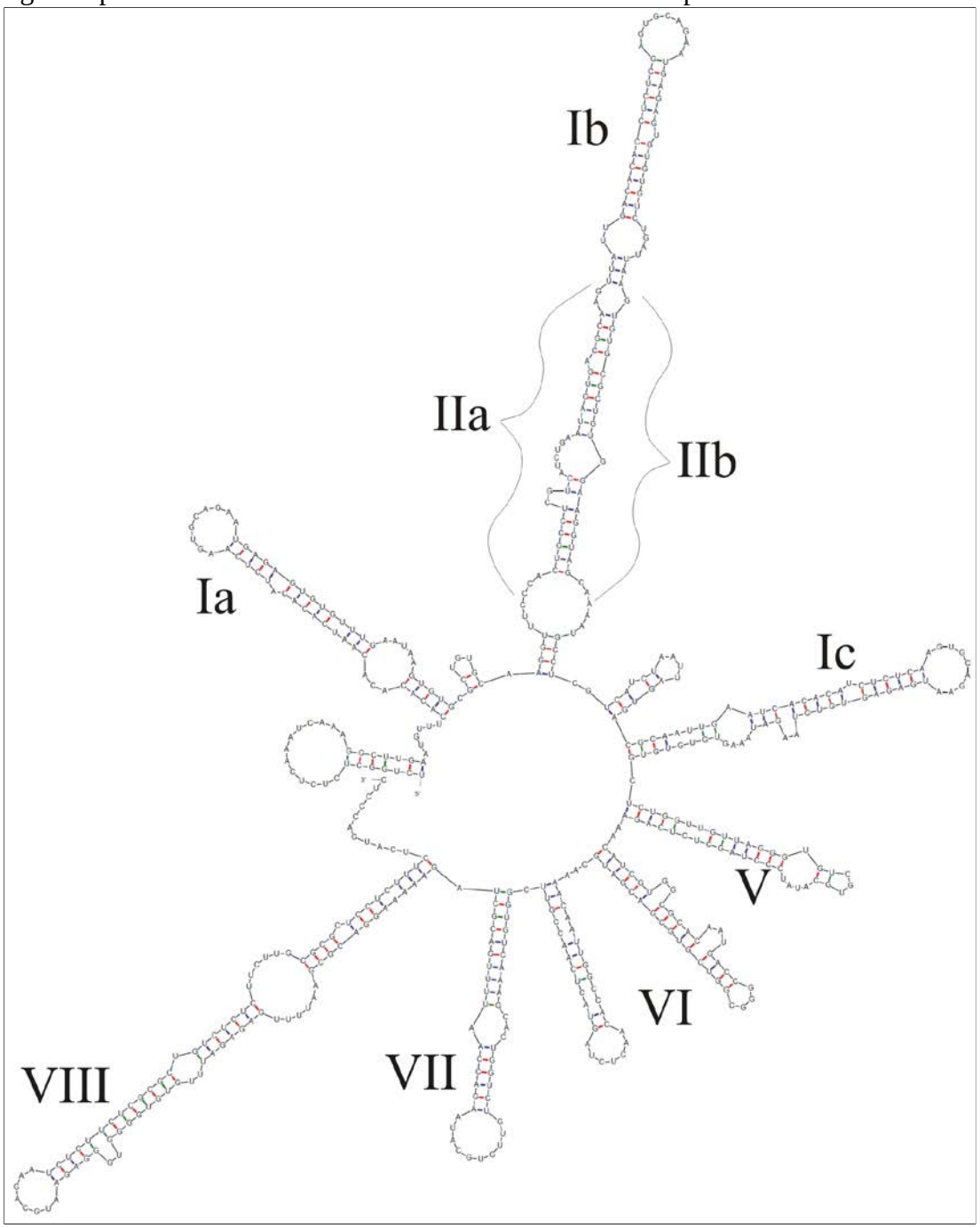


Figure 6.23: Alternative folding of the ITS1 variant 5-5 consensus sequence in species 5
The structure is derived from the structural variant 5-4 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, IIa, Ib, IIb, Ic, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of five structural variants detected in species 5.

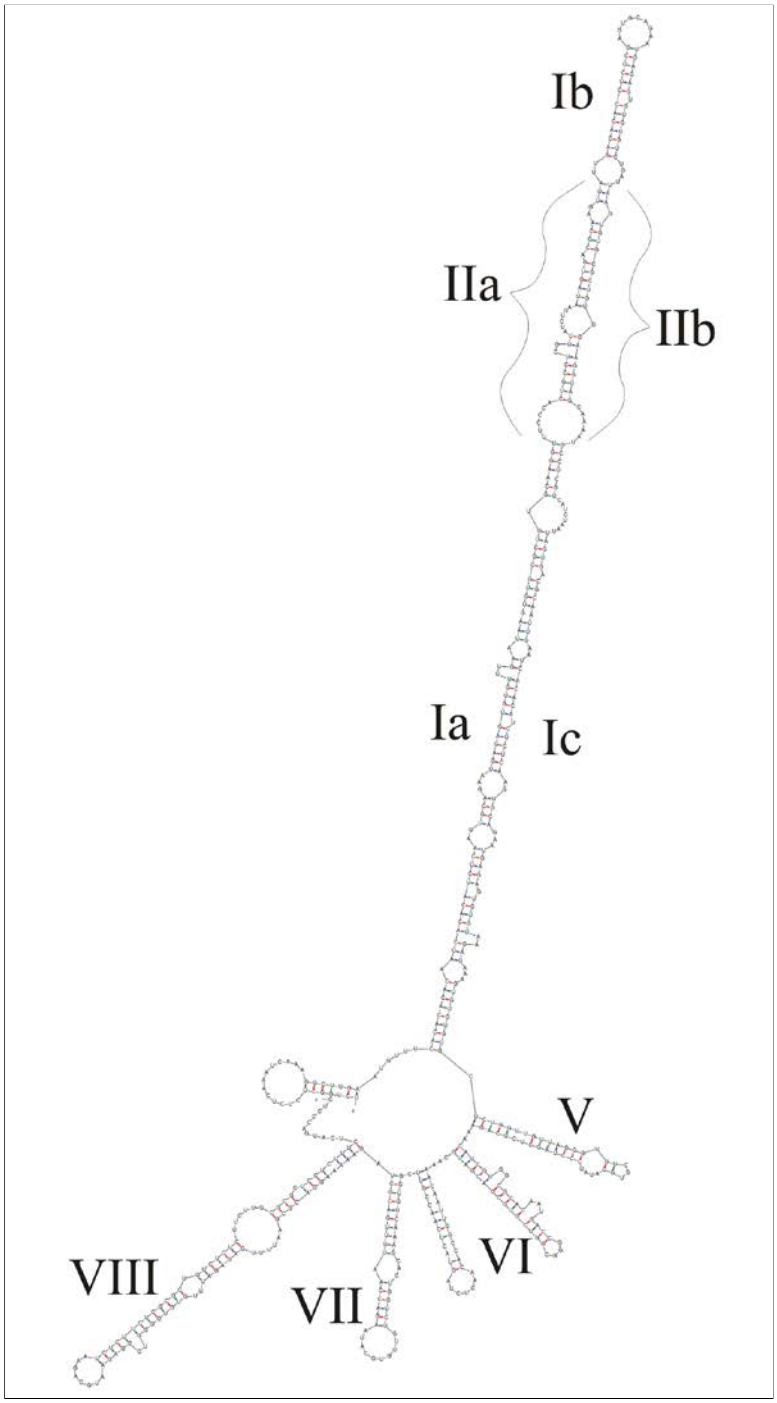


Figure 6.24: Species 5 ITS1 network with respect to ITS1 structural variants
The ITS1 network is colored according to species 5 structural variants: 5-1 (red), 5-2 (yellow), 5-3 (blue), 5-4 (purple) and 5-5 (green). This network was constructed using domains VI, VII and VIII. The sizes of the circles are proportional to the sample size and colored according to each structural variant.

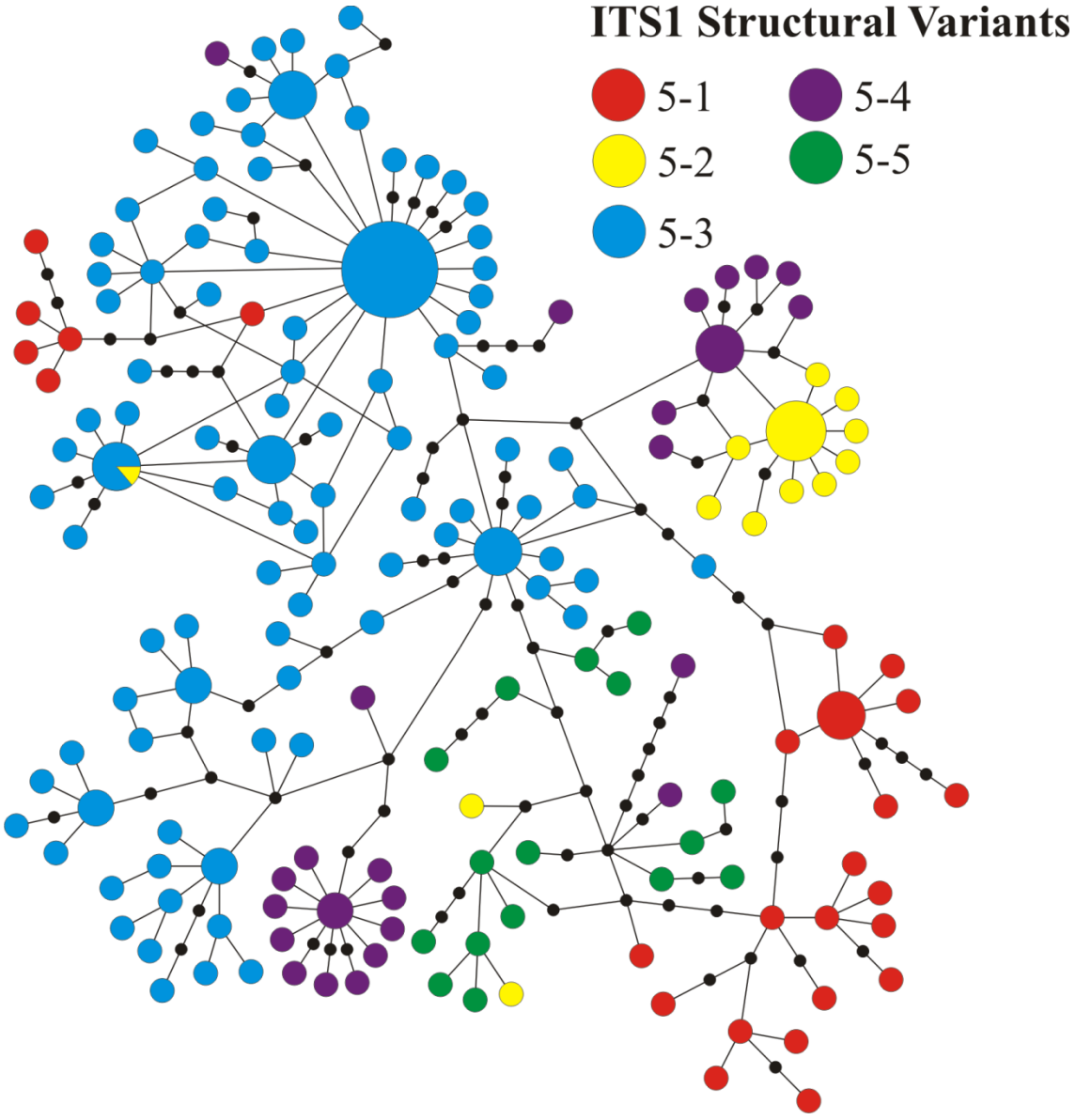
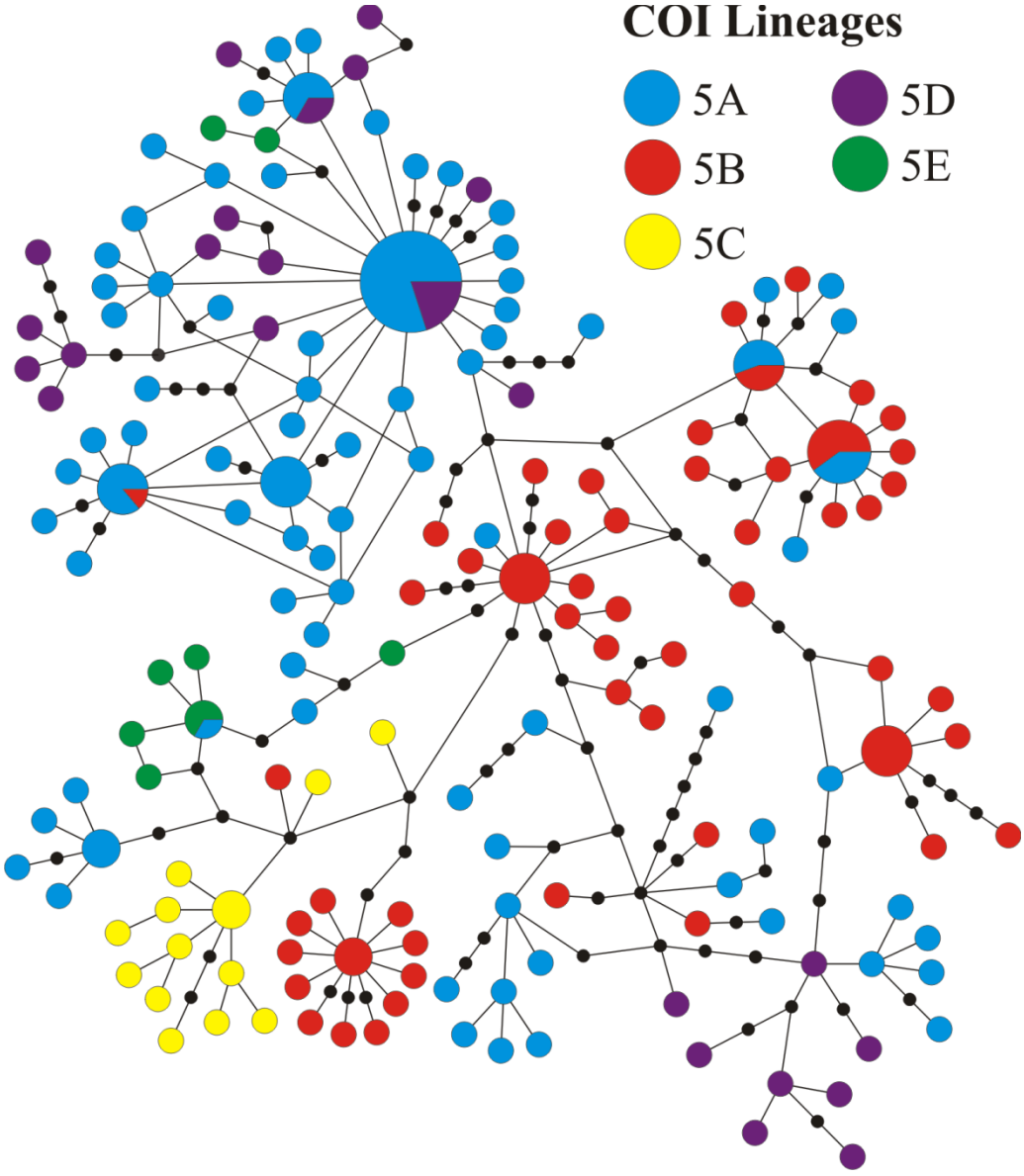


Figure 6.25: Species 5 ITS1 network with respect to COI lineage
The ITS1 network is colored according to species 5 COI lineages: 5A (blue), 5B (red), 5C (yellow), 5D (purple) and 5E (green). This network was constructed using domains VI, VII and VIII. The sizes of the circles are proportional to the sample size and colored according to each COI lineage.



6.3.13 Folding structures of the ITS1 gene within species 6 and 9

Thirty-five species 6 and seven species 9 individuals were sequenced for ITS1 resulting in 116 and 22 clones, respectively (table 6.1). The sixty-five species 6 ITS1 sequences ranged in length from 401 to 734 bases ($\pi = 0.076$). The fifteen species 9 sequences were less varied in length from 445 to 545 bases ($\pi = 0.049$). Species 6 possesses six major ITS1 structural variants (fig. 6.2) and only domains I, VI and VIII were present in all versions. These variants separated into three groups. ITS1 structural variants 6-1 (fig. 6.26) and 6-2 (fig. 6.27) both lacked domain VII and differed only in the number of repeats of domain I and II. Structural variants 6-3 (fig. 6.28), 6-4 (fig. 6.29) and 6-5 (fig. 6.30) possessed helix VII along with a variable number of domain I and II repeats. Variant 6-6 (fig. 6.31) consisted of a single domain I and the two helices upstream of domain VI were only homologous with the ITS1 sequence from species 9 (domain IX).

Species 9 possessed two copies of domain I, a single intervening domain II along with domains VI, VII and VIII (fig. 6.32). Similar to structural variant 6-6, two helices (domain IX) were upstream of domain VI.

Figure 6.26: Structure of the ITS1 variant 6-1 consensus sequence in species 6
The structure is derived from the structural variant 6-1 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, Ic, VI and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. The sequences between the domains Ia, Ib and Ic were prevented from binding with each other to force an open structure. This figure represents one of six structural variants detected in species 6.

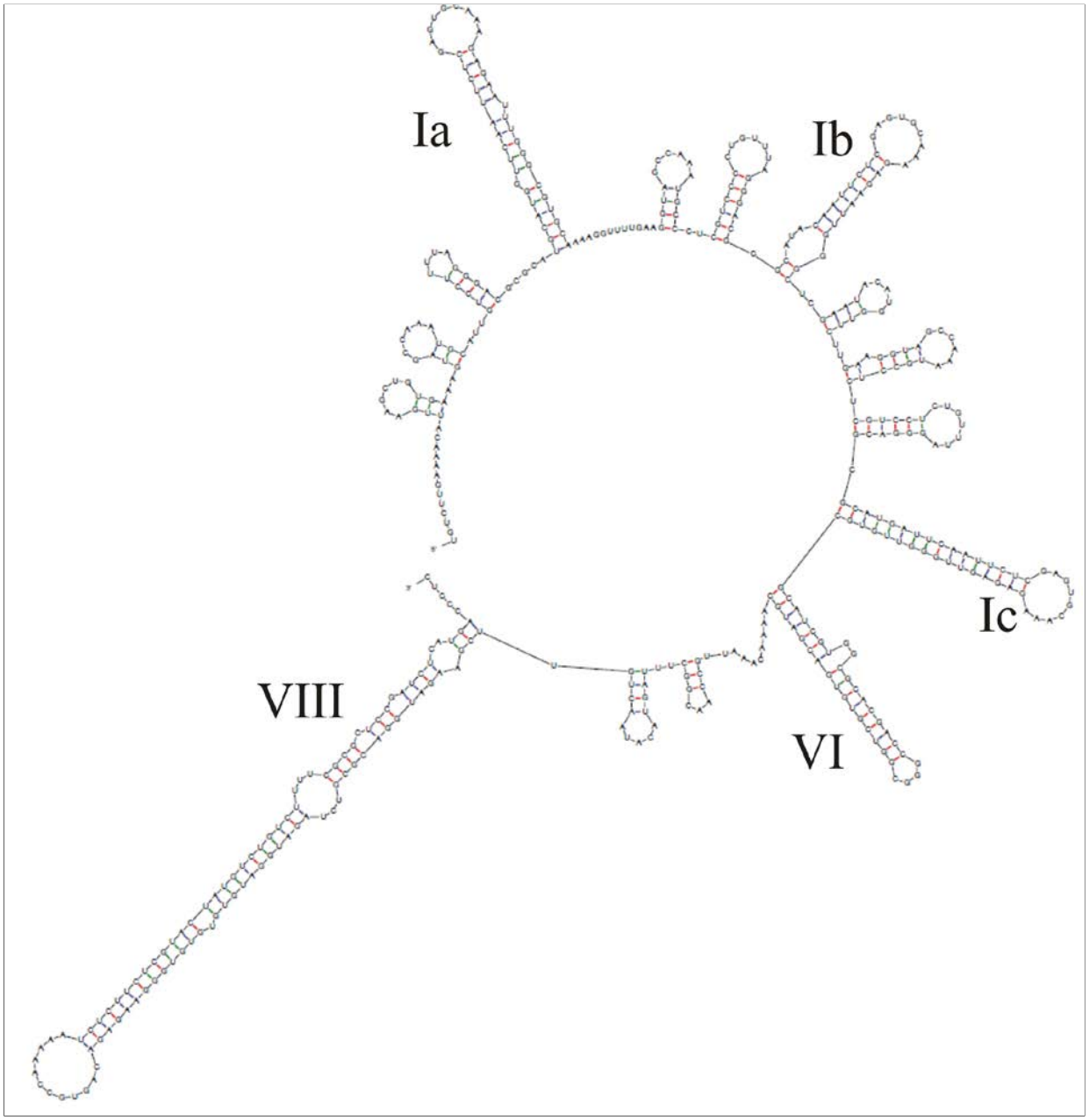


Figure 6.27: Structure of the ITS1 variant 6-2 consensus sequence in species 6
The structure is derived from the structural variant 6-2 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, Ic, Id, VI and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. The sequences between the domains Ia, Ib, Ic and Id were prevented from binding with each other to force an open structure. This figure represents one of six structural variants detected in species 6.

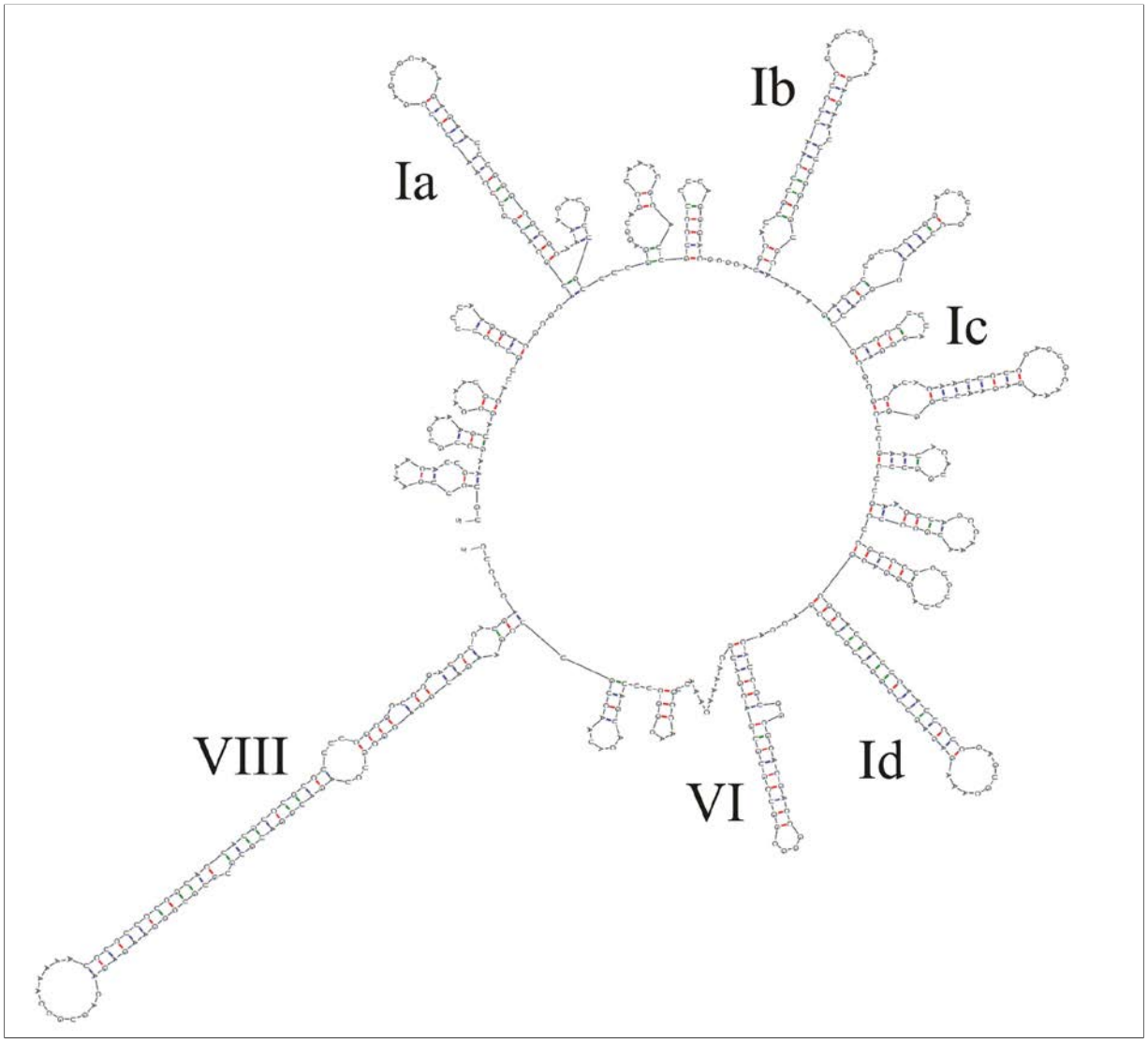


Figure 6.28: Structure of the ITS1 variant 6-3 consensus sequence in species 6
The structure is derived from the structural variant 6-3 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. The sequences between the domains Ia and Ib were prevented from binding with other areas of the sequence to force an open structure. This figure represents one of six structural variants detected in species 6.

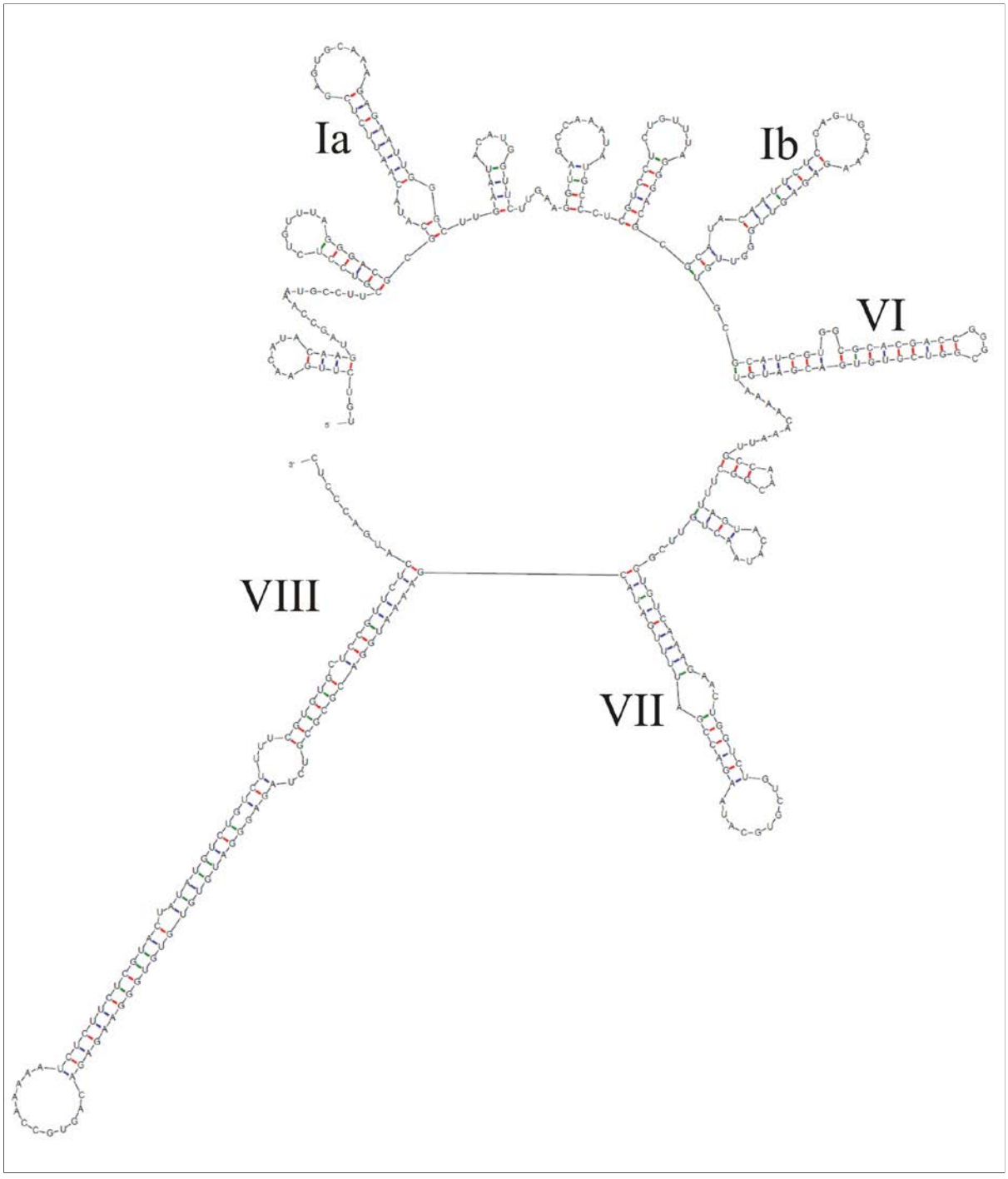


Figure 6.29: Structure of the ITS1 variant 6-4 consensus sequence in species 6
The structure is derived from the structural variant 6-4 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, Ic, IV, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. The sequences between the domains Ia, Ib and Ic were prevented from binding with other areas of the sequence to force an open structure. This figure represents one of seven six variants detected in species 6.

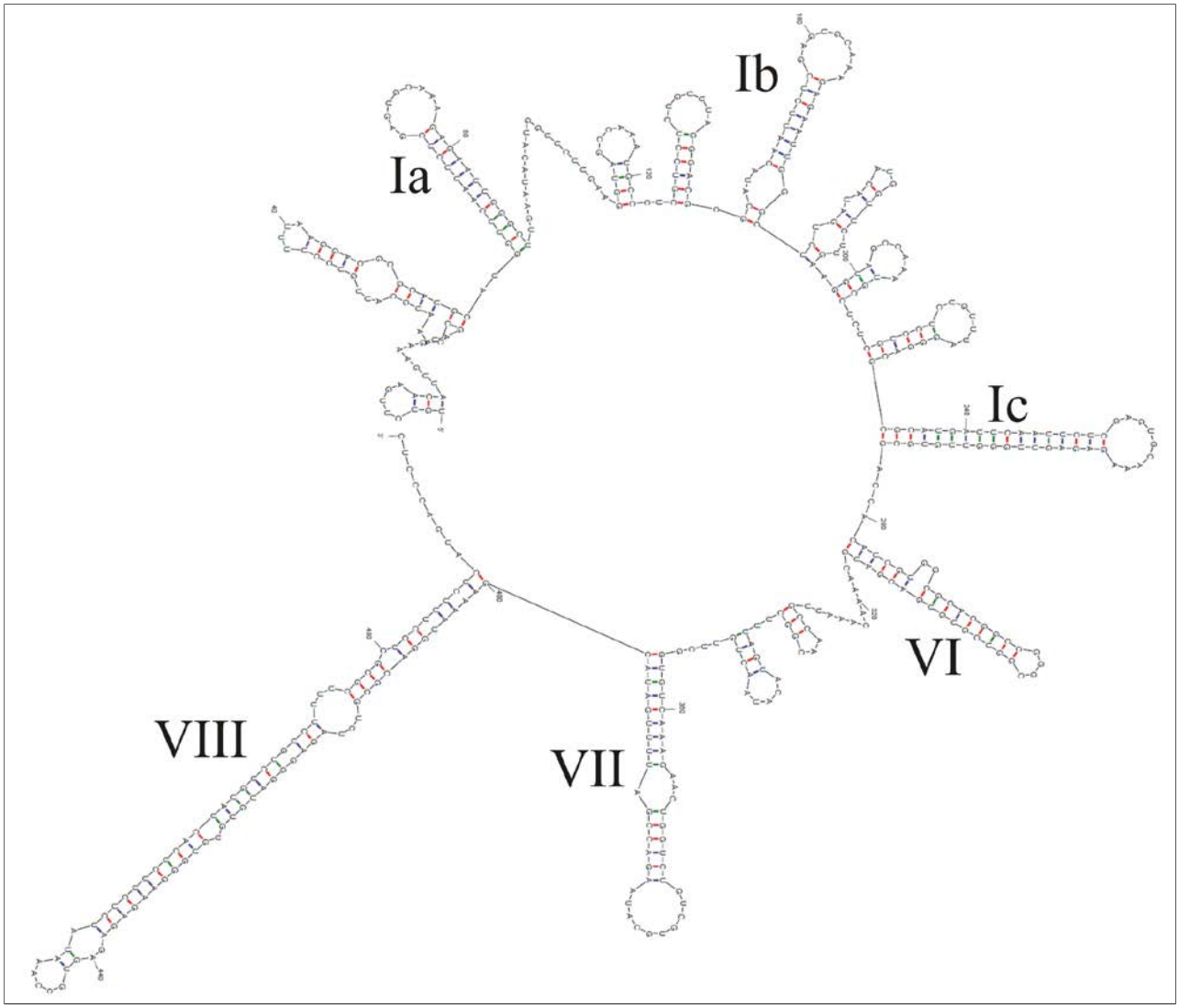


Figure 6.30: Structure of the ITS1 variant 6-5 consensus sequence in species 6
The structure is derived from the structural variant 6-5 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, Ic, Id, IV, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. The sequences between the domains Ia, Ib, Ic, and Id were prevented from binding with other areas of the sequence to force an open structure. This figure represents one of six structural variants detected in species 6.

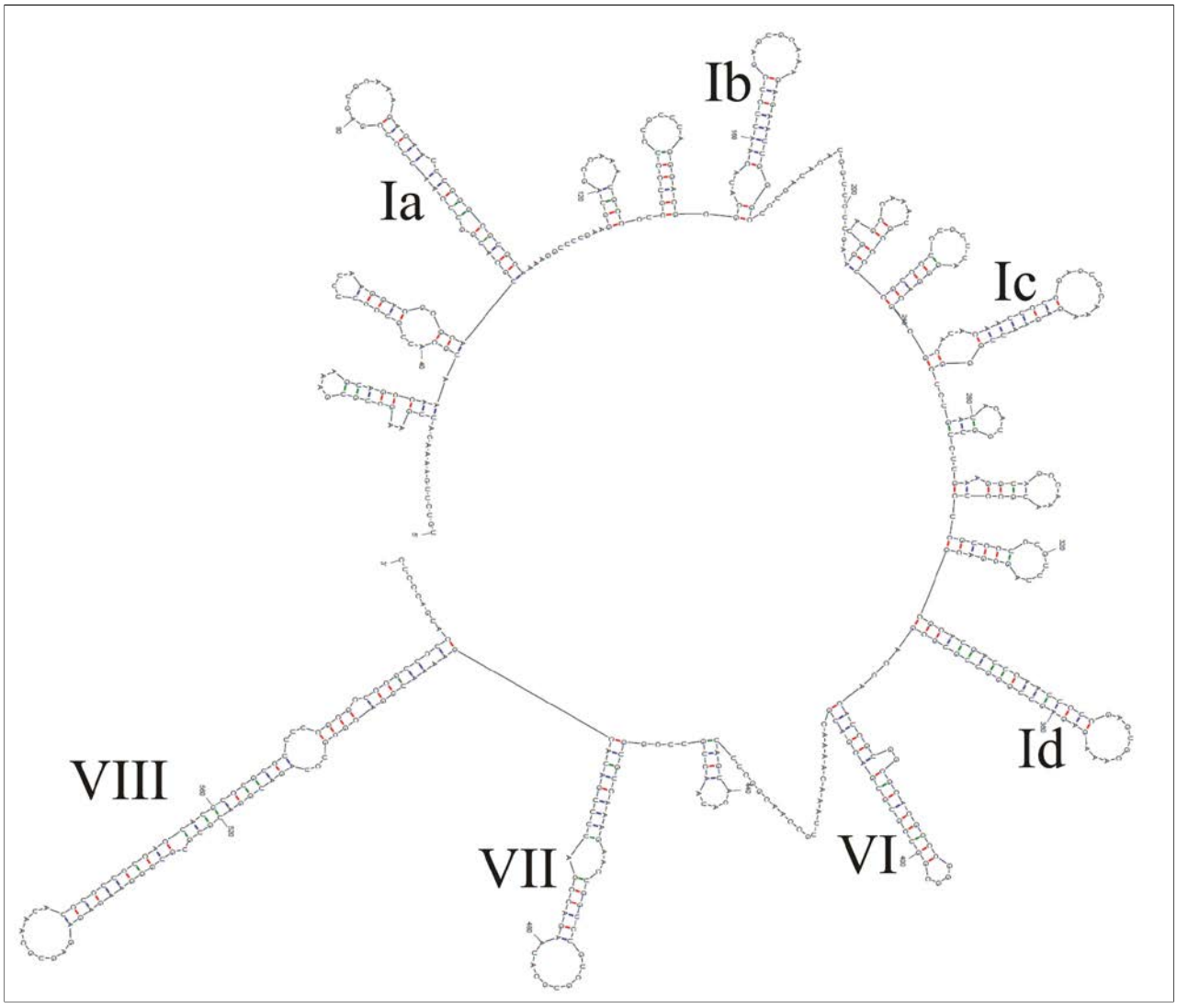


Figure 6.31: Structure of the ITS1 variant 6-6 consensus sequence in species 6
The structure is derived from the structural variant 6-6 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (I, IV, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites. This figure represents one of six structural variants detected in species 6.

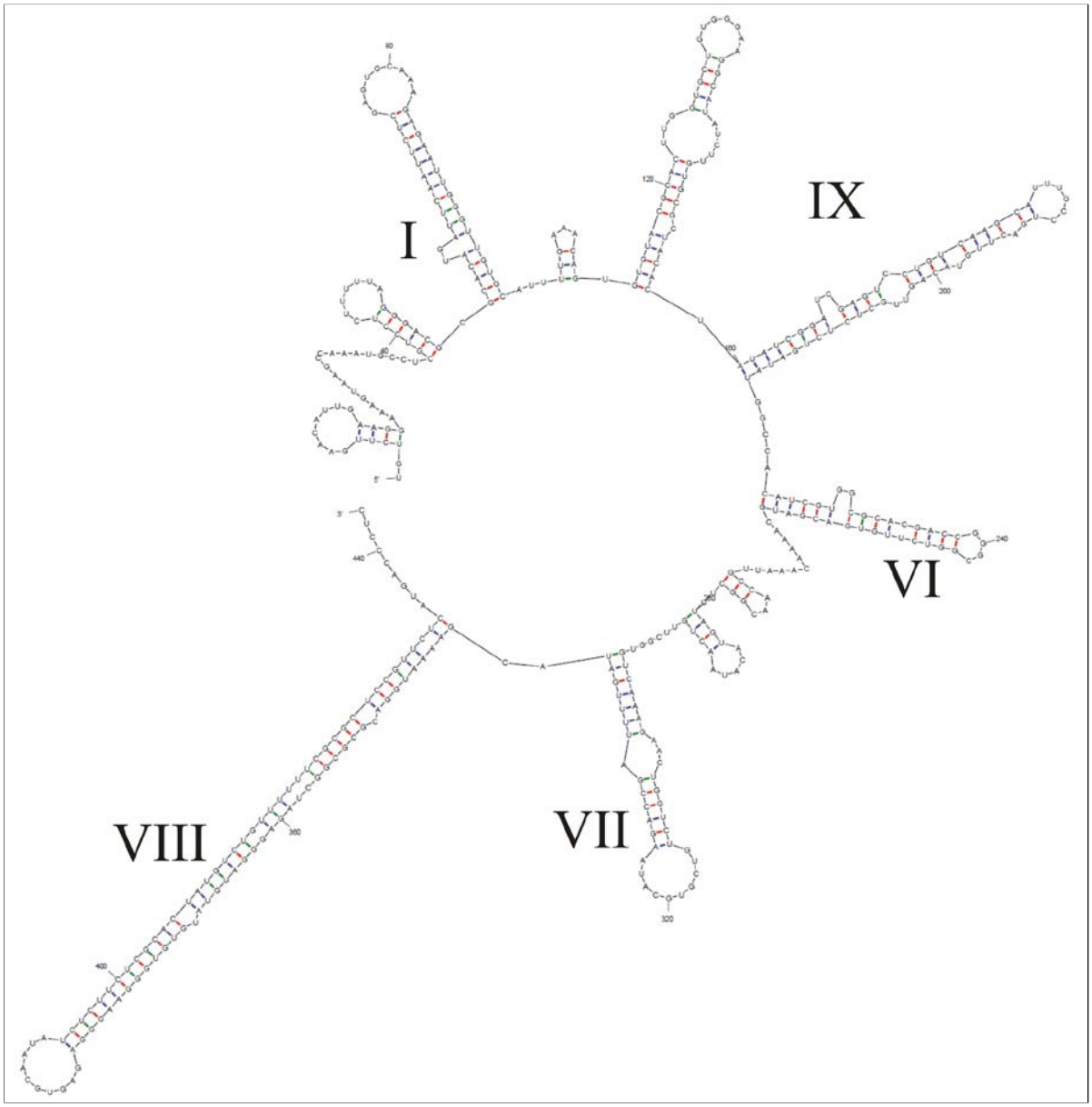
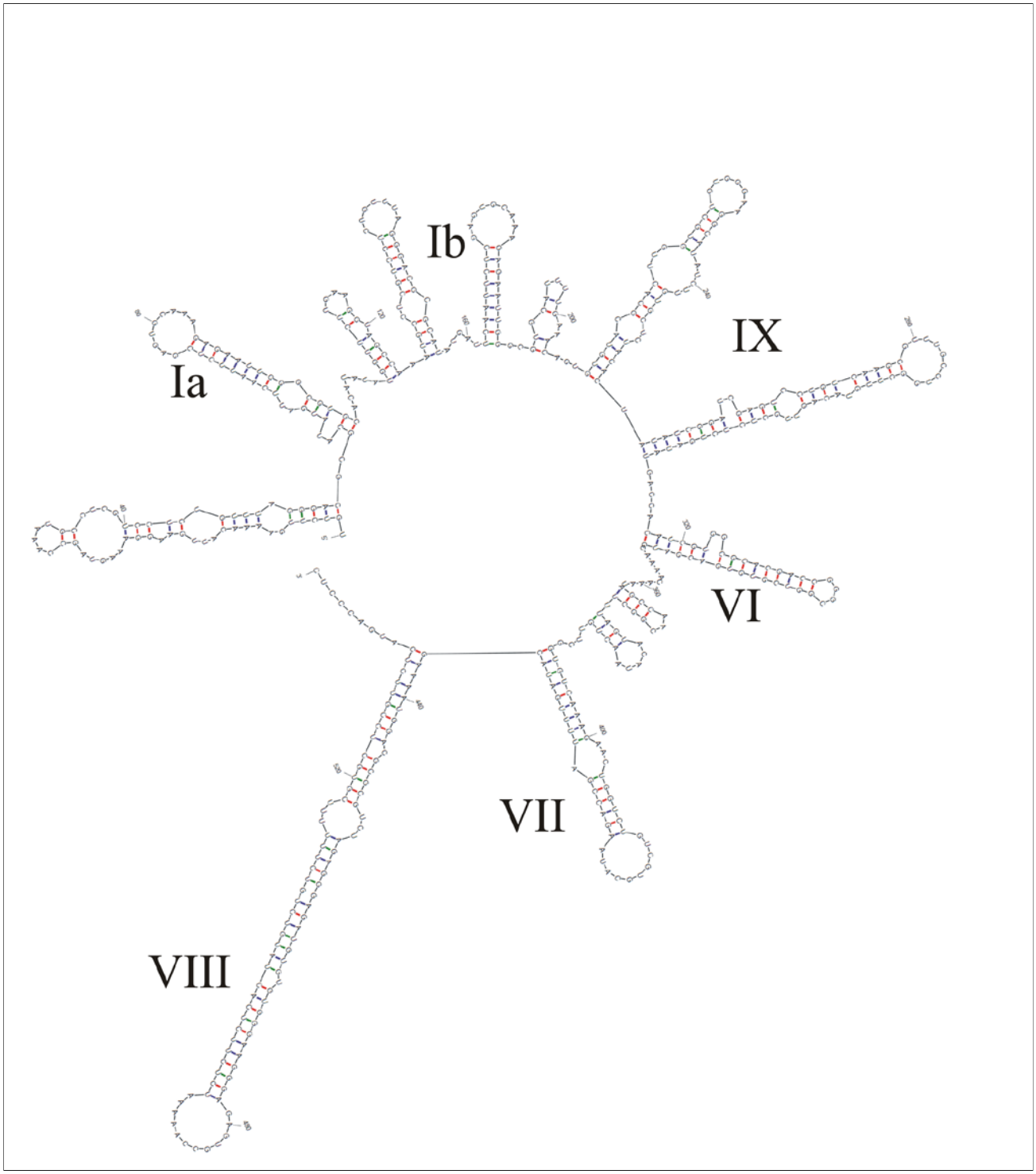


Figure 6.32: Structure of the species 9 ITS1 consensus sequence
The structure is derived from the species 9 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, IX, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites.



6.3.14 Genetic patterns of the ITS1 gene within species 6 and 9

A network of ITS1 sequences from species 6 and 9 (using domains VI, VII and VIII only) indicated a close genetic link between ITS1 structural variants 6-6 and 6-7 (fig. 6.33). Interestingly, these two groups were spatially separated by the greatest geographic distance (~8000 km). Populations possessing species 6 ITS1 structural variant 6-6 were distributed in populations along the eastern coast within the Central Appalachian, Great Lakes and Southeastern faunal provinces (fig. 6.34). Species 9 was restricted to the Cascadia faunal province and possesses a single ITS1 structural variant (6-7). ITS1 structural variant 6-5 was the longest of the variants and most common (fig. 6.35), while the rest (6-1, 6-2, 6-3 and 6-4) were rare.

The AMOVA (table 6.8) clearly indicated a strong between faunal variance component, which was consistent with spatial patterns of species 6 and 9. The 35 individuals were partitioned on the basis of faunal provinces and had six groups (Cascadia: 7 individuals, Central Appalachian: 4 individuals, Great Lakes: 7 individuals, Hudson Bay: 2 individuals, Northern Appalachian: 4 individuals and Mississippi: 13 individuals) and the between-group variance component was 31%. Although intraindividual variation was not as large a variance component as detected in other *Hyalella* species (table 6.8) it was still significant at 23.02%. This was a likely the result of the more homogenized ITS1 sequences within species 9. Species 6 individuals were more likely to have multiple structural variants in the Great Lakes and Mississippi faunal provinces.

Figure 6.33: ITS1 rRNA gene network of species 6 and 9 with respect to ITS1 structural variant

The ITS1 network is colored according to the structural variant listed on the insert: 6-1 (pink), 6-2 (yellow), 6-3 (light green), 6-4 (green), 6-5 (blue), 6-6 (dark purple) and 6-7 (red). This network was constructed using domains VI, VII and VIII. The sizes of the circles are proportional to the sample size and colored according to each structural variant.

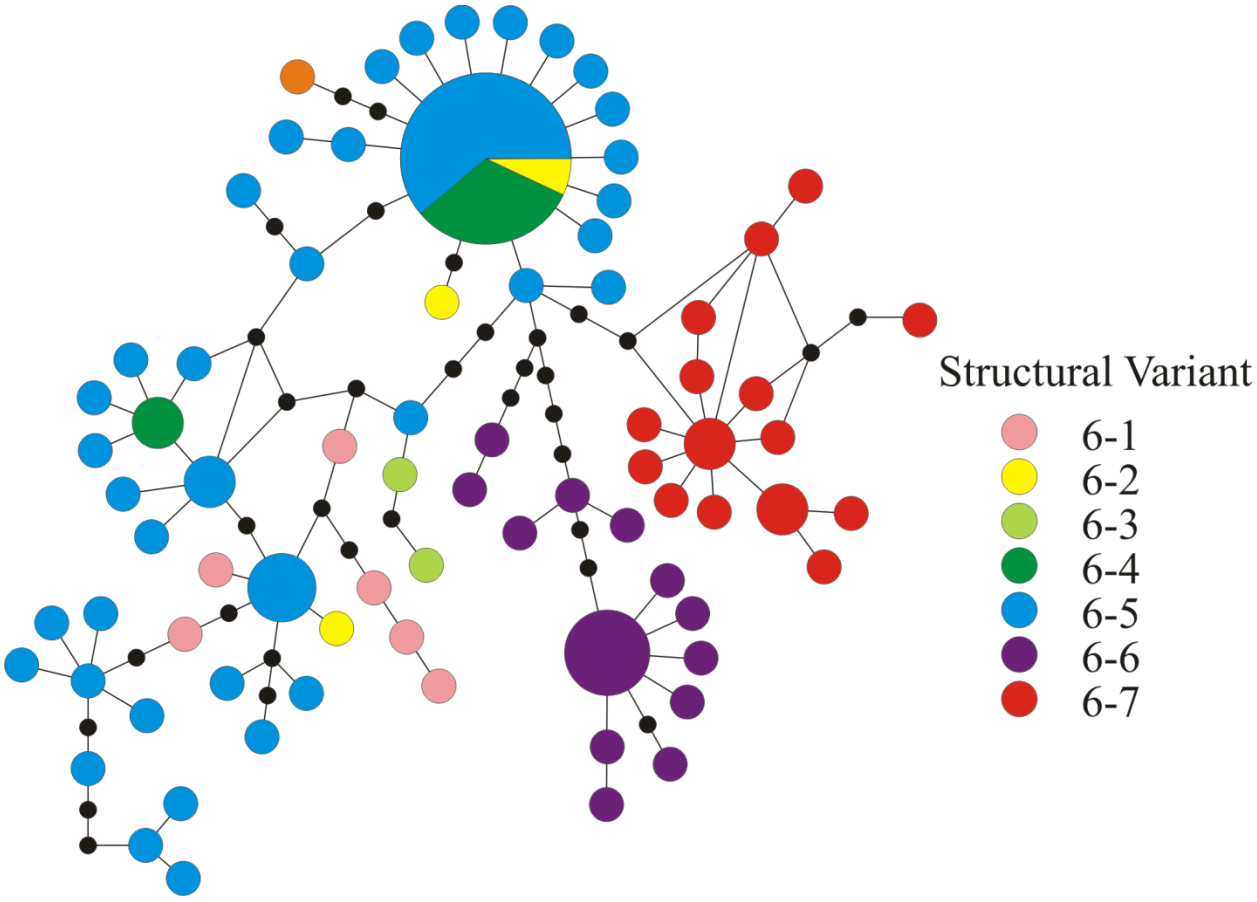
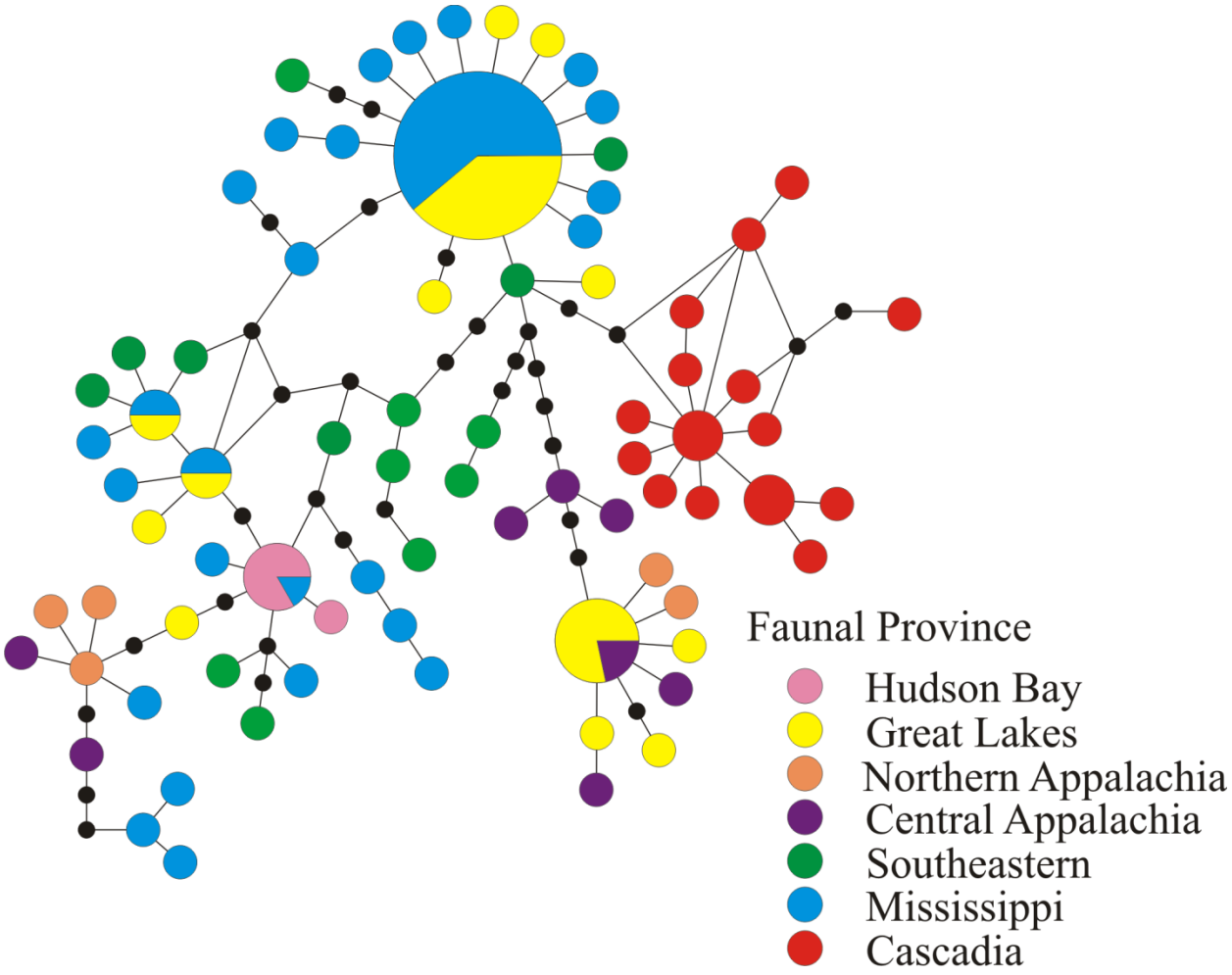


Figure 6.34: ITS1 rRNA gene network of species 6 and 9 with respect to faunal province
The ITS1 network is colored according to faunal province as listed on the insert: Hudson Bay (pink), Great Lakes (yellow), Northern Appalachia (orange), Central Appalachia (purple), Southeastern (green), Mississippi (blue) and Cascadia (red). This network was constructed using domains VI, VII and VIII. The sizes of the circles are proportional to the sample size.



6.3.15 Genetic patterns of the ITS1 gene within species 7

Nineteen species 7 individuals were selected haphazardly on the basis of the COI phylogeny and spatial distribution. Fifty-one clones were generated resulting in 31 unique sequences (π = 0.0117) with a sequence length that ranged from 462 to 492 bases (table 6.1). The folding structure (at 16°C) consisted of at least six domains. Domain I was repeated twice, although unlike other *Hyalella* species (fig. 6.1), there was no intervening domain II (fig. 6.35). The remaining domains were consistent with the folding patterns detected in species 1, 2 and 3 (domains IV, V, VI, VII and VIII).

The AMOVA using faunal provinces (Hudson Bay: 13 individuals, Mississippi: 1 population and Yukon-Mackenzie: 5 individuals) indicated that the between-group variance component was −8.84%. Since an AMOVA was an analysis of covariance, negative values were possible and should be taken as 0% (Excoffier, 1998). This strongly suggests that ITS1 variation within species 7 does not partition along faunal provinces. The intraindividual variance component was quite high (52%), which suggests the persistence of variation across large spatial distances; a pattern that suggests either multiple NORs or high migration rates (or both). The ITS1 network indicated two distinct groups (fig. 6.37), both of which were detected species 7 individuals from Green Lake and Barrier River, SK.

Figure 6.35: Structure of the species 7 ITS1 consensus sequence
The structure is derived from the species 7 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (Ia, Ib, IV, V, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites.

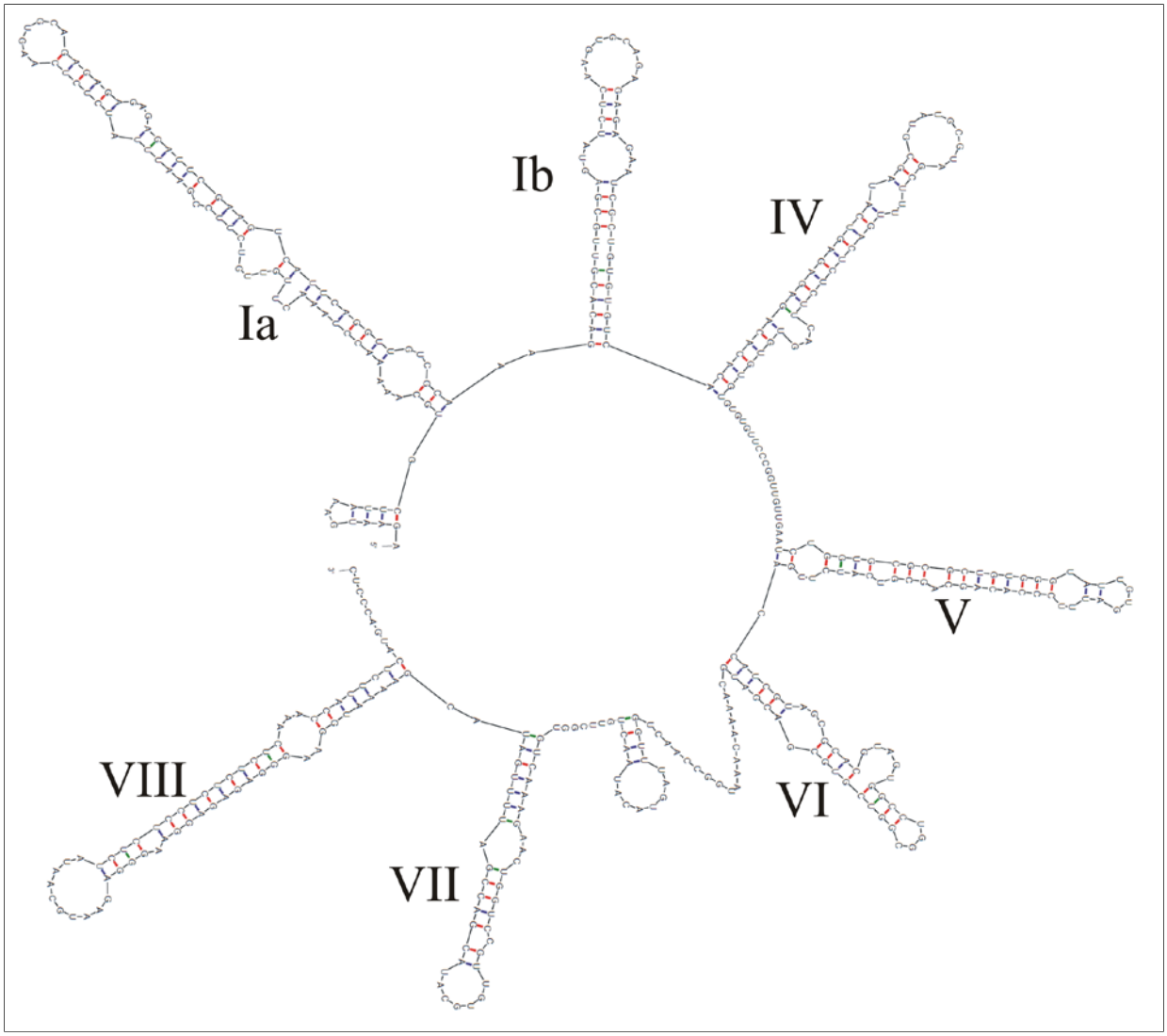
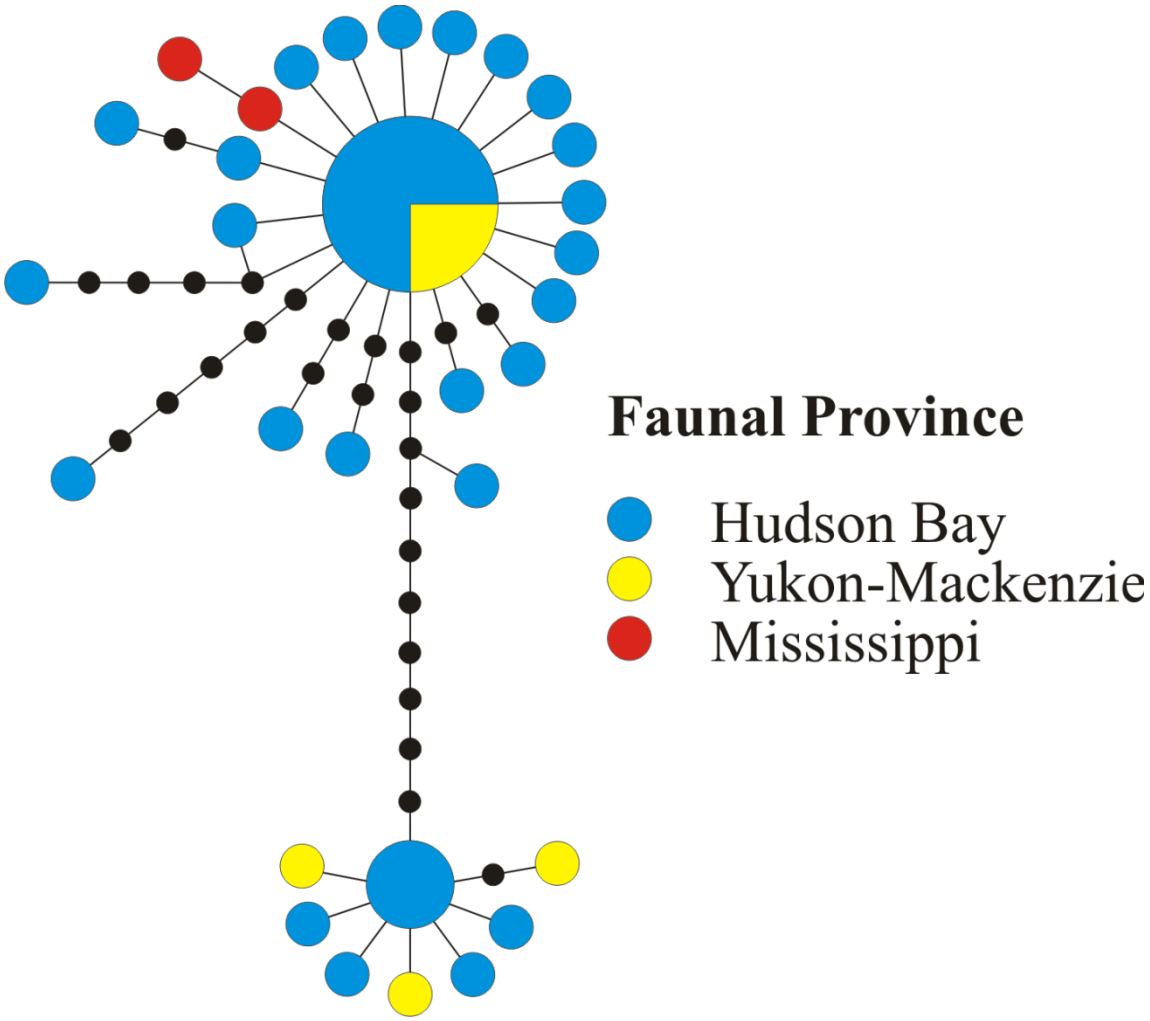


Figure 6.36: ITS1 network of species 7 with respect to faunal province
The ITS1 network is colored according to faunal province as listed on the insert: Hudson Bay (blue), Yukon-Mackenzie (yellow) and Mississippi (red). This network was constructed using all species 7 domains. The sizes of the circles are proportional to the sample size.

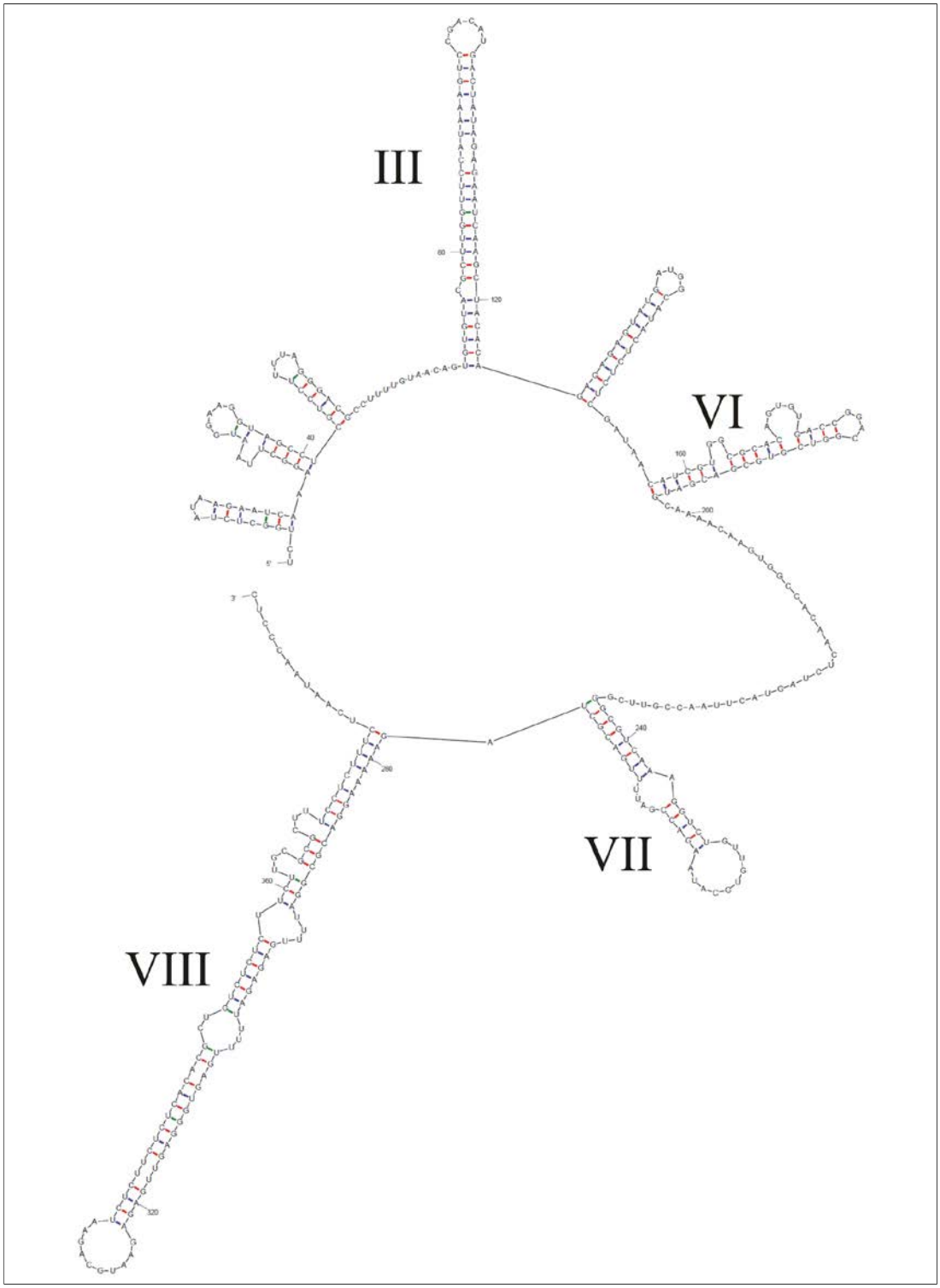


6.3.16 Genetic patterns of the ITS1 gene within species 8

Twenty-three species 8 individuals were haphazardly selected on the basis of the COI phylogeny from populations along the Florida Peninsula and Mississippi River watershed. Seventy-four clones were generated resulting in 49 unique sequences ($\pi = 0.0291$) with a highly variable ITS1 sequence length, ranging from 381 to 527 bases (table 6.1). The folding of the rRNA structure at 16°C resolved five major domains. Unlike every other *Hyalella* species analyzed in this study, species 8 was the only one with a domain I. The initial 5' domain was III followed by domains IV, VI, VII and VIII (fig. 6.39).

The AMOVA using faunal provinces (Mississippi: 2 individuals and Southeastern: 21 individuals) had a between group variance component of 54.8% (table 6.8), although the sample size was small the Mississippi group. This pattern suggests that the peninsular species 8 group was distinct from the western populations.

Figure 6.37: Structure of the species 8 ITS1 consensus sequence
The structure is derived from the species 8 consensus sequence with a folding temperature of 16°C. Base pair bindings within the helices (III, VI, VII and VIII) were enforced at the base of the stems as supported by the analysis of covarying sites.



6.4 DISCUSSION

The genetic patterns of the first internal transcribed spacer (ITS1) within the *Hyalella* species complex are intricate, with high levels of intra-individual variation and multiple structural variants. Identifying compensatory mutations was successful in detecting common domains among structurally dissimilar molecules. While the second internal transcribed spacer (ITS2) possesses a better understood structure (Coleman, 2013), it seems that ITS1 allows for the exclusion, duplication or reordering of helices. Duplication and recombination of domains within ITS1 can be problematic, because they can undermine phylogenetic analysis (Posada & Crandall, 2002). Domains VI and VIII are potentially important to the functionality of the ITS1 molecule given their conservation across all variants, while domains I and VII had nearly universal representation. Domain I was consistently duplicated in species 5 and 6, with the most conserved elements being located near the tip of the structure. The internal transcribed spacer is cleaved during assemble of ribosomal subunits by both endonucleases and exonucleases (Henras *et al.*, 2008). The ribosomal biogenesis pathways in yeast and humans have highlighted different cleavage sites (Sloan *et al.*, 2013). These provide anchor points that allow comparison between ITS1 molecules. Determination of the larger structure was not possible given the potential for reordering of domains and duplications. A broader approach will be required to determine the in situ structure of ITS1, potentially using x-ray crystallography (Cate *et al.*, 1999). This could highlight folding structures that are not detected by the analysis of compensatory mutations.

The high levels of intra-individual variation strongly suggests the presence of multiple lineages within the genome that either exist within the same nucleolus operating regions (NOR), or more likely as separate entities (possibly on multiple chromosomes). This is gaining new attention in the literature with respect to the suitability of transcribed spacers to identify species

boundaries. When multiple sequence variants are present within the same genome, analysis of cDNA libraries have suggested that they are not equally expressed (Ruiz-Estévez *et al.,* 2015). This may provide a tool to refine phylogenetic analysis of ITS1 genes in *Hyalella* species with multiple structural variants, by excluding variants that are potential pseudogenes. However, the large number of ribosomal repeats is useful in periods of rapid demographic expansion and only a portion may be active at a given time (Kobayashi, 2011). These patterns in Hyalella are consistent with levels of intra-individual ITS1 variation detected in *Daphnia* (Gießler & Englbrecht, 2009) and *Drosophila* (Averbeck & Eickbush, 2005).

All species (except 5 and 7) indicated significant covariation among faunal provinces. The greatest of which were in species 6 and 9, with distinct spatial/genetic groups at the Atlantic and Pacific coasts, or species 8, which partitioned the variation into a Southeastern and Mississippi River group. The genetic patterns within species 3 indicated a clinal shift in diversity as might be expected under a model of isolation by distance. What is further apparent is that the ITS1 genetic patterns provide a different resolution on the demographic history of the *Hyalella* species complex.

6.4.1 Constrained ITS1 variation within species 1 and 2

The degree of ITS1 sequence homogenization within species 1 and 2 makes it impossible to delineate the phylogroups. Concerted evolution will promote divergences among groups that are reproductively isolated. Given the level of ITS1 homogenization between these two “species”, either sufficient gene flow between them (including disparate refugial populations in Beringia, Cascadia and the Missouri River) has allowed for their homogenization, or insufficient time has

passed since the split of species 1 and 2 to allow for the divergence of ITS1 lineages. In comparison, the ITS1 phylogroups between species 6 and 9 (COI divergence of 12% rather than 8%) are not fully resolved. *Daphnia cucullata* and *Daphnia galeata* have an 8.5% COI divergence as well as nearly identical ITS1 variants (Gießler & Englbrecht, 2009). The authors attributed this pattern to incomplete lineage sorting in conjunction with larger effective population sizes. Vollmer & Palumbi (2004) offered likeminded explanations with respect to similar ITS1 lineages within different species of coral. However, Taylor *et al*., (2005) suggested that the nearly identical ITS1 sequences within *Daphnia galeata* and *Daphnia dentifera* are a consequence of introgression followed by fixation.

The spatial structure among faunal provinces was significant, but given the size of the area this encompasses the effect is much more muted than indicated by the biogeographical analysis of COI (chapter 3). Compensatory mutations within the transcribed spacers are often correlated with species boundaries (Wolf *et al.,* 2015). The full extent of gene flow among species 1 and 2 remains poorly understood and will likely necessitate mating trails or extensive analysis of sympatric wild populations.

6.4.2 Isolation-by-distance and gene flow within species 3

The gradual shift in diversity among and within faunal provinces suggests that species 3 individuals possess ITS1 sequences more closely related to their neighbors than distant populations along a longitudinal gradient. This is consistent with the biographical analysis of species using COI, which suggested a range expansion from Atlantic refugium into the Great

Lakes and Mississippi River basins. The analysis of ITS1 was not able to detect a distinct group consistent with the COI clade 5-1.

5.4.3 Potential hybridization among species 5 lineages

Species 5 possesses a high degree of structural and sequence variation. Populations along the coastal plain in the Southeastern faunal province and within the Mississippi River watersheds have individuals with all the structural variants. These southern regions are far more diverse than the northern populations of species 5. While COI lineage 5A individuals possess all five ITS1 structural variants, only 5-3 is widely distributed in the northern populations. As explored in chapter 4, the Great Lakes region was colonized by COI lineage 5A through proglacial waterways from Atlantic refugia within the Coastal Plain. Within the Southeastern Faunal Province are two notable populations (Dargon Pond, SC, and Mountain Island Lake, NC). Within Dargon Pond, SC, are ITS1 sequences that possess 3' domains (VI, VII and VIII) more consistent with ITS1 structural variant 5-2, but folding configurations consistent with structural variant 5-4. Similarly, in Mountain Island Lake, NC, there are sequences that are genetically similar to structural variant 5-3, but with folding structures akin to variant 5-1. The genetic patterns in COI are consistent with older lineages within species that may warrant their own classification; deep genetic divergences coupled with spatial structuring among faunal provinces. However, some COI lineage 5A sequences were detected in the Southeastern faunal province, which is predominately associated with COI lineage 5B. In conjunction with the potential hybridization among species 5 ITS1 structural variants, it may suggest recent introgression. Comparisons of parsimony network ITS1 topologies among the COI and ITS1 genes has yielded

similar patterns in the European rose chafer whereby incongruence among markers was inferred as hybridization among subspecies *Cetonia aurata aurata* and *Cetonia aurata pisana* (Ahrens *et al.*, 2013).

5.4.4 Incomplete lineage sorting among species 6 and 9

Species 6 possessed the largest number of structural variants that were largely separated on the basis of the number of domain I and II repeats. As observed in species 5, the 5' end of the ITS1 molecule in *Hyalella* appears to be a notably variable region that likely undergoes recurrent recombination with other ITS1 repeats. This would explain the variable number of domain I and II repeats as they are gained or lost over successive generations. While structural variants 6-1 through 6-5 possess little spatial structure, variant 6-6 and species 9 are closely allied to distinct coastal regions (Atlantic and Pacific, respectively). What is most fascinating is that these two groups are the most genetically closely related within the two species. On the other hand, the COI phylogeny strongly suggests that species 6 and 9 are sister taxa that have been partitioned into distinct groups. The distribution of species 6 extends into the Great Lakes and Mississippi River basin, but these regions lack *Hyalella* with the ITS1 structural variant 6-6. At one time, species 6 and 9 likely represented a single group and the paraphyletic relationship among ITS1 lineages may suggest incomplete lineage sorting. However, topological incongruence between COI and ITS1 markers may not be uncommon in invertebrates. Nekola *et al.*, (2015) detected similar incongruence in at least 15% of their gastropod samples (genus *Pupilla*) throughout Europe, North America and Asia, for which they cited mitochondrial introgression as a potential factor.

5.5 CONCLUSIONS

Nuclear genes evolve at a slower rate than their mitochondrial counterparts as a consequence of higher effective population sizes and diploid (usually) state. This is referred to as the 3x rule in relation to diploid nuclear genes (Palumbi, 1996). In the case of ITS1 and other members of the ribosomal multigene family, there can be multiple copies in several distinct NORs spread throughout the genome. As such, the genetic patterns of ITS1 offer a different resolution than those analyzed with the COI gene. While the species classifications are usually supported by ITS1, in both sequence and structural evolution, some of the delimitations of *Hyalella* species are blurred at COI divergences between 8 and 12%. Furthermore, this is consistent with the suggestions by Gießler & Englbrecht (2009) and Taylor *et al.*, (2005) that the ITS1 genome within invertebrates is particularly permeable to introgression.

Broadly distributed large-bodied *Hyalella* species (1, 2 and 7) in glaciated regions possess little structural variation, while small-bodied lineages south of the glacial limit (3, 5 and 6) tended to have a greater degree of variation. As the search for universal barcoding genes intensifies, the utility of the internal transcribed spacers will be continuously assessed. What is evident in *Hyalella* is that the structural complexities of ITS1 are not insurmountable and common elements can be aligned in spite of gross dissimilarities. If these structural components can be aligned to other invertebrate species, ITS1 may be a powerful tool in species delineation and as a nuclear marker for biogeographical analysis.

Chapter 7

Comparative Phylogeography of Small and Large-Bodied *Hyalella* Species

7.1 Conclusions regarding the comparison of small and large-bodied *Hyalella* species

Hyalella 'azteca' was once thought to be a single species with a distribution that spanned North America, but the use of more recent molecular techniques have resulted in the discovery of a morphologically cryptic complex of over eighty members. The use of molecular tools within the field of biogeography has created an opportunity to disentangle the demographic histories of these cryptic groups. This study expands our understanding of divergent *Hyalella* lineages with respect to Pleistocene vicariance and post-glacial dispersal. These broadly distributed lineages have allowed the assessment of spatial patterns on a large enough scale to include both glaciated and unglaciated regions. The genetic and spatial patterns of *Hyalella 'azteca'* are complex and have been shaped by dynamic processes that stretch back through the Pleistocene.

In this study, nine major *Hyalella* species comprised of twenty-two distinct lineages were analyzed and all but one, (species 8), had populations within both glaciated and unglaciated regions. The genetic and spatial patterns within these species are indicative of rapid demographic expansion in conjunction with invasion into previously ice covered areas. The nested clade analysis of these species generated 147 significant geographical associations. The large-bodied species associations indicated patterns of isolation by distance in 47 to 73% of cases, and 22 to 37% suggested the potential for long distance dispersal. The small-bodied species were similar, with 43 to 64% of associations indicating isolation by distance, and 9 to 38% suggesting long distance dispersal. While the nested clade analysis indicated that long distance dispersal remains an important mechanism for colonization of glaciated regions, it was

not utilized by either size morphotype to a greater extent. The slopes of the isolation by distance curves revealed little difference between the two morphotypes. If one group was more reliant on avian vectors to disperse, then we would expect to see less structuring among the genetic and spatial patterns. Both small and large-bodied species exhibited patterns identified in the nested clade analysis consistent with a stepping stone model, whereby new populations are founded from neighboring demes. However, long distance colonization events remain a salient pattern at all levels of the nested clade analysis.

The refugial areas occupied by the different morphotypes were split along the continental divide with large-bodied and small-bodied species having refuges within the west and east, respectively (except for species 9). The spatial/genetic patterns within species 1, 2 and 7 indicate that the colonization of central Canada occurred prior to expanding into the Great Lakes region. This is consistent with the movement of fish species through the Lake Agassiz – Peace River connection (Rempel & Smith, 1998). The large-bodied *Hyalella* species represent a diverse and broadly distributed group of species. While species 8 was not detected in glaciated regions, all of these species possess genetic and spatial patterns consistent with range expansions during or after the Pleistocene. These species have dynamic ranges and their distributions have been shaped by a combination of both short and long distance dispersal. However, population structure with genetic patterns of isolation by distance indicates that long distance dispersal events, though significant, are likely rarer than either active dispersal via swimming or short distance jumps via animal vectors to neighboring habitats. The small-bodied *Hyalella* species (3, 5 and 6) possess spatial patterns more consistent with re-colonization of eastern glaciated regions from populations originating in Appalachian or Atlantic refugia. Freshwater fish species that dispersed out of Atlantic refugia (April *et al.*, 2013*)* have exploited the pro-glacial lake system

that connected Lake Hall (now the Finger Lakes) through the Lake Ontario basin and into the Lake Champlain basin. The amphipod *Diporeia hoyi* (Usjak, 2010) has a similar split in regional groups within the Great Lakes; however both groups were attributed to western refugia (Mississippian and Missourian). As a glacial relict that is restricted to deep cold lake waters, this invertebrate disperses through connected waterways only.

Though the exact ecological relationship between *Hyalella* morphotypes is not fully understood, body-size is associated with a number of traits including foraging aggressiveness, metabolic rate, dispersal ability and vulnerability to predation (Blanckenhorn, 2000; Wellborn & Broughton, 2008). In order for larger species to achieve bigger sizes at maturity, they must consume resources at a greater rate in comparison to smaller species. In *Hyalella azteca*, this is a potential disadvantage. The larger species are well adapted to compete with their smaller counterparts, except in the presence of fish predation. The biogeographic patterns of *Hyalella* south of the glacial limit are consistent with the distribution of centrarchids, a family of fish predators that preferentially consume large-bodied morphotypes (Wellborn, 1994a; Wellborn & Cothran, 2007). The natural range of centrarchids is east of the continental divide, except for the Sacremento perch (Warren, 2009). In glaciated regions, these fish species have successfully colonized the Great Lakes region. *Lepomis macrochirus* dispersed out of the Mississippi River Basin (Kawamura *et al.,* 2009), while other species originated from the east or were introduced anthropogenically (Alofs *et al.*, 2014). Centrarchids that have dispersed into the Great Lakes region are associated with either the decrease in body size or extinction of prey cyprinid species (Alofs & Jackson, 2015).

The Great Lakes region is a suture zone of sympatric *Hyalella* species. If lineages were derived from a single refuge, it would be expected that the diversity within those populations

would be greater than that represented by groups within expansion regions. On the other hand, a scenario involving multiple refugia would create suture zones of high diversity where the expanding fronts of the species (or lineages) meet. With respect to *Hyalella 'azteca'*, the Great Lakes region represents a suture zone of at least six distinct species: three large-bodied (species 1, 2 and 7) and three small-bodied morphotypes (species 3, 5 and 6). This pattern is consistent with dispersal patterns in fish species with the admixture of lineages from different refugia (April *et al.*, 2013). The expansive spatial distributions possessed by large-bodied *Hyalella* species are the likely result of earlier access to proglacial lake systems. The movement of these species into the Great Lakes region originated from previously glaciated regions in western Canada and potentially occurred prior to the colonization by their small-bodied counterparts. When the large-bodied *Hyalella* species colonized the Great Lakes region, they would have competitively excluded the small-bodied morphotypes (Wellborn & Broughton, 2008). This suggests that the migrations of small-bodied *Hyalella* species likely coincided with the range expansions of centrarchids into the region (Alofs & Jackson, 2015). As these fish species migrate throughout the Great Lakes, they would provide opportunities for small-bodied *Hyalella* to disperse into habitats with established populations. The reverse scenario is unlikely, because habitats with centrarchid populations would resist colonization attempts by large-bodied *Hyalella*. The current Great Lakes suture zone is a snapshot of a dynamic process and provides a testable hypothesis for future research. Large-bodied *Hyalella* colonized the Great Lakes region prior to the small-bodied species and their populations are currently being replaced when in the presence of centrarchids. The first expectation is that the timing of the range expansion into the Great Lakes regions by large-bodied species should predate their small-bodied counterparts. While estimates can be derived from mismatch distributions (tau), the rate heterogeneity within *Hyalella* COI

makes such determinations uncertain (Witt *et al.*, 2008). Future work will depend on isolating new markers for phylogeographic analyses, which would allow the use of multilocus coalescent approaches to date the timings of range expansions (Hurt *et al.*, 2009). The second expectation is that the presence of centrarchids would result in a decline in large-bodied *Hyalella* species population density as well as extinction and replacement by small-bodied lineages. Centrarchid predation within Great Lakes region has been associated with reduction in body size and extinction of prey cyprinid species (Alofs & Jackson, 2015).

The region south of the Cordilleran ice sheet underwent dramatic changes during the Pleistocene as a consequence of repeated flooding within eastern Washington State (Hanson *et al.*, 2011). These fluvial events were the result of repeated damming of the Clark River by the Purcell Trench lobe (Cordilleran Ice Sheet). As such, the following areas have been suggested as potential freshwater refugia: the Upper Columbia River (Houston *et al.*, 2014), Lower Columbia River (McCusker *et al.*, 2000) and the Chehalis River Valley (Redenbach & Taylor, 2002). With respect to the *Hyalella* species considered in this study, species 1 lineage H, species 7 and 9 have potential refugial populations within this region; a large and small-bodied morphotype, respectively. Species 2 clade 5-3 was also detected within a single population in the Chehalis River Valley, but this group's haplotypes were derived. The species 1 lineage H individuals collected from within glaciated regions were connected to groups (clade 4-8) located in the Upper Columbia River. On the other hand, species 1 lineage H clade 4-9 was distributed throughout the Lower Columbia River and the Lake Bonneville Basin, a partition that was supported by the SAMOVA. The species 7 clade 4-3 network indicated that the interior haplotypes occurred in populations within the Upper Columbia River. The remaining 4-step clades possess spatial patterns more consistent with a Missouri refugium. More sampling within

the central United States is required to properly assess the source of species 7 clades. The spatial patterns of species 9 mimic those of species 1 lineage H, with interior haplotypes detected in populations within the Upper Columbia River system. Several derived haplotypes occurred in populations within the Chehalis River Valley (clade 4-3) or along the Pacific Coast (clade 4-1). The biogeographic patterns of *Hyalella azteca* are consistent with those detected in the minnow *Richardsonius balteatus* (Houston, *et al.,* 2014), with glaciated regions being colonized by clades from the Upper Columbia River, while groups within the Lower Columbia River and Lake Bonneville Basin remained south of the glacial limit.

The small-bodied species 4, 5 and 6 (clades 5-2 and 5-3) along with the large-bodied species 8 consist of clades that are spatially partitioned into eastern and western groups along the southeastern Coastal Plain bifurcated by the Appalachian Mountains. Soltis *et al.,* (2006) described three major discontinuities in the southeastern United States (Appalachian, Apalachicola River and Mississippi River), although the partitioning of diversity is very often unclear. The spatial patterns in species 8 indicate distinct groups within the Mississippi River Basin and another combining the Apalachicola and Appalachian discontinuities; a trend that is consistent across multiple taxa; including the gopher tortoise (Ennen *et al.*, 2012), Gulf Coast box turtle (Butler *et al.*, 2011), green anole lizard (Campbell-Staton *et al.*, 2012) and southern leopard frog (Newman & Rissler, 2011). Spatial patterns within *Hyalella* species 5 and 6 patterns support an Appalachian and Mississippi River partition, but the populations are not completely sorted. The phylogenetic split between species 5 clades on either side of the mountain range is considerable enough to warrant further investigation (~8.5% COI divergence between lineage A and B) as it is comparable to the distance between species 1 and 2. Species 5 lineage B and D occur in the Appalachian discontinuity, while lineage C and E reside within the Apalachicola

River and Mississippi River discontinuities, respectively. Species 5 lineage A is a complex group that appears to have historically occurred within the Appalachian (clades 3-2 and 3-3) and the Apalachicola River discontinuities (clade 4-3), with a range expansion into the Great Lakes and northern Mississippi River Basin from an Atlantic refuge. The distribution of species 5 lineage B possesses a clear spatial structure, but some hybridization among ITS1 sequences indicate potential introgression among populations in which species 5 lineage A co-occurs. Species 6 clades 5-2 and 5-3 are distributed within the Mississippi River and Appalachian discontinuities, respectively. Species 6 clade 5-1 is similar to species 5 lineage A, both having a range expansion into the Great Lakes and Mississippi River; a pattern also detected in species 3. In terrestrial animals, the discontinuities suggest the existence of multiple Gulf Coast refugia, with the current distributions resulting from northwardly migrations or admixture of lineages (Soltis *et al.,* 2006). However, *Hyalella* groups within these discontinuities are denoted in the haplotype networks with longer branch lengths between haplotypes and no clear patterns associated with demographic expansions (i.e. large interior haplotypes surrounded by numerous unique, but rare tip haplotypes). These patterns are further supported by ragged, multimodal mismatch distributions and the population genetic summary statistics. The segments of species 5 and 6 that were undergoing range expansion appear to be restricted to either Atlantic or northern Appalachian groups that have successfully recolonized glaciated regions. While range expansions may be prevalent in terrestrial groups, the same pattern does not seem to be apparent in *Hyalella*, with these groups representing long standing communities with little evidence of expansion.

The *Hyalella 'azteca'* species within the Great Basin possess remarkable COI divergences among habitats with well established ages (Witt *et al.,* 2008). The degree of

endemism within this region is in contrast to the much more broadly distributed *Hyalella* species characterized in this study. The rate heterogeneity within the COI gene seems to be common within *Hyalella* '*azteca*' (chapter 2) the magnitude of divergence rates seem to be far greater within the Great Basin species. Witt *et al.*, (2008) suggest that the higher temperature within these habitats, and subsequent higher metabolic rates, could result in an increase in mutation rate. Fascinatingly, the high level of *Hyalella* endemism within the Great Basin was not common throughout North America.

The first internal transcribed spacer (ITS1) within *Hyalella azteca* is a complex molecule with extreme molecular structural variation, but with spatial patterns sufficient to suggest its phylogeographic utility. The degree of sequence and structural variation within species was highly variable across *Hyalella* taxa. Species 1 and 2 shared identical ITS1 sequences and maintained a single ITS1 structural variant. Taylor *et al.*, (2005) suggested that the nearly identical ITS1 sequences among invertebrate species could be a consequence of introgression followed by fixation. The ordering and number of structural domains within species 1, 2, 3 and 7 were most readily identifiable within these groups, allowing for the alignment of homologous regions supported by an analysis of covarying mutations. Species 5 and 6 represented the other end of the spectrum, with multiple ITS1 structural variants being maintained in conjunction with a high degree variation, even at the intra-individual level. These distinct rRNA molecules possessed numerous copies and reorderings of structural domains, potentially indicating substantial recombination between distinct ITS1 lineages, possibly maintained at separate genomic nucleolus operating regions (NOR). The variability of the 5' end was a distinct characteristic of the ITS1 molecule, with the structural domains at the 3' end remaining more conserved and easier to align homologous regions. The mapping of ITS1 haplotype networks

according to their spatial distributions indicate that local groups tend to be more related to each other than distant ones. This pattern is consistent across taxa, suggesting some degree of isolation by distance that is not always consistent with biogeographic patterns identified in their mitochondrial cytochrome c oxidase I (COI) counterparts. Incongruence among the spatial patterns of COI and ITS1 markers may not be uncommon in invertebrates (Nekola *et al.,* 2015).

While the species classifications are usually supported by ITS1, in both sequence and structural evolution, some of the delimitations of *Hyalella* species are blurred. The structural analysis of species 5 ITS1 variants suggests the possibility of hybrids that are consistent with the introgression of species 5 COI lineage A haplotypes into lineage B populations. It would seem that despite high levels of COI divergence among species 5 lineages (~12%) and between species 1 and 2 (~8%), the ITS1 gene is far more susceptible to introgression and homogenization. While populations within the coastal plain of the southeastern United States and those in the Mississippi River possessed all the species 5 structural variants, the southern regions tended to be more diverse. Populations north of the glacial limit were far less diverse than their southern counterparts, but this is a likely consequence of the greater variability within species 5 and 6 or the consequence of successive founder effects. The potential for incomplete lineage sorting among groups cannot be discounted. Nuclear genes, specifically ITS1, evolve at a slower rate than their mitochondrial counterparts as a consequence of higher effective population sizes and multiple loci. This genetic pattern is evident in the species 6 and 9 (sister taxa), separated by 12% COI divergences. The distribution of species 6 extends into the Great Lakes and Mississippi River basin, and these populations possess several ITS1 structural variants that are the product of repeated duplication of the 5’ domains. However, the populations within the Appalachian discontinuity possess an ITS1 variant with sequence and structure more akin to species 9. The

grouping of some species 6 ITS1 haplotypes with species 9 represents a paraphyletic relationship that might be indicative of incomplete lineage sorting between species.

What is evident in *Hyalella* is that the structural complexities of ITS1 are the result of recombination and duplication of homologous regions. Species 5 and 6 possessed multiple duplications of the 5' domain and several other domains were reordered throughout the ITS1 gene. The possibility of multiple structural variants maintained at the intra-individual level is fascinating, as is the ability of varied structures to all meet the demands required of the ITS1 molecule. The analysis of the *Hyalella* genome is still in its infancy (Vergilino *et al.*, 2012); however, it remains a crucial step in the understanding ITS1 variation. Many of the ITS1 structural variants (within species 5 and 6) possess broad geographical distributions, but this is no guarantee that they are not unexpressed pseudogenes. It is potentially possible that these small-bodied species maintain multiple NORs in distinct segments of the genome that exhibit little intra-genomic recombination. Additionally, a better understanding of the genome may provide access to more phylogentically informative markers. The incongruence among lineages and species of *Hyalella* needs to be placed in the context of multiple molecular markers.

Hyalella 'azteca' represents a broadly distributed complex of morphologically cryptic taxa. This has allowed the identification of several biogeographic patterns that are well supported in the literature, including the Great Lakes suture zone, the Appalachian/Apalachicola River/Mississippi River discontinuities along with broad ecological patterns as indicated by the relationship between the distributions of morphotypes. The future of *Hyalella 'azteca'* research should focus on resolving some of the questions raised by this study. While the refugial sources of small-bodied species distributions were clearly identified, the origins of their large-bodied counterparts require further investigation; specifically, the watersheds of the Missouri River need

to be fully explored. There is a pressing need to increase the number of molecular markers used in these forms of analysis. While this approach may be prohibitive at larger scales, it is necessary to confirm phylogeographical conclusions with the patterns of multiple molecular markers. The effect of the Pleistocene glaciations on the distribution of *Hyalella 'azteca'* cannot be understated. Vicariance, short distance and long distance dispersal are all extremely important processes that have helped shaped nearly all of the lineages. The degree of clade and spatial diversity within the *Hyalella 'azteca'* species complex demonstrates the need to promote the discipline of biogeography, in order to characterize hidden diversity before it is lost to extinction.

REFERENCES

Adamowicz, S. J., P. D. N. Hebert, and M. C. Marinone. (2004). Species diversity and endemism in the *Daphnia* of Argentina: a genetic investigation. *Zoological Journal of the Linnean Society, 140*, 171–205.

Adamowicz, S. J., A. Pertrisek, J. Colbourne, P. D. N. Hebert, and J. D. S. Witt. (2009). The scale of divergence: A phylogenetic appraisal of intercontinental allopatric speciation in a passively dispersed freshwater zooplankton genus. *Molecular Phylogenetics and Evolution, 50*(3), 423-436.

Agapow, P., O. R. P. Bininda-Emonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall and A. Purvis. (2004). The Impact of Species Concept on Biodiversity Studies. *The Quarterly Review of Biology, 79*(2), 161-179.

Ahrens, D., S. Fabrizi, P. Šípek, and P. Lago. (2013). Integrative analysis of DNA phylogeography and morphology of the European rose chafer (Cetonia aurata) to infer species taxonomy and patterns of postglacial colonisation in Europe. *Molecular Phylogenetics and Evolution, 69*(1), 83-94.

Aldenhoven, J. T., M. A. Miller, P. S. Corneli, and M. D. Shapiro. (2010). Phylogeography of ninespine sticklebacks (*Pungitius pungitius*) in North America: glacial refugia and the origins of adaptive traits. *Molecular Ecology, 19*(18), 4061-4076.

Alofs, K. M., D. A. Jackson, and N. P. Lester. (2014). Ontario freshwater fishes demonstrate differing range-boundary shifts in a warming climate. *Diversity and Distributions, 20*(2), 123-136.

Alofs, K. M., and D. A. Jackson. (2015). The vulnerability of species to range expansions by predators can be predicted using historical species associations and body size. *Proceedings of the Royal Society B: Biological Sciences, 282*(1812), 1-9.

Alonso-Garcia, M., F. J. Sierro, M. Kucera, J. A. Flores, I. Cacho, and N. Anderson. (2011). Ocean circulation, ice sheet growth and interhemispheric coupling of millennial climate variability during the mid-Pleistocene (ca 800e400 ka). *Quat. Sci. Reviews, 30*, 3234-3247.

Alsos, I. G., P. B. Eidesen, D. Ehrich, I. Skrede, K. Westergaard, G. H. Jacobsen, J. Y. Landvik, P. Taberlet, and C. Brochmann. (2007). Frequent Long-Distance Plant Colonization in the Changing Arctic. *Science, 316*(5831), 1606-1609.

Anteau, M. J., A. D. Afton, A. C. E. Anteau, and E. B. Moser. (2011). Fish and land use influence *Gammarus lacustris* and *Hyalella azteca* (Amphipoda) densities in large wetlands across the upper Midwest. *Hydrobiologia, 664*(1), 69-80.

Appeltans, W., S. T. Ahyong, G. Anderson, M. V. Angel, T. Artois, N. Bailly, R. Bamber, A. Barber, I. Bartsch, A. Berta, M. Błazewicz-Paszkowycz, P. Bock, G. Boxshall, C. B. Boyko, S. N. Brandao, R. A. Bray, N. L. Bruce, S. D. Cairns, T. Chan, L. Cheng, A. G. Collins, T. Cribb, M. Curini-Galletti, F. Dahdouh-Guebas, P. J. F. Davie, M. N. Dawson, O. De Clerck, W. Decock, S. De Grave, N. J. de Voogd, D. P. Domning, C. C. Emig, C. Erseus, W. Eschmeyer, K. Fauchald, D. G. Fautin, S. W. Feist, C. H. J. M. Fransen, H. Furuya, O. Garcia-Alvarez, S. Gerken, D. Gibson, A. Gittenberger, S. Gofas, L. Gomez-Daglio, D. P. Gordon, M. D. Guiry, F. Hernendez, B. W. Hoeksema, R. R. Hopcroft, D. Jaume, P. Kirk, N. Koedam, S. Koenemann, J. B. Kolb, R. M. Kristensen, A. Kroh, G. Lambert, D. B. Lazarus, R. Lemaitre, M. Longshaw, J. Lowry, E. Macpherson, L. P. Madin, C. Mah, G. Mapstone, P. A. McLaughlin, J. Mees, K. Meland, C. G. Messing, C. E. Mills, T. N. Molodtsova, R. Mooi, B. Neuhaus, P. K. L. Ng, C. Nielsen, J. Norenburg, D. M. Opresko, M. Osawa, G. Paulay, W. Perrin, J. F. Pilger, G. C. B. Poore, P. Pugh, G. B. Read, J. D. Reimer, M. Rius, R. M. Rocha, J. I. Saiz-Salinas, V. Scarabino, B. Schierwater, A. Schmidt-Rhasea, K. E. Schnabel, M. Schotte, P. Schuchert, E. Schwabe, H. Segers, C. Self-Sullivan, N. Shenkar, V. Siegel, W. Sterrer, S. Stöhr, B. Swalla, M. L. Tasker, E. V. Theusen, T. Timm, M. A. Todaro, X. Turon, S. Tyler, P. Uetz, J. van der Land, B. Vanhoorne, L. P. van Ofwegen, R. W. M. van Soest, J. Vananverbeke, G. Walker-Smith, T. C. Walter, A. Warren, G. C. Williams, S. P. Wilson, and M. J. Costello. (2012). The Magnitude of Global Marine Species Diversity. *Current Biology, 22*, 2189-2202.

April, J., R. H. Hanner, A. Dion-Cote, and L. Bernatchez. (2013). Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. *Molecular Ecology, 22*(2), 409-422.

Armbruster, G. F. J. (2001). Temperature-based variation of rRNA secondary structure models: a case study in the insect Drosophila simulans, the land snail Isabellaria adriani, and the crustacean Daphnia pulex. *Canadian Journal of Zoology*, *79*, 334–345.

Averbeck, K. T., and T. H. Eickbush. (2005). Monitoring the Mode and Tempo of Concerted Evolution in the Drosophila melanogaster rDNA Locus. *Genetics, 171*(4), 1837–1846.

Avise, J. C., W. S. Nelson, and H. Sugita. (1994). A Speciational History of "Living Fossils": Molecular Evolutionary Patterns in Horseshoe Crabs. *Evolution, 48*(6), 1986-2001.

Avise, J. C. (2000). *Phylogeography: the history and formation of species*. Cambridge, MA: Harvard university press.

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography, 36*(1), 3-15.

Avise. J. C., C. G.-D., J. Laerm, J. C. Patton, and R. A. Lansman. (1979). Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. *Proceedings of the National Academy of Sciences, 76*(12), 6694-6698.

Bagley, J. C., M. Sandel, J. Travis, M. de Lourdes Lozano-Vilano, and J. B. Johnson. (2013). Paleoclimatic modeling and phylogeography of least killifish, *Heterandria formosa*: insights into Pleistocene expansion-contraction dynamics and evolutionary history of North American Coastal Plain freshwater biota. *BMC Evolutionary Biology, 13*(223).

Baldinger, A. J., W. D. Shepard, and D. L. Threloff. (2000). Two new species of Hyalella (Crustacea : Amphipoda : Hyalellidae) from Death Valley National Park, California, USA. *Proceedings of the Biological Society of Washington, 113*(2), 443-457.

Baldinger, A. J. (2004). A new species of Hyalella (Crustacea : Amphipoda : Hyalellidae) from Ash Springs, Lincoln County, Nevada, USA, with a key to the species of the genus in North America and the Caribbean region. *Journal of Natural History, 38*(9), 1087-1096.

Ballard, J. W., and M. C. Whitlock. (2004). The incomplete natural history of mitochondria. *Molecular Ecology, 13*(4), 729-744.

Barluenga, M., K. N. Stölting, W. Salzburger, M. Muschick, and A. Meyer. (2006). Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature, 439*(7077), 719-723.

Barrowclough, G. F., and R. M. Zink. (2009). Funds enough, and time: mtDNA, nuDNA and the discovery of divergence. *Molecular Ecology, 18*(14), 2934-2936.

Bate, C. S. (1862). *Catalogue of the Specimens of Amphiopodous Crustacea in the Collection of the British Museum*. London: British Museum, Natural History.

Bateson, P. (2002). William Bateson: a biologist ahead of his time. *Journal of Genetics, 81*(2), 49-58.

Beaumont, M. A., R. Nielsen, C. Robert, J. Hey, O. Gaggiotti, L. Knowles, A. Estoup, M. Panchal, J. Corander, M. Hickerson, S. A. Sisson, N. Fagundes, L. Chikhi, P. Beerli, R. Vitalis, J. Cornuet, J. Huelsenbeck, M. Foll, Z. Yang, F. Rousset, D. Balding, and L. Excoffier. In defence of model-based inference in Phylogeography (2010). *Molecular Biology 19*, 436-446.

Berendzen, P. B., J. F. Dugan, and T. Gamble. (2010). Post-glacial expansion into the Paleozoic Plateau: evidence of an Ozarkian refugium for the Ozark minnow *Notropis nubilus* (Teleostei: Cypriniformes). *Journal of fish biology, 77*, 1114-1136.

Bermingham, E., and J. C. Avise. (1986). Molecular Zoogeography of Fresh-water Fishes in the Southeastern United-States. *Genetics, 113*(4), 939-965.

Bernatchez, L., and C. C. Wilson. (1998). Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular Ecology, 7*(4), 431–452.

Bernatchez, L. (2001). The evolutionary history of brown trout (*Salmo trutta L.*) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation. *Evolution, 55*(2), 351-379.

Bialozyt, R., B. Ziegenhagen, and R. J. Petit. (2006). Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. *Journal of Evolutionary Biology, 19*(1), 12-20.

Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution, 22*(3), 148-155.

Bilton, D. T., J. R. Freeland, and B. Okamura. (2001). Dispersal in Freshwater Invertebrates. *Annual Review of Ecology, Evolution, and Systematics, 32,* 159–181.

Bird, C. E., D. J. Skillings, I. Fernandez-Silva, and R. J. Toonen. (2012). Sympatric Speciation in the Post "Modern Synthesis" Era of Evolutionary Biology. *Evolutionary Biology, 39*(2), 158-180.

Blanckenhorn, W. U. (2000). The Evolution of Body Size: What Keeps Organisms Small? *Quaterly Review of Biology, 75*(4), 385-407.

Bloomquist, E. W., P. Lemey, and M. A. Suchard. (2010). Three roads diverged? Routes to phylogeographic inference. *Trends in Ecology & Evolution, 25*(11), 626-632.

Bohonak, A. J. (2002). IBD (Isolation by Distance): A Program for Analyses of Isolation by Distance. *Journal of Heredity, 93*(2), 153-154.

Borden, W. C., and R. A. Krebs. (2009). Phylogeography and postglacial dispersal of smallmouth bass (*Micropterus dolomieu*) into the Great Lakes. *Canadian Journal of Fisheries and Aquatic Sciences, 66*(12), 2142-2156.

Bossu, C. M., J. M. Beaulieu, P. A. Ceas, and T. J. Near. (2013). Explicit tests of palaeodrainage connections of southeastern North America and the historical biogeography of Orangethroat Darters (Percidae: *Etheostoma: Ceasia*). *Molecular Ecology, 22*(21), 5397-5417.

Bousfield, E. L. (1996). A contribution to the reclassification of Neotropical freshwater hyalellid amphipods (Crustacea: Gammaridea, Talitroidea). *Bollettino Del Museo Civico Di Storia Naturale Di Verona, 20,* 175–224.

Bradford, T., M. Adams, W. F. Humphreys, A. D. Austin, and S. J. B. Cooper. (2010). DNA barcoding of stygofauna uncovers cryptic amphipod diversity in a calcrete aquifer in Western Australia's arid zone. *Molecular Ecology Resources, 10*(1), 41-50.

Bromilow, S. M., and F. A. H. Sperling. (2011). Phylogeographic signal variation in mitochondrial DNA among geographically isolated grassland butterflies. *Journal of Biogeography, 38*(2), 299-310.

Brown, D. D., P. C. Wensink, and E. Jordan. (1972). A Comparison of the Ribosomal DNA's of Xenopus Zaevis and Xenopus mulleri: the Evolution of Tandem Genes. *Journal of Molecular Biology, 63*(1), 57-73.

Brown, W. M., M. George, jr., and A. C. Wilson. (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences, 76*(4), 1967-1971.

Buhay, J. E. (2009). "COI-LIKE" Sequences are Becoming Problematic in Molecular Systematic and DNA BArcoding Studies. *Journal of Crustacean Biology, 29*(1), 96-110.

Bulycheva, A. I. (1957). Morskie bloximorej SSSR i sopredel 'nyx vod (Amphipoda-Talitroidea). *Zoological Institute Academy of Sciences USSR, 65*, 1-188.

Burton, O. J., and J. M. J Travis. (2008). The frequency of fitness peak shifts is increased at expanding range margins due to mutation surfing. *Genetics, 179*(2), 941-950.

Butler, J. M., C. K. Dodd, jr., M. Aresco, and J. D. Austin. (2011). Morphological and molecular evidence indicates that the Gulf Coast box turtle (*Terrapene carolina major*) is not a distinct evolutionary lineage in the Florida Panhandle. *Biological Journal of the Linnean Society, 102*(4), 889-901.

Campbell, M. A., and J. A. López. (2014). Mitochondrial phylogeography of a Beringian relict: the endemic freshwater genus of blackfish Dallia (Esociformes). *Journal of fish biology, 84*(2), 523-538.

Campbell-Staton, S. C., R. M. Goodman, N. Backström, S. V Edwards, J. B Losos, and J. J. Kolbe. (2012). Out of Florida: mtDNA reveals patterns of migration and Pleistocene range expansion of the Green Anole lizard (*Anolis carolinensis*). *Ecology and Evolution, 2*(9), 2274-2284.

Cassens, I., P. Mardulyn, and M. C. Milinkovitch. (2005). Evaluating Intraspecific "Network" Construction Methods Using Simulated Sequence Data: Do Existing Algorithms Outperform the Global Maximum Parsimony Approach? *Systematic Biology, 54*(3), 363-372.

Cate, J. H., M. M. Yusupov, G. Zh. Yusupova, T. N. Earnest, and H. F. Noller. (1999). X-ray Crystal Structures of 70S Ribosome Functional Complexes. *Science, 285*(5436), 2095-2104.

Charlesworth, B. (2012). The Role of Background Selection in Shaping Patterns of Molecular Evolution and Variation: Evidence from Variability on the Drosophila X Chromosome. *Genetics, 191*(1), 233-246.

Chung, Y., and C. Ané. (2011). Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. *Systematic Biology, 60*(3), 261-275.

Clark, P. U., A. S. Dyke, J. D. Shakun, A. E. Carlson, J. Clark, B. Wohlfarth, J. X. Mitrovica, S. W. Hostetler, and A. M. McCabe. (2009). The Last Glacial Maximum. *Science, 325*, 710-714.

Clement, M., D. Posada, and K. A. Crandall. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology, 9*(10), 1657-1659.

Cole, G. A., and R. L. Watkins. (1977). *Hyalella-montezuma,* a new species (Crustacea-Amphipoda) from Montezuma Well, Arizona. *Hydrobiologia, 52*(2-3), 175-184.

Coleman, A. (2013). Analysis of Mammalian rDNA Internal Transcribed Spacers. *PLoS One, 8*(11), 1-10.

Cooper, W. E. (1965). Dynamics and production of a natural population of a fresh-water amphipod, *Hyalella azteca*. *Ecological Monographs, 35*, 377-394.

Costello, M. J., S. Wilson, and B Houlding. (2012). Predicting Total Global Species Richness Using Rates of Species Description and Estimates of Taxonomic Effort. *Systematic Biology, 61*(5), 871-883.

Costello, M. J., R. M. May, and N. E. Stork. (2013). Can We Name Earth's Species Before They Go Extinct? *Science, 339*(6118), 413-416.

Cothran, R. D., and P. D. Jeyasingh. (2010). Condition dependence of a sexually selected trait in a crustacean species complex: importance of the ecological context. *Evolution, 64*(9), 2535-2546.

Cothran, R. D., K. A. Henderson, D. Schmidenberg, and R. A. Relyea. (2013). Phenotypically similar but ecologically distinct: differences in competitive ability and predation risk among amphipods. *Oikos, 122*, 1429-1440.

Cox, A. J., and P. D. N. Hebert. (2001). Colonization, extinction, and phylogeographic patterning in a freshwater crustacean. *Molecular Ecology, 10*(2), 371-386.

Coyne, J. A., and H. A. Orr. (2004). *Speciation*. Sunderland, Massachusetts: Sinauer Associates.

Coyne, J. A. (2011). Speciation in a small space. *Proceedings of the National Academy of Sciences, 108*(32), 12975-12976.

Crease, T. J., and J. K. Colbourne. (1998). The Unusually Long Small-Subunit Ribosomal RNA of the Crustacean *Daphnia pulex:* Sequence and Predicted Secondary Structure. *Journal of Molecular Evolution 46*, 307-313.

Crease, T. J., A. R. Omilian, K. S. Costanzo, and D. J. Taylor. (2012). Transcontinental phylogeography of the Daphnia pulex species complex. *PLoS One, 7*(10).

Cristescua, M. E. A., and P. D. N. Hebert. (2005). The" Crustacean Seas" an evolutionary perspective on the Ponto Caspian peracarids. *Canadian Journal of Fisheries and Aquatic Sciences, 62*(3), 505-517.

Cunha, A. F., G. N. Genzano, A. C. Marques. (2015). Reassessment of Morphological Diagnostic Characters and Species Boundaries Requires Taxonomical Changes for the Genus Orthopyxis L. Agassiz, 1862 (Campanulariidae, Hydrozoa) and Some Related Campanulariids. *PLoS One, 10*(2).

Currat, M., M. Ruedi, R. Pétit, and L. Excoffier. (2008). The hidden side of invasions: Massive introgression by local genes. *Evolution, 62*(8), 1908-1920.

De Meester, L., A. Gomez, B. Okamura, and K. Schwenk. (2002). The Monopolization Hypothesis and the dispersal–gene flow paradox in aquatic organisms. *Acta Oecologica, 23*, 121-135.

de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences, 102*(Supplement 1), 6600-6607.

de Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology, 56*(6), 879–886.

De Schepper, S., E. I. Fischer, J. Groeneveld, M. J. Head, and J. Matthiessen. (2011). Deciphering the palaeoecology of Late Pliocene and Early Pleistocene dinoflagellate cysts. *Palaeogeography, Palaeoclimatology, Palaeoecology, 309*, 17-32.

Degnan, J. H., and L. A. Salter. (2005). Gene tree distributions under the coalescent process. *Evolution, 59*(1), 24–37.

Dennenmoser, S., S. M. Rogers, and S. M. Vamosi. (2014). Genetic population structure in prickly sculpin (*Cottus asper*) reflects isolation-by-environment between two life-history ecotypes. *Biological Journal of the Linnean Society, 113*(4), 943-957.

Diniz-Filho, J. A., T. N. Soares, J. S. Lima, R. Dobrovolski, V. L. Landeiro, M. P. de Campos Telles, T. F. Rangel, and L. M. Bini. (2013). Mantel test in population genetics. *Genetics and Molecular Biology, 36*(4), 475-485.

Dooh, R. T., S. J. Adamowicz, and P. D. N. Hebert. (2006). Comparative phylogeography of two North American ‘glacial relict’ crustaceans. *Molecular Ecology, 15*(14), 4459-4475.

Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. *Nature, 299*(5879), 111-117.

Dupanloup, I., S. Schneider, and L. Excoffier. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology, 11*(12), 2571-2581.

Dyke, A. S., and Prest, V. K. (1987). Late Wisconsinan and Holocene History of the Laurentide Ice Sheet. *Géographie physique et Quaternaire, 41*(2), 237-263.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza. (2004). Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences, 101*(4), 975-979.

Edwards, S., and Bensch, S. (2009). Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Molecular Ecology, 18*(14), 2930-2933.

Eickbush, T. H. and D. G. Eickbush (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics, 175*(2), 477-485.

Elderkin, C. L., A. D. Christian, J. L. Metcalfe-Smith, and D. J. Berg. (2008). Population genetics and phylogeography of freshwater mussels in North America, *Elliptio dilatata* and *Actinonaias ligamentina* (Bivalvia: Unionidae). *Molecular Ecology, 17*(9), 2149-2163.

Engels, S., and L. C. Cwynar. (2011). Changes in fossil chironomid remains along a depth gradient: evidence for common faunal thresholds within lakes. *Hydrobiologia, 665*(1), 15-38.

Ennen, J. R., B. R. Kreiser, C. P. Qualls, D. Gallard, M. Aresco. (2012). Mitochondrial DNA Assessment of the Phylogeography of the Gopher Tortoise. *Journal of Fish and Wildlife Management, 3*(1), 110-122.

Eo, S. H., and A. J. DeWoody. (2010). Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. *Proceedings of the Royal Society B: Biological Sciences, 277*(1700), 3587-3592.

Evans, K. M., A. H. Wortley, and D. G. Mann. (2007). An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). *Protist, 158*(3), 349-364.

Excoffier, L., P. E. Smouse, and J. M. Quattro. (1992). Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data *Genetics, 131*, 479-491.

Excoffier, L. (2004). Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology, 13*(4), 853-864.

Excoffier, L., G. Laval, and S. Schneider. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online, 1*, 47–50.

Excoffier, L., and N. Ray. (2008). Excoffier L, Ray N.. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution, 23*(7), 347-351.

Excoffier, L., M. Foll, and R. J. Petit. (2009). Genetic Consequences of Range Expansions. *Annual Review of Ecology Evolution and Systematics, 40*, 481-501.

Excoffier, L., and H. E. L. Lischer. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources, 10*(3), 564-567.

Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences, 104*(45), 17614-17619.

Figuerola, J., and A. J. Green. (2002). Dispersal of aquatic organisms by waterbirds: a review of past research and priorities for future studies. *Freshwater Biology, 47*(3), 483-494.

Folmer, O., M. Black, R. Hoeh, R. A. Lutz, and R. Vrijenhoek. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology, 3*(5), 294-299.

Foote, C. J., J. W. Clayton, C. C. Lindsey, and R. A. Bodaly. (1992). Evolution of Lake Whitefish (*Coregonus clupeaformis*) in North America during the Pleistocene: Evidence for a Nahanni Glacial Refuge Race in the Northern Cordillera Region. *Canadian Journal of Fisheries and Aquatic Sciences, 49*(4), 760-768.

Forró, L., N. M. Korovchinsky, and A. A. Kotov. (2008). Global diversity of cladocerans (Cladocera; Crustacea) in freshwater. *Hydrobiologia, 595*, 177-184.

Frankham, R., J. D. Ballou,, M. R. Dudash, M. D.B. Eldridge, C. B. Fenster, R. C. Lacy, J. R. Mendelson III, I. J. Porton, K. Ralls, and O. A. Ryder. (2012). Implications of different species concepts for conserving biodiversity. *Biological Conservation, 153*, 25-31.

Frogley, M. R., H. I. Griffiths, and T. H. E. Heaton. (2001). Historical biogeography and Late Quaternary environmental change of Lake Pamvotis, Ioannina (north‐western Greece): evidence from ostracods. *Journal of Biogeography, 28*(6), 745-756.

Fu, Y. (1997). Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics, 147*(2), 915-925.

Fujita, Y., H. Furuta, and Y. Ikawa. (2009). Tailoring RNA modular units on a common scaffold: A modular ribozyme with a catalytic unit for β-nicotinamide mononucleotide-activated RNA ligation. *RNA- A Publication Of The RNA Society, 15*(5), 877-888.

Galtier, N., B. Nabholz, S. Glémin, and G. D. D. Hurst. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology, 18*(22), 4541–4550.

Gandolfi, A., P. Bonilauri, V. Rossi, and P. Menozzi. (2001). Intraindividual and intraspecies variability of ITS1 sequences in the ancient asexual Darwinula stevensoni (Crustacea: Ostracoda). *Heredity, 87,* 449-455.

Ganley, A. R. D. a. T. K. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research, 17*(2), 184-191.

Geisler, S. F. S. (1944). Studies on the Postembryonic Development of *Hyalella azteca* (Saussure). *The Biological Bulletin, 86*(1), 6-22.

Gießler, S., and C. C. Englbrecht. (2009). Dynamic reticulate evolution in a Daphnia multispecies complex. *Journal of Experimental Zoology, 311A*(7), 530-548.

Gillespie, R. (2004). Community Assembly Through Adaptive Radiation in Hawaiian Spiders. *Science, 303*(5656), 356-359.

Gonzalez, E. R., and L. Watling. (2002). Redescription of Hyalella Azteca from Its Type Locality, Vera Cruz, Mexico (Amphipoda : Hyalellidae). *Journal of Crustacean Biology, 22*(1), 173-183.

Green, A. J., J. Figuerola, and M. I. Sanchez. (2002). Implications of waterbird ecology for the dispersal of aquatic organisms. *Acta Oecologica, 23*(3), 177-189.

Gutell, R. R., J. C. Lee, and J. J. Cannone. (2002). The Accuracy of Ribosomal RNA Comparative Structure Models. *Current Opinion in Structural Biology, 12*(3), 301-310.

Hallatschek, O., and D. R. Nelson. (2010). Life at the front of an expanding population. *Evolution, 64*(1), 193-206.

Hamilton, A. J., V. Novotný, E. K. Waters, Y. Basset, K. K. Benke, P. S. Grimbacher, S. E. Miller, G. A. Samuelson, G. D. Weiblen, J. D. L. Yen, and N. E. Stork. (2013). Estimating global arthropod species richness: refining probabilistic models using probability bounds analysis. *Oecologia, 171*(2), 357-365.

Hampe, A., and R. J. Petit. (2005). Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters, 8*(5), 461-467.

Hanson, M. A., O. B. Lian, and J. J. Clague. (2011). The sequence and timing of large late Pleistocene floods from glacial Lake Missoula. *Quaternary Science Reviews, 31*, 67-81.

Hardin, G. (1960). The Competitive Exclusion Principle. *Science, 131*(3409), 1292-1297.

Hare, M. (2001). Prospects for nuclear gene phylogeography. *Trends in Ecology & Evolution, 16*(12), 700-706.

Harnik, P. G., H. K. Lotze, S. C. Anderson, Z. V. Finkel, S. Finnegan, D. R. Lindberg, L. H. Liow, R. Lockwood, C. R. McClain, J. L. McGuire, A. O'Dea, J. M. Pandolfi, C. Simpson, and D. P. Tittensor. (2012). Extinctions in ancient and modern seas. *Trends in Ecology & Evolution, 27*(11), 608-617.

Harpending, H., and A. Rogers. (2000). Genetic Perspectives on Human Origins and Defferentiation. *Annual Review of Genomics and Human Genetics, 1*, 361-385.

Harpending, H. C., S. T. Sherry, A. R. Rogers, and M. Stoneking. (1993). The Genetic Structure of Ancient Human Populations. *Current Anthropology, 34*(4), 483-496.

Harris, D. J., and K. A. Crandall. (2000). Intragenomic Variation Within ITS1 and ITS2 of Freshwater Crayfishes (Decapoda: Cambaridae): Implications for Phylogenetic and Microsatellite Studies. *Molecular Biology and Evolution, 17*(2), 284-291.

Harris, L. N., and E. B. Taylor. (2010). Genetic population structure of broad whitefish, Coregonus nasus, from the Mackenzie River, Northwest Territories: implications for subsistence fishery management. *Canadian Journal of Fisheries and Aquatic Sciences, 67*(6), 905-918.

Hart, M. (2011). The species concept as an emergent property of population biology. *Evolution, 65*(3), 613-616.

Hasegawa, M., H. Kishino, and T. Yano. (1985). Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA *Journal of Molecular Evolution, 22*(2), 160-174.

Hausdorf, B. (2011). Progress toward a general species concept. *Evolution, 65*(4), 923-931.

Hays, J. D., J. Imbrie, and N. J. Shackleton. (1976). Variations in the Earth's orbit: pacemaker of the ice ages. *Science, 194*(4270), 1121-1132.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences, 270*(1512), 313-321.

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

Heim, N. A., M. L. Knope, E. K. Schaal, S. C. Wang, and J. L. Payne. (2015). Cope's rule in the evolution of marine animals. *Science, 347*(6224), 867-870.

Hellberg, M. E. (1994). Relationships Between Inferred Levels of Gene Flow and Geographic Distance in a Philopatric Coral *Balanophyllia elegans*. *Evolution, 48*(6), 1829-1854.

Henras, A. K., J. Soudet, M. Gérus, S. Lebaron, M. Caizergues-Ferrer, A. Mougin and Y. Henry. (2008). The post-transcriptional steps of eukaryotic ribosome biogenesis. *Cellular and Molecular Life Sciences, 65*(15), 2334-2359.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature, 405*(6789), 907-913.

Hewitt, G. (2001). Speciation, hybrid zones and phylogeography — or seeing genes in space and time. *Molecular Ecology, 10*(3), 537-549.

Hickerson, M. J., B. C. Carsten, J. Cavender-Bares, K. A. Crandall, C .H. Graham, J. B. Johnson, L. Rissler, P. F. Victoriano, and A. D. Yoder. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution, 54*(1), 291-301.

Hillis, D. M., M. T. Dixon, and L. K. Ammerman. (1991). The relationships of the coelacanth Latimeria chalumnae: evidence from sequences of vertebrate 28S ribosomal RNA genes. *Environmental Biology of Fishes, 32*(1-4), 119-130.

Hollingsworth, P. M., S. W. Graham, and D. P. Little. (2011). Choosing and Using a Plant DNA Barcode. *PLoS One, 6*(5), 1-13.

Hou, Z., Fu, J., and Li, S. (2007). A molecular phylogeny of the genus *Gammarus* (Crustacea: Amphipoda) based on mitochondrial and nuclear gene sequences. *Molecular Phylogenetics and Evolution, 45*(2).

Houston, D. D., D. K. Shiozawa, B. T. Smith, and B. R Riddle. (2014). Investigating the effects of Pleistocene events on genetic divergence within Richardsonius balteatus, a widely distributed western North American minnow. *BMC Evolutionary Biology, 14*(111), 1-18.

Huang, H., Q. He, L. S. Kubatko, and L. L. Knowles. (2010). Sources of Error Inherent in Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing among Different Methods. *Systematic Biology, 59*(5), 573-583.

Huelsenbeck, J. P., and J. J. Bull. (1996). A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology, 45*(1), 92-98.

Hurt, C., A. Anker, and N. Knowlton (2009). A multilocus test of simultaneous divergence across the Isthmus of Panama using snapping shrimp in the genus *Alpheus*. *Evolution,* 63(2), 514-530.

Imbrie, J., A. Berger, E. A. Boyle, S. C. Clemens, A. Duffy, W. R. Howard, G. Kukla, J. Kutzbach, D. G. Martinson, A. McIntyre, A. C. Mix, B. Molfino, J. J. Morley, L. C. Peterson, N. G. Pisias, W. L. Prell, M. E. Raymo, N. J. Shackleton, and J. R. Toggweiler. (1993). On the structure and origin of major glaciation cycles 2. The 100,000-year cycle. *Paleoceanography, 8*(6), 691–875.

Inoue, K., E. M. Monroe, C. L. Elderkin, and D. J. Berg. (2014). Phylogeographic and population genetic analyses reveal Pleistocene isolation followed by high gene flow in a wide ranging, but endangered, freshwater mussel. *Heredity, 112*, 282-290.

Ishida, S., and D. J. Taylor. (2007). Quaternary diversification in a sexual Holarctic zooplankter, Daphnia galeata. *Molecular Ecology, 16*(3), 569-582.

Jackson, S. T., J. T. Overpeck, T. Webb III, S. E. Keattch, and K. H. Anderson. (1997). Mapped plant-macrofossil and pollen records of late quaternary vegetation change in Eastern North America. *Quaternary Science Reviews, 16*(1), 1-70.

Joyce, D. A., D. H. Lunt, M. J. Genner, G. F. Turner, R. Bills, and O. Seehausen. (2011). Repeated colonization and hybridization in Lake Malawi cichlids. *Current Biology, 21*(3).

Kawamura, K., R. Yonekura, O. Katano, Y. Taniguchi, and K. Saitoh. (2009). Phylogeography of the Bluegill Sunfish, *Lepomis macrochirus*, in the Mississippi River Basin. *Zoological Science, 26*, 24-34.

Keppel, G., K. P. Van Niel, G. W. Wardell-Johnason, C. J. Yates, M. Byrne, L. Mucina, A. G. T. Schut, S. D. Hopper, and S. E. Franklin. (2012). Refugia: identifying and understanding safe havens for biodiversity under climate change. *Global Ecology and Biogeography, 21*, 393-404.

Kimura, M., and J. F. Crow. (1963). The measurement of effecive population number. *Evolution*, 279-288.

Klassen, R. W. (1994). Late Wisconsinan and Holocene history of southwestern Saskatchewan. *Canadian Journal of Earth Sciences, 31*(12), 1822-1837.

Klopfstein, S., M. Currat, and L. Excoffier. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution, 23*(3), 482–490.

Knowles, L. L. (2009). Statistical Phylogeography. *Annual Review of Ecology, Evolution, and Systematics, 40*, 593–612.

Knowles, L. L., and P. B. Klimov. (2011). Estimating phylogenetic relationships despite discordant gene trees across loci: the species tree of a diverse species group of feather mites (Acari: Proctophyllodidae). *Parasitology, 138*(13), 1750-1759.

Knowles, L. L., H. C. Lanier, P. B. Klimov, and Q. He. (2012). Full modeling versus summarizing gene-tree uncertainty: Method choice and species-tree accuracy. *Molecular Phylogenetics and Evolution, 65*(2), 501-509.

Knowlton, N., L. A. Weight, L. A. Solorzano, D. K. Mills, and E. Bermingham. (1993). Mitochondrial DNA, and Reproductive Compatibility Across the Isthmus of Panama. *Science, 260*(5114), 1629-1632.

Knowlton, N., and L. A. Weigt. (1998). New dates and new rates for divergence across the Isthmus of Panama. *Proceedings of the Royal Society B: Biological Sciences, 265*(1412), 2257-2263.

Kobayashi, T. (2011). Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cellular and Molecular Life Sciences, 68*(8), 1395-1403.

Kornobis, E., and S. Pálsson. (2011). Discordance in Variation of the ITS Region and the Mitochondrial COI Gene in the Subterranean Amphipod Crangonyx islandicus. *Journal of Molecular Evolution, 73*(1-2), 34-44.

Kubatko, L. S., and J. H. Degnan. (2007). Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology, 56*(1), 17–24.

Kuo, C., and J. C. Avise. (2005). Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees. *Genetica, 124*(2-3), 179–186.

Latch, E. K., E. M. Kierepka, J. R. Heffelfinger, and O. E. Rhodes, Jr. (2011). Hybrid swarm between divergent lineages of mule deer (*Odocoileus hemionus*). *Molecular Ecology, 20*, 5265-5279.

Leese, F., and C. Held. (2011). *Analysing intraspecific genetic variation: A practical guide using mitochondrial DNA and microsatellites* (Vol. 19). Boca Raton, FL: CRC Press.

Lefebure, T., C. J. Douady, M. Gouy, and J. Gibert. (2006). Relationship between morphological taxonomy and molecular divergence within Crustacea: Proposal of a molecular threshold to help species delimitation. *Molecular Phylogenetics and Evolution, 40*(2), 435-447.

Librado, P. and Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25**:** 1451-1452.

Lieberman., B. S. (2012). Adaptive Radiations in the Context of Macroevolutionary Theory: A Paleontological Perspective. *Evolutionary Biology, 39*(2), 181-191.

MacArthur, R. H., and E. O. Wilson. (1967). *The theory of island biogeography*. Princeton, New Jersey: Princeton University Press.

MacDonald III, K. S., Yampolsky, L., and J. E. Duffy. (2005). Molecular and morphological evolution of the amphipod radiation of Lake Baikal. *Molecular Phylogenetics and Evolution, 35,* 323-343.

Major, K., D. Soucek, R. Giordano, M. Wetzel, and F. Soto-Adames (2013). The common ecotoxicology laboratory strain of *Hyalella azteca* is genetically distinct from most wild strains sampled in eastern North America. *Environmental Toxicology and Chemistry, 32*(11), 2637-2647.

Mallet, J. (1995). A species definition for the Modern Synthesis. *Trends in Ecology & Evolution, 10*(7), 294-299.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution, 20*(5).

Mallet, J. (2007). *Species, Concepts Of. Encyclopedia of Biodiversity*

Mallet, J. (2010). Why was Darwin's view of species rejected by twentieth century biologists? *Biology & Philosophy, 25*(4), 497-527.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research, 27*(2), 209-220.

Martinson, D. G., N. G. Pisias, J. D. Hays, J. Imbrie, T. C. Moore, jr., and N. J. Shackleton. (1987). Age Dating and the Orbital Theory of the Ice Ages: Development of a High-Resolution on 0 to 300,000-Year Chronostratigraphy. *Quaternary Research, 27,* 1-29.

Mayr, E. (1963). *Animal Species and Evolution*. Cambridge: Belknap Press.

Mayr, E. (1992). A Local Flora and the Biological Species Concept. *American Journal of Botany, 79*(2), 222-238.

Mayr, E. (2001). The Philosophical Foundations of Darwinism. *Proceedings of the American Philosophical Society, 145*(4), 488-495.

McAllister, D. E., S. P. Platania, F. W. Schueler, M. E. Baldwin, and D. S. Lee. (1986). *Ichthyofaunal patterns on a geographic grid*. New York: John Wiley and Sons.

McCusker, M. R., E. Parkinson, and E. B. Taylor. (2000). Mitochondrial DNA variation in rainbow trout (Oncorhynchus mykiss) across its native range: testing biogeographical hypotheses and their relevance to conservation. *Molecular Ecology, 9*(12), 2089-2108.

Meirmans, P. G., J. Goudet, Intrabio Div Consortium, and O. E. Gaggiotti. (2011). Ecology and life history affect different aspects of the population structure of 27 high-alpine plants. *Molecular Ecology, 20*(15), 3144–3155.

Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology, 21*(12), 2839-2846.

Meng, C., and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology, 75*(1), 35-45.

Mickelson, D. M., and P. M. Colgan. (2003). The southern Laurentide Ice Sheet. *Development in Quaternary Science, 1*, 1-16.

Millette, K. L., S. Xu, J. D. S. Witt, and M. E. Cristescu. (2011). Pleistocene-driven diversification in freshwater zooplankton: Genetic patterns of refugial isolation and postglacial recolonization in *Leptodora kindtii* (Crustacea, Cladocera). *Limnology and Oceanography, 56*(5), 1725-1736.

Moore, A. C., J. B. Burch, and T. F. Duda Jr. (2015). Recognition of a highly restricted freshwater snail lineage (Physidae: Physella) in southeastern Oregon: convergent evolution, historical context, and conservation considerations. *Conservation Genetics, 16*(1), 113-123.

Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. (2011). How Many Species Are There on Earth and in the Ocean? *PloS Biology, 9*(8), 1-8.

Muller, J. (2000). Mitochondrial DNA Variation and the Evolutionary History of Cryptic Gammarus fossarum Types. *Molecular Phylogenetics and Evolution, 15*(2), 260-268.

Murdoch M. H., a. P. D. N. H. (1997). Mitochondrial DNA evidence of distinct glacial refugia for brown bullhead (Ameiurus nebulosus) in the Great Lakes. *Canadian Journal of Fisheries and Aquatic Sciences, 54*(7), 1450-1460.

Murphy, N. P., M. Adams, M. T. Guzik, and A. D. Austin. (2013). Extraordinary micro-endemism in Australian desert spring amphipods. *Molecular Phylogenetics and Evolution, 66*(3), 645-653.

Murton, J. B., M. D. Bateman, S. R. Dallimore, J. T. Teller, and Z. Yang. (2010). Identification of Younger Dryas outburst flood path from Lake Agassiz to the Arctic Ocean. *Nature, 464*(7289), 740-743.

Nassonova, E., A. Smirnov, J. Fahrni, and J. Pawlowski. (2010). Barcoding Amoebae: Comparison of SSU, ITS and COI Genes as Tools for Molecular Identification of Naked Lobose Amoebae. *Protist, 161*, 102-115.

Nei, M., and A. P. Rooney. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics, 39*, 121-152.

Nekola, J. C., B. F. Coles, and M. Horsák. (2015). Species assignment in *Pupilla* (Gastropoda: Pulmonata: Pupillidae): integration of DNA-sequence data and conchology. *Journal of Molluscan Studies, 81,* 196–216.

Newman, C. E., and L. J. Rissler. (2011). Phylogeographic analyses of the southern leopard frog: the impact of geography and climate on the distribution of genetic lineages vs. subspecies. *Molecular Ecology, 20*(24), 5295-5312.

Nylander, J. A. A., J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics, 24*(4), 581-583.

Osada, N., and H. Akashi. (2012). Mitochondrial–nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome c oxidase complex. *Molecular Biology and Evolution, 29*(1), 337-346.

Oswood, M. W., J. B. Reynolds, J. G. Irons III, and A. M. Milner. (2000). Distributions of freshwater fishes in ecoregions and hydroregions of Alaska. *Journal of the North American Benthological Society, 19*(3), 405-418.

Palumbi, S. R., F. Cipriano, and M. P. Hare. (2001). Predicting nuclear gene coalescence from mitochondrial data: The three-times rule. *Evolution, 55*(5), 859–868.

Panchal, M., and M. A. Beaumont. (2007). The automation and evaluation of nested clade phylogeographic analysis. *Evolution, 61*(6), 1466-1480.

Papadopulos A. S. T., W. J. B., D. Crayn, R. K. Butlin, R. G. Kynast, I. Hutton, and V. Savolainen. (2011). Speciation with gene flow on Lord Howe Island. *Proceedings of the National Academy of Sciences, 108*(32), 13188-13193.

Peck, S. B. (1975). Amphipod dispersal in the fur of aquatic mammals. *The Canadian Field-Naturalist, 89*(2), 181-182.

Penton, E. H., P. D. N. Hebert, and T. J. Crease. (2004). Mitochondrial DNA variation in North American populations of *Daphnia obtusa*: continentalism or cryptic endemism? *Molecular Ecology, 13*(1), 97-107.

Petit, R., I. Aguinagalde, J. de Beauliew, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, S. Fineshi, D. Grivet, M. Lascoux, A. Mohanty, G. Müller-Starck, B. Demesure-Musch, A. Palmé, J. P. Martin, S. Rendell, and G. G. Vendramin. (2003). Glacial Refugia: Hotspots But Not Melting Pots of Genetic Diversity. *Science, 300*(5625), 1563-1565.

Petit, R. J. (2008). The coup de grâce for the nested clade phylogeographic analysis? *Molecular Ecology, 17*(2), 516-518.

Petit, R. J., and L. Excoffier. (2009). Gene flow and species delimitation. *Trends in Ecology & Evolution, 24*(7), 386-393.

Pfenninger, M., and K. Schwenk. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical *BMC Evolutionary Biology, 7*(121), 1-6.

Pielou, E. C. (1991). *After the Ice Age: the return of life to glaciated North America*. Chicago, IL: The University of Chicago Press.

Pierce, A. A., M. P. Zalucki, M. Bangura, M. Udawatta, M. R. Kronforst, S. Altizer, J. Ferna´ndez Haeger, and J. C. de Roode. (2014). Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies. *Proceedings of the Royal Society B: Biological Sciences, 281*, 1979.

Posada, D., and K. A. Crandall. (1998). MODELTEST: testing the model of DNA substition. *Bioinformatics, 14*(9), 817-818.

Posada, D., K. A. Crandall, and A. R. Templeton. (2000). GeoDis: A program for the Cladistic Nested Analysis of the Geographical Distribution of Genetic Haplotypes. *Molecular Ecology, 9*(4), 487-488.

Posada, D., and T. R. Buckley. (2004). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology, 53*(3), 793-808.

Posada, D. (2008). jModeltest: phylogenetic model averaging. *Molecular Biology and Evolution, 25*(7), 1253-1256.

Pothoven S. A., D. W. H., and T. F. Nalepa. (2011). Declines in deepwater sculpin *Myoxocephalus thompsonii* energy density associated with the disappearance of *Diporeia* spp. in lakes Huron and Michigan. *Ecology of Freshwater Fish, 20*, 14-22.

Pruett, C. L., S. M. Tanksley, M. F. Small, J. F. Taylor, and M. R. J. Forstner. (2011). The Effects of Range Expansion on the Population Genetics of White-Winged Doves in Texas. *The American Midland Naturalist, 166*(2), 415-425.

Puissant, S., and J. Sueur. (2010). A hotspot for Mediterranean cicadas (Insecta: Hemiptera: Cicadidae): New genera, species and songs from southern Spain. *Systematics and Biodiversity, 8*(4), 555-574.

Rabosky, D. L. (2009). Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. *Ecology Letters, 12*(8), 735–743.

Radulovici, A. E., B. Sainte-Marie, and F. Dufresne. (2009). DNA barcoding of marine crustaceans from the Estuary and Gulf of St Lawrence: a regional-scale approach. *Molecular Ecology Resources, 9*(Supplement 1), 181-187.

Rambaut, A., and A. Drummond. (2009). FigTree v1 3.1.

Ramirez-Soriano, A., S. E. Ramos-Onsins, J. Rozas, F. Calafell, and A. Navarro. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics, 179*(1), 555-567.

Ramos-Onsins, S. E., and J. Rozas. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution, 19*(12), 2092-2100.

Ratnasingham, S., and P. D. N. Hebert. (2007). BOLD: The Barcode of Life Data System. *Molecular Ecology Notes, 7*(3), 355-364.

Ray, N., and L. Excoffier. (2010). A first step towards inferring levels of long-distance dispersal during past expansions. *Molecular Ecology Resources, 10*(5), 902-914.

Raymo, M. E. (1997). The Timing of Major Climate Terminations. *Paleoceanography, 12*(4), 577-585.

Redenbach, Z., and E. B. Taylore. (2002). Evidence for historical introgression along a contact zone between two species of char (Pisces: salmonidae) in northwestern North America. *Evolution, 56*(5), 1021-1035.

Remigio, E. A., P. D. N. Hebert. (2000). Affinities among Anostracan (Crustacea: Branchiopoda) Families Inferred from Phylogenetic Analyses of Multiple Gene Sequences. *Molecular Phylogenetics and Evolution, 17*(1), 117-128.

Rempel, L. L., and D. G. Smith. (1998). Postglacial fish dispersal from the Mississippi refuge to the Mackenzie River basin. *Canadian Journal of Fisheries and Aquatic Sciences, 55*(4), 893-899.

Rissler, L. J., and W. H. Smith. (2010). Mapping amphibian contact zones and phylogeographical break hotspots across the United States. *Molecular Ecology, 19*, 5404-5416.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature, 425*(6960), 798-804.

Ronquist, F., and J. P. Huelsenbeck. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics, 19*(12), 1572-1574.

Rosine, W. N. (1956). On the transport of the common amphipod, *Hyalella azteca.* South Dakota by the mallard duck. *Proceedings of the South Dakota Academy of Science, 35*, 203.

Rosine, W. N. (1962). New evidence of the transport of freshwater amphipoda by waterfowl. *Proceedings of the South Dakota Academy of Science, 41*, 198.

Rowe, C. L., S. J. Adamowicz, and P. D. N. Hebert. (2007). Three new cryptic species of the freshwater zooplankton genus Holopedium (Crustacea : Branchiopoda : Ctenopoda), revealed by genetic methods. *Zootaxa*, 1-49.

Ruddiman, W. F. (2006). Orbital changes and climate. *Quaternary Science Reviews, 25*(23-24).

Ruiz-Estévez, M., F. J. Ruiz-Ruano, J. Cabrero, M. Bakkali, F. Perfectti, M. D. López-León and J. P. M. Camacho. (2015). Non-random expression of ribosomal DNA units in a grasshopper showing high intragenomic variation for the ITS2 region. *Insect Molecular Biology, 24*(3), 319–330.

Saussure, H. (1858). Memoire sur divers crustaces nouveaux des Antilles et du Mexique. *Memoires de la Société de physique et d'histoire naturelle de Genève, 14*, 417–496.

Schoch, C. L., K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, and Fungal Barcoding Consortiuma. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences, 109*(16), 6241–6246.

Schwentner, M., S. Clavier, M. Fritsch, J. Olesen, S. Padhye, B. V. Timms, and S. Richter. (2013). Cyclestheria hislopi (Crustacea: Branchiopoda): A group of morphologically cryptic species with origins in the Cretaceous. *Molecular Phylogenetics and Evolution, 66*(3), 800-810.

Seidel, R. A., B. K. Lang, and D. J. Berg. (2009). Phylogeographic analysis reveals multiple cryptic species of amphipods (Crustacea: Amphipoda) in Chihuahuan Desert springs. *Biological Conservation, 142*(10).

Serejo, C. S. (2004). Cladistic revision of talitroidean amphipods (Crustacea, Gammaridea), with a proposal of a new classification. *Zoologica Scripta, 33*(6), 551-586.

Shackleton, N. J., Berger, A., and Peltier, W. R. (1990). An alternative astronomical calibration of the lower Pleistocene timescale based on ODP Site 677. *Transactions of the Royal Societ of Edinburgh - Earth Sciences, 81*, 251-261.

Shafer, A. B. A., C. I. Cullingham, S. D. Côté, and D. W. Coltman. (2010). Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology, 19*(21), 4589–4621.

Sims, P. A., D. G. Mann, and L. K. Medlin. (2006). Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia, 45*(4), 361-402.

Slatkin, M., and L. Excoffier. (2012). Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics, 191*(1), 171-181.

Sloan, K. E., S. Mattijssen, S. Lebaron, D. Tollervey, G. J. M. Pruijn, and N. J. Watkins. (2013). Both endonucleolytic and exonucleolytic cleavage mediate ITS1 removal during human ribosomal RNA processing. *Journal of Cell Biology, 200*(5), 577-588.

Slowinski, J. B., and R. D. M. Page. (1999). How Should Species Phylogenies Be Inferred from Sequence Data? *Systematic Biology, 48*(4), 814– 825.

Smith, S. (1875) Report on Amphipod Crustaceans. Annual Report of the United States Geological and Geographical Survey of the territories, Embracing Colorado. The Exploration for the year 1873, by F.V.Hayden, Washington, Goverment Printing Office, 1874: 608-611.

Smith, D. G. (2001). *Pennak's freshwater invertebrates of the United States*. New York: John Wiley & Sons.

Smith, K. L., L. J. Harmon, L. P. Shoo, and J. Melville. (2011). Evidence of constrained phenotypic evolution in a cryptic species complex of agamid lizards. *Evolution, 65*(4), 976-992.

Sokal, R. R., and T. J. Crovello. (1970). The biological species concept: a critical evaluation. *American Naturalist, 104*(936), 127-153.

Soltis, D. E., A. B. Morris, J. S. McLachlan, P. S. Manos, and P. S. Soltis. (2006). Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology, 15*(14), 4261-4293.

Soltis, P. S., and D. E. Soltis. (2009). The Role of Hybridization in Plant Speciation. *Annual Review of Plant Biology, 60*, 561-588.

Soucek, D. J., E. A. Lazo-Wasem, C. A. Taylor, and K. M. Major (2015). Description of two new species of *Hyalella* (Amphipoda: Hyalellidae) from eastern North America with a revised key to North American members of the genus. *Journal of crustacean biology, 35*(6), 814-829.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics, 22*(21), 2688-2690.

Stamford, M. D., and E. B. Taylor. (2004). Phylogeographical lineages of Arctic grayling (*Thymallus arcticus*) in North America: divergence, origins and affinities with Eurasian *Thymallus Molecular Ecology, 13*(6), 1533–1549.

Stentiford, G. D., S. W. Feist, D. M. Stone, E. J. Peeler, and D. Bass. (2014). Policy, phylogeny, and the parasite. *Trends in Parasitology, 30*(6), 274-281.

Stepien, C. A., D. J. Murphy, R. N. Lohner, O. J. Sepulveda-Villet, and A. E. Haponski. (2009). Signatures of vicariance, postglacial dispersal, and spawning philopatry: Population genetics and biogeography of the walleye *Sander vitreus*. *Molecular Ecology, 18*(16), 3411-3428.

Stevenson, M. M., and Peden, A. E. (1973). Description and Ecology of *Hyalella texana* n. sp. (Crustacea: Amphipoda) from the Edwards Plateau of Texas. *The American Midland Naturalist, 89*(2), 426-436.

Stewart, J. R., and A. M. Lister. (2001). Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology & Evolution, 16*(11), 608-613.

Stuvier, M., P. J. Reimer, E. Bard, J. W. Beck, G. S. Burr, K. A. Hughen, B. Kromer, G. McCormac, J. van der Plicht, and M. Spurk. (1998). INTCAL98 Radiocarbon Age Calibration, 24,000-0 cal BP. *Radiocarbon, 40*(3), 1041-1083.

Swanson, G. A. (1984). Dissemination of Amphipods of Waterfowl. *Journal of Wildlife Management, 48*(3), 988-991.

Swofford, D. L. (2003). {PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.}.

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics, 123*(3), 585-595.

Tamura, K., and M. Nei. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution, 10*(3), 512-526.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. (2011). MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution, 28*(10), 2731-2739.

Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution, 30*(12), 2725-2729.

Taylor, D. J., H. L. Sprenger and S. Ishida. (2005). Geographic and phylogenetic evidence for dispersed nuclear introgression in a daphniid with sexual propagules. *Molecular Ecology, 14*(2), 525–537.

Templeton, A. R. (1998). Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology, 7*(4), 381-397.

Templeton, A. R. (2004). Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology, 13*, 789-809.

Templeton, A. R. (2005). Haplotype Trees and Modern Human Origins. *American journal of physical anthropology, Suppl 41*, 33-59.

Templeton, A. R. (2008). Nested Clade Analysis: an extensively validated method for strong phylogeographic inference. *Molecular Ecology, 17*(8), 1877-1880.

Templeton, A. R. (2009a). Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology, 18*(2), 319-331.

Templeton, A. R. (2009b). Why does a method that fails continue to be used: the answer. *Evolution, 63*, 807-812.

Templeton, A. R., K. A. Crandall., and C. F. Sing. (1992). A Cladistic Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction Endonuclease Mapping and DNA Sequence Data. III. Cladogram Estimation. *Genetics, 132*(2), 619-633.

Templeton, A. R. a. C. F. Sing. (1993). A Cladistic-analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping. 4. Nested Analyses with Cladogram Uncertainty and Recombination. *Genetics, 134*(2), 659-669.

Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. F. de Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. van Jaarsveld, G. F. Midgley, L. Miles, M. A. Ortega-Huerta, A. T. Peterson, O. L. Phillips, and S. E. Williams. (2004). Extinction risk from climate change. *Nature, 427*(6970), 145-148.

Travis, J. M. J., T. Münkemüller, O. J. Burton, A. Best, C. Dytham, and K. Johst. (2007). Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution, 24*(10), 2334-2343.

Trontelj, P., and Cene Fiser. (2009). Cryptic species diversity should not be trivialised. *Systematics and Biodiversity, 7*(1), 1-3.

Turelli, M., N. H. Barton, and J. A.Coyne. (2001). Theory and speciation. *Trends in Ecology & Evolution, 16*(7), 330-343.

Turgeon, J., and L. Bernatches. (2001). Blackwell Science, Ltd Mitochondrial DNA phylogeography of lake cisco (*Coregonus artedi*): evidence supporting extensive secondary contacts between two glacial races. *Molecular Ecology, 10*(4), 987-1001.

Uit de Weerd, D. R., W. H. Piel, and E. Gittenberger. (2004). Widespread polyphyly among Alopiinae snail genera: when phylogeny mirrors biogeography more closely than morphology. *Molecular Phylogenetics and Evolution, 33*(3), 533–548.

Usjak, S. (2010). *Comparative Phylogeography of North American Diporeia hoyi and Gammarus lacustris (order: Amphipoda).* University of Waterloo, Waterloo.

Väinölä, R., J. D. S. Witt, M. Grabowski, J. H. Bradbury, K. Jazdzewski, and B. Sket. (2008). Global diversity of amphipods (Amphipoda; Crustacea) in freshwater. *Hydrobiologia, 595*, 241–255.

Valentini, A., F. Pompanon, and P. Taberlet. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution, 24*(2), 110-117.

Van Houdt, J. K., B. Hellemans, and F. A. M. Volckaert. (2003). Phylogenetic relationships among Palearctic and Nearctic burbot (Lota lota): Pleistocene extinctions and recolonization. *Molecular Phylogenetics and Evolution, 29*(3), 599-612.

Venton, D. (2013). Phylogeography. *Proceedings of the National Academy of Sciences, 110*(11), 4158.

Vergilino, R., K. Dionne, C. Nozais, F. Dufresne, and C. Belzile. . (2012). Genome size differences in *Hyalella* cryptic species. *Genome, 55*(2), 134-139.

Verheyen E., W. S., J. Snoeks, and A. Meyer. (2003). Origin of the Superflock of Cichlid Fishes from Lake Victoria, East Africa. *Science, 300*(5617), 325-329.

Vié, J. C., C. Hilton-Taylor, and S. N. Stuart. (eds). (2009). *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species*. Gland, Switzerland: IUCN.

Vollmer, S. V., and S. R. Palumbi. (2004). Testing the utility of internally transcribed spacer sequences in coral phylogenetics. *Molecular Ecology*, *13*(9), 2763–2772.

Wang, X., C. Liu, L. Huang., J. Bengtsson-Palme, H. Chen, J. Zhang, D. Cai and J. Li. (2015). ITS1: a DNA barcode better than ITS2 in eukaryotes? *Molecular Ecology Resources, 15*, 573–586.

Waples, R., and O. Gaggiotti. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology, 15*(6), 1419-1439.

Warren jr., M. L. (2009). *Centrarchid identification and natural history*. West Sussex, UK: Blackwell Publishing.

Washietl, S., I. L. Hofacker, P. F. Stadler and M. Kellis. (2012). RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Research, 40*(10), 4261-4272.

Webb III, T., and P. J. Bartlein. (1992). Global Changes During the Last 3 Million Years: Climatic Controls and Biotic Responses. *Annual Reviews of Ecology and Systematics, 23*, 141-173.

Weckel, A. D. (1907). The fresh-water amphipoda of North America. *Proceedings of the United States National Museum, 32*, 54-57.

Weider, L. J., J. J. Elser, T. J. Crease, M. Mateos, J. B. Cotner, and T. A. Markow. (2005). The Functional Significance of Ribosomal (r)DNA Variation: Impacts on the Evolutionary Ecology of Organisms. *Annual Review of Ecology, Evolution, and Systematics, 36*, 219-242.

Wellborn, G. A. (1994a). Size-biased predation and the evolution of prey life histories: a comparative study of freshwater amphipod populations. *Ecology, 75*(7), 2104-2117.

Wellborn, G. A. (1994b). The mechanistic basis of body size differences between *Hyalella* (Amphipoda) species. *Journal of Freshwater Ecology, 19*(2), 159-168.

Wellborn, G. A. (2002). Trade-off Between Competitive Ability and Antipredator Adaptation in a Freshwater Amphipod Species Complex. *Ecology, 83*(1), 129-136.

Wellborn, G. A., and S. E. Bartholf. (2005). Ecological context and the importance of body and gnathopod size for pairing success in two amphipod ecomorphs. *Oecologia*, *143*, 308–316.

Wellborn, G. A., R. Cothran, and S. E. Bartholf. (2005). Life history and allozyme diversification in regional ecomorphs of the *Hyalella azteca* (Crustacea: Amphipoda) species complex. *Biological Journal of the Linnean Society, 84*, 161–175.

Wellborn, G. A., and R. D. Cothran. (2007). Niche diversity in crustacean cryptic species: complementarity in spatial distribution and predation risk. *Oecologia, 154*(1), 175-183.

Wellborn, G. A., and R. E. Broughton. (2008). Diversification on an ecologically constrained adaptive landscape. *Molecular Ecology, 17*(12), 2927–2936.

Wellborn, G. A., and E. Capps. (2013). Establishment of a new population by a single gravid female: implications for Hyalella biogeography and speciation. *Evolutionary Ecology, 27*(1), 101-115.

Werren, J. H. (1997). Biology of Wolbachia. *Annual Review of Entomology, 42*, 587-609.

Westhof, E., B. Masquida, and F. Jossinet. (2011). Predicting and Modeling RNA Architecture. *Cold Spring Harbor perspectives in biology, 3*(2), 1-14.

Westram, A. M., J. Jokela, and I. Keller. (2013). Hidden biodiversity in an ecologically important freshwater amphipod: differences in genetic structure between two cryptic species. *PLoS One, 8*(8).

Wislon, C. C., and P. D. N. Hebert. Phylogeography and postglacial dispersal of lake trout (*Salvelinus namaycush*) in North America. *Canadian Journal of Fisheries and Aquatic Sciences, 55(4),* 1010-1024.

Witt, J. D. S., and P. D. N. Hebert. (2000). Cryptic species diversity and evolution in the amphipod genus *Hyalella* within central glaciated North America: a molecular phylogenetic approach. *Canadian Journal of Fisheries and Aquatic Sciences, 57*(4), 687-698.

Witt, J. D. S., D. W. Blinn, and P. D. N. Hebert. (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. *Molecular Ecology, 12*(2), 405-413.

Witt, J. D. S., D. L. Threloff, and P. D. N. Hebert. (2006). DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology, 15*(10), 3073-3082.

Witt, J. D. S., D. L. Threloff, and P. D. N. Hebert. (2008). *Genetic zoogeography of the Hyalella azteca species complex in the Great Basin: rapid rates of molecular diversification in desert springs.* (Vol. 439): Geological Society of America Special Paper.

Witt, J. D. S., E. B.Taylor, R. and Zemlak. (2011). Phylogeography and the origins of range disjunctions in a north temperate fish, the pygmy whitefish (*Prosopium coulterii*), inferred from mitochondrial and nuclear DNA sequence analysis. *Journal of Biogeography, 38*, 1557-1569.

Wolf, M., S. Chen, J. Song, M. Ankenbrand, and T. Müller. (2013). Compensatory Base Changes in ITS2 Secondary Structures Correlate with the Biological Species Concept Despite Intragenomic Variability in ITS2 Sequences – A Proof of Concept. *PLoS One, 8*(6), 1-5.

Yang, Z., and B. Rannala. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences, 107*(20), 9264-9269.

Yeh, W. B. M. J. T., N. Chang, S.Y. Wu and Y.S. Tsai. (2015). Agronomically important thrips: development of species-specific primers in multiplex PCR and microarray assay using internal transcribed spacer 1 (ITS1) sequences for identification. *Bulletin of Entomological Research, 105*(1), 52-59.

Zagoskin, M. V., V. I. Lazareva, A. K. Grishanin, and D. V. Mukha. (2014). Phylogenetic Information Content of Copepoda Ribosomal DNA Repeat Units: ITS1 and ITS2 Impact. *BioMed Research International, 2014*, 1-15.

Zink, R. M., G. F. Barrowclough. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology, 17*(9), 2107-2121.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research, 31*(13), 3406-3415.

Appendix A

Table A.1 List of locations and sample data
This list includes all samples sites in alphabetical order according to province or state abbreviation. Each row contains the location code used in this study, location name, province/state code, latitude, longitude (in decimal degrees), species (lineage) and haplotypes.

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
AB001	Buffalo Lake	AB	52.45	-113.06	7	S7H001, S7H004, S7H033
AB002	Cold Lake	AB	54.47	-110.11	2, 7	S2H092, S2H126, S2H132, S7H004, S7H010, S7H027, S7H047, S7H048
AB003	Crandell Lake	AB	49.09	-113.97	1(F), 7	S1H184, S1H186, S1H187, S7H046
AB004	East Crane Lake	AB	54.50	-110.54	2, 7	S2H056, S7H004, S7H006, S7H039, S7H053
AB005	Edith Lake	AB	52.91	-118.05	7	S7H012, S7H044, S7H045
AB006	Fork Lake	AB	54.45	-111.54	7	SH004
AB007	Grand Prairie 2 pond	AB	55.23	-118.31	1(C), 1(G), 2	S1H139, S1H235, S1H249, S2H132, S2H187
AB008	Gull Lake	AB	52.46	-113.97	7	S7H004, S7H033
AB009	Hanmore Lake	AB	54.30	-112.55	2, 7	S2H102, S2H112, S2H122, S2H189, S7H002, S7H004, S7H095
AB010	HCR	AB	53.82	-112.08	2	S2H132, S2H191
AB011	Highway 21 Pond	AB	52.03	-113.24	7	S7H004, S7H005, S7H0015, S7H092
AB012	Highway 28 Creek	AB	54.10	-112.78	2	S2H126, S2H139, S2H163, S2H187, S2H195
AB013	Highway 855 Bog	AB	54.40	-112.49	2	S2H132, S2H135, S2H186
AB014	Highway 855 Creek	AB	54.13	-112.48	2	S2H132, S2H194, S2H216
AB015	Highway 855 River	AB	54.61	-112.42	2	S2H129, S2140, S2H162, S2H218
AB016	Honeymoon Lake	AB	52.56	-117.68	2	S2H121
AB017	Jack Fish Creek	AB	54.44	-110.69	2	S2132, S2H162, S2H187, S2H192, S2H218
AB018	Lac la Biche	AB	54.77	-111.98	2, 7	S2H081, S2H126, S2H162, S7H004, S7H014, S7H034, S7H050, S7H051
AB019	Lake Newell	AB	50.45	-111.91	2, 7	S2H132, S2H162, S2H165, S2H187, S7H018, S7H019, S7H039
AB020	Lake St. Anne	AB	53.72	-114.34	2	S2H056, S2H118, S2H119, S2H197
AB021	Lesser Slave Lake	AB	55.33	-115.62	7	S7H004, S7H059
AB022	Little Akamina Lake	AB	49.02	-114.04	1(F)	S1H205
AB023	Lower Waterton Lake	AB	49.09	-113.86	7	S7H004
AB024	Marie Creek	AB	54.52	-110.32	2	S2H092, S2H095, S2H124, S2H132
AB025	Mildred Lake	AB	52.89	-118.05	1(F)	S1H188
AB026	North Buck Lake	AB	54.66	-112.51	2	S2H057, S2H080, S2H187

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
AB027	Pyramid Lake	AB	52.92	-118.09	2, 7	S2H126, S7H099
AB028	Sturgeon Lake	AB	55.07	-117.50	2, 7	S2H162, S2H184, S7H039
AB029	Ward Pond	AB	50.57	-111.88	7	S7H004, S7H007, S7H089
AB030	Waterton Lake	AB	49.05	-113.90	1(G), 7	S1H207, S1H208, S7H004, S7H054, S7H059
AK001	31.6 mile Pond	AK	65.14	-147.44	1(F), 1(G)	S1H192, S1H232, S1H234
AK002	Beach Lake	AK	61.40	-149.57	1(I)	S1H345, S1H347, S1H348
AK003	Birch Lake	AK	64.31	-146.65	1(I)	S1H316, S1H317
AK004	Bonnie Lake	AK	61.81	-148.30	1(I)	S1H326, D1H340
AK005	Cheney Lake	AK	61.20	-149.76	1(I)	S1H320, S1H321, S1H328, S1H339
AK006	Cottenwood Lake	AK	61.60	-149.36	1(I)	S1H339, S1H342
AK007	Glen Hwy Pond	AK	62.40	-145.12	1(I)	S1H330, S1H331
AK008	Goose Lake	AK	61.20	-149.82	1(I)	S1H324
AK009	Highway 3 pond	AK	64.84	-148.01	1(F), 1(G)	S1H194, S1H213, S1H234, S1H237
AK010	Jan Lake	AK	63.57	-143.92	2	S2H116
AK011	Jerome Lake	AK	60.55	-149.58	1(I)	S1H329
AK012	Jewel Lake	AK	61.14	-149.97	1(I)	S1H332, S1H333
AK013	Johnson Lake	AK	60.29	-151.27	1(I)	S1H327
AK014	Long Lake	AK	61.80	-148.24	1(I), 5(A)	S1H343, S1H344, S5H008
AK015	Lower Fire Lake	AK	61.35	-149.55	1(I)	S1H339, S1H341, S1H345, S1H346
AK016	Moon lake	AK	63.38	-143.54	7	S7H072
AK017	Nenana CIty Pond	AK	64.55	-149.09	1(F)	S1H193, S1H195
AK018	Potter Marsh	AK	61.06	-149.80	1(C), 1(I)	S1H140, S1H321, S1H325
AK019	Quartz Lake	AK	64.20	-145.83	1(I), 7	S1H317, S1H318, S7H069, S7H072, S7H073
AK020	Ryans Lake	AK	62.03	-146.64	1(I)	S1H322, S1H323, S1H331
AK021	Seward Hwy Pond	AK	60.92	-149.13	1(C), 1(I)	S1H139, S1H140, S1H349
AK022	Watson Lake	AK	60.54	-150.46	1(I)	S1H350, S1H351, S1H352, S1H353, S1H354
AK023	Yarger Lake	AK	62.96	-141.64	7	S7H060

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
AL001	Coffeeville Lake	AL	31.75	-86.53	5(A), 8	S5H008, S5H071, S8H087, S8H089, S8H093
AL002	Concecuh River	AL	31.35	-86.53	4, 5(A)	S4H023, S4H024, S5H072, S5H073, S5H075, S5H076
AL003	Guntersville Lake	AL	34.36	-86.34	3	S3H004, S3H019, S3H023, S3H024, S3H118
AL004	Mulberry Fork	AL	33.81	-87.14	3, 6	S3H020, S3H022, S3H118, S6H122
AR001	Arkansas River	AR	34.02	-91.34	4, 8	S4H016, S4H038, S4H039, S8H087, S8H088, S8H089
AR002	Bear Creek Lake	AR	34.71	-90.69	5(A)	S5H085
AR003	Buck Lake	AR	34.40	-91.12	5(A)	S5H085
AR004	Hwy 65	AR	33.63	-91.39	8	S8H091. S8H094
AR005	Lake Enterprise	AR	33.06	-91.57	5(A)	S5H088, S5H089, S5H090
BC010	Andy Bailey Lake	BC	58.55	-122.51	2, 7	S2H002, S2H011, S2H012, S2H013, S2H014, S7H087, S7H088
BC011	Attlin Warm Spring	BC	59.40	-133.58	7	S7H067, S7H068
BC012	Babine Pond	BC	54.77	-126.18	2	S2H117, S2H120
BC013	Beaton River	BC	57.08	-122.90	2	S2H030
BC014	Boya Lake	BC	59.37	-129.10	2, 7	S2H100, S2H101, S7H016
BC015	Centurion Creek	BC	55.69	-121.61	2	S2H002, S2H016
BC016	Charlie Lake	BC	59.30	-120.99	2, 7	S2H016, S7H025, S7H026
BC017	Chuchi Lake	BC	55.19	-124.51	1(C), 2	S1H139, S1H141, S2H148, S1H350, S2H149, S2H180
BC018	Cluculz Pond	BC	53.91	-123.64	1(G), 2	S1H210, S2H12
BC019	Condo Pond	BC	49.27	-123.16	9	S9H040
BC020	Crooked River	BC	54.79	-122.83	1(G)	S1H227, S1H243
BC021	Dawson's Lake	BC	54.77	-120.26	1(G), 2	S1H233, S2H031, S2H132, S2H137
BC022	Dina Lake	BC	55.53	-123.31	1(G), 7	S1H242, S7H083, S7H084
BC023	Drywilliam Lake	BC	54.06	-124.69	1(F), 1(G), 2	S1H183, S1H224, S2H096
BC024	Elk Lake	BC	48.53	-123.39	1(H), 9	S1H281, S1H284, S1H286, S9H002, S9H033
BC025	Fuller Lake	BC	48.91	-123.72	1(I)	S1H369, S1H370, S1H371, S1H372, S2H138
BC026	Heather Lake	BC	55.50	-123.25	1(D)	S1H146

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
BC027	Hodder Lake	BC	56.73	-129.78	1(G)	S1H223
BC028	Huble Lake	BC	54.27	-122.62	1(D), 1(G), 2	S1H147, S1H148, S1H238, S2H181
BC029	Hwy 97 bog pond	BC	54.49	-122.68	1(G)	S1H243, S1H244
BC030	Kidd Lake	BC	49.92	-120.63	1(G), 1(H), 7	S1H233, S1H264, S7H055
BC031	Labour Day Lake	BC	49.13	-124.46	1(I)	S1H366, S1H367, S1H368
BC032	Lake Okanagan	BC	50.26	-119.44	1(H), 2	S1H255, S1H256, S1H259, S1H261, S1H264, S2H187
BC033	Liard River	BC	59.43	-126.10	7	S7H028, S7H029, S7H030, S7H032
BC034	Lightning Lake	BC	49.06	-120.83	1(H), 7	S1H282, S7H097, S7H098
BC035	Lower Tacheeda	BC	54.72	-122.52	1(F), 1(G)	S1H189, S1H244
BC036	Macleod River	BC	54.91	-122.94	1(D)	S1H142, S1H143
BC037	Maine Lake, QI	BC	50.19	-125.17	1(I), 9	S1H355, S1H356, S1H357, S1H360, S1H361, S9H023
BC038	Mill Creek	BC	58.67	-123.95	2	S2H002, S2H006, S2H033
BC039	Oosoyos Lake	BC	49.04	-119.46	1(H)	S1H262, S1H264, S1H265, S1H266, S1H267, S2186, S9H021, S9H022, S9H024
BC040	Palmer Lake	BC	59.44	-133.59	1(F), 1(G), 2, 6, 7	S1H205, S1H220, S1H221, S2H200, S2H212, S3H104, S6H073, S6H115, S7H065
BC041	Parker Lake	BC	52.41	-121.85	2	S2H032
BC042	Pond by Kitwancool Lake	BC	55.34	-128.09	2	S2H106
BC043	Purden Lake	BC	53.92	-121.91	1(D), 1(G)	S1H149, S1H150, S1H233, S1H246
BC044	Ross Lake	BC	49.01	-121.08	1(C)	S1H138, S1H136, S1H139
BC045	Skagit River	BC	49.12	-121.17	1(H)	S1H283, S7H097
BC046	Smith River	BC	59.76	-126.46	2	S2H097, S2H180
BC047	Spider Lake, QI	BC	49.34	-124.62	1(H)	S1H289, S1H290
BC048	Swan Lake	BC	59.76	-126.46	1(H), 1(I)	S1H257, S1H259, S1H295, S1H373
BC049	Toad River	BC	58.85	-125.23	1(G)	S1H233, S1H236
BC050	Tuqueluit Lake	BC	59.20	-119.52	1(G), 1(H)	S1H239, S1H240, S1H258, S1H268, S1H292
BC051	Upper Tacheeda	BC	54.69	-122.55	1(F), 1(G)	S1H183, S1H191, S1H244, S1H245

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
BC052	Vaseaux Lake	BC	49.30	-119.53	1(H)	S1H264, S1H292, S1H293, S1H294
BC053	Village Bay, QI	BC	50.18	-125.19	1(H)	S1H355, S1H358, S1H359, S1H365
CT001	Pearson Pond	CT	41.61	-72.65	5, 6	S5H058, S6H155, S6H161
FL001	Apalachicola River	FL	30.12	-85.17	5	S5H093, S5H094, S5H095, S5H096
FL002	Juniper's Lake	FL	30.76	-86.13	4	S4H033
FL003	Lake Istokpoga	FL	27.44	-81.29	8	S8H034, S8H035, S8H064, S8H065, S8H066
FL004	Lake Munroe	FL	28.86	-81.25	8	S8H044, S8H047, S8H048, S8H055, S8H071
FL005	Lake Panasoffkee	FL	28.80	-82.15	8	S8H011, S8H016, S8H017, S8H023, S8H070
FL006	Lake Pasadena	FL	28.32	-82.15	8	S8H018, S8H019, S8H025, SH026, S8H027
FL007	Lake Placid	FL	27.25	-81.38	8	S8H063, S8H068
FL008	Lake Tohopekaliga	FL	28.29	-81.41	8	S8H032, S8H033, S8H037, S8H038, S8H040, S8H041, S8H049, S8H050
FL009	Little Black Creek	FL	30.10	-81.80	8	S8H007, S8H039, S8H043, S8H067, S8H096
FL010	Little Orange Lake	FL	29.56	-82.07	8	S8H082, S8H083, S8H084, S8H085, S8H086
FL011	Lochloosa Lake	FL	29.51	-82.10	8	S8H002, S8H003, S8H005, S8H042, S8H056, S8H081
FL012	Mosquito Creek	FL	30.70	-84.83	5	S5H083
FL013	Ocean Pond	FL	30.21	-82.43	8	S8H022, S8H092
FL014	Ochlockonee River	FL	30.47	-84.41	5	S5H083
FL015	Okeechobee	FL	27.20	-80.83	8	S8H020, S8H024, S8H051, S8H051, S8H054, S8H058, S8H061, S8H105
FL016	Oklawaha River	FL	29.21	-81.99	8	S8H001, S8H004
FL017	Palestine Lake	FL	30.10	-82.40	4, 8	S4H011, S4H012, S4H029, S8H006, S8H010
FL018	Peace River	FL	27.75	-81.78	8	S8H015, S8H029, S8H030, S8H031, S8H062
FL019	Penholloway River	FL	30.06	-83.59	8	S8H072, S8H075, S8H076, S8H077, S8H078
FL020	Steinhatchee River	FL	29.77	-83.32	8	S8H012, S8H073, S8H074, S8H079, S8H080
FL021	Suwanee River	FL	29.63	-82.96	8	S8H008, S8H021, S8H057, S8H060
FL022	Wacissa River	FL	30.34	-83.99	8	S8H009, S8H013, S8H014, S8H059
GA001	Flint River	GA	30.79	-84.67	5	S5H006, S5H077, S5H079, S5H080, S5H081, S5H082
GA002	Lake Seminole	GA	30.74	-84.85	5	S5H091, S5H092, S5H097

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
GA003	Lindsey Grace Lake	GA	31.58	-82.05	4	S4H002, S4H003, S4H007, S4H014, S4H015
GA004	Ogeechee River	GA	32.49	-81.56	4, 6	S4H010, S6H150, S6H151, S6H153, S6H154
GA005	Pendleton Creek	GA	32.43	-82.51	4	S4H002, S4H004, S4H006, S4H026, S4H027
GA006	Rays Lake	GA	30.84	-84.85	8	S8H036, S8H045, S8H046
GA007	Satilla River	GA	31.30	-82.45	4	S4H004, S4H005, S4H006, S4H013, S4H027
GA008	Watson Lake	GA	32.55	-82.13	4	S4H002, S4H006, S4H009, S4H028, S4H030
IA001	Briggs Woods Lake	IA	42.43	-93.80	2	S2H021, S2H036, S2H050, S2H066
IA002	Brushy Creek	IA	42.42	-93.99	3	S3H043, S3H044, S3H046
IA003	Calhoun County Wildlife Area	IA	42.52	-94.45	3	S3H040, S3H043, S3H045
IA004	Hendrick's Lake	IA	43.37	-92.53	2	S2H015, S2H036, S2H066, S2H070, S2H071
IA005	Hickory Grove Lake	IA	41.99	-93.36	2, 6	S2H036, S2H046, S2H050, S2H088, S3H045, S6H007, S6H023, S6H131
IA006	Lake Macbride	IA	41.79	-91.55	2, 3	S2H002, S2H069, S3H043, S3H045, S3H064
IA007	Little Sioux River	IA	42.45	-95.80	3	S3H039, S3H045
IA008	Melrose Lake	IA	40.89	-93.03	2	S2H002, S2H072
IA009	Rock Creek Lake	IA	41.76	-92.84	2, 6	S2H062, S3H045, S3H047, S3H059, S6H001, S6H006
IA010	Twin Lakes	IA	42.50	-94.61	2	S2H022, S2H023, S2H024, S2H050
IA011	West Lake	IA	41.52	-90.68	5, 6	S5H003, S6H005, S6H018, S6H023, S6H037
IL001	Altamont Reservoir	IL	39.03	-88.71	6	S6H023, S6H033
IL002	Coffeen Lake	IL	39.09	-89.41	6	S6H024, S6H143, S6H144
IL003	Evergreen Lake	IL	40.64	-89.03	6	S6H009, S6H023
IL004	Horseshow Lake	IL	38.70	-90.15	5	S5H001, S5H002, S5H008
IL005	JC Murphey Lake	IL	40.97	-87.52	1(B), 3, 6	S1H102, S1H103, S3H025, S3H029, S6H005
IL006	Lake Bloomington	IL	40.66	-88.94	3	S3H016
IL007	Lake Nellie	IL	39.05	-88.85	1, 6	S5H001, S6H023
IL008	Lake Sara	IL	39.14	-88.64	6	S6H019, S6H033, S6H034
IL009	Lake Vandalia	IL	39.01	-89.14	5	S5H008, S5H011, S5H012
IL010	Staunton Reservoir	IL	39.04	-89.78	6	S6H004, S6H006, S6H021

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
IL011	Storey Lake	IL	40.99	-90.40	3, 6	S3H017, S3H028, S6H019, S6H025, S6H039
IL012	Illinois Creek	IL	38.23	-90.09	6	S6H120
IN001	Fox Lake	IN	41.62	-85.02	3, 5	S3H141, S3H147, S3H169, S3H185, S5H007
IN002	Geist Reservoir	IN	39.95	-85.89	3	S3H002
IN003	Lake Sullivan	IN	39.10	-87.39	5	S5H005, S5H008
IN004	Point Mill Lake	IN	39.33	-87.37	3	S3H021, S3H025, S3H026
IN005	St. Joseph River	IN	41.22	-85.01	3, 6	S3H002, S3H050, S3H062, S6H008, S6H010, S6H053
IN006	Wabash River	IN	39.06	-87.57	5, 6	S5H008, S6H013
KS001	Farnum Creek	KS	39.15	-96.90	2, 3	S2H039, S2H040, S2H142, S2H144, S2H145, S3H042
KS002	Kill Creek	KS	38.91	-94.97	6	S6H001, S6H002, S6H130
KS003	Logan County Lake	KS	39.63	-99.58	3	S3H027, S3H082
KS004	Ottawa State Lake	KS	39.11	-97.57	8	S8H102, S8H103, S8H104
KS005	Perry Lake	KS	39.12	-95.46	2, 3	S2H036, S2H049, S2H050, S2H060, S2H143, S3H041, S3H081
KS006	Plainsville Township Lake	KS	39.23	-99.33	3	S3H080, S3H081, S3H091, S3H092
KS007	Pottawatomie Lake	KS	39.23	-96.53	3	S3H070, S3H081, S3H116
KS008	Rock Creek	KS	38.88	-95.40	2	S2H060
KS009	Wilson Lake	KS	38.96	-98.49	3	S3H027, S3H081, S3H082
KY001	Green River	KY	37.23	-85.34	3, 6	S3H012, S6H031, S6H035
KY002	Lake Malone	KY	37.08	-87.03	6	S6H144
KY003	North Elkhorn Creek	KY	38.22	-84.60	6	S6H011, S6H012, S6H032
KY004	Stream off of Lake Malone	KY	37.08	-87.03	6	S6H144
LA001	Cheniere Lake	LA	32.45	-92.19	4, 5	S4H017, S4H018, S4H034, S4H035, S4H036, S4H040, S5H086, S5H087
MB001	Chruchill River Site 1	MB	58.68	-94.17	2, 6	S2H174, S2H187, S2H196, S6H074
MB002	Churchill River Site 2	MB	58.63	-94.23	2	S2H173, S2H174, S2H185, S2H187, S2H190, S2H193
MB003	Cutt Lake	MB	52.90	-118.32	2, 7	S2H162, S7H093, S7H097
MB004	Fairford River	MB	51.59	-98.73	1(B)	S1H117, S2H020, S2H087, S2H093, S2H166
MB005	Goose Creek	MB	58.66	-94.17	2, 7	S2H187, S2H188, S7H003, S7H017, S7H035, S7H057
MB006	Hyw 10 Pond	MB	55.14	-98.14	2	S2H155, S2H218

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
MB007	Lake Manitoba	MB	50.22	-98.53	2	S2H133
MB008	Lake Wekusko	MB	54.80	-99.97	2	S2H166, S2H187, S2H223
MB009	Lake William	MB	53.91	-99.32	1(A), 2, 6, 7	S1H078, S1H091, S2H084, S2H162, S6H103, S6H105, S7H082
MB010	Lake Winnipeg	MB	50.65	-96.92	2	S2H083, S2H162
MB011	Little Saskatchewan River	MB	50.25	-99.84	2	S2H052, S2H131
MB012	Minnedosa Pond	MB	50.24	-99.86	2	S2H075, S2H130, S2H151
MB013	Odei River	MB	56.10	-96.91	2	S2H003, S2H007, S2H213
MB014	Paint Lake	MB	55.49	-98.03	2, 3	S2H180, S2H187, S3H104
MB015	Pond	MB	50.20	-99.93	2	S2H075, S2H216, S2H218
MB016	Pond	MB	50.30	-99.00	7	S7H020
MB017	Pond	MB	50.63	-99.97	2	S2H075
MB018	Pond	MB	50.61	-99.89	2	S2H053, S2H066
MB019	South Indian Lake	MB	56.78	-98.93	2	S2H002, S2H017, S2H221
MB020	Strong Lake Creek	MB	52.24	-98.91	2, 7	S2H085, S2H093, S7H004
MB021	Tributary to Burntwood River	MB	44.81	-78.64	2	S2H010, S2H132, S2H211
ME001	191 Bog Lake	ME	45.01	-67.42	5, 6	S5H030, S5H042, S6H164, S6H165, S6H166
ME002	Bog in Maine	ME	45.23	-67.22	2, 5, 6	S2H132, S2H194, S2H216, S5H045, S6H164
ME003	Cathance Lake	ME	44.94	-67.43	5	S5H022, S5H045, S5H050, S5H051
ME004	Gardner Lake	ME	44.74	-67.33	5, 6	S5H030, S6H166
ME005	Graham Lake	ME	44.61	-68.39	5, 6	S5H027, S5H039, S5H045, S6H179
ME006	Green Lake	ME	44.64	-68.48	5	S5H021, S5H027, S5H038, S5H049
MI001	Cedar Lake	MI	42.31	-84.09	3, 5	S3H033, S3H051, S3H123, S3H162, S3H163, S5H017, S5H019, S5H020
MI002	Coldwater River	MI	41.84	-85.00	3, 5, 6	S3H148, S3H150, S3H151, S5H008, S6H047
MI003	Eagle Lake	MI	42.22	-85.73	1(B)	S1H110, S5H008, S5H010
MI004	Unnamed Lake	MI	42.33	-85.33	1(B)	S1H099, S1H103, S1H109
MI005	Loon Lake Channel	MI	41.87	-84.94	3, 6	S3H002, S3H053, S3H149, S3H161, S3H188, S6H060, S6H061, S6H069
MI006	Michigan Bog	MI	45.99	-86.12	1(B)	S1H119
MI007	Mill Lake	MI	42.26	-84.95	1(B), 6	S1H100, S1H103, S6H032, S6H053, S6H055

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
MI008	Reynolds Lake	MI	42.20	-85.99	1(B), 3	S1H104, S1H105, S1H106, S3H048, S3H049, S3H054, S3H056, S3H067, S3H068, S3H072, S5H014, S5H016, S6H032
MI009	Upper Brace Lake	MI	42.18	-84.94	3, 5, 6	S3H124, S3H136, S3H137, S3H138, S3H171, S5H008, S6H005
MN001	Boulder Lake	MN	47.05	-92.20	2, 3, 6	S2H209, S2H222, S3H038, S3H065, S3H109, S3H195, S6H067
MN002	Little Fork River	MN	48.39	-93.56	2	S2H203
MN003	Oak Lake	MN	46.39	-92.59	2, 6	S2H074, S2H078, S6H112, S6H113, S6H114
MN004	Pelican Lake	MN	46.58	-94.20	2, 3, 6	S2H200, S2H212, S3H104, S3H135, S6H068, S6H073, S6H076, S6H115
MN005	Perch Lake	MN	46.68	-92.68	3, 6	S3H088, S3H101, S3H108, S6H095
MN006	Upper Nemadjii River	MN	46.49	-92.47	2, 6	S2H043, S2H044, S2H225, S6H078
MN007	Vermillion Lake	MN	47.84	-92.36	2, 3	S2H171, S2H203, S2H213, S3H102, S3H103, S3H110
MO001	Binder Lake	MO	38.59	-92.31	6	S6H001, S6H017, S6H130, S6H134
MO002	Haysler A Poague	MO	38.42	-93.83	6	S6H001, S6H002, S6H130, S6H132
MO003	Lake Luna	MO	38.66	-94.34	3, 6	S3H045, S6H001, S6H002
MO004	Little Dixie Lake	MO	38.90	-92.13	6	S6H001, S6H017, S6H052
MO005	Longview Lake	MO	38.90	-94.47	3, 6	S3H042, S3H045, S3H081, S3H115, S6H142
MS001	American Legion Lake	MS	33.51	-89.78	6	S6H124, S6H135
MS002	Crystal Creek	MS	32.38	-89.82	6	S6H139, S6H140, S6H141
MS003	English Lake	MS	33.03	-89.92	5	S5H008, S5H009
MS004	Lake Lamar Bruce	MS	34.39	-88.67	6	S6H137, S6H138
MS005	Old Natchez Trace Lake	MS	34.25	-88.89	6	S6H123, SH136
MS006	Ross Barnett Reservoir	MS	32.46	-90.05	8	S8H089, S8H091, S8H097, S8H098
MS007	Simpson Legion Lake	MS	31.91	-89.79	6	S6H127, S6H128, S6H129
MS008	Thompson Creek	MS	31.75	-88.90	4	S4H019, S4H031, S6H032
NB001	630 Bog	NB	45.66	-67.45	5, 6	S5H052, S6H180
NB002	Digdoguash River	NB	45.31	-67.07	4, 5, 6	S4H037, S5H029, S5H031, S5H032, S6H167, S6H180
NB003	Eel River	NB	48.01	-66.42	5	S5H046
NB004	Glebe road pond	NB	45.13	-67.07	6	S6H174, S6H175, S6H176, S6H177
NB005	Grand Lake	NB	45.82	-67.78	1(A), 5, 6	S1H065, S1H082, S5H021, S5H045, S5H056, S5H059, S5H064, S6H166
NB006	Lake Utopia	NB	45.13	-66.77	5, 6	S5H031, S5H034, S6H162, S6H163, S6H180

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
NB007	North Lake	NB	45.83	-67.73	1(A), 5, 6	S1H070, S5H040, S5H041, S5H046, S5H047, S5H048, S5H057, S5H065, S6H166
NB008	Pond near Chamcook Lake	NB	45.13	-67.07	6	S6H181
NB009	Second Eel Lake	NB	45.80	-67.64	1(A), 5	S1H096, S5H041, S5H046, S5H052
NB010	St. John River	NB	45.36	-66.23	1(A), 6	S1H083, S1H095, S6H178
NB011	Unnamed Stream	NB	45.86	-67.49	1(A), 5, 6	S1H076, S1H081, S5H063, S5H0143
NB012	Upper Skiff Lake	NB	45.83	-67.55	1(A)	S1H064, S1H097, S1H098
NC001	B. Everett Jordan Lake	NC	35.64	-79.09	5	S5H118, S5H124
NC002	Cape Fear River	NC	35.40	-78.77	5	S5H119, S5H120, S5H124
NC003	Fall's Lake	NC	36.02	-78.69	5	S5H067, S5H104, S5H123
NC004	Harris Reservoir	NC	35.64	-79.09	6	S6H158, S6H159, S6H160
NC005	Lake Rogers	NC	36.13	-78.70	5	S5H121, S5H123
NC006	Lake Wallace	NC	34.63	-79.68	4	S4H001, S4H002, S4H023, S4H025
NC007	Mountain Island Lake	NC	35.37	-80.93	5	S5H098, S5H099
NE001	Atkinson Lake	NE	42.54	-99.00	2, 3	S2H064, S2H145, S3H042, S3H060
NE002	Bowman Lake	NE	41.28	-98.99	3, 6	S3H045, S3H061, S3H196, S6H038
NE003	Calamus River	NE	41.79	-99.13	2	S2H038, S2H050, S2H073, S2H075
NE004	Harlan County Lake	NE	40.10	-99.36	8	S8H099, S8H100, S8H102
NE005	Swan Lake	NE	42.17	-99.03	2	S2H038, S2H050, S2H064, S2H075
NH001	Bog in New Hampshire	NH	43.29	-71.00	1(A), 2	S1H074, S2H085
NH002	Pond off of Connecticut River	NH	44.50	-71.58	1(A), 6	S1H068, S1H085, S6H166
NL001	Bauline	NL	47.72	-52.83	5, 6	S5H060, S6H166
NL002	Duck Pond	NL	47.70	-52.81	6	S6H166
NL003	Holyrood	NL	47.39	-53.13	6	S6H170
NL004	Long Pond	NL	47.77	-52.80	5	S5H028, S5H045, S5H061, S5H062
NL005	New World Island	NL	49.65	-54.75	6	S6H172, S6H173
NL006	Pond South of St. Johns	NL	47.46	-52.76	5	S5H026
NS001	Neuton Mills Pond	NS	45.24	-62.91	6	S6H168, S6H169
NT001	Boot Lake	NT	68.35	-133.70	1(G)	S1H252

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
NT002	Crossly Lakes 15	NT	67.86	-131.47	1(F)	S1H196, S1H197
NT003	Eskimo Lake 5	NT	68.56	-133.75	1(G)	S1H253, S1H254
NT004	Ferry Pond	NT	67.33	-134.92	1(G)	S1H233
NT005	Mackenzie Delta 22	NT	68.37	-133.77	7	S7H064
NT006	Unnamed Pond	NT	68.07	-133.62	1(G)	S1H251
NT007	Unamed Lake	NT	62.46	-114.38	1(F), 2	S1H185, S1H198, S1H199, S2H001, S2H002, S2H177, S2H178
NV001	Side Hill Spring	NV	38.25	-116.68	7	S7H022, S7H023
NV002	W11	NV	37.65	-114.50	1(H)	S1H300, S1H308
NV003	W6	NV	38.97	-115.74	1(H)	S1H309, S1H310
NY001	Canadadrago Lake	NY	42.82	-75.01	6	S6H116, S6H117, S6H118
NY002	Canadaiagua Lake	NY	42.88	-77.28	3, 6	S3H030, S3H031, S3H070, S3H178, S6H040, S6H048
NY003	Chittnang Pond	NY	42.91	-75.32	1(A), 3, 6	S1H032, S1H033, S1H071, S3H001, S3H0032, S3H178, S6H083, S6H084
NY004	Jamesville Reservoir	NY	42.97	-76.07	3, 6	S3H070, S3H072, S3H074, S6H036, S6H041
NY005	Nathaniel Cole County Lake	NY	42.98	-75.70	5	S5H111, S5H112
NY006	Oneida Lake	NY	43.19	-76.04	3	S3H075, S3H090, S3H111, S3H172, S3H176, S3H183, S3H189, S3H193
NY007	Pembroke Creek	NY	43.00	-78.31	1(B), 3	S1H103, S1H107, S3H002, S3H030, S3H070, S3H105
OH001	Alum Creek Lake	OH	40.20	-82.95	3, 6	S2H006, S3H010, S3H070, S3H193, S3H194
OH002	Blanchard River	OH	41.01	-83.56	3, 6	S3H070, S3H111, S6H029, S6H030, S6H032
OH003	Buckyrus Reservoir	OH	40.83	-82.94	6	S6H032
OH004	Lake Snowden	OH	39.24	-82.19	3	S3H009, S3H112, S3H113
OH005	New Lexington Reservoir	OH	39.74	-82.22	5	S5H010
OH006	Olentangy River	OH	40.35	-83.07	3, 6	S3H003, S3H005, S3H013, S6H023, S6H046
OH007	Rush Creek Lake	OH	39.78	-82.41	5	S5H010
OK001	Boomer Lake	OK	49.55	-87.76	6	S6H133
OK002	Lake Sanborn	OK	36.15	-97.08	6	S6H133
ON001	Batchewana River	ON	46.92	-84.52	1(B), 1(E), 2, 3	S1H118, S1H172, S2H179, S3H089
ON002	Beaver River	ON	44.22	-79.06	1(A), 1(E)	S1H018, S1H051, S1H056, S1H161, S2H206
ON003	Blue Springs Creek	ON	43.57	-80.15	6	S6H070, S6H071, S6H087

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
ON004	Unnamed Bog Stream	ON	46.90	-84.38	1(B), 1(C), 1(E)	S1H101, S1H137, S1H154
ON005	Bridgeland River	ON	46.30	-83.55	1(B), 1(D), 1(E)	S1H133, S1H165, S1H179
ON006	Bright Lake	ON	46.25	-83.26	3, 6	S3H152, S3H192, S6H107, S6H108
ON007	Burbot Lake	ON	46.26	-78.90	1(A), 3	S1H039, S1H041, S1H042, S3H143, S3H143
ON008	Butterfly Lake	ON	45.08	-79.63	5	S5H036
ON009	Credit River by Belfountain	ON	43.80	-80.01	1(A), 1(E), 2	S1H026, S1H041, S1H160, S1H174, S1H176, S2H027, S2H029, S2H055
ON010	Deer Creek Resevoir	ON	42.70	-80.56	6	S6H053
ON011	Ear Falls	ON	50.63	-93.22	2	S2H204, S2H205, S2H207, S2H210
ON012	Eramosa River	ON	43.61	-80.15	1(B)	S1H107
ON013	Fushimi Lake	ON	49.84	-83.92	1(A), 1(E)	S1H050, S1H154, S1H162, S1H163
ON014	Georgian Bay at Parry Sound	ON	45.35	-80.05	3	S3H070, S3H128, S3H145
ON015	Harp Lake	ON	45.38	-79.13	2, 5	S2H025, S5H054, S5H055
ON016	Helen Lake	ON	49.04	-88.25	1(F), 2	S1H204, S2H183
ON017	Hwy 625 Pond	ON	49.77	-86.33	2	S2H208, S2H224
ON018	Ice Lake	ON	45.90	-82.40	1(A), 1(B), 3, 6	S1H090, S1H111, S3H126, S6H062, S6H072
ON019	Klotz Lake	ON	49.79	-85.93	1(A), 1(D), 2, 3	S1H044, S1H047, S1H145, S2H076, S2H147, S3H070, S3H154
ON020	Lac Gilman	ON	49.91	-74.22	1(A), 1(E), 6	S1H048, S1H173, S6H093
ON021	Lac Seul	ON	50.29	-92.29	2, 3	S2H041, S2H042, S2H051, S3H071, S3H104
ON022	Lake Erie at Long Point Prov Parl	ON	42.58	-80.39	3, 6	S3H014, S3H070, S3H181, S6H051, S6H066
ON023	Lake Joseph	ON	45.20	-79.79	5	S5H037
ON024	Lake Nipigon	ON	49.47	-88.33	2, 3	S2H203, S3H084, S3H114
ON025	Lake of the Woods	ON	48.85	-94.70	1(H), 2, 3, 6	S1H255, S2H203, S3H069, S3H104, S6H110, S6H111
ON026	Lake Ontario at Brockville	ON	44.58	-75.69	6	S6H085, S6H096
ON027	Mattagami River	ON	50.15	-82.21	1(A), 1(E), 3	S1H023, S1H026, S1H154, S3H131, S3H132, S3H167, S3H168
ON028	McCauly Creek	ON	45.52	-78.07	2, 3, 5	S2H199, S3H015, S3H085, S5H054
ON029	Moose Lake	ON	48.11	-90.07	1(E), 3	S1H154, S1H158, S3H094
ON030	Opasitika River	ON	49.52	-82.84	1(A), 1(E)	S1H017, S1H030, S1H031, S1H168
ON031	Pond near Guelph	ON	43.50	-80.16	1(A), 1(E), 2	S1H026, S1H176, S1H178, S2H027, S2H028

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
ON032	Pond south of Belfountain	ON	43.79	-80.02	1(A)	S1H041
ON033	Rainy Lake	ON	48.69	-93.18	2, 3	S2H170, S3H100, S3H104, S3H191
ON034	Rouge Park, Markham	ON	43.87	-79.25	1(A), 1(D), 6	S1H026, S1H151, S1H152, S1H153, S1H057, S6H053
ON035	Saugeen River, Walkerton	ON	44.14	-81.16	1(A), 1(E), 6	S1H054, S1H154, S6H106
ON036	Saugine River	ON	44.42	-81.35	1(E), 2	S1H167, S1H170, S2H035
ON037	Severn River	ON	44.89	-79.61	3, 6	S3H133, S3H134, S3H170, S3H177, S3H184, S6H045, S6H063
ON038	Somehwere in Guelph	ON	43.67	-80.33	1(A), 1(B), 6	S1H051, S1H108, S6H056
ON039	South of Beardmore	ON	49.59	-87.96	2	S2H146, S2H172, S2H213
ON040	Speed River	ON	43.55	-80.24	1(B), 6	S1H107, S6H049, S6H050, S6H064, S6H065
ON041	Stretch Lake	ON	46.11	-78.65	1(B)	S1H134, S1H135
ON042	Sudbury Pond	ON	46.48	-81.02	1(A)	S1H045
ON043	SWAN	ON	48.24	-80.26	1(A), 3	S1H024, S3H122, S3H165, S3H166
ON044	Teeswater River	ON	44.00	-81.29	3	S3H144
ON045	Vic Pond Guelph	ON	43.53	-80.20	2	S2H027
ON046	Wabigoon Lake	ON	49.76	-92.83	3	S3H018, S3H037, S3H063, S3H187
ON047	Walker Lake	ON	48.84	-87.31	5	S5H033
ON048	West Credit	ON	43.79	-80.02	1(E), 2	S1H175, S1H177, S2H018, S2H019
ON049	Crocker's Lake	ON	48.62	-85.37	7	S7H085
OR001	Carter Lake	OR	43.86	-124.14	9	S9H005, S9H008
OR002	Crump Lake	OR	42.26	-119.88	1(H)	S1H371
OR003	Dexter Reservoir	OR	43.92	-122.82	9	S9H016
OR004	Eel Lake	OR	43.60	-124.18	9	S9H006, S9H007
OR005	Heart Lake	OR	42.25	-120.84	1(H)	S1H298, S1H299, S1H307
OR006	lake of the Woods	OR	42.39	-122.21	1(H)	S1H272, S1H277, S1H312
OR007	Lofton Pond	OR	42.27	-120.83	1(H)	S1H311
OR008	Lost Lake	OR	45.50	-121.82	1(H)	S1H273, S1H276, S1H277
OR009	Lost River	OR	44.18	-121.60	9	S9H013

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
OR010	Mt. Wellington Pond	OR	44.42	-121.83	1(H)	S1H275, S1H304
OR011	Odell Lake	OR	43.58	-122.05	1(H)	S1H297, S1H303, S1H306
OR012	Outflow of Ten Mile Lake	OR	43.57	-124.17	9	S9H009
OR013	Saunders Lake	OR	43.53	-124.22	9	S9H001, S9H019
OR014	Siltcoos lake	OR	43.88	-124.11	9	S9H008
OR015	Suttle Lake	OR	44.41	-121.76	1(H)	S1H276
OR016	Tahkemitch Lake	OR	43.80	-124.15	9	S9H005, S9H008
PA001	Kyle Lake	PA	41.16	-78.84	6	S6H026, S6H027, S6H082
PA002	Quaker Lake	PA	41.97	-75.93	3	S3H070, S3H073, S3H076, S3H190
PA003	Rose Valley Lake	PA	41.39	-76.98	5, 6	S5H113, S5H114, S5H115, S5H116, S6H088
PA004	Shawnee Lake	PA	40.04	-78.65	6	S6H059
PA005	Sylvan Lake	PA	41.26	-76.16	5, 6	S5H110, S5H111. S6H156, S6H157
PA006	Twin Lakes	PA	40.32	-79.47	3, 6	S3H025, S6H042, S6H057, S6H058
QC001	Blue Sea Lake	QC	46.21	-76.05	1(A), 1(B), 6	S1H038, S1H115, S6H082, S6H101, S6H102
QC002	Cameron Lake	QC	46.18	-75.92	1(A), 3, 6	S1H015, S3H174, S3H180, S6H080, S6H097
QC003	Hwy 113 Bridge Creek	QC	49.79	-74.58	1(A)	S1H011, S1H046
QC004	Hwy 113 Creek	QC	48.12	-77.39	1(A)	S1H020, S1H023, S1H026
QC005	Hwy 167 Creek	QC	49.98	-74.20	1(A), 1(E), 6	S1H058, S1H059, S1H060, S1H061, S1H157, S1H169, S1H173, S6H089, S6H090, S6H091
QC006	Hwy 167 Marsh	QC	50.50	-73.58	1(A), 1(B), 1(E)	S1H013, S1H015, S1H112, S1H154
QC007	Lac Achen	QC	49.35	-73.98	1(A), 1(E)	S1H016, S1H052, S1H159
QC008	Lac Albanel	QC	51.07	-73.02	1(A), 1(E)	S1H012, S1H015, S1H166, S1H180
QC009	Lac Anne-Marie	QC	49.17	-73.65	5	S5H035
QC010	Lac Aylmer	QC	45.83	-71.38	1(A)	S1H066, S1H075, S1H078
QC011	Lac Blovin	QC	48.18	-77.79	1(A)	S1H023
QC012	Lac Charron	QC	49.51	-74.17	1(A), 1(E), 6, 7	S1H004, S1H012, S1H014, S1H016, S1H181, S1H182, S6H093, S7H058
QC013	Lac Chigoubiche	QC	49.14	-73.59	1(A), 1(E), 5, 6	S1H003, S1H004, S1H005, S1H049, S1H052, S1H156, S1H164,
QC014	Lac Cummings	QC	49.95	-74.22	1(E), 7	S1H173, S7H080, S7H081
QC015	Lac du Mileu	QC	48.82	-72.80	5	S5H025

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
QC016	Lac Gilman	QC	49.91	-74.22	1(A), 1(E)	S1H048, S1H173
QC017	Lac Jupiter	QC	47.79	-71.23	1(A)	S1H064
QC018	Lac Mistassini	QC	50.41	-73.87	1(B), 1(E)	S1H112, S1H157
QC019	Lac Parent	QC	48.79	-77.11	1(A)	S1H026, S1H034, S1H035, S1H059
QC020	Lac Renault	QC	48.18	-79.27	1(A), 3	S1H019, S1H021, S1H023, S1H026, S1H038, S3H129, S3H164
QC021	Lac Saint Jean	QC	48.43	-71.87	1(A), 1(E)	S1H001, S1H006, S1H008, S1H009, S1H154, S1H155
QC022	Lac Tiblemont	QC	48.21	-77.36	1(A), 6	S1H020, S1H026, S1H034, S6H093
QC023	Lac Vert	QC	48.36	-71.64	1(A)	S1H007, S1H009, S1H010, S1H053, S1H072
QC024	Lac-des-iles	QC	46.41	-75.52	1(A),3, 6	S1H039, S3H155, S3H179, S3H180, S6H043
QC025	Lake Elgin	QC	45.76	-71.35	1(A)	S1H066, S1H079
QC026	Lake Papineau	QC	46.13	-74.32	5	S5H023
QC027	Lake Pemichangan	QC	46.08	-75.86	1(A), 3	S1H037, S1H043, S3H119, S3H182
QC028	Lake St. Remi	QC	46.01	-74.76	3	S3H156, S3H157, S3H158, S3H159, S3H160
QC029	Lake Tapani	QC	46.91	-75.32	1(A), 6	S1H026, S1H038, S6H092, S6H099
QC030	Lake Temiscouta	QC	47.61	-68.78	1(A), 6	S1H069, S1H073, S1H077, S6H098, S6H100, S6H103
QC031	Lake Timiskaming	QC	47.48	-79.67	3	S3H079, S3H120, S3H121, S3H164, S3H186
QC032	Lake William	QC	46.12	-71.58	1(A), 6	S1H078, S1H091, S6H103, S6H105
QC033	Mathews Pond	QC	50.99	-72.99	1(E)	S1H154, S1H171
QC034	Missassauga Lake	QC	44.71	-78.33	1(A)	S1H026, S1H028, S1H029, S6H104
QC035	Murry Lake	QC	46.97	-74.01	1(A), 6	S1H062, S1H063, S6H081, S6H103, S6H119
QC036	Petit Lac a L'epaule	QC	47.29	-71.19	1(A)	S1H055, S1H086, S1H087
QC037	River Du Nord	QC	45.88	-74.09	3, 5, 6	S3H070, S3H173, S5H024, S6H086
QC038	Riviere Audet	QC	49.65	-74.33	1(A), 1(D), 1(E), 7	S1H002, S1H004, S1H144, S1H154, S7H078
QC039	Riviere Des Deux Origneaux	QC	49.82	-75.33	1(A)	S1H025, S1H026
QC040	Senneterre River	QC	48.39	-77.23	1(A), 5	S1H023, S1H034, S1H035, S1H038, S1H040, S5H044
QC041	Small Pond in QC	QC	46.05	-75.28	1(A)	S1H067, S1H080, S1H092
QC042	Sutton Creek	QC	47.56	-79.60	1(A)	S1H026
QC043	Thetford Pond	QC	46.13	-71.32	1(A)	S1H084

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
QC044	Unamed Bog Quebec (melania)	QC	46.14	-73.50	1(A)	S1H088, S1H089
QC045	Waswanipi River	QC	49.69	-75.97	1(A), 1(B)	S1H022, S1H026, S1H113
SC001	Dargon's Pond	SC	34.33	-79.72	5	S5H066, S5H105, S5H106, S5H107
SC002	Edisto River	SC	33.59	-81.10	5, 6	S5H068, S5H069, S6H152
SC003	Lake Brown	SC	33.25	-81.37	5	S5H070
SC004	Lake Marion	SC	33.63	-80.54	5	S5H098
SC005	Saluda River	SC	34.05	-81.19	5	S5H100, S5H101, S5H102
SC006	Watertree Creek	SC	34.47	-80.92	5	S5H100
SD001	Creek in Iroquois	SD	44.37	-97.85	1(C), 2	S1H139, S2H067, S2H075, S2H079, S2H086, S2H088, S2H168
SD002	Duck Pond in Arlington	SD	44.36	-97.13	2	S2H036, S2H050, S2H088, S2H089, S2H160
SD003	James River	SD	44.37	-98.20	2	S2H048, S2H061, S2H066, S2H079
SD004	Lake Albert	SD	44.53	-97.13	7	S7H039, S7H086
SD005	Lake Andes	SD	43.15	-98.53	1(C), 2	S1H139, S2H038, S2H045, S2H075, S2H077, S2H090, S2H091, S2H141
SD006	Lake Louise	SD	44.62	-99.14	2	S2H038, S2H050, S2H075, S2H079, S2H088
SD007	Lake Thompson	SD	44.32	-97.44	1(C), 2	S1H137, S2H037, S2H050, S2H089, S2H169
SD008	Pond off HWY 81	SD	43.97	-97.07	7	S7H039
SK001	Amisk Lake	SK	54.68	-102.08	1(F)	S1H199, S1H201, S1H202
SK002	Arm River Reservoir	SK	51.06	-105.80	7	S7H004, S7H020, S7H025
SK003	Arm River Site 1	SK	50.84	-105.44	2, 7	S2H150, S2H151, S7H025
SK004	Arm River Site 2	SK	51.06	-105.80	2	S2H068, S2H128, S2H151
SK005	Barrier River	SK	52.57	-104.03	2, 7	S2H117, S7H035, S7H037, S7H039, S7H096, S7H097
SK006	Beaver River	SK	54.28	-109.23	2	S2H054, S2H092, S2H187, S2H198
SK007	Cowen Lake	SK	53.83	-107.04	2, 7	S2H005, S7H004, S7H008, S7H035
SK008	Des chambat Lake	SK	54.90	-103.37	7	S7H094
SK009	Green Lake	SK	54.28	-107.81	1(A), 1(F), 2, 7	S1H064, S1H200, S1H203, S2H054, S2H058, S2H132, S2H220, S7H004, S7H011, S7H039, S7H043
SK010	Greenwater Lake	SK	52.50	-103.52	2	S2H054, S2H167, S2H218
SK011	Kipabiskau Lake	SK	52.57	-104.19	1(C), 2, 7	S1H139, S2H060, S2H082, S2H092, S7H020, S7H059
SK012	Lac Des Iles	SK	54.41	-109.41	2	S2H107, S2H111, S2H167

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
SK013	Lac La Ronge	SK	55.10	-105.27	2, 6	S2H109, S2H110, S2H162, S6H074, S6H075
SK014	Lake Tobin	SK	53.51	-103.81	2, 7	S2H134, S2H151, S2H162, S2H215, S7H027, S7H039, S7H091
SK015	Last Mountain Lake	SK	51.23	-105.30	7	S7H004, S7H025, S7H052, S7H059
SK016	Makwa River	SK	54.08	-109.20	2	S2H094, S2H107, S2H159, S2H162, S2H218
SK017	Meadow River	SK	54.14	-108.39	7	S7H095
SK018	Prairie pothole	SK	53.13	-106.08	2	S2H151, S2H219
SK019	Reed Lake	SK	50.41	-107.04	7	S7H004, S7H018, S7H021, S7H025, S7H040
SK020	Saskatchewan River	SK	50.66	-107.97	1(C), 1(G), 2	S1H139, S2H066, S2H151, S2H153, S2H158
SK021	Shell River	SK	53.25	-106.39	2, 7	S2H005, S2H054, S2H214, S2H217, S7H035
SK022	Stalwalt Bird Pond	SK	51.24	-105.42	7	S7H004, S7H024, S7H039, S7H041
SK023	Stocked Trout Pond	SK	53.39	-104.01	2	S2H151, S2H161, S2H175, S2H215
SK024	Stoney Lake	SK	52.16	-105.15	7	S7H004, S7H020, S7H035, S7H038
SK025	Sturgeon Lake	SK	53.42	-106.16	2, 7	S2H151, S2H152, S1H154, S1H164, S7H004
TN001	Beech Lake	TN	35.71	-88.42	6	S6H126
TN002	Carroll Lake	TN	36.10	-88.49	5	S5H013
TN003	Cumberland River	TN	36.30	-86.27	6, 8	S6H014, S6H015, S6H016, S8H091
TN004	David Crockette Lake	TN	35.27	-87.36	3, 5	S3H117, S5H015
TN005	Dyers Creek	TN	36.50	-87.83	8	S8H069
TN006	J Perey Priest Lake	TN	35.98	-86.42	6	S6H121, S6H182
TN007	Kentucky Lake Tennessee River	TN	36.45	-88.05	8	S8H069, S8H090, S8H095, S8H096
TN008	Lake Graham	TN	35.63	-88.72	6	S6H124, S6H125, S6H126
VA001	Brunswick County Lake	VA	36.79	-77.74	5	S5H126, S5H127, S5H128, S5H129
VA002	Great Creek	VA	36.78	-77.89	5	S5H104, S5H121, S5H122
VA003	James River	VA	37.53	-79.68	3	S3H003
VA004	Lake Orange	VA	38.23	-78.02	5	S5H129, S5H130
VA005	Lake Pelham	VA	38.46	-78.05	6	S6H147, S6H148, S6H149
VA006	New River	VA	37.05	-80.67	3, 5	S3H005, S3H011, S5H108, S5H109
VA007	Shenandoah River	VA	38.85	-78.53	3	S3H008

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
VA008	Swift Creek Lake	VA	37.39	-77.58	6	S6H145, S6H146
VT001	Lake Champlain	VT	44.77	-73.29	1 (A)	S1H026, S1H027, S1H038
WA001	Banks Lake	WA	47.62	-119.29	9	S9H010, S9H014, S9H032, S9H037
WA002	Beaver Lake	WA	48.11	-124.25	1(H)	S1H279, S1H301, S1H302
WA003	Brown Lake	WA	48.46	-119.62	1(G), 7	S1H248, S7H033
WA004	Cooper Lake	WA	47.43	-121.07	1 (COOP)	S1G313, S1H314, S1H315
WA005	Fish Lake	WA	48.61	-119.70	1 (H)	S1H278
WA006	Goose Lake	WA	45.94	-121.76	7	S7H003, S7H057
WA007	Hoh River	WA	47.81	-124.13	9	S9H025, S9H026
WA008	Hwy 113	WA	48.11	-124.25	9	S9H034, S9H035
WA009	Lake Cavanaugh	WA	48.31	-121.99	9	S9H029, S9H030, S9H031
WA010	Lake Crescent	WA	48.09	-123.74	1 (H), 1(I)	S1H285, S1H334, S1H335
WA011	Lake Entiat	WA	47.54	-120.28	9	S9H011. S9H012, S9H015, S9H038, S9H039
WA012	Lake Terrell	WA	48.86	-122.69	9	S9H003, S9H004, S9H030
WA013	Link Lake	WA	47.92	-120.20	1 (H)	S1H271, S1H274
WA014	McMurray Lake	WA	48.31	-122.23	1 (H), 9	S1H280, S9H027, S9H028
WA015	Quinault Lake	WA	47.48	-123.83	1 (I)	S1H336, S1H337, S1H338
WA016	Stan Coffin Lake	WA	47.15	-119.92	1 (H)	S1H263, S1H269, S1H270
WA017	Wapata Lake	WA	47.92	-120.16	1 (H)	S1H371, S1H291
WI001	Big Muskellunge Lake, WI1	WI	46.02	-89.62	1 (B), 7	S1H128, S7H085
WI002	Green Lake	WI	43.84	-88.96	1 (B), 3	S1H120, S1H122, S3H034, S3H-35, S3H036, S3H078
WI003	Mecan Pond (pond near Mecan	WI	43.84	-89.15	1 (B)	S1H121, S1H125, S1H127
WI004	Pike Lake	WI	44.82	-89.35	3, 6	S3H086, S3H093, S6H077, S6H109
WI005	Plum Lake	WI	46.04	-89.47	1 (B), 2, 3, 7	S1H128, S1H129, S2H002, S2H008, S2H025, S2H055, S3H095, S3H096, S3H097, S3H098, S3H099, S7H090
WI006	Razorback Lake	WI	46.02	-89.53	2	S2H025
WI007	St. Croix River	WI	44.94	-92.75	1 (A), 2, 3	S1H017, S1H030, S2H113, S3H052, S3H055, S3H057, S3H058, S3H066, S3H083, S3H087, S3H106, S3H109
WI008	Stream near Princeton	WI	43.85	-89.13	1 (B)	S1H123, S1H124, S1H126
WI009	Unnamed inflow to Turtle Lake, \	WI	46.23	-89.62	1 (B), 2, 6	S1H116, S1H128, S1H130, S1H131, S1H132, S1H133, S1H134, S2H009, S2H203, S6H054

Code	Location	Province /State	Latitude (DD)	Longitude (DD)	Species (lineage)	Haplotypes
WV001	Elk Fork	WV	38.80	-81.61	3	S3H070, S3H111, S3H139, S3H140
WV002	Tygart Lake	WV	39.27	-80.02		S6H044
YT001	Alaska Hwy Pond	YT	61.77	-139.99	1(G), 2, 7	S1H214, S1H250, S7H052
YT002	Gravel Lake	YT	63.81	-137.90	2, 7	S2H114, S2H126, S7H061
YT003	Horse Creek	YT	60.98	-135.39	1(G)	S1H225
YT004	Klusha Pond	YT	61.74	-136.00	2	S2H125
YT005	Lhutsaw Lake	YT	62.67	-136.75	7	S7H071
YT006	Old Crow 12	YT	67.57	-139.82	1(G)	S1H230, S1H249,
YT007	Old Crow 22	YT	67.57	-139.76	1(G), 7	S1H229, S1H231, S7H061, S7H062
YT008	Pine Lake	YT	60.80	-137.48	1(G)	S1H219, S1H228
YT009	Pond in Dawson City	YT	64.04	-139.36	2	S2H123, S2H126
YT010	Pond on Alaska Highway	YT	60.45	-133.59	1(G)	S1H228
YT011	Rancheria Pond	YT	60.34	-129.04	1(G), 7	S1H212, S1H222, S7H056, S7H097
YT012	Rocky Island Lake	YT	62.72	-136.70	2, 7	S2H115, S7H045
YT013	Schwatka Lake	YT	60.67	-135.02	1(F)	S1H190, S1H191
YT014	Squanga Lake	YT	60.45	-133.60	1(F), 1(I), 7	S1H183, S1H189, S1H362, S7H066
YT015	Steephill Pond	YT	61.67	-135.91	7	S7H061, S7H063
YT016	Teslin Lake	YT	60.11	-132.55	1(G)	S1H209, S1H211, S1H215
YT017	Twin Lakes	YT	61.70	-135.93	1(D)	S1H145
YT018	Yukon Pond	YT	60.01	-128.56	2, 7	S2H097, S2H103, S2H104, S7H076, S7H077, S7H100

Appendix B

Table B.1: Species 1 lineage A Nested Clade Analysis

The table includes all species 1 lineage A clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-8 (nested within clade 2-4)					
Tip	S1H017	377.35	(0.794, 0.206)	**883.4**^{L}	(1.000, <**0.001**)
Tip	S1H018	0	(0.054, 1.000)	328.84	(0.629, 0.371)
Tip	S1H019	0	(1.000, 1.000)	113.64	(0.189, 0.934)
Tip	S1H020	**4.28**^{S}	(<**0.001,** 1.000)	159.85	(0.019, 0.981)
Tip	S1H021	0	(1.000, 1.000)	113.64	(0.175, 0.960)
Tip	S1H022	0	(1.000, 1.000)	355.20	(0.652, 0.371)
Tip	S1H023	148.56	(0.028, 0.972)	211.95	(0.046, 0.954)
Tip	S1H024	0	(0.055, 1.000)	152.11	(0.192, 0.809)
Tip	S1H025	0	(0.045, 1.000)	396.98	(0.817, 0.185)
Interior	S1H026	289.86	(0.224, 0.776)	299.72	(0.281, 0.719)
	I-T	176.59	(0.946, 0.0540)	-28.08	(0.303, 0.697)
Inference Key			1-2-5-6-15-NO		
Conclusion			Isolation by Distance with Long Distance Dispersal		
Clade 1-33 (nested in clade 2-10)					
Interior	S1H064	127.54	(0.155, 0.845)	208.17	(0.722, 0.278)
Tip	S1H066	**4.70**^{S}	(<**0.001**, 1.000)	**70.67**^{S}	(<**0.001**, 1.000)
Tip	S1H067	0	(1.000, 1.000)	65.65	(0.042, 1.000)
Tip	S1H068	0	(1.000, 1.000)	187.07	(0.478, 0.568)
Tip	S1H069	0	(1.000, 1.000)	221.28	(0.956, 0.100)
Tip	S1H070	0	(1.000, 1.000)	215.21	(0.883, 0.163)
Tip	S1H071	0	(1.000, 1.000)	513.19	(1.000, 0.051)
	I-T	124.8	(0.823, 0.177)	66.75	(0.806, 0.194)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-4 (nested within clade 3-2)					
Interior	1-8	316.03	(0.708, 0.292)	319.53	(0.796, 0.204)
Tip	1-9	0	(0.043, 1.000)	458.11	(0.904, 0.100)
Tip	1-10	0	(0.050, 1.000)	248.86	(0.410, 0.591)
Tip	1-11	318.13	(0.808, 0.192)	600.25	(0.965, 0.035)
Tip	1-12	0	(0.046, 1.000)	505.90	(0.903, 0.097)
Tip	1-13	**23.01**^{S}	(<**0.001**, 1.000)	**183.98**^{S}	(**0.001**, 0.999)
Tip	1-14	**5.81**^{S}	(**0.020**, 0.998)	204.86	(0.147, 0.853)
Tip	1-15	**79.12**^{S}	(**0.001**, 0.999)	217.56	(0.092, 0.908)
	I-T	**258.86**^{L}	(1.000, <**0.001**)	46.99	(0.845, 0.155)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-10 (nested within clade 3-4)					
Interior	1-33	184.25	(0.457, 0.543)	191.39	(0.478, 0.522)
Tip	1-34	0	(1.000, 1.000)	304.59	(0.980, 0.046)
Tip	1-35	0	(1.000, 1.000)	218.94	(0.659, 0.403)
Tip	1-36	0	(1.000, 1.000)	292.36	(0.934, 0.191)
Tip	1-37	0	(1.000, 1.000)	102.22	(0.298, 0.894)
Tip	1-38	0	(1.000, 1.000)	198.84	(0.591, 0.458)
Tip	1-39	0	(1.000, 1.000)	218.94	(0.660, 0.415)
Tip	1-40	**10.98**^{S}	(**0.008**, 0.012)	103.96	(0.012, 0.988)
Tip	1-41	140.41	(0.741, 0.267)	157.00	(0.316, 0.692)
Tip	1-42	62.85	(0.248, 0.753)	251.94	(0.863, 0.138)
Tip	1-43	0	(0.008, 1.000)	102.68	(0.042, 0.958)
Tip	1-44	**0**^{S}	(**<0.001**, 1.000)	292.36	(0.9730, 0.027)
Tip	1-45	150.91	(0.493, 0.507)	180.07	(0.424, 0.576)
	I-T	143.7	(0.989, 0.011)	-3.21	(0.443, 0.557)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-7 (nested within clade 3-3)					
Interior	1-22	38.16	(0.158, 0.875)	274.51	(0.041, 0.992)
Interior	1-23	11.32	(0.023, 0.982)	505.63	(0.732, 0.273)
Tip	1-24	0	(1.000, 1.000)	553.20	(1.000, 0.074)
Tip	1-25	0	(1.000, 1.000)	318.22	(0.321, 0.748)
Tip	1-26	**0**^{S}	(**0.003**, 1.000)	530.61	(0.979, 0.024)
Tip	1-27	0	(1.000, 1.000)	183.01	(0.165, 1.000)
	I-T	22.06	(0.388, 0.612)	-40.63	(0.396, 0.604)
Inference Key			1-19-20-21-YES		
Conclusion			Range Expansion with Long Distance Colonization		
Clade 3-1 (nested within clade 4-1)					
Tip	2-1	**45.31**^{S}	(**0.008**, 0.992)	**63.44**^{S}	(**0.008**, 0.992)
Tip	2-2	**9.25**^{S}	(<**0.001**, 1.000)	**174.84**^{L}	(1.000, <**0.001**)
Interior	2-3	74.14	(0.140, 0.860)	89.97	(0.338, 0.662)
	I-T	44.85	(0.935, 0.065)	-22.99	(0.158, 0.842)
Inference Key			1-2-3-5-6-(TooFew)-7-YES		
Conclusion			Isolation by Distance, Contiguous Range Expansion or Long Distance Colonization		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-2 (nested within clade 4-1)					
Tip	2-4	298.49	(0.963, 0.037)	**303.41**^{L}	(0.994, **0.006**)
Interior	2-5	169.74	(0.040, 0.960)	199.17	(0.061, 0.939)
	I-T	-128.75	(0.040, 0.960)	-104.24	(0.047, 0.953)
Inference Key			1-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-4 (nested clade within clade 4-2)					
Interior	2-10	198.00	(0.158, 0.842)	196.47	(0.149, 0.851)
Tip	2-11	343.50	(0.962, 0.038)	561.18	(0.956, 0.044)
Tip	2-12	**18.72**^{S}	(<**0.001**, 1.000)	**65.81**^{S}	(<**0.001**, 1.000)
Tip	2-13	53.00	(0.210, 0.791)	275.68	(0.858, 0.143)
Tip	2-14	0	(0.052, 1.000)	244.54	(0.764, 0.237)
Tip	I-T	116.44	(0.778, 0.222)	-10.80	(0.286, 0.714)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-1 (nested within clade 5-1)					
Tip	3-1	**95.36**^{S}	(<**0.001**, 1.000)	**388.98**^{L}	(0.985, **0.015**)
Interior	3-2	**286.17**^{S}	(**0.006**, 0.994)	**289.32**^{S}	(**0.014**, 0.986)
	I-T	**190.81**^{L}	(1.000, <**0.001**)	**-99.66**^{S}	(**0.014**, 0.986)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		
Clade 4-2 (nested within clade 5-2)					
Interior	3-3	436.10	(0.760, 0.240)	**494.78**^{L}	(0.995, **0.005**)
Tip	3-4	**239.70**^{S}	(<**0.001**, 1.000)	**345.60**^{S}	(**0.011**, 0.989)
	I-T	**196.40**^{L}	(0.994, **0.006**)	**149.18**^{L}	(0.993, **0.007**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 5-1 (Total Cladogram)					
Interior	4-1	**311.07**^{S}	(<**0.001**, 1.000)	**328.63**^{S}	(<**0.001**, 1.000)
Tip	4-2	410.80	(0.886, 0.114)	**466.26**^{L}	(1.000, <**0.001**)
	I-T	**-100.13**^{S}	(**0.002**, 0.998)	**-139.31**^{S}	(<**0.001**, 1.000)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Table B.2: Species 1 lineage B Nested Clade Analysis

The table includes all species 1 lineage B clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-15 (nested within clade 3-5)					
Interior	1-55	230.53	(0.088, 0.912)	304.36	(0.719, 0.281)
Tip	1-54	**0**^{S}	(<**0.001**, 1.000)	252.42	(0.254, 0.746)
Tip	1-56	**0**^{S}	(<**0.001**, 1.000)	274.53	(0.354, 0.646)
	I-T	**199.76**^{L}	(0.992, **0.008**)	39.04	(0.772, 0.228)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-23 (nested within clade 3-8)					
Interior	1-68	6.96	(0.080, 0.928)	**10.91**^{S}	(**0.008**, 1.000)
Interior	1-69	0	(0.057, 1.000)	13.72	(1.000, 0.057)
Tip	1-70	0	(0.454, 1.000)	13.72	(1.000, 0.424)
	I-T	3.48	(0.530, 0.491)	-1.40	(0.424, 0.597)
Inference Key			1-2-11-17-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-8 (nested within clade 4-4)					
Interior	2-21	**6.44**^{S}	(<**0.001**, 1.000)	**124.78**^{S}	(**0.002**, 0.998)
Tip	2-22	**12.08**^{S}	(<**0.001**, 1.000)	131.95	(0.955, 0.045)
	I-T	-5.64	(0.035, 0.965)	**-7.17**^{S}	(**0.020**, 0.980)
Inference Key			1-19-20-NO		
Conclusion			Inadequate Geographical Sampling		
Clade 4-3 (nested within clade 5-2)					
Tip	3-5	**283.09**^{S}	(<**0.001**, 1.000)	**334.27**^{S}	(**0.001**, 0.999)
Interior	3-6	391.51	(0.540, 0.460)	**561.76**^{L}	(0.984, **0.016**)
	I-T	108.42	(0.901, 0.099)	**227.49**^{L}	(0.991, **0.009**)
Inference Key			1-19-20-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-4 (nested within clade 5-2)					
Interior	3-7	475.97	(0.844, 0.156)	444.90	(0.822, 0.178)
Tip	3-8	**126.87**^{S}	(<**0.001**, 1.000)	**180.43**^{S}	(<**0.001**, 1.000)
Tip	3-9	188.65	(0.233, 0.767)	538.83	(0.905, 0.095)
	I-T	337.14	(0.970, 0.030)	190.79	(0.837, 0.163)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 5-2 (Total Cladogram)					
Tip	4-3	405.58	(0.186, 0.814)	440.38	(0.436, 0.564)
Interior	4-4	**307.85**^{S}	(<**0.001**, 1.000)	438.06	(0.632, 0.368)
	I-T	-97.72	(0.117, 0.883)	21.92	(0.598, 0.402)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Table B.3: Species 1 lineage D Nested Clade Analysis

The table includes all species 1 lineage D clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-25 (nested within clade 3-10)					
Tip	1-74	0	(0.168, 1.000)	**763.44**^{S}	(**0.014**, 1.000)
Interior	1-75	**1671.03**^{L}	(1.000, **0.001**)	**1690.6**^{L}	(1.000, **0.001**)
Tip	1-76	0	(0.0164, 1.000)	780.89	(0.205, 0.807)
	I-T	**1671.03**^{L}	(1.000, **0.001**)	**918.44**^{L}	(1.000, **0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-10 (nested within clade 4-5)					
Interior	2-25	1384.46	(0.744, 0.258)	1356.17	(0.885, 0.117)
Tip	2-26	**0**^{S}	(**0.019**, 1.000)	**619.30**^{S}	(**0.001**, 1.000)
	I-T	**1384.46**^{L}	(0.979, **0.023**)	**736.87**^{L}	(0.991, **0.011**)
Inference Key			1-2-3-5-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		

Table B.4: Species 1 lineage E Nested Clade Analysis

The table includes all species 1 lineage E clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-81 (nested within Clade 2-28)					
Interior	C1H154	426.4	(0.582, 0.417)	428.74	(0.572, 0.428)
Tip	C1H155	0	(1.000, 1.000)	450.51	(0.824, 0.288)
Tip	C1H156	0	(1.000, 1.000)	312.97	(0.246, 0.817)
Tip	C1H157	**14.50**^{S}	(**0.001**, 1.000)	308.27	(0.060, 0.941)
Tip	C1H158	0	(1.000, 1.000)	900.98	(1.000, 0.125)
Tip	C1H159	0	(1.000, 1.000)	283.77	(0.175, 0.885)
	I-T	419.15	(0.968, 0.032)	31.08	(0.623, 0.377)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-29 (nested within clade 3-12)					
Interior	1-90	**468.54**^{L}	(1.000, **0.001**)	**505.39**^{L}	(1.000, **0.001**)
Tip	1-91	**2.81**^{S}	(**0.001**, 1.000)	**173.21**^{S}	(**0.001**, 1.000)
	I-T	**465.73**^{L}	(1.000, **0.001**)	**332.18**^{L}	(1.000, **0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-12 (nested within clade 4-6)					
Interior	2-28	419.37	(0.262, 0.738)	428.27	(0.255, 0.745)
Tip	2-29	313.28	(0.042, 0.958)	404.08	(0.067, 0.933)
Tip	2-30	**61.29**^{S}	(<**0.001**, 1.000)	**546.41**^{L}	(0.999, **0.001**)
	I-T	**212.70**^{S}	(1.000, <**0.001**)	-36.03	(0.156, 0.844)
Inference Key			1-2-3-5-6-7-8-NO		
Conclusion			Isolation by Distance with either Short Distance Dispersal or Long Distance Dispersal		

Table B.5: Species 1 lineage F Nested Clade Analysis

The table includes all species 1 lineage F clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-105 (nested within clade 2-33)					
Interior	C1H192	**0**^{S}	(**0.001**, 1.000)	47.89	(0.501, 0.500)
Tip	C1H193	0	(1.000, 1.000)	54.12	(1.000, 0.416)
Tip	C1H194	0	(1.000, 1.000)	8.04	(0.083, 1.000)
Tip	C1H195	0	(0.039, 1.000)	**54.12**^{L}	(1.000, **0.010**)
	I-T	0	(0.276, 0.724)	1.46	(0.500, 0.500)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 2-33 (nested within clade 3-14)					
Interior	1-99	618.02	(0.009, 0.991)	804.80	(0.124, 0.876)
Tip	1-100	0	(0.007, 1.000)	**1439.62**^{L}	(0.998, **0.003**)
Tip	1-101	0	(1.000, 1.000)	933.02	(0.476, 0.547)
Tip	1-102	455.14	(0.149, 0.851)	673.55	(0.122, 0.879)
Tip	1-103	0	(1.000, 1.000)	587.12	(0.296, 0.772)
Tip	1-104	0	(1.000, 1.000)	406.82	(0.046, 0.977)
Tip	1-105	**36.68**^{S}	(<**0.001**, 1.000)	**1216.59**^{L}	(0.997, **0.003**)
Tip	1-106	0	(0.007, 1.000)	1037.76	(0.597, 0.403)
Tip	1-107	0	(0.007, 1.000)	908.54	(0.400, 0.603)
	I-T	537.28	(0.989, 0.012)	-244.42	(0.077, 0.924)
Inference Key			1-2-11-12-13-YES		
Conclusion			Range Expansion or Past Fragmentation		
Clade 3-14 (Total Cladogram)					
Tip	2-33	918.22	(0.097, 0.903)	915.56	(0.068, 0.933)
Interior	2-34	582.6	(0.048, 0.952)	**1710.84**^{L}	(0.995, **0.005**)
Tip	2-35	0	(0.064, 1.000)	658.45	(0.154, 0.848)
	I-T	-295.71	(0.208, 0.792)	**806.46**^{L}	(0.994, **0.006**)
Inference Key			1-19-20-2-11-17-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Table B.6: Species 1 lineage G Nested Clade Analysis
The table includes all species 1 lineage G clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-127 (nested within clade 2-40)					
Interior	S1H233	642.97	(0.054, 0.946)	736.12	(0.148, 0.852)
Tip	S1H234	18.99	(0.136, 0.890)	1254.38	(0.973, 0.053)
Tip	S1H235	0	(1.000, 1.000)	696.84	(0.571, 0.477)
Tip	S1H236	0	(1.000, 1.000)	183.34	(0.246, 1.000)
Tip	S1H237	**0**^{S}	(**0.010**, 1.000)	**1260.29**^{L}	(1.000, **0.001**)
	I-T	637.55	(0.916, 0.084)	-288.14	(0.026, 0.974)
Inference Key			1-2-3-5-15-NO		
Conclusion			Past Fragmentation and/or Long Distance Colonization		
Clade 2-40 (nested within clade 3-15)					
Interior	1-127	825.09	(0.875, 0.125)	809.75	(0.854, 0.146)
Tip	1-128	**0**^{S}	(<**0.001**, 1.000)	491.80	(0.043, 0.957)
	I-T	**825.09**^{L}	(1.000, <**0.001**)	317.94	(0.942, 0.058)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-43 (nested within clade 3-15)					
Interior	1-131	47.06	(0.644, 0.356)	64.92	(0.973, 0.027)
Tip	1-132	**9.48**^{S}	(<**0.001**, 1.000)	**16.25**^{S}	(<**0.001**, 1.000)
Tip	1-133	0	(1.000, 1.000)	107.25	(1.000, 0.069)
	I-T	38.63	(0.956, 0.044)	38.55	(0.969, 0.031)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-15 (nested within clade 4-6)					
Tip	2-36	205.83	(0.167, 0.833)	545.78	(0.539, 0.461)
Interior	2-37	0	(0.050, 1.000)	**1521.23**^{L}	(1.000, <**0.001**)
Interior	2-38	**448.93**^{S}	(<**0.001**, 1.000)	**444.40**^{S}	(<**0.001**, 1.000)
Tip	2-39	149.26	(0.061, 0.939)	**1087.50**^{L}	(1.000, **0.003**)
	I-T	254.05	(0.797, 0.203)	**-359.66**^{S}	(**0.010**, 0.990)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-16 (nested within clade 4-6)					
Interior	2-40	**763.22**^{L}	(0.997, **0.003**)	**688.89**^{L}	(0.991, **0.009**)
Tip	2-41	0	(0.053, 1.000)	451.14	(0.561, 0.439)
Tip	2-42	436.90	(0.722, 0.279)	1013.87	(0.908, 0.093)
Tip	2-43	**42.50**^{S}	(<**0.001**, 1.000)	**121.87**^{S}	(<**0.001**, 1.000)
	I-T	**687.20**^{L}	(1.000, <**0.001**)	**450.73**^{L}	(0.999, **0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-17 (nested within clade 4-6)					
Interior	2-44	**761.49**^{L}	(0.995, **0.006**)	**758.41**^{L}	(1.000, **0.001**)
Tip	2-45	**18.85**^{S}	(**0.002**, 0.999)	**281.36**^{S}	(**0.001**, 1.000)
	I-T	**742.63**^{L}	(1.000, **0.001**)	**477.05**^{L}	(1.000, **0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-6 (Total Cladogram)					
Interior	3-15	**571.82**^{S}	(**0.004**, 0.996)	**591.43**^{S}	(**0.004**, 0.996)
Interior	3-16	**533.88**^{S}	(<**0.001**, 1.000)	761.25	(0.725, 0.275)
Tip	3-17	472.18	(0.051, 0.949)	**1035.88**^{L}	(0.994, **0.006**)
	I-T	81.27	(0.588, 0.412)	**-362.22**^{S}	(**0.005**, 0.995)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Table B.7: Species 1 lineage H Nested Clade Analysis
The table includes all species 1 lineage H clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	No	Dc	Prob	Dn	Prob
Clade 2-46 (nested within clade 3-18)					
Tip	1-141	528.18	(0.773, 0.228)	**653.52**^{L}	(0.995, **0.005**)
Interior	1-142	**133.90**^{S}	(**0.011**, 0.989)	314.63	(0.026, 0.974)
Tip	1-143	555.52	(0.830, 0.173)	533.48	(0.816, 0.188)
	I-T	**-402.69**^{S}	(**0.012**, 0.988)	**-301.95**^{S}	(**0.007**, 0.994)
Inference Key			1-2-11-12-13-YES		
Conclusion			Range Expansion with Long Distance Colonization		
Clade 2-47 (nested within clade 3-18)					
Interior	1-144	0	(0.133, 1.000)	184.92	(0.132, 0.934)
Tip	1-145	0	(1.000. 1.000)	184.92	(0.408, 0.898)
Tip	1-146	305.61	(0.974, 0.035)	304.90	(0.974, 0.035)
	I-T	**-267.41**^{S}	(**0.017**, 0.992)	-104.98	(0.116, 0.892)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-19 (nested within clade 4-8)					
Tip	2-48	**52.95**^{S}	(<**0.001**, 1.000)	201.66	(0.026, 0.974)
Interior	2-49	111.61	(0.059, 0.942)	237.48	(0.845, 0.155)
	I-T	**58.66**^{L}	(0.999, **0.002**)	35.82	(0.929, 0.071)
Inference Key			1-29-20-2-3-4-9-10-NO		
Conclusion			Past Fragmentation or Isolation by Distance		
Clade 3-20 (nested within clade 4-8)					
Interior	2-50	103.68	(0.478, 0.524)	121.81	(0.863, 0.137)
Tip	2-51	**0**^{S}	(<**0.001**, 1.000)	**57.44**^{S}	(<**0.001**, 1.000)
Tip	2-52	**0**^{S}	(**0.010**, 1.000)	**147.46**^{L}	(1.000, <**0.001**)
	I-T	**103.68**^{L}	(1.000, <**0.001**)	**32.21**^{L}	(0.995, **0.005**)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		
Clade 4-8 (nested within clade 5-3)					
Tip	3-18	397.91	(0.903, 0.097)	415.25	(0.960, 0.040)
Interior	3-19	219.44	(0.126, 0.874)	286.07	(0.346, 0.654)
Tip	3-20	**115.35**^{S}	(<**0.001**, 1.000)	**140.82**^{S}	(**0.001**, 0.999)
Tip	3-21	521.41	(0.893, 0.107)	463.38	(0.823, 0.172)
	I-T	-97.38	(0.214, 0.786)	-42.81	(0.372, 0.627)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 5-3 (Total Cladogram)					
Interior	4-8	**326.62**^{S}	(<**0.001**, 1.000)	**367.38**^{S}	(<**0.001**, 0.999)
Tip	4-9	320.24	(0.096, 0.902)	**625.77**^{L}	(0.988, **0.012**)
	I-T	6.38	(0.472, 0.528)	**-258.39**^{S}	(**0.007**, 0.993)
Inference Key		1-2-11-12-13-YES			
Conclusion		Range Expansion with Long Distance Colonization			

Table B.8: Species 1 lineage I Nested Clade Analysis

The table includes all species 1 lineage I clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-175 (nested within clade 2-63)					
Interior	S1H321	7.82	(0.167, 0.833)	52.33	(0.042, 0.958)
Tip	S1H322	0	(0.233, 1.000)	141.52	(1.000, 0.016)
Tip	S1H323	0	(1.000, 1.000)	141.52	(1.000, 0.152)
Tip	S1H324	**0**^{S}	(**0.001**, 1.000)	50.49	(0.017, 0.983)
Tip	S1H325	**0**^{S}	(**0.009**, 1.000)	57.63	(0.487, 0.519)
	I-T	7.83	(0.332, 0.668)	-23.29	(0.193, 0.806)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-63 (nested wihin clade 3-24)					
Interior	1-175	**76.75**^{S}	(**0.008**, 0.992)	**231.88**^{S}	(**0.007**, 0.993)
Tip	1-176	0	(0.009, 1.000)	229.24	(0.262, 0.740)
Tip	1-177	**0**^{S}	(<**0.001**, 1.000)	286.07	(0.635, 0.365)
Tip	1-178	0	(1.000, 1.000)	236.66	(0.540, 0.593)
Tip	1-179	194.39	(0.580, 0.420)	258.03	(0.548, 0.452)
Tip	1-180	42.6	(0.127, 0.874)	249.69	(0.545, 0.456)
Tip	1-181	**0**^{S}	(<**0.001**, 1.000)	243.09	(0.419, 0.581)
Tip	1-182	**0**^{S}	(**0.007**, 1.000)	**1954.76**^{L}	(1.000, <**0.001**)
	I-T	25.33	(0.460, 0.540)	-220.13	(0.161, 0.839)
Inference Key			1-2-3-5-6-13-YES		
Conclusion			Coupled with Subsequent Extinction in Some Intermediate Geographical Areas, or Past Range Expansion		
Clade 3-26 (nested within clade 4-10)					
Interior	2-64	36.69	(0.799, 0.202)	36.34	(0.791, 0.209)
Tip	2-65	**2.76**^{S}	(<**0.001**, 1.000)	29.58	(0.230, 0.770)
	I-T	**33.93**^{L}	(1.000, <**0.001**)	6.76	(0.781, 0.219)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-26 (nested within clade 4-10)					
Interior	2-64	36.69	(0.799, 0.202)	36.34	(0.791, 0.209)
Tip	2-65	**2.76**^{S}	(<**0.001**, 1.000)	29.58	(0.230, 0.770)
	I-T	**33.93**^{L}	(1.000, <**0.001**)	6.76	(0.781, 0.219)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-28 (nested within clade 4-10)					
Interior	2-69	0	(0.251, 1.000)	**51.40**^{L}	(1.000, **0.012**)
Tip	2-70	**0**^{S}	(**0.012**, 1.000)	**32.27**^{S}	(**0.012**, 1.000)
	I-T	0	(0.786, 0.761)	**19.13**^{L}	(1.000, **0.012**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-10 (Total Cladogram)					
Node	3-24	**378.06**^{S}	(<**0.001**, 1.000)	**744.45**^{S}	(**0.001**, 0.999)
Tip	3-25	**0**^{S}	(<**0.001**, 1.000)	**1559.44**^{L}	(1.000, <**0.001**)
Tip	3-26	**34.21**^{S}	(<**0.001**, 1.000)	674.45	(0.010, 0.990)
Tip	3-27	**39.63**^{S}	(<**0.001**, 1.000)	664.94	(0.053, 0.948)
Tip	3-28	**0.73**^{S}	(<**0.001**, 1.000)	**1265.64**^{L}	(0.999, <**0.001**)
Tip	3-29	697.28	(0.201, 0.800)	1027.55	(0.925, 0.075)
	I-T	158.62	(0.773, 0.227)	**-222.73**^{S}	(**0.004**, 0.996)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		

Table B.9: Species 2 Nested Clade Analysis

The table includes all species 2 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-26 (nested within clade 2-9)					
Tip	S2H214	0	(0.050, 1.000)	105.98	(0.139, 0.877)
Tip	S2H215	9.31	(0.061, 0.949)	**74.19**^{S}	(**0.010**, 1.000)
Tip	S2H216	479.82	(0.966, 0.044)	506.07	(0.854, 0.056)
Tip	S2H217	0	(1.000, 1.000)	105.98	(0.374, 0.782)
Interior	S2H218	332.9	(0.808, 0.192)	332.27	(0.770, 0.230)
	I-T	193.14	(0.793, 0.207)	121.07	(0.881, 0.119)
Inference Key			1-2-11-17-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-40 (nested within clade 2-15)					
Tip	S2H045	**0**^{S}	(**0.003**, 1.000)	**156.63**^{S}	(**0.001**, 1.000)
Tip	S2H046	0	(1.000, 1.000)	296.43	(0.810, 0.306)
Tip	S2H047	0	(1.000, 1.000)	236.04	(0.644, 0.511)
Tip	S2H048	0	(1.000, 1.000)	244.16	(0.677, 0.370)
Tip	S2H049	0	(1.000, 1.000)	382.25	(0.951, 0.149)
Interior	S2H050	243.63	(0.552, 0.448)	243.59	(0.428, 0.572)
	I-T	**243.63**^{L}	(0.996, **0.004**)	10.91	(0.615, 0.385)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-42 (nested within clade 2-16)					
Tip	S2H038	**133.47**^{S}	(**0.013**, 1.000)	186.98	(0.158, 0.855)
Interior	S2H039	0	(0.124, 1.000)	301.03	(0.728, 0.359)
Tip	S2H040	0	(1.000, 1.000)	301.23	(0.742, 0.608)
	I-T	-111.22	(0.234, 0.779)	95.04	(0.712, 0.301)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-48 (nested within clade 2-17)					
Tip	S1H001	**0**^{S}	(**0.003**, 1.000)	1063.08	(0.354, 0.646)
Interior	S1H002	1040.18	(0.510, 0.490)	988.77	(0.404, 0.596)
Tip	S1H003	0	(1.000, 1.000)	453.67	(0.168, 0.878)
Tip	S1H016	**198.88**^{S}	(<**0.00**1, 1.000)	1154.58	(0.579, 0.421)
Tip	S1H017	0	(1.000, 1.000)	376.88	(0.123, 0.962)
Tip	S1H019	0	(1.000, 1.000)	2080.68	(1.000, 0.056)
	I-T	933.09	(0.971, 0.029)	-102.19	(0.468, 0.532)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-64 (nested within clade 2-22)					
Interior	S2H025	**404.07**^{L}	(1.000, <**0.001**)	435.85	(0.966, 0.034)
Interior	S2H027	**14.34**^{S}	(<**0.00**1, 1.000)	**182.76**^{S}	(<**0.00**1, 1.000)
Tip	S2H028	0	(1.000, 1.000)	186.71	(0.747, 0.486)
	I-T	**116.01**^{S}	(<**0.00**1, 1.000)	62.07	(0.253, 0.747)
Inference Key			1-2-3-5-6-too few-7-8		
Conclusion			Isolation by Distance with Long Distance Dispersal or Range		
Clade 1-75 (nested within clade 2-28)					
Tip	S2H074	0	(1.000, 1.000)	277.97	(0.119, 1.000)
Interior	S2H075	405.70	(0.052, 0.948)	489.95	(0.598, 0.402)
Tip	S2H076	**0**^{S}	(<**0.001**, 1.000)	**796.99**^{L}	(1.000, <**0.001**)
Tip	S2H077	0	(1.000, 1.000)	491.13	(0.718, 0.416)
Tip	S2H078	0	(0.068, 1.000)	**277.97**^{S}	(**0.004**, 1.000)
Tip	S2H079	**38.24**^{S}	(**0.002**, 0.998)	356.07	(0.048, 0.952)
	I-T	**382.95**^{L}	(0.997, **0.003**)	-4.82	(0.497, 0.503)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		
Clade 1-88 (nested within clade 2-33)					
Tip	S2H116	**0**^{S}	(<**0.001**, 1.000)	1714.69	(0.854, 0.146)
Tip	S2H120	0	(0.078, 1.000)	814.75	(0.245, 0.757)
Tip	S2H121	0	(0.068, 1.000)	632.09	(0.210, 0.795)
Tip	S2H122	0	(1.000, 1.000)	411.46	(0.135, 0.954)
Tip	S2H123	0	(0.008, 1.000)	1506.00	(0.853, 0.150)
Tip	S2H124	0	(1.000, 1.000)	443.84	(0.168, 0.864)
Tip	S2H125	0	(0.009, 1.000)	1284.17	(0.765, 0.235)
Interior	S2H126	2213.40	(0.826, 0.174)	1144.9	(0.573, 0.427)
	I-T	**2213.40**^{L}	(1.000, <**0.001**)	-72.33	(0.539, 0.461)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-89 (nested within clade 2-33)					
Interior	S2H097	**43.69**^{S}	(**0.012**, 1.000)	80.46	(0.944, 0.068)
Interior	S2H100	0	(0.322, 1.000)	66.91	(0.202, 0.898)
Tip	S2H101	0	(1.000, 1.000)	66.91	(0.511, 0.870)
	I-T	**31.20**^{S}	(**0.012**, 1.000)	9.68	(0.870, 0.142)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-1 (nested within clade 3-1)					
Interior	1-1	**503.48**^{L}	(1.000, <**0.001**)	**511.32**^{L}	(1.000, <**0.001**)
Tip	1-2	0	(1.000, 1.000)	593.15	(0.883, 0.134)
Tip	1-3	0	(1.000, 1.000)	217.39	(0.345, 0.691)
Tip	1-4	**164.97**^{S}	(<**0.001**, 1.000)	**194.51**^{S}	(<**0.001**, 1.000)
	I-T	**351.71**^{L}	(1.000, <**0.001**)	**299.95**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-2 (nested within clade 3-1)					
Interior	1-5	333.46	(0.065, 0.935)	**344.71**^{S}	(**0.013**, 0.987)
Tip	1-6	**1035.17**^{L}	(0.992, **0.008**)	**1164.16**^{L}	(1.000, <**0.001**)
Tip	1-7	14.21	(0.165, 0.846)	631.08	(0.342, 0.669)
	I-T	-446.47	(0.079, 0.921)	**-686.18**^{S}	(**0.001**, 0.999)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 2-7 (nested within clade 3-3)					
Tip	1-19	0	(1.000, 1.000)	251.36	(0.741, 0.392)
Tip	1-20	0	(1.000, 1.000)	398.99	(1.000, 0.052)
Interior	1-21	192.82	(0.081, 0.919)	217.35	(0.104, 0.896)
Tip	1-22	0	(1.000, 1.000)	236.37	(0.630, 0.615)
Tip	1-23	**0**^{S}	(**0.010**, 1.000)	236.67	(0.597, 0.409)
	I-T	192.83	(0.908, 0.092)	-48.62	(0.099, 0.901)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-16 (nested within clade 3-8)					
Tip	1-41	0	(1.000, 1.000)	757.52	(0.679, 0.427)
Interior	1-42	**217.60**^{S}	(**0.009**, 1.000)	723.94	(0.616, 0.393)
Tip	1-43	0	(0.084, 1.000)	655.32	(0.132, 0.887)
	I-T	217.6	(0.430, 0.573)	34.56	(0.888, 0.115)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-17 (nested within clade 3-9)					
Tip	1-44	878.95	(0.680, 0.328)	1305.18	(0.679, 0.329)
Tip	1-45	545.78	(0.284, 0.716)	813.21	(0.188, 0.812)
Tip	1-46	0	(1.000, 1.000)	2082.41	(0.977, 0.075)
Tip	1-47	114.01	(0.014, 0.986)	1174.66	(0.501, 0.499)
Interior	1-48	1043.17	(0.379, 0.621)	1082.28	(0.502, 0.498)
Tip	1-49	0	(1.000, 1.000)	1443.09	(0.788, 0.236)
Tip	1-50	0	(1.000, 1.000)	1954.87	(0.926, 0.205)
Tip	1-51	0	(1.000, 1.000)	2095.75	(1.000, 0.021)
Tip	1-52	**32.50**^{S}	(<**0.001**, 1.000)	**168.89**^{S}	(**0.010**,0.999)
	I-T	801.47	(0.937, 0.063)	30.67	(0.647, 0.353)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-21 (nested within clade 3-9)					
Interior	1-60	**166.48**^{S}	(**0.001**, 1.000)	398.8	(0.460, 0.541)
Tip	1-61	**17.84**^{S}	(**0.003**, 1.000)	**297.15**^{S}	(**0.010**, 0.993)
Tip	1-62	0	(1.000, 1.000)	738.51	(1.000, 0.090)
	I-T	152.21	(0.551, 0.449)	13.38	(0.495, 0.505)
Inference Key			1-19-20-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 2-25 (nested within clade 3-11)					
Tip	1-69	0	(1.000, 1.000)	478.94	(0.496, 0.607)
Interior	1-70	530.52	(0.941, 0.059)	525.06	(0.931, 0.069)
Tip	1-71	**136.38**^{S}	(**0.016**, 0.985)	**290.60**^{S}	(**0.007**, 0.994)
	I-T	416.86	(0.981, 0.019)	203.08	(0.963, 0.037)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-28 (nested within clade 3-13)					
Interior	1-75	484.73	(0.033, 0.967)	511.40	(0.096, 0.904)
Tip	1-76	**0**^{S}	(<**0.001**, 1.000)	**208.22**^{S}	(**0.006**, 0.994)
Tip	1-77	**226.24**^{S}	(**0.010**, 0.990)	**967.56**^{L}	(0.997, **0.003**)
Tip	1-78	0	(1.000, 1.000)	531.01	(0.389, 0.636)
	I-T	226.95	(0.753, 0.247)	**-426.71**^{S}	(**0.003**, 0.997)
Inference Key			1-2-3-5-15-NO		
Conclusion			Past Fragmentation and/or Long Distance Colonization		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-33 (nested within clade 3-15)					
Interior	1-88	1139.2	(0.738, 0.262)	1045.61	(0.969, 0.031)
Tip	1-89	**73.85**^{S}	(<**0.001**, 1.000)	606.81	(0.087, 0.913)
Tip	1-90	**0**^{S}	(**0.002**, 1.000)	950.36	(0.777, 0.223)
Tip	1-91	**0**^{S}	(<**0.001**, 1.000)	605.62	(0.104, 0.896)
Tip	1-92	**0**^{S}	(<**0.001**, 1.000)	**466.32**^{S}	(<**0.001**, 1.000)
Tip	1-93	**0**^{S}	(**0.001**, 1.000)	526.49	(0.107, 0.893)
	I-T	**1114.59**^{L}	(1.000, <**0.001**)	**429.61**^{L}	(0.993, **0.007**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-1 (nested clade within 4-1)					
Tip	2-1	**397.35**^{S}	(**0.004**, 0.996)	**411.95**^{S}	(**0.004**, 0.995)
Interior	2-2	635.03	(0.908, 0.092)	**888.88**^{L}	(1.000, <**0.001**)
	I-T	237.69	(0.956, 0.044)	**476.92**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-3 (nested within clade 4-2)					
Interior	2-5	**555.79**^{L}	(0.992, **0.012**)	704.00	(0.897, 0.107)
Tip	2-6	**235.13**^{S}	(**0.016**, 0.984)	**336.90**^{S}	(**0.002**, 0.998)
Interior	2-7	360.25	(0.905, 0.126)	766.77	(0.971, 0.060)
	I-T	**222.89**^{L}	(0.997, **0.003**)	**398.49**^{L}	(0.998, **0.002**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-4 (nested within clade 4-2)					
Interior	2-8	417.21	(0.303, 0.697)	878.24	(0.961, 0.039)
Tip	2-9	**293.99**^{S}	(<**0.001**, 1.000)	**483.96**^{S}	(<**0.001**, 1.000)
	I-T	123.21	(0.781, 0.219)	**394.29**^{L}	(0.990, **0.010**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-8 (nested within clade 4-4)					
Tip	2-15	**300.45**^{S}	(**0.006**, 0.994)	**328.87**^{S}	(<**0.001**, 1.000)
Interior	2-16	**713.01**^{L}	(0.999, **0.001**)	**714.72**^{L}	(1.000, <**0.001**)
	I-T	**412.55**^{L}	(1.000, <**0.001**)	**385.84**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-9 (nested within clade 4-4)					
Interior	2-17	1084.11	(0.093, 0.907)	1247.49	(0.725, 0.275)
Tip	2-18	212.69	(0.158, 0.843)	1004.44	(0.332, 0.669)
Tip	2-19	**0**^{S}	(<**0.001**, 1.000)	**1689.33**^{L}	(0.998, **0.002**)
Tip	2-20	523.84	(0.098, 0.902)	**492.90**^{S}	(**0.003**, 0.997)
Tip	2-21	**415.05**^{S}	(<**0.001**, 1.000)	**447.78**^{S}	(<**0.001**, 1.000)
Tip	2-22	**310.61**^{S}	(<**0.001**, 1.000)	**1457.84**^{L}	(0.997, **0.003**)
	I-T	**768.24**^{L}	(1.000, <**0.001**)	96.31	(0.784, 0.216)
Inference Key			1-2-3-5-6-13-YES		
Conclusion			Long Distance Dispersal or Allopatric Fragmentation		
Clade 3-11 (nested clade within 4-4)					1-2-
Tip	2-24	330.73	(0.123, 0.877)	**1231.66**^{L}	(1.000, <**0.001**)
Interior	2-25	**489.78**^{S}	(<**0.001**, 1.000)	**591.59**^{S}	(<**0.001**, 1.000)
	I-T	159.05	(0.600, 0.400)	**-640.07**^{S}	(<**0.001**, 1.000)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-15 (nested within clade 4-7)					
Interior	2-33	**866.20**^{S}	(**0.002**, 0.998)	**1252.4**^{S}	(**0.002**, 0.998)
Tip	2-34	4505.21	(0.976, 0.024)	2638.96	(0.983, 0.017)
Tip	2-35	426.54	(0.154, 0.846)	1016.03	(0.178, 0.822)
	I-T	-2007.54	(0.055, 0.945)	-737.05	(0.064, 0.936)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-16 (nested within clade 4-7)					
Interior	2-36	**133.13**^{S}	(**0.001**, 0.999)	**532.03**^{S}	(<**0.001**, 1.000)
Tip	2-37	**71.98**^{S}	(<**0.001**, 1.000)	**881.13**^{L}	(1.000, <**0.001**)
	I-T	61.15	(0.712, 0.288)	**-349.09**^{S}	(<**0.001**, 1.000)
Inference Key			1-19-20-2-11-12-13-YES		
Conclusion			Long Distance Dispersal or Allopatric Fragmentation		
Clade 4-1 (nested within clade 5-1)					
Tip	3-1	**506.76**^{S}	(<**0.001**, 1.000)	**518.47**^{S}	(<**0.001**, 1.000)
Interior	3-2	**939.19**^{L}	(0.997, **0.003**)	**925.37**^{L}	(0.999, **0.001**)
	I-T	**432.43**^{L}	(0.998, **0.002**)	**406.89**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 4-2 (nested within clade 5-1)					
Interior	3-3	**385.30**^{S}	(**0.006**, 0.994)	478.54	(0.080, 0.920)
Tip	3-4	631.78	(0.646, 0.354)	745.56	(0.969, 0.031)
	I-T	-246.48	(0.060, 0.940)	-267.02	(0.057, 0.943)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 4-3 (nested within clade 5-1)					
Interior	3-5	1185.97	(0.371, 0.629)	1330.75	(0.649, 0.351)
Tip	3-6	60.02	(0.089, 0.928)	1330.60	(0.571, 0.446)
Tip	3-7	**219.71**^{S}	(<**0.001**, 1.000)	1210.54	(0.136, 0.864)
	I-T	**1001.75**^{L}	(1.000, <**0.001**)	93.53	(0.762, 0.238)
Inference Key			1-19-20-2-3-4-9-NO		
Conclusion			Allopatric Fragmentation		
Clade 4-4 (nested within clade 5-2)					
Tip	3-8	**468.50**^{S}	(<**0.001**, 1.000)	**823.43**^{S}	(<**0.001**, 1.000)
Interior	3-9	1218.91	(0.968, 0.032)	**1270.00**^{L}	(1.000, <**0.001**)
	I-T	**750.41**^{L}	(1.000, <**0.001**)	**446.57**^{L}	(1.000, <**0.001**)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 4-7 (nested within clade 5-3)					
Tip	3-14	**581.67**^{S}	(<**0.001**, 1.000)	**618.36**^{S}	(<**0.001**, 1.000)
Interior	3-15	1463.5	(0.853, 0.147)	**1460.96**^{L}	(1.000, <**0.001**)
Tip	3-16	613.23	(0.063, 0.937)	**553.78**^{S}	(**0.006**, 0.994)
	I-T	**872.54**^{L}	(0.985, **0.015**)	**861.60**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal or Range		
Clade 5-1 (nested within clade 6-1)					
Interior	4-1	**692.70**^{S}	(**0.004**, 0.996)	745.94	(0.022, 0.978)
Tip	4-2	**615.98**^{S}	(**0.009**, 0.991)	778.44	(0.304, 0.696)
Interior	4-3	**1308.35**^{L}	(1.000, <**0.001**)	**1287.94**^{L}	(1.000, <**0.001**)
	I-T	175.85	(0.926, 0.074)	54.77	(0.673, 0.327)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal or Range Expansion with Long Distance Colonization		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 5-2 (nested within clade 6-1)					
Tip	4-4	**1119.23**^{L}	(0.994, **0.006**)	**1120.55**^{L}	(0.998, **0.002**)
Interior	4-5	**737.00**^{S}	(<**0.001**, 1.000)	**832.87**^{S}	(**0.002**, 0.998)
	I-T	**-382.23**^{S}	(<**0.001**, 1.000)	**-287.68**^{S}	(**0.002**, 0.998)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 5-3 (nested within clade 6-1)					
Interior	4-6	**681.59**^{S}	(<**0.001**, 1.000)	**817.80**^{S}	(**0.006**, 0.994)
Tip	4-7	1109.14	(0.639, 0.361)	**1204.84**^{L}	(0.996, **0.004**)
	I-T	**-427.54**^{S}	(**0.013**, 0.987)	**-387.33**^{S}	(**0.005**, 0.995)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 6-1 (Total Cladogram)					
Tip	5-1	**831.62**^{S}	(<**0.001**, 1.000)	**836.78**^{S}	(<**0.001**, 1.000)
Interior	5-2	1048.74	(0.473, 0.527)	1132.36	(0.932, 0.068)
Tip	5-3	1083.40	(0.707, 0.293)	**1216.86**^{L}	(0.997, **0.003**)
	I-T	-138.72	(0.055, 0.945)	**-226.17**^{S}	(**0.006**, 0.994)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal or Range Expansion with Long Distance Colonization		

Table B.10: Species 3 Nested Clade Analysis

The table includes all species 3 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-47 (nested within clade 2-21)					
Tip	S3H178	**38.87**^{S}	(<**0.001**, 1.000)	**306.89**^{L}	(1.000, <**0.001**)
Tip	S3H179	0	(1.000, 1.000)	93.93	(0.416, 0.929)
Interior	S3H180	**74.01**^{S}	(<**0.001**, 1.000)	**130.27**^{S}	(<**0.001**, 1.000)
	I-T	39.99	(0.578, 0.422)	**-150.01**^{S}	(<**0.001**, 1.000)
Inference Key			1-19-20-2-11-12-13-14-YES		
Conclusion			Contiguous Range Expansion, Long Distance Colonization or Past Fragmentation		
Clade 1-50 (nested within clade 2-23)					
Interior	S3H117	0	(0.308, 1.000)	**890.97**^{L}	(1.000, **0.013**)
Tip	S3H118	**46.70**^{S}	(**0.013**, 1.000)	**734.92**^{S}	(**0.013**, 1.000)
	I-T	**-46.70**^{L}	(1.000, **0.013**)	**156.05**^{L}	(1.000, **0.013**)
Inference Key			1-19-20-NO		
Conclusion			Incomplete Geographically Sampling		
Clade 1-68 (nested withn clade 2-30)					
Tip	S3H027	57.43	(0.446, 0.583)	570.54	(1.000, 0.029)
Interior	S3H028	0	(1.000, 1.000)	197.97	(0.105, 1.000)
Tip	S3H029	**0**^{S}	(**0.008**, 1.000)	433.24	(0.684, 0.324)
	I-T	**-14.36**^{L}	(1.000, **0.004**)	-269.59	(0.105, 0.324)
Inference Key			1-19-20-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-90 (nested within clade 2-39)					
Tip	S3H016	**0**^{S}	(**0.004**, 1.000)	244.01	(0.056, 0.948)
Tip	S3H017	0	(1.000, 1.000)	187.07	(0.105, 1.000)
Interior	S3H018	0	(0.450, 1.000)	815.01	(1.000, 0.020)
	I-T	0	(0.618, 0.383)	579.13	(0.981, 0.020)
Inference Key			1-19-20-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-95 (nested within clade 2-43)					
Tip	S3H001	0	(1.000, 1.000)	708.65	(1.000, 0.052)
Interior	S3H002	**147.16**^{S}	(<**0.001**, 1.000)	299.23	(0.621, 0.379)
Tip	S3H003	**146.90**^{S}	(**0.003**, 0.998)	293.54	(0.520, 0.481)
Tip	S3H005	0	(1.000, 1.000)	83.22	(0.091, 1.000)
	I-T	29.64	(0.336, 0.664)	-14.81	(0.458, 0.542)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-14 (nested within clade 3-6)					
Interior	1-34	211.52	(0.524, 0.477)	222.39	(0.581, 0.420)
Tip	1-35	0	(1.000, 1.000)	303.58	(1.000, 0.183)
Tip	1-36	**0**^{S}	(<**0.001**, 1.000)	**166.11**^{S}	(**0.019**, 0.981)
	I-T	**211.52**^{L}	(1.000, <**0.001**)	44.82	(0.877, 0.123)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-24 (nested within clade 3-12)					
Interior	1-49	603.65	(0.055, 0.945)	636.77	(0.036, 0.964)
Tip	1-50	0	(0.056, 1.000)	400.24	(0.033, 0.967)
Tip	1-51	**0**^{S}	(**0.006**, 1.000)	**1704.94**^{L}	(1.000, <**0.001**)
Tip	1-52	0	(1.000, 1.000)	1543.14	(0.953, 0.061)
Tip	1-53	68.65	(0.141, 0.859)	680.76	(0.569, 0.431)
Tip	1-54	180.11	(0.321, 0.681)	451.23	(0.065, 0.937)
Tip	1-73	**171.99**^{S}	(**0.009**, 0.991)	627.33	(0.391, 0.609)
	I-T	508.06	(0.976, 0.024)	-206.13	(0.053, 0.947)
Inference Key			1-2-3-5-6-7-8-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance or Lond		
Clade 2-37 (nested within clade 3-12)					
Tip	1-83	**56.32**^{S}	(**0.009**, 0.992)	294.53	(0.402, 0.599)
Interior	1-84	370.73	(0.617, 0.384)	345.13	(0.582, 0.419)
	I-T	314.42	(0.951, 0.050)	50.61	(0.594, 0.407)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-42 (nested within clade 3-14)					
Tip	1-90	**190.50**^{S}	(<**0.001**, 1.000)	**226.37**^{S}	(**0.004**, 0.996)
Interior	1-91	203.61	(0.197, 0.803)	277.68	(0.947, 0.053)
	I-T	13.12	(0.780, 0.220)	51.30	(0.969, 0.031)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-43 (nested within clade 3-15)					
Interior	1-92	291.62	(0.059, 0.941)	306.11	(0.212, 0.788)
Tip	1-93	**0**^{S}	(<**0.001**, 1.000)	343.41	(0.706, 0.294)
Tip	1-94	0	(1.000, 1.000)	623.27	(0.974, 0.051)
Tip	1-95	**0**^{S}	(<**0.001**, 1.000)	320.29	(0.555, 0.445)
Tip	1-96	193.01	(0.393, 0.620)	262.66	(0.172, 0.841)
	I-T	**256.53**^{L}	(1.000, <**0.001**)	-39.65	(0.176, 0.824)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-44 (nested within clade 3-15)					
Interior	1-97	45.32	(0.207, 0.808)	**176.58**^{L}	(1.000, **0.015**)
Tip	1-98	**0**^{S}	(**0.015**, 1.000)	**130.31**^{S}	(**0.015**, 1.000)
	I-T	**45.32**^{L}	(1.000, **0.015**)	**46.27**^{L}	(1.000, **0.015**)
Inference Key			1-19-20-2-3-4-9-NO		
Conclusion			Allopatric Fragmentation		
Clade 3-1 (nested within clade 4-1)					
Tip	2-1	**689.94**^{L}	(0.993, **0.023**)	**814.10**^{L}	(1.000, **0.016**)
Interior	2-2	319.33	(0.089, 0.927)	**392.76**^{S}	(**0.023**, 0.993)
	I-T	-370.61	(**0.023**, 0.993)	**-421.34**^{S}	(**0.016**, 1.000)
Inference Key			1-2-11-17-NO		
Conclusion			Inconclusive Outcome		
Clade 3-2 (nested within clade 4-1)					
Tip	2-3	0	(1.000, 1.000)	735.52	(1.000, 0.132)
Interior	2-4	53.94	(0.095, 0.961)	519.20	(0.056, 1.000)
Interior	2-5	612.13	(0.622, 0.408)	649.49	(0.774, 0.256)
	I-T	426.07	(0.240, 0.770)	**-129.46**^{S}	(**0.010**, 1.000)
Inference Key			1-2-11-17-NO		
Conclusion			Inconclusive Outcome		
Clade 3-5 (nested within clade 4-3)					
Interior	2-10	**161.18**^{S}	(**0.016**, 0.991)	348.55	(0.542, 0.465)
Tip	2-11	236.16	(0.364, 0.653)	**247.19**^{S}	(**0.017**, 1.000)
Interior	2-12	413.13	(0.733, 0.292)	565.71	(0.899, 0.126)
	I-T	33.00	(0.342, 0.659)	**194.43**^{L}	(0.986, **0.015**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-6 (nested within clade 4-3)					
Tip	2-13	**215.35**^{S}	(**0.012**, 0.988)	**224.19**^{S}	(**0.001**, 0.999)
Tip	2-14	**204.76**^{S}	(<**0.001**, 1.000)	547.72	(0.907, 0.093)
Interior	2-15	501.86	(0.785, 0.215)	541.00	(0.897, 0.103)
	I-T	**293.57**^{L}	(0.996, **0.004**)	101.13	(0.820, 0.180)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-12 (nested within clade 4-4)					
Interior	2-24	722.92	(0.628, 0.372)	**823.13**^{L}	(0.997, **0.003**)
Tip	2-25	**41.93**^{S}	(<**0.001**, 1.000)	875.28	(0.838, 0.162)
Tip	2-26	**86.96**^{S}	(<**0.001**, 1.000)	416.72	(0.022, 0.978)
Tip	2-27	351.69	(0.012, 0.988)	**358.70**^{S}	(**0.004**, 0.996)
Tip	2-28	557.20	(0.751, 0.250)	554.81	(0.281, 0.720)
Tip	2-29	**132.99**^{S}	(<**0.001**, 1.000)	897.05	(0.920, 0.080)
Tip	2-30	446.15	(0.005, 0.995)	668.84	(0.352, 0.965)
Tip	2-31	**82.91**^{S}	(<**0.001**, 1.000)	397.55	(0.035, 0.965)
Tip	2-32	**0**^{S}	(<**0.001**, 1.000)	**204.19**^{S}	(<**0.001**, 1.000)
Tip	2-33	292.98	(0.026, 0.974)	654.63	(0.409, 0.591)
Tip	2-34	**31.65**^{S}	(<**0.001**, 1.000)	599.05	(0.266, 0.734)
Tip	2-35	649.96	(0.483, 0.517)	658.63	(0.350, 0.650)
Tip	2-36	341.03	(0.467, 0.533)	676.90	(0.501, 0.499)
Tip	2-37	**332.72**^{S}	(<**0.001**, 1.000)	759.46	(0.688, 0.312)
Tip	2-38	401.88	(0.125, 0.875)	355.88	(0.016, 0.984)
	I-T	**434.48**^{L}	(1.000, <**0.001**)	**235.35**^{L}	(1.000, <**0.001**)
Inference Key			1-2-11-17-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-13 (nested within clade 4-4)					
Interior	2-39	349.24	(0.053, 0.972)	430.65	(0.053, 0.972)
Tip	2-40	127.48	(0.415, 0.610)	**951.77**^{L}	(1.000, **0.025**)
	I-T	221.75	(0.610, 0.415)	**-521.12**^{S}	(**0.025**, 1.000)
Inference Key			1-19-20-2-11-17-NO		
Conclusion			Inconclusive Outcome		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-17 (nested within clade 4-4)					
Interior	2-48	268.57	(0.230, 0.773)	373.98	(0.916, 0.087)
Tip	2-49	**165.65**^{S}	(**0.003**, 1.000)	**322.80**^{S}	(**0.003**, 0.998)
	I-T	**102.92**^{L}	(1.000, **0.003**)	51.18	(0.937, 0.066)
Inference Key			1-19-20-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-3 (nested within clade 5-2)					
Tip	3-5	385.53	(0.150, 0.850)	488.29	(0.337, 0.663)
Interior	3-6	453.32	(0.056, 0.944)	458.09	(0.027, 0.973)
Tip	3-7	**16.07**^{S}	(<**0.001**, 1.000)	501.53	(0.422, 0.578)
Tip	3-8	**37.56**^{S}	(**0.001**, 1.000)	391.65	(0.134, 0.865)
Tip	3-9	289.00	(0.117, 0.883)	416.40	(0.217, 0.783)
Tip	3-10	**176.54**^{S}	(<**0.001**, 1.000)	472.83	(0.152, 0.848)
Interior	3-11	**80.65**^{S}	(<**0.001**, 1.000)	**1210.45**^{L}	(1.000, <**0.001**)
	I-T	184.32	(0.977, 0.023)	152.52	(0.979, 0.021)
Inference Key			1-19-20-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-4 (nested within clade 5-2)					
Interior	3-12	710.75	(0.892, 0.108)	**732.97**^{L}	(0.992, **0.008**)
Tip	3-13	531.93	(0.177, 0.823)	504.69	(0.036, 0.964)
Tip	3-14	**245.10**^{S}	(<**0.001**, 1.000)	576.82	(0.027, 0.973)
Tip	3-15	**289.11**^{S}	(<**0.001**, 1.000)	711.17	(0.675, 0.325)
Tip	3-16	323.10	(0.027, 0.973)	782.18	(0.788, 0.212)
Tip	3-17	**346.80**^{S}	(**0.001**, 0.999)	**344.67**^{S}	(<**0.001**, 1.000)
	I-T	**401.79**^{L}	(1.000, <**0.001**)	**129.53**^{L}	(0.997, **0.003**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 5-2 (nested within clade 6-1)					
Tip	4-3	**547.10**^{S}	(<**0.001**, 1.000)	726.37	(0.663, 0.337)
Interior	4-4	688.03	(0.041, 0.959)	708.88	(0.353, 0.647)
	I-T	**140.93**^{L}	(0.999, **0.001**)	-17.49	(0.341, 0.659)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Table B.11: Species 4 Nested Clade Analysis

The table includes all species 4 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	No	Dc	Prob	Dn	Prob
Clade 3-3 (nested within clade 4-2)					
Interior	2-7	327.71	(0.867, 0.150)	372.88	(0.983, 0.034)
Tip	2-8	62.08	(0.071, 0.933)	94.32	(0.028, 0.976)
Tip	2-9	0	(1.000, 1.000)	374.16	(1.000, 0.199)
Tip	2-10	0	(0.108, 1.000)	130.82	(0.222, 0.799)
	I-T	**292.24**^{L}	(1.000, **0.001**)	**228.15**^{L}	(0.995, **0.005**)
Inference Key			1-2-11-17-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-1 (nested within clade 5-1)					
Tip	3-1	**126.11**^{S}	(<**0.001**, 1.000)	**480.68**^{L}	(1.000, <**0.001**)
Interior	3-2	**146.07**^{S}	(**0.017**, 0.983)	311.32	(0.114, 0.886)
	I-T	19.95	(0.711, 0.289)	**-169.36**^{S}	(**0.012**, 0.988)
Inference Key			1-2-11-12-13-YES		
Conclusion			Contiguous Range Expansion, Long Distance Colonization or Past Range Expansion		

Table B.12: Species 5 lineage A Nested Clade Analysis

The table includes all species 5 lineage A clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-27 (nested within clade 2-13)					
Tip	S5H005	0	(1.000, 1.000)	120.16	(0.163, 1.000)
Tip	S5H006	0	(1.000, 1.000)	1056.67	(0.946, 0.082)
Tip	S5H007	0	(1.000, 1.000)	231.25	(0.440, 0.587)
Interior	S5H008	532.78	(0.645, 0.355)	523.50	(0.599, 0.401)
Tip	S5H009	0	(0.100, 1.000)	832.68	(0.883, 0.131)
Tip	S5H013	**0**^{S}	(<**0.001**, 1.000)	467.12	(0.539, 0.460)
	I-T	**532.78**^{L}	(0.995, **0.005**)	-10.70	(0.490, 0.510)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-1 (nested within clade 3-1)					
Interior	1-1	455.25	(0.687, 0.313)	438.26	(0.619, 0.381)
Tip	1-2	0	(0.104, 1.000)	664.84	(0.929, 0.072)
Tip	1-3	**0**^{S}	(<**0.001**, 1.000)	174.00	(0.351, 0.650)
Tip	1-4	4.15	(0.180, 0.840)	136.78	(0.205, 0.812)
Tip	1-5	**3.86**^{S}	(**0.006**, 0.994)	131.52	|(0.009, 0.992)
Tip	1-6	42.03	(0.224, 0.776)	717.56	(0.926, 0.074)
Tip	1-7	0	(0.104, 1.000)	174.00	(0.652, 0.398)
Tip	1-8	26.51	(0.112, 0.891)	142.00	(0.147, 0.855)
	I-T	444.33	(0.923, 0.077)	156.07	(0.678, 0.322)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-13 (nested within clade 3-5)					
Interior	1-27	493.3	(0.332, 0.668)	516.69	(0.395, 0.605)
Tip	1-28	0	(1.000, 1.000)	162.75	(0.170, 0.861)
Tip	1-29	**0**^{S}	(<**0.001**, 1.000)	333.47	(0.320, 0.681)
Tip	1-30	0	|(0.067, 1.000)	322.32	(0.400, 0.601)
Tip	1-31	1544.51	(0.976, 0.024)	1335.88	(0.970, 0.030)
Tip	1-32	**162.95**^{S}	(<**0.001**, 1.000)	335.45	(0.078, 0.922)
	I-T	84.44	(0.614, 0.386)	-18.61	(0.507, 0.493)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-4 (nested within clade 4-1)					
Interior	2-9	**29.74**^{S}	(<**0.001**, 1.000)	545.44	(0.371, 0.630)
Tip	2-10	460.90	(0.544, 0.456)	550.37	(0.619, 0.381)
	I-T	**-431.16**^{S}	(<**0.001**, 1.000)	-4.93	(0.379, 0.621)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-5 (netsed within clade 4-1)					
Interior	2-11	**84.24**^{S}	(**0.001**, 0.999)	227.50	(0.025, 0.975)
Tip	2-12	**0**^{S}	(<**0.001**, 1.000)	658.89	(0.710, 0.290)
Interior	2-13	560.44	(0.839, 0.161)	552.25	(0.813, 0.187)
	I-T	**-431.72**^{S}	(**0.007**, 0.993)	-333.21	(0.028, 0.972)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-6 (nested within clade 4-2)					
Tip	2-14	**0**^{S}	(**0.018**, 1.000)	**820.42**^{S}	(**0.018**, 1.000)
Interior	2-15	6.95	(0.197, 0.820)	**1015.20**^{L}	(1.000, **0.018**)
	I-T	6.95	(0.376, 0.641)	**194.79**^{L}	(1.000, **0.018**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-8 (nested within clade 4-3)					
Interior	2-18	**0**^{S}	(**0.023**, 1.000)	**16.29**^{S}	(**0.021**, 1.000)
Tip	2-19	**23.45**^{L}	(1.000, **0.021**)	23.33	(0.844, 0.177)
	I-T	**-23.45**^{S}	(**0.021**, 1.000)	**-7.04**^{S}	(**0.021**, 1.000)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 4-1 (nested within clade 5-1)					
Interior	3-1	**409.62**^{S}	(<**0.001**, 1.000)	**651.03**^{S}	(<**0.001**, 1.000)
Tip	3-2	1420.39	(0.953, 0.047)	1566.78	(0.952, 0.048)
Tip	3-3	104.91	(0.033, 0.967)	998.32	(0.635, 0.365)
Tip	3-4	**548.41**^{S}	(**0.005**, 0.996)	667.83	(0.022, 0.978)
Interior	3-5	**687.74**^{S}	(<**0.001**, 1.000)	**1129.09**^{L}	(1.000, <**0.001**)
	I-T	-262.69	(0.010, 0.990)	**-400.10**^{S}	(<**0.001**, 1.000)
Inference Key			1-2-11-12-13-14-YES		
Conclusion			Contiguous Range Expansion, Long Distance Colonization or		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 4-3 (nested within clade 5-1)					
Interior	3-8	**31.92**^{S}	(**0.010**, 1.000)	**26.19**^{S}	(**0.010**, 1.000)
Tip	3-9	0	(0.236, 1.000)	**172.43**^{L}	(1.000, **0.010**)
	I-T	21.92	(0.256, 0.754)	**-146.24**^{S}	(**0.010**, 1.000)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 5-1 (Total Cladogram)					
Tip	4-1	**928.66**^{S}	(**0.006**, 0.994)	952.59	(0.025, 0.975)
Interior	4-2	833.08	(0.343, 0.657)	1181.52	(0.818, 0.182)
Tip	4-3	**44.14**^{S}	(<**0.001**, 1.000)	**1492.57**^{L}	(0.995, **0.005**)
	I-T	-6.54	(0.656, 0.344)	174.57	(0.799, 0.201)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		

Table B.13: Species 6 Nested Clade Analysis

The table includes all species 6 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-8 (nested within clade 2-2)					
Tip	S6H004	0	(1.000, 1.000)	203.84	(0.507, 0.656)
Interior	S6H005	200.82	(0.401, 0.599)	238.13	(0.592, 0.408)
Tip	S6H006	0	(1.000, 1.000)	203.84	(0.507, 0.656)
Tip	S6H007	0	(1.000, 1.000)	431.39	(1.000, 0.035)
Tip	S6H008	0	(1.000, 1.000)	302.81	(0.696, 0.3530
Tip	S6H009	**0**^{S}	(<**0.001**, 1.000)	**42.90**^{S}	(<**0.001**, 1.000)
Tip	S6H013	0	(1.000, 1.000)	189.43	(0.346, 0.692)
Tip	S6H017	18.20	(0.075, 0.931)	373.68	(0.977, 0.029)
Tip	S6H018	0	(1.000, 1.000)	207.58	(0.643, 0.517)
	I-T	198.02	(0.933, 0.067)	33.38	(0.654, 0.346)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-13 (nested within clade 2-3)					
Tip	S6H029	0	(1.000, 1.000)	69.73	(0.245, 1.000)
Tip	S6H030	0	(1.000, 1.000)	69.72	(0.233, 1.000)
Tip	S6H031	**0**^{S}	(**0.008**, 1.000)	**376.01**^{L}	(1.000, <**0.001**)
Interior	S6H032	133.35	(0.214, 0.786)	137.07	(0.199, 0.801)
	I-T	133.35	(0.838, 0.162)	-116.79	(0.052, 0.948)
Inference Key			1-2-3-5-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 2-2 (nested within clade 3-1)					
Tip	1-4	**5.62**^{S}	(**0.003**, 0.997)	127.21	(**0.005**, 0.995)
Tip	1-5	0	(1.000, 1.000)	748.55	(0.908, 0.104)
Tip	1-6	**162.86**^{S}	(<**0.001**, 1.000)	507.53	**(0.928, 0.072)**
Tip	1-7	**0**^{S}	(<**0.001**, 1.000)	455.87	**(0.719, 0.281)**
Interior	1-8	**232.06**^{S}	(<**0.001**, 1.000)	**237.21**^{S}	(<**0.001**, 1.000)
Tip	1-9	0	(1.000, 1.000)	283.33	(0.327, 0.711)
Tip	1-10	**0**^{S}	(<**0.001**, 1.000)	**784.77**^{L}	(0.994, **0.006**)
Tip	1-11	493.39	(0.916, 0.086)	822.23	(0.0975, 0.027)
Tip	1-12	0	(0.031, 1.000)	365.78	(0.419, 0.582)
	I-T	128.93	(0.794, 0.206)	**-270.91**^{S}	(<**0.001**, 1.000)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-4 (nested within clade 3-2)					
Tip	1-15	0	(0.050, 1.000)	188.47	(0.097, 0.906)
Tip	1-16	0	(0.034, 1.000)	359.03	(0.657, 0.347)
Interior	1-17	338.32	(0.535, 0.465)	336.50	(0.444, 0.556)
Tip	1-18	**0**^{S}	(<**0.001**, 1.000)	**649.47**^{L}	(0.997, **0.003**)
	I-T	**338.32**^{S}	(0.998, **0.002**)	-71.08	(0.198, 0.802)
Inference Key			1-2-3-5-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 2-6 (nested within clade 3-3)					
Interior	1-22	**0**^{S}	(**0.003**, 1.000)	**122.37**^{S}	(**0.003**, 1.000)
Tip	1-23	0	(0.320, 1.000)	132.94	(0.630, 0.434)
Tip	1-24	144.47	(0.553, 0.463)	217.41	(0.970, 0.046)
	I-T	-86.68	(0.123, 0.879)	-61.25	(0.130, 0.872)
Inference Key			1-19-20-2-3-5-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 2-14 (nested within clade 3-5)					
Tip	1-34	0	(1.000, 1.000)	354.53	(0.716, 0.425)
Tip	1-35	0	(0.077, 1.000)	1280.73	(1.000, **0.005**)
Tip	1-36	**0**^{S}	(**0.002**, 1.000)	275.97	(0.091, 0.911)
Interior	1-37	134.34	(0.017, 0.984)	277.29	(0.072, 0.929)
Tip	1-38	0	(1.000, 1.000)	557.15	(0.854, 0.216)
Tip	1-39	0	(1.000, 1.000)	304.34	(0.546, 0.526)
	I-T	134.34	(0.460, 0.540)	-265.08	(0.078, 0.922)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-16 (nested within clade 3-5)					
Interior	1-42	**149.46**^{S}	(**0.020**, 1.000)	**223.95**^{S}	(**0.020**, 1.000)
Tip	1-43	**0**^{S}	(**0.020**, 1.000)	**655.72**^{L}	(1.000, **0.020**)
	I-T	149.46	(0.821, 0.199)	**-431.77**^{S}	(**0.020**, 1.000)
Inference Key			1-19-20-2-11-12-13-YES		
Conclusion			Long Distance Colonization, Past larger Range followed by		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-31 (nested within clade 3-13)					
Tip	1-72	**5.74**^{S}	(**0.004**, 0.996)	50.06	(0.223, 0.777)
Tip	1-73	**0**^{S}	(**0.002**, 1.000)	**26.38**^{S}	(**0.002**, 1.000)
Interior	1-74	72.91	(0.858, 0.142)	**90.44**^{L}	(1.000, <**0.001**)
	I-T	**70.04**^{L}	(1.000, <**0.001**)	**52.22**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-1 (nested within clade 4-1)					
Tip	2-1	**173.33**^{S}	(<**0.001**, 1.000)	**232.64**^{S}	(<**0.001**, 1.000)
Interior	2-3	402.27	(0.970, 0.030)	**410.01**^{L}	(0.987, **0.013**)
Tip	2-4	**240.93**^{S}	(**0.013**, 0.987)	367.34	(0.576, 0.424)
	I-T	**199.55**^{L}	(1.000, <**0.001**)	**118.50**^{L}	(0.995, **0.005**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-2 (nested within clade 4-1)					
interior	2-5	**338.63**^{S}	(**0.015**, 0.985)	**347.76**^{S}	(**0.015**, 0.985)
Tip	2-6	**722.46**^{L}	(1.000, <**0.001**)	**776.76**^{L}	(0.999, **0.001**)
	I-T	**-383.83**^{S}	(<**0.001**, 1.000)	**-429.99**^{S}	(**0.001**, 0.999)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-5 (nested within clade 4-2)					
Tip	2-12	490.46	(0.016, 0.984)	**1455.46**^{L}	(1.000, <**0.001**)
Tip	2-13	0	(1.000, 1.000)	590.67	(0.189, 0.852)
Tip	2-14	0	(0.058, 1.000)	594.68	(0.108, 0.892)
Interior	2-15	**438.63**^{S}	(<**0.001**, 1.000)	779.89	(0.120, 0.880)
Tip	2-16	**0**^{S}	(**0.004**, 1.000)	**444.80**^{S}	(**0.001**, 0.999)
	I-T	158.15	(0.523, 0.477)	-274.28	(0.053, 0.947)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-6 (nested within clade 4-2)					
Interior	2-17	319.9	(0.742, 0.258)	338.75	(0.770, 0.230)
Tip	2-18	**0**^{S}	(**0.002**, 1.000)	335.97	(0.655, 0.345)
Tip	2-19	255.19	(0.080, 0.920)	262.58	(0.084, 0.916)
	I-T	112.56	(0.947, 0.053)	62.41	(0.799, 0.201)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-9 (nested within clade 4-3)					
Interior	2-23	**283.04**^{S}	(**0.011**, 0.990)	**272.07**^{S}	(**0.002**, 0.999)
Tip	2-24	71.39	(0.229, 0.800)	638.75	(0.688, 0.341)
Tip	2-25	0	(0.018, 1.000)	772.78	(0.976, 0.031)
	I-T	254.48	(0.832, 0.168)	**-447.09**^{S}	(**0.004**, 0.996)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-9 (nested within clade 4-3)					
Interior	2-23	**283.04**^{S}	(**0.011**, 0.990)	**272.07**^{S}	(**0.002**, 0.999)
Tip	2-24	71.39	(0.229, 0.800)	638.75	(0.688, 0.341)
Tip	2-25	0	(0.018, 1.000)	772.78	(0.976, 0.031)
	I-T	254.48	(0.832, 0.168)	**-447.09**^{S}	(**0.004**, 0.996)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 3-11 (nested within clade 4-4)					
Tip	2-27	**0**^{S}	(<**0.001**, 1.000)	433.38	(0.944, 0.056)
Interior	2-28	277.03	(0.330, 0.670)	311.84	(0.342, 0.658)
Tip	2-29	**0**^{S}	(<**0.001**, 1.000)	**218.65**^{S}	(<**0.001**, 1.000)
	I-T	**277.03**^{L}	(1.000, <**0.001**)	-4.41	(0.348, 0.652)
Inference Key			1-19-20-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-13 (nested within clade 4-5)					
Interior	2-31	**0**^{S}	(**0.001**, 1.000)	**305.98**^{L}	(1.000, <**0.001**)
Tip	2-32	**61.23**^{S}	(<**0.001**, 1.000)	**115.10**^{S}	(<**0.001**, 1.000)
	I-T	-61.23	(0.314, 0.686)	**190.89**^{L}	(1.000, <**0.001**)
Inference Key			1-19-20-2-11-12-13-YES		
Conclusion			Long Distance Colonization, Past larger Range followed by		
Clade 3-14 (nested within clade 4-5)					
Interior	2-33	**142.90**^{S}	(**0.001**, 1.000)	**288.21**^{S}	(**0.001**, 1.000)
Tip	2-34	**835.01**^{L}	(1.000, **0.001**)	838.44	(0.911, 0.090)
	I-T	**-692.11**^{S}	(**0.001**, 1.000)	**-578.61**^{S}	(**0.001**, 1.000)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 3-19 (nested within clade 4-7)					
Interior	2-42	512.16	(0.733, 0.267)	505.99	(0.759, 0.241)
Tip	2-43	**0**^{S}	(<**0.001**, 1.000)	354.06	(0.131, 0.869)
	I-T	**512.16**^{L}	(0.998, 0.002)	151.94	(0.863, 0.137)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-1 (nested within clade 5-1)					
Tip	3-1	**358.97**^{S}	(<**0.001**, 1.000)	**393.78**^{S}	(<**0.001**, 1.000)
Interior	3-2	424.33	(0.334, 0.666)	488.14	(0.820, 0.180)
Tip	3-3	**245.09**^{S}	(**0.008**, 0.992)	**683.26**^{L}	(0.997, **0.003**)
	I-T	77.46	(0.928, 0.072)	63.62	(0.866, 0.134)
Inference Key			1-2-3-5-6-7-8-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		
Clade 4-2 (nested within clade 5-1)					
Interior	3-4	353.94	(0.119, 0.881)	865.68	(0.779, 0.221)
Interior	3-5	**866.76**^{L}	(0.988, **0.012**)	846.05	(0.983, 0.017)
Tip	3-6	**2996.14**^{S}	(<**0.001**, 1.000)	**551.54**^{S}	(<**0.001**, 1.000)
Interior	3-7	0	(0.020, 1.000)	1294.68	(0.966, 0.034)
	I-T	**447.83**^{L}	(1.000, <**0.001**)	**322.94**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-3 (nested within clade 5-1)					
Tip	3-8	**0**^{S}	(<**0.001**, 1.000)	559.34	(0.625, 0.375)
Interior	3-9	487.99	(0.452, 0.548)	490.94	(0.465, 0.535)
	I-T	**487.99**^{L}	(1.000, <**0.001**)	68.41	(0.379, 0.621)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-5 (nested within clade 5-2)					
Interior	3-12	0	(1.000, 1.000)	498.05	(0.742, 0.286)
Tip	3-13	**160.55**^{S}	(<**0.001**, 1.000)	**341.91**^{S}	(**0.004**, 0.996)
Tip	3-14	552.97	(0.603, 0.397)	677.52	(0.918, 0.082)
	I-T	**-305.12**^{L}	(1.000, <**0.001**)	32.50	(0.741, 0.259)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 4-6 (nested within clade 5-3)					
Interior	3-15	275.68	(0.258, 0.742)	465.96	(0.983, 0.017)
Tip	3-16	**0**^{S}	(<**0.001**, 1.000)	175.27	(0.033, 0.967)
Tip	3-17	358.89	(0.622, 0.378)	**506.59**^{L}	(0.999, **0.001**)
Tip	3-18	**62.90**^{S}	(<**0.001**, 1.000)	**167.58**^{S}	(<**0.001**, 1.000)
	I-T	103.84	(0.932, 0.068)	**158.84**^{L}	(0.989, **0.011**)
Inference Key			1-2-3-5-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 4-7 (nested within clade 5-3)					
Interior	3-18	494.29	(0.778, 0.222)	485.32	(0.757, 0.243)
Tip	3-19	**17.62**^{S}	(**0.003**, 0.997)	341.15	(0.124, 0.876)
	I-T	**476.67**^{L}	(0.993, **0.007**)	144.17	(0.855, 0.145)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 5-1 (nested within clade 6-1)					
Tip	4-1	**473.71**^{S}	(<**0.001**, 1.000)	**557.62**^{S}	(<**0.001**, 1.000)
Interior	4-2	722.86	(0.925, 0.075)	**848.39**^{L}	(1.000, <**0.001**)
Interior	4-3	503.19	(0.131, 0.869)	582.05	(0.220, 0.780)
	I-T	**205.22**^{L}	(1.000, <**0.001**)	**237.50**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 5-3 (nested within clade 6-1)					
Interior	4-6	**366.52**^{S}	(<**0.001**, 1.000)	**1163.77**^{L}	(1.000, <**0.001**)
Tip	4-7	**473.63**^{S}	(<**0.001**, 1.000)	**582.96**^{S}	(**0.001**, 0.999)
	I-T	-107.1	(0.281, 0.719)	**580.81**^{L}	(1.000, <**0.001**)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 6-1 (Total Cladogram)					
Tip	5-1	**651.72**^{S}	(<**0.001**, 1.000)	**659.72**^{S}	(<**0.001**, 1.000)
Interior	5-2	**423.63**^{S}	(<**0.001**, 1.000)	**1063.76**^{L}	(0.998, **0.002**)
Tip	5-3	757.01	(0.131, 0.869)	**1228.21**^{L}	(1.000, <**0.001**)
	I-T	**-255.95**^{S}	(<**0.001**, 1.000)	**253.56**^{L}	(0.997, **0.003**)
Inference Key			1-2-11-12-13-YES		
Conclusion			Long Distance Colonization, Past larger Range followed by Fragmentation or Past Range Expansions		

Table B.14: Species 7 Nested Clade Analysis

The table includes all species 7 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 1-14 (nested within clade 2-5)					
Tip	S7H043	0	(0.026, 1.000)	**1045.17**^{L}	(1.000, **0.018**)
Interior	S7H045	680.22	(0.054, 0.954)	818.85	(0.104, 0.904)
Tip	S7H047	0	(1.000, 1.000)	905.32	(0.324, 0.787)
	I-T	680.22	(0.957, 0.046)	**-191.36**^{S}	(**0.013**, 0.990)
Inference Key			1-2-11-17-NO		
Conclusion			Inconclusive Outcome		
Clade 1-22 (nested within clade 2-8)					
Tip	S7H054	0	(1.000, 1.000)	465.77	(0.618, 0.544)
Tip	S7H055	0	(1.000, 1.000)	741.05	(0.679, 0.391)
Tip	S7H056	0	(1.000, 1.000)	1376.64	(1.000, 0.076)
Tip	S7H057	0	(1.000, 1.000)	1089.01	(0.845, 0.234)
Tip	S7H058	0	(1.000, 1.000)	1240.50	(0.930, 0.141)
Interior	S7H059	**377.44**^{S}	(**0.003**, 1.000)	**519.30**^{S}	(**0.008**, 0.995)
	I-T	**377.44**^{S}	(**0.003**, 1.000)	**-463.29**^{S}	(**0.008**, 0.995)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 1-25 (nested within clade 2-10)					
Tip	S7H060	**0**^{S}	(**0.001**, 1.000)	273.38	(0.476, 0.525)
Interior	S7H061	218.57	(0.277, 0.725)	227.82	(0.092, 0.910)
Tip	S7H063	0	(0.199, 1.000)	345.99	(0.851, 0.207)
Tip	S7H064	0	(1.000, 1.000)	454.27	(1.000, 0.074)
	I-T	227.79	(0.944, 0.056)	-85.44	(0.150, 0.850)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 1-30 (nested within clade 2-11)					
Tip	S7H071	**0**^{S}	(<**0.001**, 1.000)	224.23	(0.888, 0.112)
Interior	S7H072	142.42	(0.206, 0.802)	155.00	(0.077, 0.931)
Tip	S7H073	0	(1.000, 1.000)	273.61	(1.000, 0.178)
	I-T	**142.42**^{L}	(1.000, <**0.001**)	-77.47	(0.029, 0.971)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-2 (nested within clade 3-1)					
Tip	1-2	544.68	(0.846, 0.154)	687.12	(0.906, 0.094)
Tip	1-3	0	(1.000, 1.000)	199.62	(0.023, 1.000)
Tip	1-4	0	(1.000, 1.000)	266.55	(0.332, 0.725)
Tip	1-5	**0**^{S}	(**0.002**, 1.000)	**1163.89**^{L}	(1.000, <**0.001**)
Tip	1-6	482.06	(0.807, 0.193)	522.93	(0.915, 0.085)
Interior	1-7	369.01	(0.010, 0.990)	**372.30**^{S}	(0.013, 0.987)
	I-T	-43.59	(0.105, 0.895)	**-208.00**^{S}	(**0.001**, 0.999)
Inference Key			1-2-3-5-6-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 2-3 (nested within clade 3-2)					
Tip	1-8	**483.32**^{S}	(**0.004**, 0.996)	548.73	(0.032, 0.968)
Interior	1-9	**0**^{S}	(**0.002**, 1.000)	**1056.04**^{L}	(0.984, **0.018**)
Interior	1-10	**0**^{S}	(<**0.001**, 1.000)	**1056.34**^{L}	(0.988, **0.012**)
	I-T	**-483.32**^{S}	(**0.001**, 0.999)	**507.31**^{L}	(0.999, **0.001**)
Inference Key			1-2-3-5-6-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 2-4 (nested within clade 3-2)					
Tip	1-11	0	(1.000, 1.000)	143.88	(0.154, 0.951)
Interior	1-12	733.55	(0.964, 0.043)	729.16	(0.964, 0.043)
Tip	1-13	43.04	(0.169, 0.842)	282.27	(0.076, 0.935)
	I-T	**704.85**^{L}	(0.987, **0.017**)	**493.02**^{L}	(0.986, **0.018**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 2-8 (nested within clade 3-2)					
Tip	1-20	157.81	(0.137, 0.869)	**181.21**^{S}	(**0.006**, 1.000)
Tip	1-21	0	(0.041, 1.000)	1114.17	(0.855, 0.153)
Interior	1-22	831.42	(0.391, 0.609)	854.05	(0.467, 0.533)
	I-T	752.52	(0.978, 0.022)	206.36	(0.729, 0.271)
Inference Key			1-2-11-17-NO		
Conclusion			Inconclusive Outcome		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-10 (nested within clade 3-4)					
Interior	1-25	296.56	(0.287, 0.713)	290.09	(0.160, 0.840)
Tip	1-26	**0**^{S}	(<**0.001**, 1.000)	**447.61**^{L}	(0.990, **0.010**)
Tip	1-27	0	(1.000, 1.000)	399.88	(0.713, 0.424)
	I-T	**296.55**^{L}	(0.999, **0.001**)	-149.57	(0.032, 0.968)
Inference Key			1-2-3-5-6-15-NO		
Conclusion			Past Fragmentation or Long Distance Colonization		
Clade 2-11 (nested within clade 3-4)					
Tip	1-27	**0**^{S}	(<**0.001**, 1.000)	424.77	(0.807, 0.193)
Tip	1-28	770.19	(1.000, 0.0210)	770.51	(0.965, 0.056)
Interior	1-29	**184.66**^{S}	(<**0.001**, 1.000)	**203.38**^{S}	(**0.004**, 0.996)
	I-T	-35.39	(0.325, 0.675)	**-320.18**^{S}	(**0.001**, 0.999)
Clade 3-1 (nested within clade 4-1)					
Tip	2-1	0	(0.028, 1.000)	**1092.01**^{L}	(0.998, **0.002**)
Interior	2-2	**440.50**^{S}	(**0.015**, 0.985)	**441.13**^{S}	(**0.017**, 0.983)
	I-T	440.50	(0.951, 0.049)	**-650.87**^{S}	(**0.002**, 0.998)
Inference Key			1-2-11-12-13-YES		
Conclusion			Long Distance Colonization, Past Larger Range Coupled		
Clade 3-5 (nested within clade 4-2)					
Interior	2-12	1154.94	(0.088, 0.912)	1205.41	(0.197, 0.803)
Tip	2-13	0	(0.035, 1.000)	262.29	(0.041, 0.961)
Tip	2-14	**162.53**^{S}	(<**0.001**, 1.000)	**2080.03**^{L}	(0.996, **0.004**)
Tip	2-15	522.37	(0.072, 0.928)	1076.85	(0.276, 0.724)
	I-T	**918.13**^{L}	(1.000, <**0.001**)	-364.73	(0.109, 0.891)
Inference Key			1-2-3-5-15-1-6-18-NO		
Conclusion			Geographical Sampling Scheme Inadequate to Discriminate		
Clade 4-1 (nested within clade 5-1)					
Tip	3-1	**445.64**^{S}	(<**0.001**, 1.000)	**445.92**^{S}	(<**0.001**, 1.000)
Interior	3-2	**732.72**^{L}	(1.000, <**0.001**)	**727.97**^{L}	(1.000, <**0.001**)
	I-T	**287.08**^{L}	(1.000, <**0.001**)	**282.05**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 4-2 (nested within clade 5-1)					
Interior	3-3	849.99	(0.031, 0.969)	**1182.03**^{L}	(0.971, 0.029)
Tip	3-4	**327.26**^{S}	(<**0.001**, 1.000)	**914.90**^{S}	(**0.024**, 0.976)
	I-T	**522.73**^{L}	(1.000, <**0.001**)	**267.13**^{L}	(0.974, 0.026)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 5-1 (Total Cladogram)					
Tip	4-1	**534.50**^{S}	(<**0.001**, 1.000)	**623.10**^{S}	(<**0.001**, 1.000)
Interior	4-2	1052.06	(0.627, 0.373)	**1268.18**^{L}	(0.997, **0.003**)
Tip	4-3	**1298.47**^{L}	(0.986, **0.014**)	**1391.67**^{L}	(1.000, <**0.001**)
	I-T	**335.06**^{L}	(1.000, <**0.001**)	**461.47**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Table B.15: Species 8 Nested Clade Analysis

The table includes all species 8 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 2-5 (nested within clade 3-1)					
Interior	1-8	0	(1.000, 1.000)	203.01	(0.3563, 1.000)
Tip	1-9	0	(0.032, 1.000)	**695.60^{L}**	(1.000, **0.001**)
Tip	1-10	0	(0.170, 1.000)	611.67	(0.933, 0.074)
Tip	1-11	**0^{S}**	(**0.001**, 1.000)	**203.01^{S}**	(**0.001**, 1.000)
Tip	1-12	**8.69^{S}**	(**0.001**, 0.999)	488.78	(0.847, 0.153)
Tip	1-13	0	(1.000, 1.000)	402.78	(0.413, 0.648)
	I-T	**-2.71^{L}**	(1.000, <**0.001**)	-245.00	(0.036, 0.964)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 3-11 (nested within clade 4-3)					
Interior	2-24	107.69	(0.910, 0.105)	109.99	(0.833, 0.182)
Tip	2-25	**0^{S}**	(**0.012**, 1.000)	**65.84^{S}**	(**0.002**, 1.000)
Interior	2-26	140.59	(0.499, 0.502)	133.79	(0.499, 0.502)
	I-T	132.36	(0.980, 0.020)	**62.00^{L}**	(0.999, **0.001**)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-1 (nested within clade 5-1)					
Interior	3-1	426.51	(0.317, 0.683)	468.22	(0.771, 0.229)
Tip	3-2	54.75	(0.126, 0.875)	668.33	(0.986, 0.015)
Tip	3-3	0	(1.000, 1.000)	264.18	(0.169, 0.846)
Tip	3-4	351.10	(0.149, 0.852)	429.64	(0.314, 0.686)
Tip	3-5	304.56	(0.022, 0.978)	436.64	(0.351, 0.649)
Tip	3-6	**54.95^{S}**	(<**0.001**, 1.000)	440.78	(0.396, 0.604)
	I-T	**211.54^{L}**	(1.000, <**0.001**)	24.71	(0.722, 0.278)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-3 (nested within clade 5-1)					
Tip	3-9	86.86	(0.671, 0.330)	171.12	(0.923, 0.078)
Tip	3-10	**49.01^{S}**	(**0.012**, 0.988)	**53.06^{S}**	(<**0.001**, 1.000)
Interior	3-11	117.7	(0.711, 0.290)	131.37	(0.942, 0.058)
	I-T	59.95	(0.897, 0.103)	50.07	(0.941, 0.060)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 4-4 (nested within clade 5-1)					
Tip	3-12	**15.40**^{S}	(<**0.001**, 1.000)	122.57	(0.770, 0.230)
Tip	3-13	39.04	(0.165, 0.837)	90.29	(0.116, 0.886)
Interior	3-14	**71.89**^{S}	(**0.007**, 0.993)	117.06	(0.566, 0.434)
	I-T	**51.42**^{L}	(0.996, **0.004**)	1.4	(0.537, 0.463)
Inference Key			1-2-3-4-NO		
Conclusion			Restricted Gene Flow with Isolation by Distance		
Clade 4-5 (nested within clade 5-1)					
Tip	3-15	83.71	(0.868, 0.135)	**109.86**^{L}	(0.985, **0.018**)
Interior	3-16	48.73	(0.027, 0.976)	55.06	(0.050, 0.954)
	I-T	-35.79	(0.076, 0.927)	-54.81	(0.028, 0.976)
Inference Key			1-19-20-2-11-17-NO		
Conclusion			Inconclusive Outcome		
Clade 4-6 (nested within clade 5-1)					
Interior	3-17	0	(0.367, 1.000)	**1303.62**^{L}	(1.000, **0.005**)
Tip	3-18	**94.30**^{S}	(<**0.001**, 1.000)	**368.66**^{S}	(<**0.001**, 1.000)
Interior	3-19	0	(0.374, 1.000)	727.01	(0.815, 0.190)
	I-T	-94.3	(0.831, 0.169)	**646.66**^{L}	(1.000, <**0.001**)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		
Clade 5-1 (Total Cladogram)					
Interior	4-1	449.74	(0.209, 0.791)	449.56	(0.080, 0.920)
Tip	4-2	156.83	(0.098, 0.902)	409.47	(0.277, 0.723)
Tip	4-3	**108.47**^{S}	(<**0.001**, 1.000)	473.31	(0.392, 0.608)
Tip	4-4	**115.64**^{S}	(<**0.001**, 1.000)	**326.60**^{S}	(**0.001**, 0.999)
Tip	4-5	**67.66**^{S}	(<**0.001**, 1.000)	**306.03**^{S}	(**0.005**, 0.995)
Tip	4-6	549.18	(0.769, 0.231)	**1195.26**^{L}	(1.000, <**0.001**)
	I-T	**235.20**^{L}	(0.997, **0.003**)	**-130.07**^{S}	(**0.006**, 0.994)
Inference Key			1-2-3-5-6-7-YES		
Conclusion			Isolation by Distance with Long Distance Dispersal		

Table B.16: Species 9 Nested Clade Analysis

The table includes all species 9 clades that had significant geographical associations with respect to the nested clade analysis. The clades sorted from lowest tier (1-step clades) up to the total cladogram. Each clade includes the tip and interior clades/haplotypes as well as the difference between interior and tip value (I-T). The clade and nested clade distances are listed (Dc and Dn, respectively) and if they possess a significant geographical association the value is bolded. The observed clade and nested clade distances are assessed if they are significantly smaller or larger than the expected values. The p-values for these permutations are listed under the Prob headings (smaller, larger) and bolded if significant. Additionally, the clade and nested clade distance values (under the Dc and Dn headings) are tagged with a superscript (S = smaller, L = larger). The decision path and final conclusion of the inference key are listed in the Inference Key and Conclusion rows, respectively.

Tip/Node	Clade	Dc	Prob	Dn	Prob
Clade 4-2 (nested within clade 5-1)					
Interior	3-3	34.04	(0.028, 0.972)	**59.97**^{S}	(**0.022**, 0.979)
Tip	3-4	**283.53**^{L}	(1.000, <**0.001**)	**250.28**^{L}	(0.996, **0.005**)
Tip	I-T	**-249.50**^{S}	(**0.002**, 0.999)	**-190.31**^{S}	(**0.005**, 0.999)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 4-3 (nested within clade 5-1)					
Interior	3-5	152.61	(0.910, 0.105)	153.68	(0.833, 0.182)
Tip	3-6	**288.67**^{L}	(0.993, **0.007**)	269.78	(0.973, 0.027)
	I-T	**-136.06**^{S}	(**0.002**, 0.983)	-116.10	(0.0.39, 0.962)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		
Clade 5-1 (Total Cladogram)					
Tip	4-1	**37.24**^{S}	(**0.004,** 0.996)	269.82	(0.230, 0.770)
Interior	4-2	**99.62**^{S}	(<**0.001**, 1.000)	304.58	(0.746, 0.255)
Tip	4-3	**182.44**^{S}	(<**0.001**, 1.000)	287.97	(0.337, 0.663)
	I-T	**-60.14**^{S}	(**0.009,** 0.991)	19.44	(0.762, 0.238)
Inference Key			1-2-11-12-NO		
Conclusion			Contiguous Range Expansion		

Appendix C

Figure C.1: The nested COI haplotype network of species 1 lineage A.
The entire cladogram of COI haployptes formed a 5-step clade (5-1). The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and sites (lower). The location of haplotypes listed under the (#) and (*) icons are indicated within clade 2-10 in the network using corresponding symbols. Their samples sizes and most recent ancestor are identical to that of the adjacent ancestor. The haplotypes denoted by (#) are directly attached to haplotype S1H064, whereas the haplotypes denoted by (*) are separated by an inferred haplotype. The root of the network is denoted by (R).

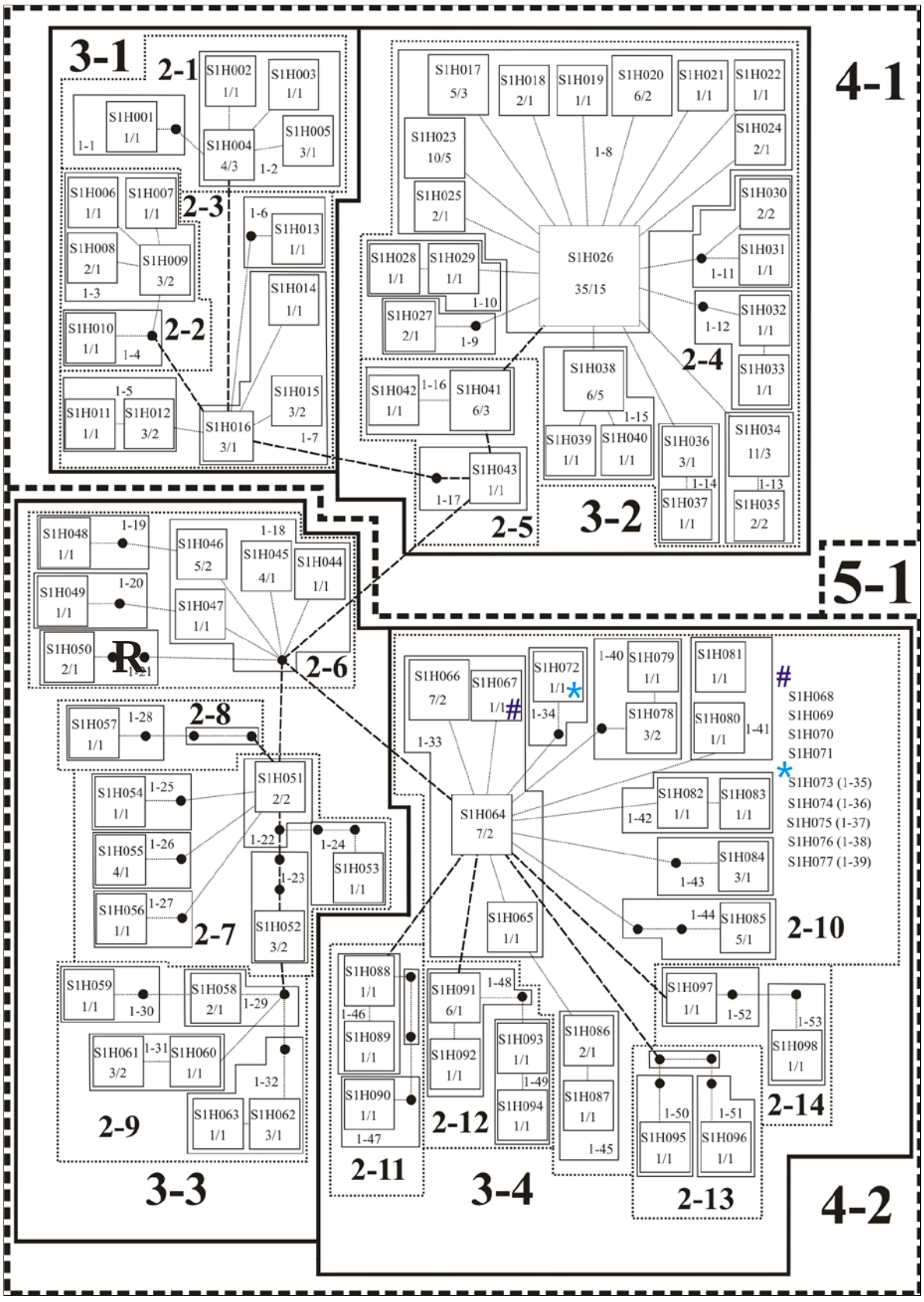

3-1
2-1
S1H001 1/1
1-1
S1H002 1/1
S1H003 1/1
S1H004 4/3
1-2
S1H005 3/1
2-3
S1H006 1/1
S1H007 1/1
S1H008 2/1
1-3
S1H009 3/2
1-6
S1H013 1/1
S1H014 1/1
2-2
S1H010 1/1
1-4
1-5
S1H011 1/1
S1H012 3/2
S1H016 3/1
S1H015 3/2
1-7
4-1
S1H017 5/3
S1H018 2/1
S1H019 1/1
S1H020 6/2
S1H021 1/1
S1H022 1/1
S1H024 2/1
S1H023 10/5
1-8
S1H025 2/1
S1H030 2/2
S1H031 1/1
1-11
S1H026 35/15
S1H028 1/1
S1H029 1/1
1-10
S1H027 2/1
1-9
1-12
S1H032 1/1
2-4
S1H033 1/1
S1H042 1/1
1-16
S1H041 6/3
S1H038 6/5
1-15
S1H039 1/1
S1H040 1/1
S1H036 3/1
1-14
S1H037 1/1
S1H034 11/3
1-13
S1H035 2/2
1-17
S1H043 1/1
2-5
3-2
5-1
S1H048 1/1
1-19
S1H046 5/2
1-18
S1H045 4/1
S1H044 1/1
S1H049 1/1
1-20
S1H047 1/1
S1H050 2/1
1-21
2-6
S1H066 7/2
S1H067 1/1 #
S1H072 1/1 *
1-40
S1H079 1/1
S1H081 1/1
#
S1H068
S1H069
S1H070
S1H071
1-34
S1H078 3/2
S1H080 1/1
1-41
*
S1H073 (1-35)
S1H074 (1-36)
S1H075 (1-37)
S1H076 (1-38)
S1H077 (1-39)
1-33
S1H064 7/2
S1H082 1/1
S1H083 1/1
1-42
1-43
S1H084 3/1
1-44
S1H085 5/1
2-10
S1H065 1/1
S1H057 1/1
1-28
2-8
S1H051 2/2
S1H054 1/1
1-25
S1H055 4/1
1-26
S1H056 1/1
1-27
2-7
1-22
1-24
S1H053 1/1
1-23
S1H052 3/2
S1H059 1/1
1-30
S1H058 2/1
1-29
S1H061 3/2
1-31
S1H060 1/1
1-32
S1H063 1/1
S1H062 3/1
2-9
3-3
S1H088 1/1
1-46
S1H089 1/1
S1H090 1/1
1-47
2-11
S1H091 6/1
1-48
S1H092 1/1
S1H093 1/1
1-49
S1H094 1/1
2-12
3-4
S1H086 2/1
S1H087 1/1
1-45
S1H097 1/1
1-52
1-53
S1H098 1/1
2-14
1-50
S1H095 1/1
1-51
S1H096 1/1
2-13
4-2

Figure C.2: Nested COI Haplotype Network of species 1 lineage B
The entire cladogram formed a 5-step clade (5-2). The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted by (R).

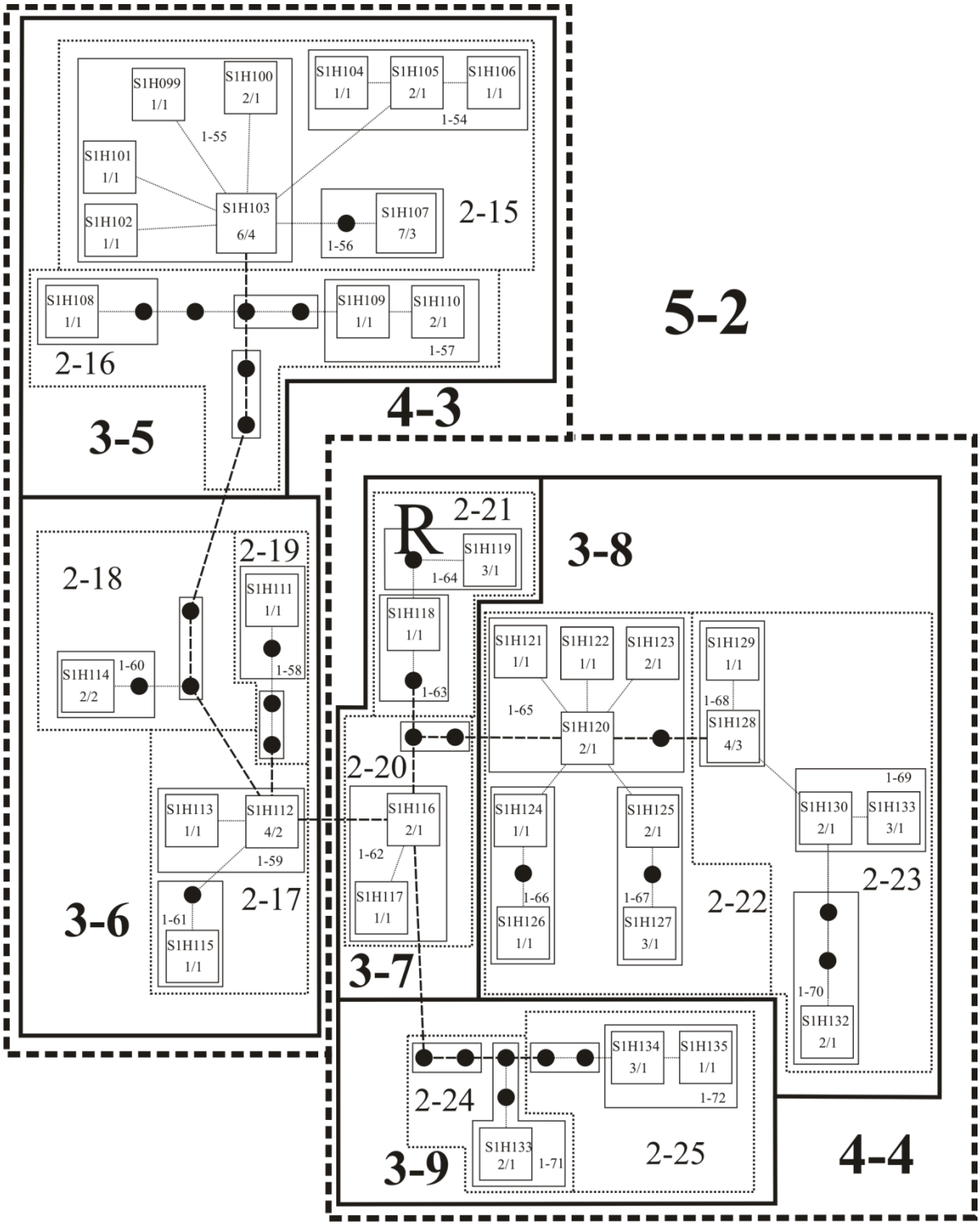


Figure C.3: Nested COI haplotype network of species 1 lineage C
The entire cladogram formed a 1-step clade (1-73). The numbers associated with each node gives species and haplotype (upper), and sample size and number of sites (lower).

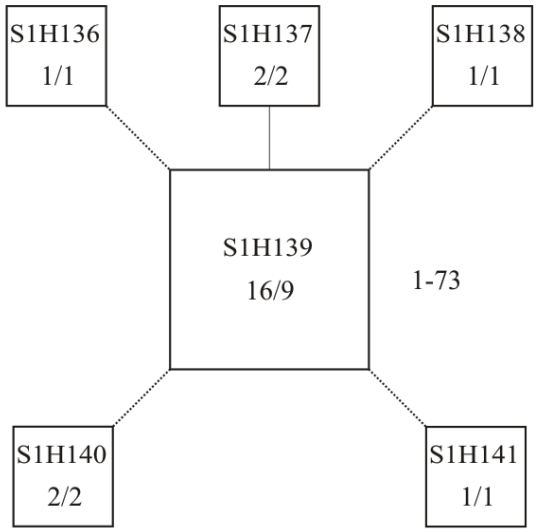


Figure C.4: Nested statistical parsimony network of COI haplotypes in species 1 lineage D The entire cladogram formed a 4-step clade (4-5). The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted by (R).

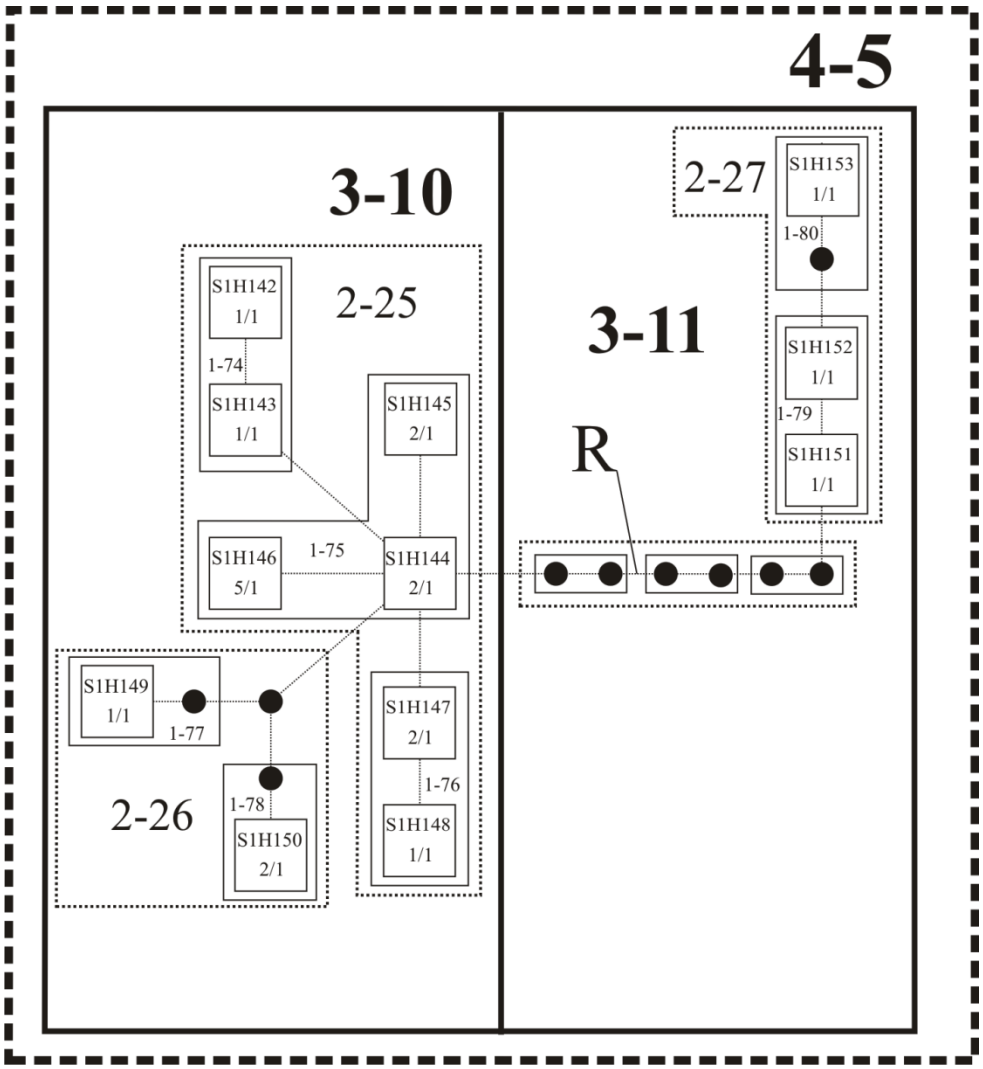


Figure C.5: Nested COI haplotype network of species 1 lineage E
The entire cladogram formed a 4-step clade (4-6). The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and sites (lower). The root is denoted by (R).

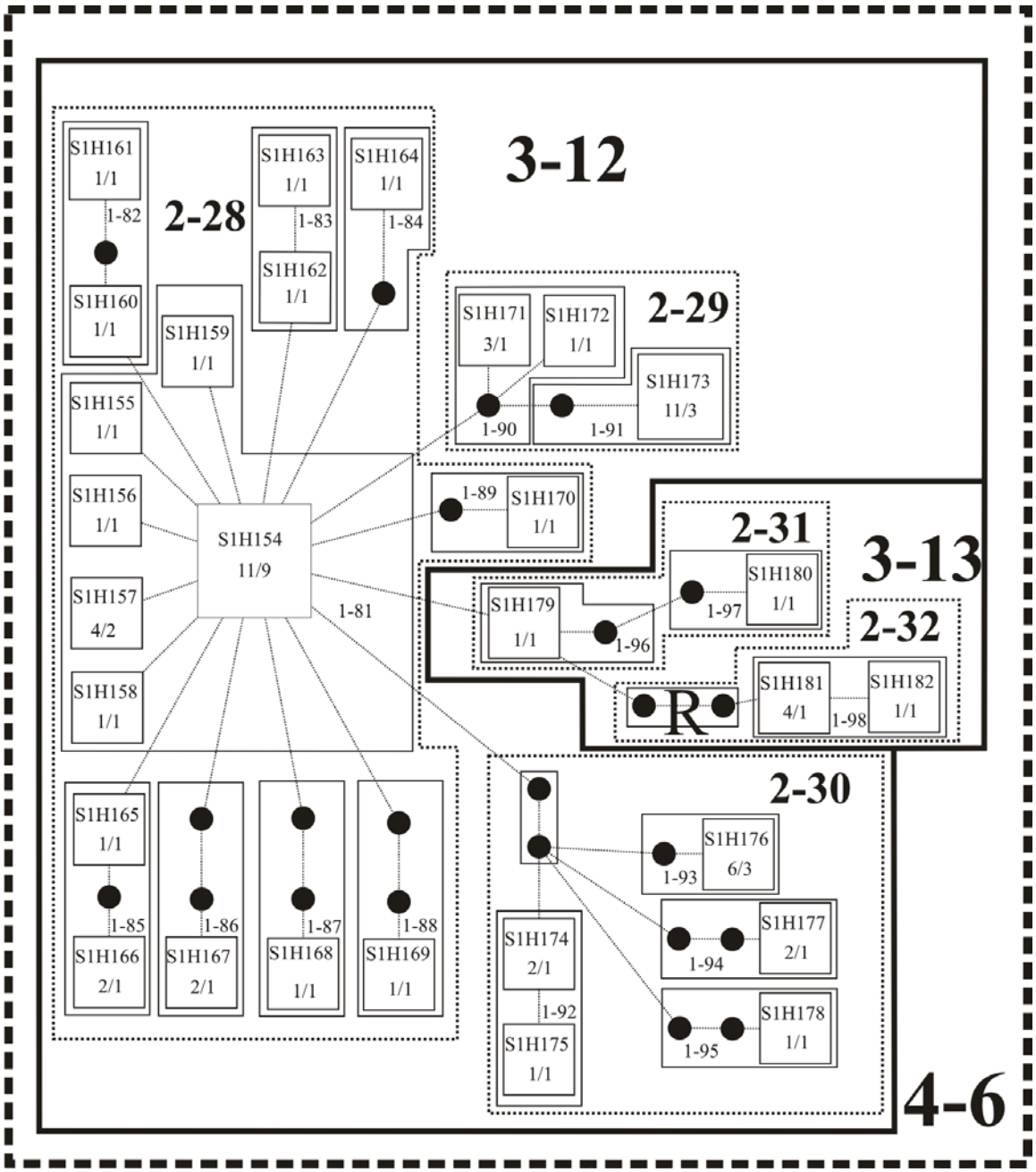


Figure C.6: Nested COI haplotype network of species 1 lineage F
The entire cladogram formed a 3-step clade (3-14). The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). Root is denoted by (R).

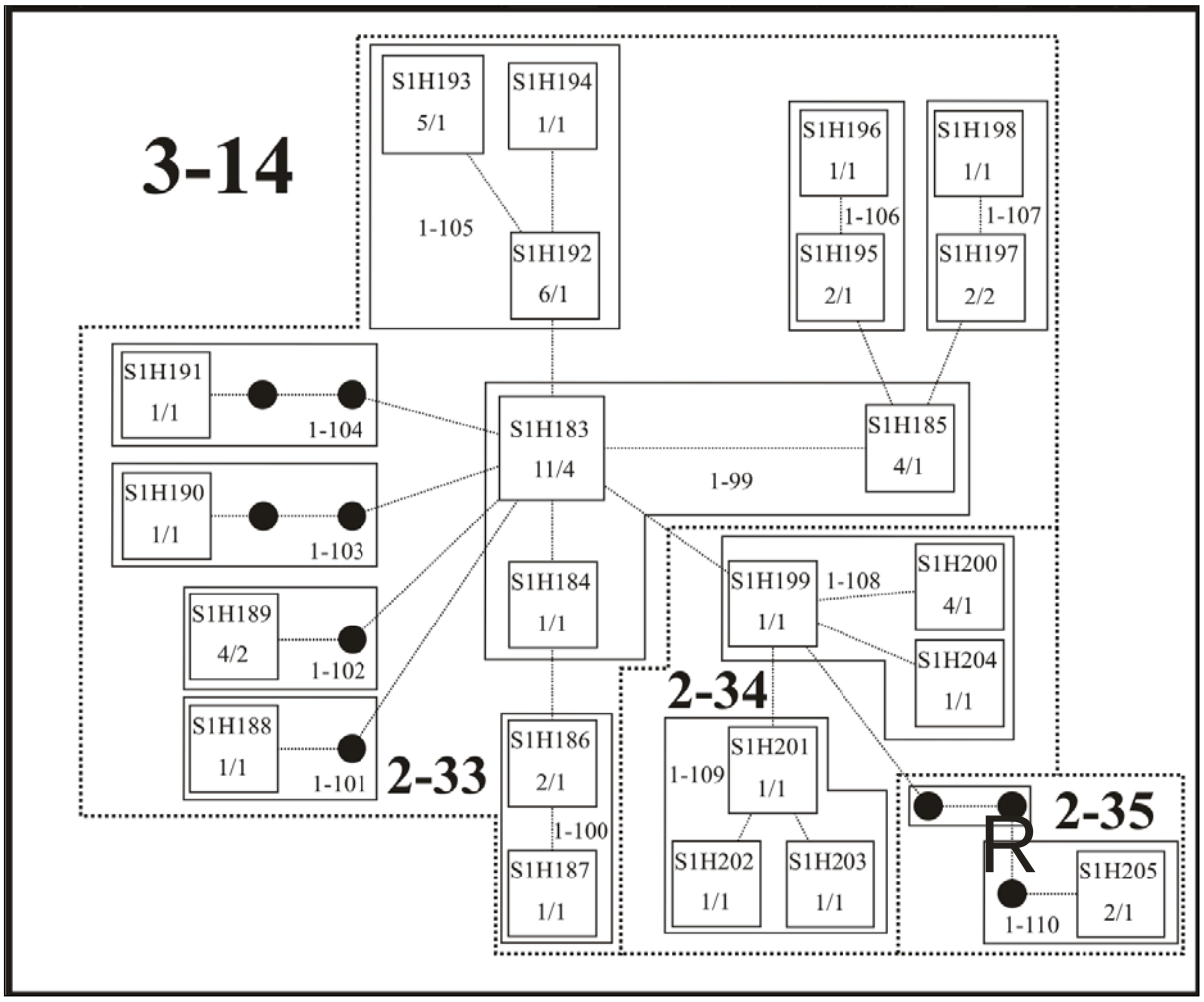


Figure C.7: Nested Haplotype Network of species 1 lineage G
The entire cladogram formed a 4-step (4-7) clade. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted by (R).

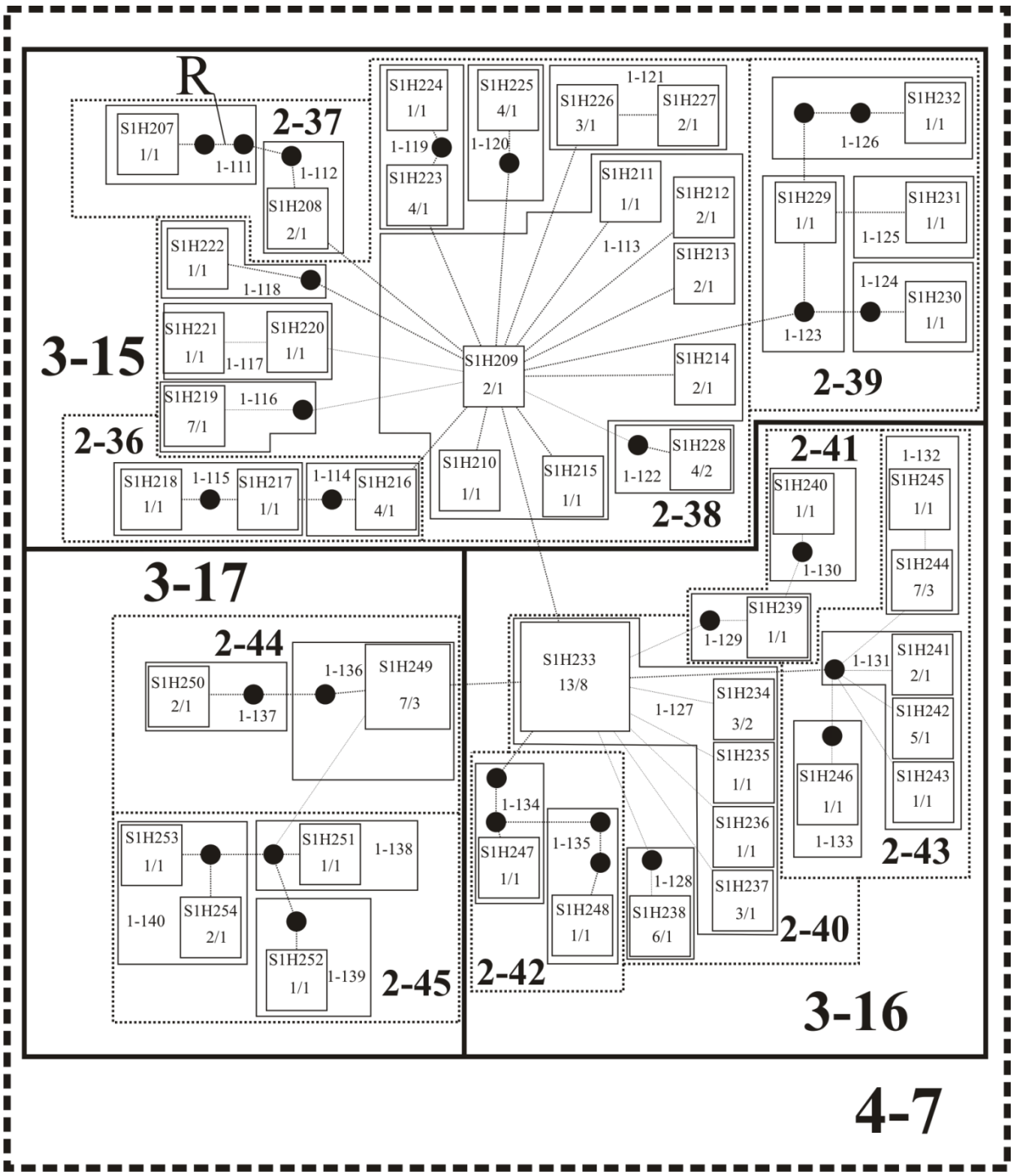


Figure C.8: Nested COI haplotype network of species 1 lineage H
The entire cladogram formed a 5-step (5-3) clade. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root of the network is denoted by (R).

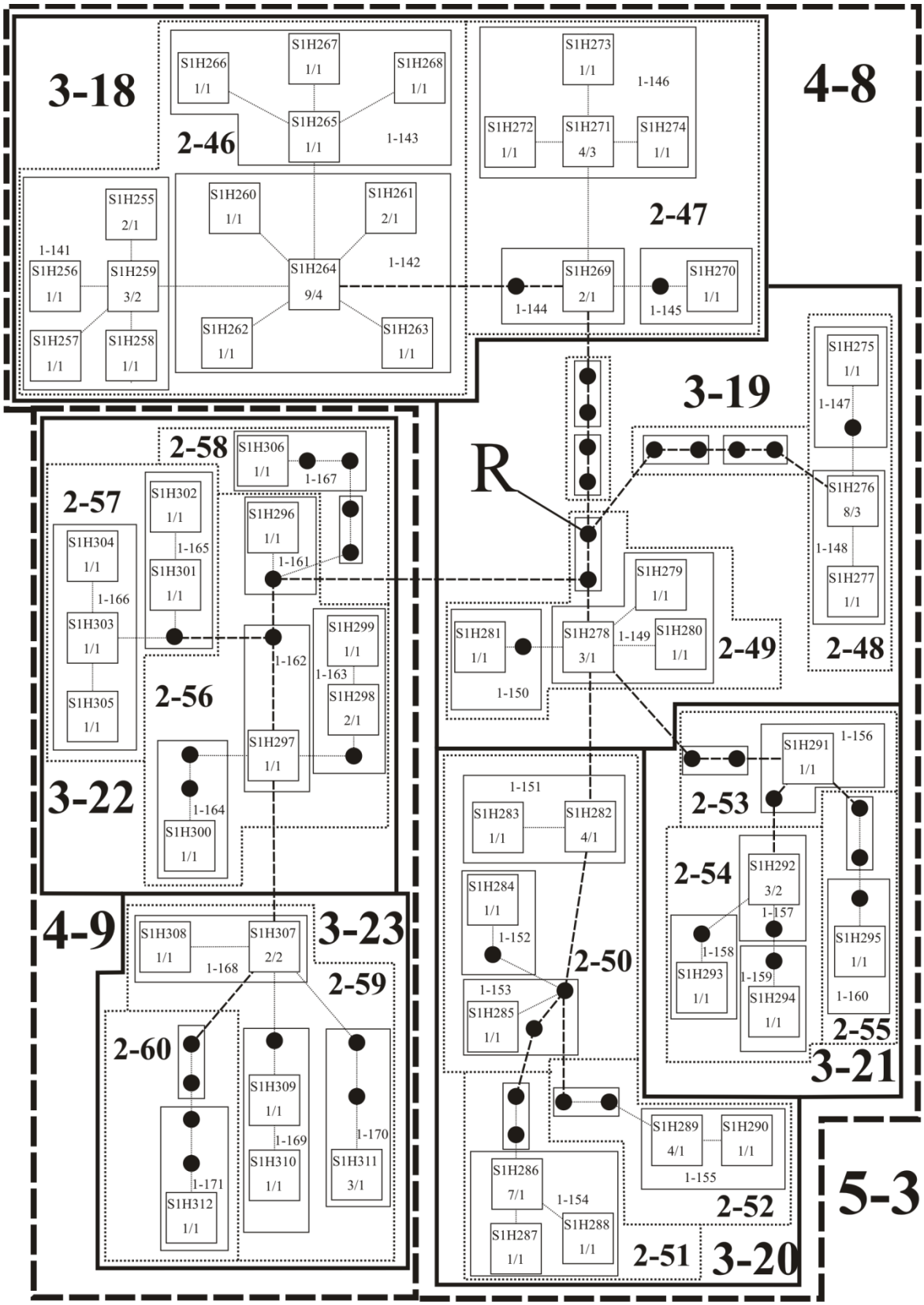

3-18
4-8
2-46
S1H266 1/1
S1H267 1/1
S1H268 1/1
S1H265 1/1
1-143
S1H273 1/1
1-146
S1H272 1/1
S1H271 4/3
S1H274 1/1
2-47
S1H255 2/1
1-141
S1H256 1/1
S1H259 3/2
S1H257 1/1
S1H258 1/1
S1H260 1/1
S1H261 2/1
S1H264 9/4
1-142
S1H262 1/1
S1H263 1/1
1-144
S1H269 2/1
1-145
S1H270 1/1
3-19
S1H275 1/1
1-147
S1H276 8/3
1-148
S1H277 1/1
2-48
R
2-58
S1H306 1/1
1-167
2-57
S1H302 1/1
1-165
S1H301 1/1
S1H304 1/1
1-166
S1H303 1/1
S1H305 1/1
S1H296 1/1
1-161
S1H299 1/1
1-163
S1H298 2/1
1-162
2-56
S1H297 1/1
1-164
S1H300 1/1
3-22
S1H279 1/1
S1H281 1/1
1-150
S1H278 3/1
1-149
S1H280 1/1
2-49
S1H291 1/1
1-156
2-53
1-151
S1H283 1/1
S1H282 4/1
2-54
S1H292 3/2
1-157
S1H284 1/1
1-152
2-50
1-153
S1H285 1/1
1-158
S1H293 1/1
1-159
S1H294 1/1
S1H295 1/1
1-160
2-55
3-21
4-9
S1H308 1/1
1-168
S1H307 2/2
3-23
2-59
2-60
S1H309 1/1
1-169
S1H310 1/1
1-170
S1H311 3/1
1-171
S1H312 1/1
S1H289 4/1
S1H290 1/1
1-155
2-52
S1H286 7/1
1-154
S1H287 1/1
S1H288 1/1
2-51
3-20
5-3

Figure C.9: Nested COI haplotype network of species 1 lineage I
The entire cladogram formed a 4-step (4-9) clade. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted by (R).

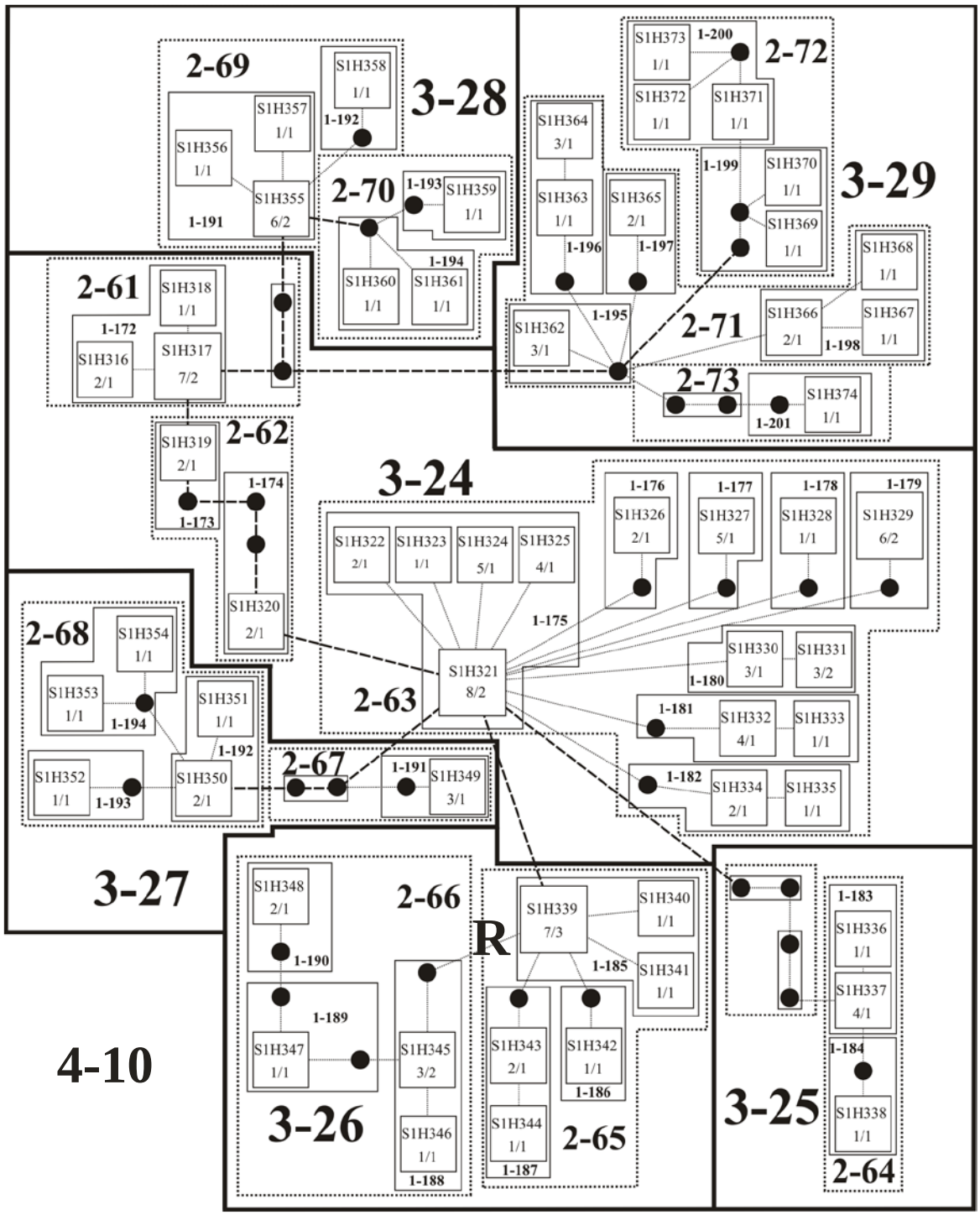
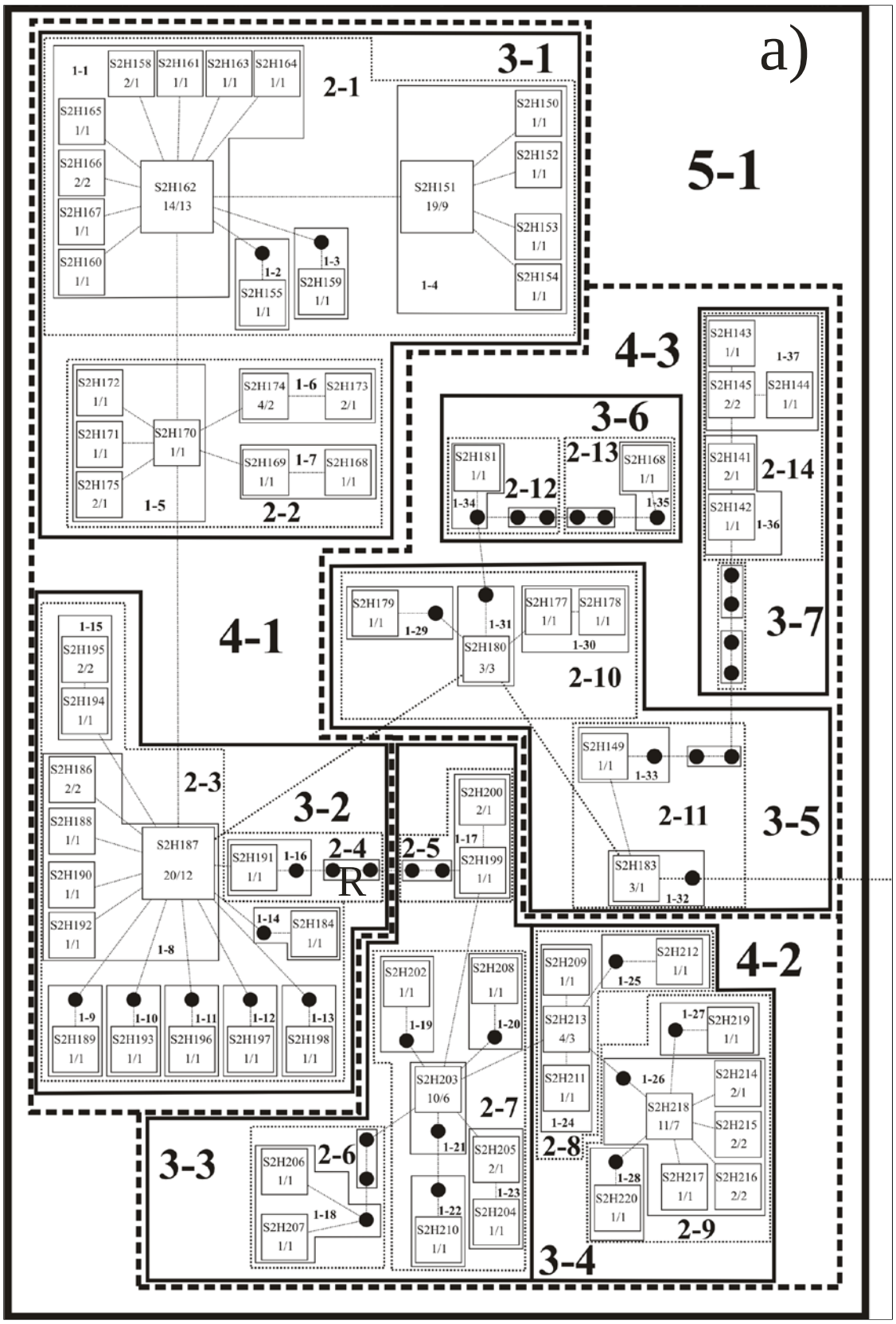
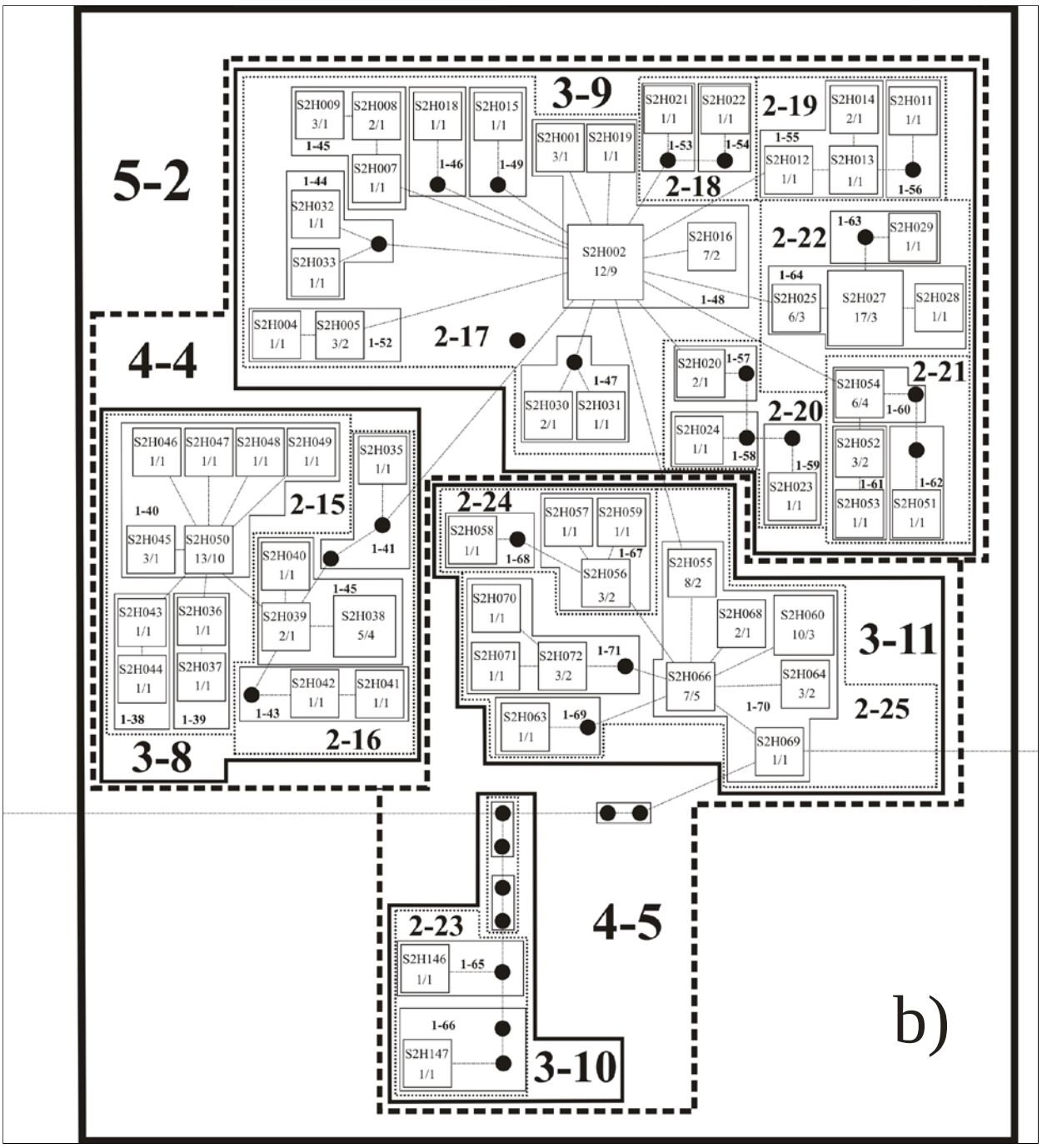


Figure C.10: Nested COI haplotype network of species 2
The entire cladogram formed a 6-step clade with three 5-step clades spread across three pages: a) clade 5-1, b) clade 5-2 and c) clade 5-3. The lines running off the page connect to the next portion of the network on the following page. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted as (R) as determined by the maximum parsimony phylogeny.


a)
5-1
3-1
2-1
1-1
S2H158 2/1
S2H161 1/1
S2H163 1/1
S2H164 1/1
S2H165 1/1
S2H166 2/2
S2H167 1/1
S2H160 1/1
S2H162 14/13
1-2
S2H155 1/1
1-3
S2H159 1/1
S2H151 19/9
1-4
S2H150 1/1
S2H152 1/1
S2H153 1/1
S2H154 1/1
4-3
S2H172 1/1
S2H171 1/1
S2H175 2/1
1-5
S2H170 1/1
S2H174 4/2
1-6
S2H173 2/1
S2H169 1/1
1-7
S2H168 1/1
2-2
3-6
S2H181 1/1
2-12
1-34
2-13
S2H168 1/1
1-35
S2H143 1/1
1-37
S2H145 2/2
S2H144 1/1
S2H141 2/1
2-14
S2H142 1/1
1-36
3-7
4-1
S2H179 1/1
1-29
1-31
S2H180 3/3
S2H177 1/1
S2H178 1/1
1-30
2-10
1-15
S2H195 2/2
S2H194 1/1
S2H186 2/2
S2H188 1/1
S2H190 1/1
S2H192 1/1
2-3
S2H187 20/12
1-8
3-2
S2H191 1/1
1-16
2-4
R
2-5
S2H200 2/1
1-17
S2H199 1/1
S2H149 1/1
1-33
2-11
3-5
S2H183 3/1
1-32
1-14
S2H184 1/1
1-9
S2H189 1/1
1-10
S2H193 1/1
1-11
S2H196 1/1
1-12
S2H197 1/1
1-13
S2H198 1/1
S2H202 1/1
1-19
S2H208 1/1
1-20
S2H203 10/6
2-7
1-21
S2H205 2/1
1-23
1-22
S2H210 1/1
S2H204 1/1
3-3
2-6
S2H206 1/1
S2H207 1/1
1-18
S2H209 1/1
1-25
S2H212 1/1
4-2
S2H213 4/3
S2H211 1/1
1-24
2-8
1-27
S2H219 1/1
1-26
S2H214 2/1
S2H218 11/7
S2H215 2/2
1-28
S2H220 1/1
S2H217 1/1
S2H216 2/2
2-9
3-4


5-2
4-4
3-8
4-5
3-9
2-19
2-18
2-22
2-21
2-20
2-17
2-15
2-16
2-24
3-11
2-25
2-23
3-10
S2H009 3/1
S2H008 2/1
S2H018 1/1
S2H015 1/1
S2H001 3/1
S2H019 1/1
S2H021 1/1
S2H022 1/1
S2H014 2/1
S2H011 1/1
S2H012 1/1
S2H013 1/1
S2H007 1/1
S2H032 1/1
S2H033 1/1
S2H002 12/9
S2H016 7/2
S2H029 1/1
S2H025 6/3
S2H027 17/3
S2H028 1/1
S2H004 1/1
S2H005 3/2
S2H030 2/1
S2H031 1/1
S2H020 2/1
S2H054 6/4
S2H024 1/1
S2H052 3/2
S2H023 1/1
S2H053 1/1
S2H051 1/1
S2H046 1/1
S2H047 1/1
S2H048 1/1
S2H049 1/1
S2H035 1/1
S2H045 3/1
S2H050 13/10
S2H040 1/1
S2H039 2/1
S2H038 5/4
S2H043 1/1
S2H036 1/1
S2H044 1/1
S2H037 1/1
S2H042 1/1
S2H041 1/1
S2H058 1/1
S2H057 1/1
S2H059 1/1
S2H056 3/2
S2H055 8/2
S2H070 1/1
S2H071 1/1
S2H072 3/2
S2H068 2/1
S2H060 10/3
S2H066 7/5
S2H064 3/2
S2H063 1/1
S2H069 1/1
S2H146 1/1
S2H147 1/1
1-45
1-44
1-46
1-49
1-53
1-54
1-55
1-56
1-63
1-64
1-48
1-52
1-47
1-57
1-60
1-58
1-59
1-61
1-62
1-40
1-41
1-45
1-38
1-39
1-43
1-68
1-67
1-71
1-70
1-69
1-65
1-66
b)

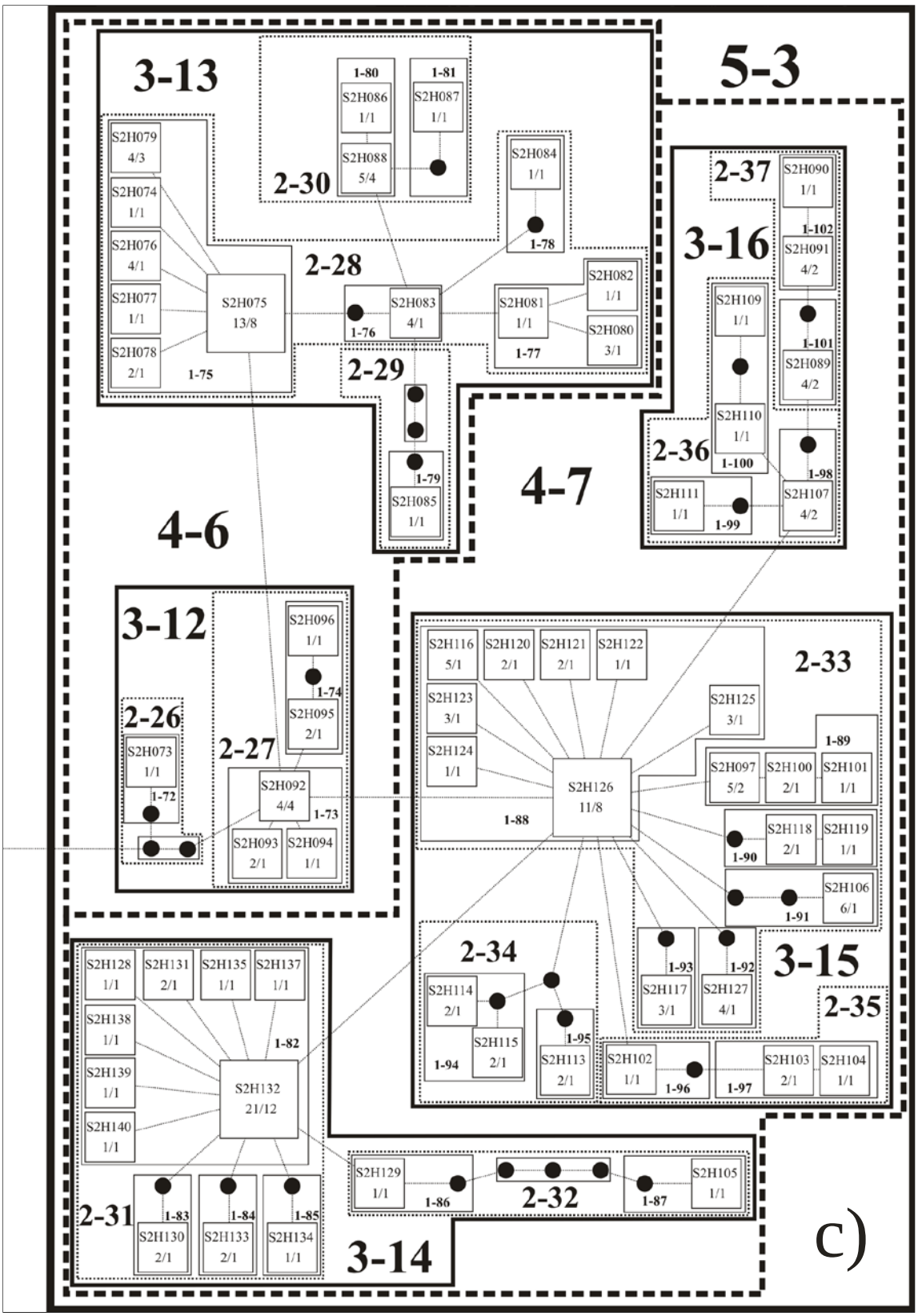
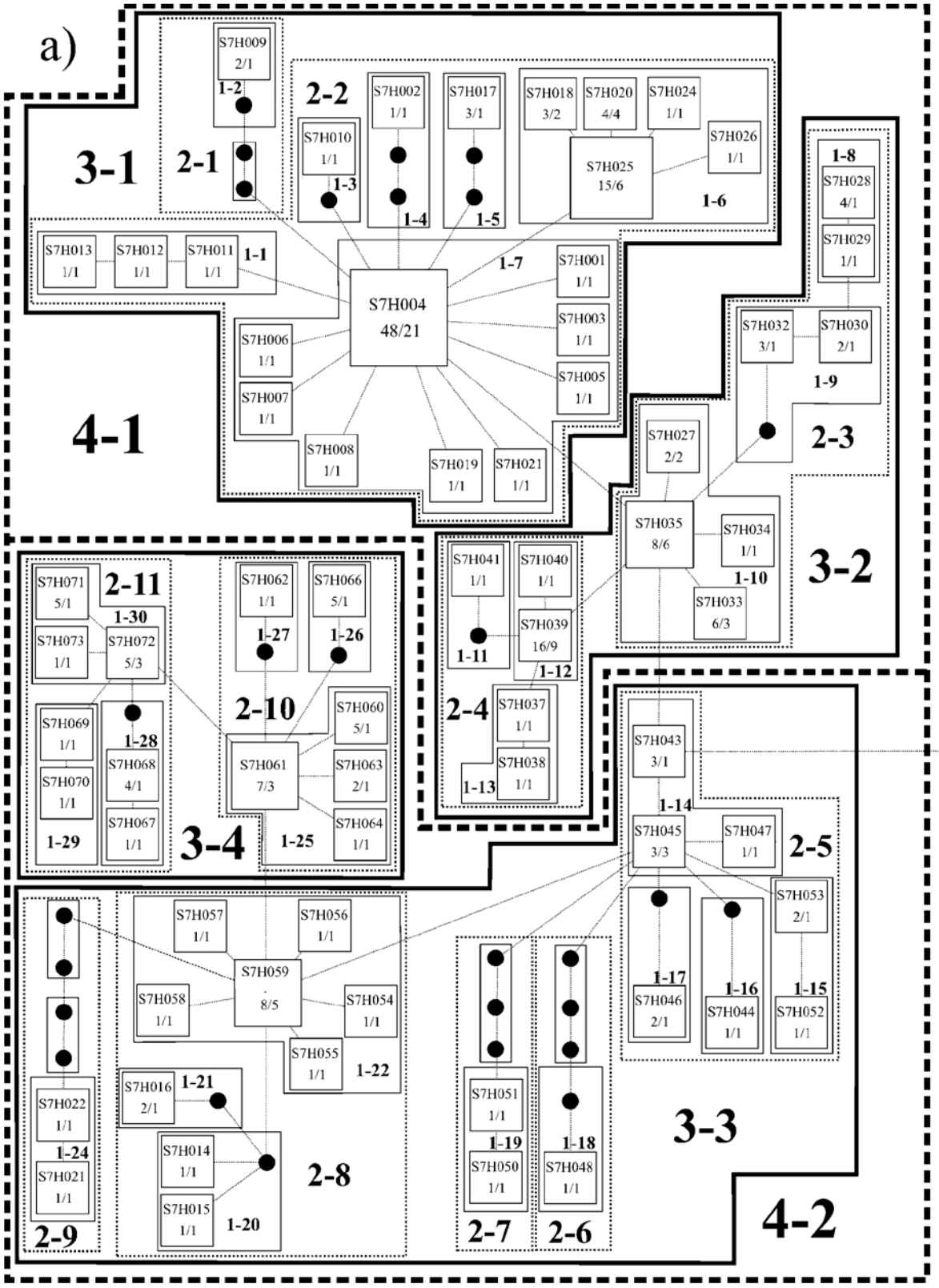

3-13
5-3
2-30
1-80
S2H086
1/1
1-81
S2H087
1/1
S2H088
5/4
S2H084
1/1
1-78
S2H079
4/3
S2H074
1/1
S2H076
4/1
S2H077
1/1
S2H078
2/1
S2H075
13/8
1-75
2-28
S2H083
4/1
1-76
S2H082
1/1
S2H081
1/1
S2H080
3/1
1-77
2-29
1-79
S2H085
1/1
2-37
S2H090
1/1
1-102
3-16
S2H091
4/2
S2H109
1/1
1-101
S2H089
4/2
S2H110
1/1
2-36
1-100
1-98
S2H111
1/1
1-99
S2H107
4/2
4-7
4-6
3-12
S2H096
1/1
1-74
S2H095
2/1
2-26
S2H073
1/1
1-72
2-27
S2H092
4/4
1-73
S2H093
2/1
S2H094
1/1
S2H116
5/1
S2H120
2/1
S2H121
2/1
S2H122
1/1
2-33
S2H123
3/1
S2H124
1/1
S2H125
3/1
1-89
S2H126
11/8
1-88
S2H097
5/2
S2H100
2/1
S2H101
1/1
S2H118
2/1
S2H119
1/1
1-90
S2H106
6/1
1-91
2-34
S2H114
2/1
S2H115
2/1
1-94
1-95
S2H113
2/1
1-93
S2H117
3/1
1-92
S2H127
4/1
3-15
2-35
S2H102
1/1
1-96
1-97
S2H103
2/1
S2H104
1/1
S2H128
1/1
S2H131
2/1
S2H135
1/1
S2H137
1/1
S2H138
1/1
S2H139
1/1
S2H140
1/1
1-82
S2H132
21/12
S2H129
1/1
1-86
2-32
1-87
S2H105
1/1
2-31
1-83
S2H130
2/1
1-84
S2H133
2/1
1-85
S2H134
1/1
3-14
c)

Figure C.11: Nested haplotype network of species 7
The entire cladogram formed a 5-step clade with three 4-step clades (spread across two pages): a) clade 4-1 and 4-2, b) clade 4-3. The lines running off the page connect to the next portion of the network on the following page. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted by (R).


a)
S7H009 2/1
1-2
2-2
S7H002 1/1
S7H017 3/1
S7H018 3/2
S7H020 4/4
S7H024 1/1
S7H010 1/1
1-3
1-4
1-5
S7H025 15/6
S7H026 1/1
1-6
1-8
S7H028 4/1
S7H029 1/1
3-1
2-1
S7H013 1/1
S7H012 1/1
S7H011 1/1
1-1
1-7
S7H004 48/21
S7H001 1/1
S7H003 1/1
S7H005 1/1
S7H006 1/1
S7H007 1/1
S7H008 1/1
S7H019 1/1
S7H021 1/1
S7H032 3/1
S7H030 2/1
1-9
2-3
4-1
S7H027 2/2
S7H035 8/6
S7H034 1/1
1-10
S7H033 6/3
3-2
S7H041 1/1
S7H040 1/1
1-11
S7H039 16/9
1-12
2-4
S7H037 1/1
S7H038 1/1
1-13
S7H071 5/1
2-11
1-30
S7H073 1/1
S7H072 5/3
S7H062 1/1
S7H066 5/1
1-27
1-26
2-10
S7H060 5/1
S7H069 1/1
1-28
S7H061 7/3
S7H063 2/1
S7H070 1/1
S7H068 4/1
S7H067 1/1
1-29
3-4
1-25
S7H064 1/1
S7H043 3/1
1-14
S7H045 3/3
S7H047 1/1
2-5
S7H053 2/1
1-17
1-16
1-15
S7H046 2/1
S7H044 1/1
S7H052 1/1
S7H057 1/1
S7H056 1/1
S7H059 8/5
S7H058 1/1
S7H054 1/1
S7H055 1/1
1-22
S7H016 2/1
1-21
S7H022 1/1
1-24
S7H014 1/1
S7H021 1/1
S7H015 1/1
1-20
2-8
2-9
S7H051 1/1
1-19
S7H050 1/1
1-18
S7H048 1/1
3-3
2-7
2-6
4-2

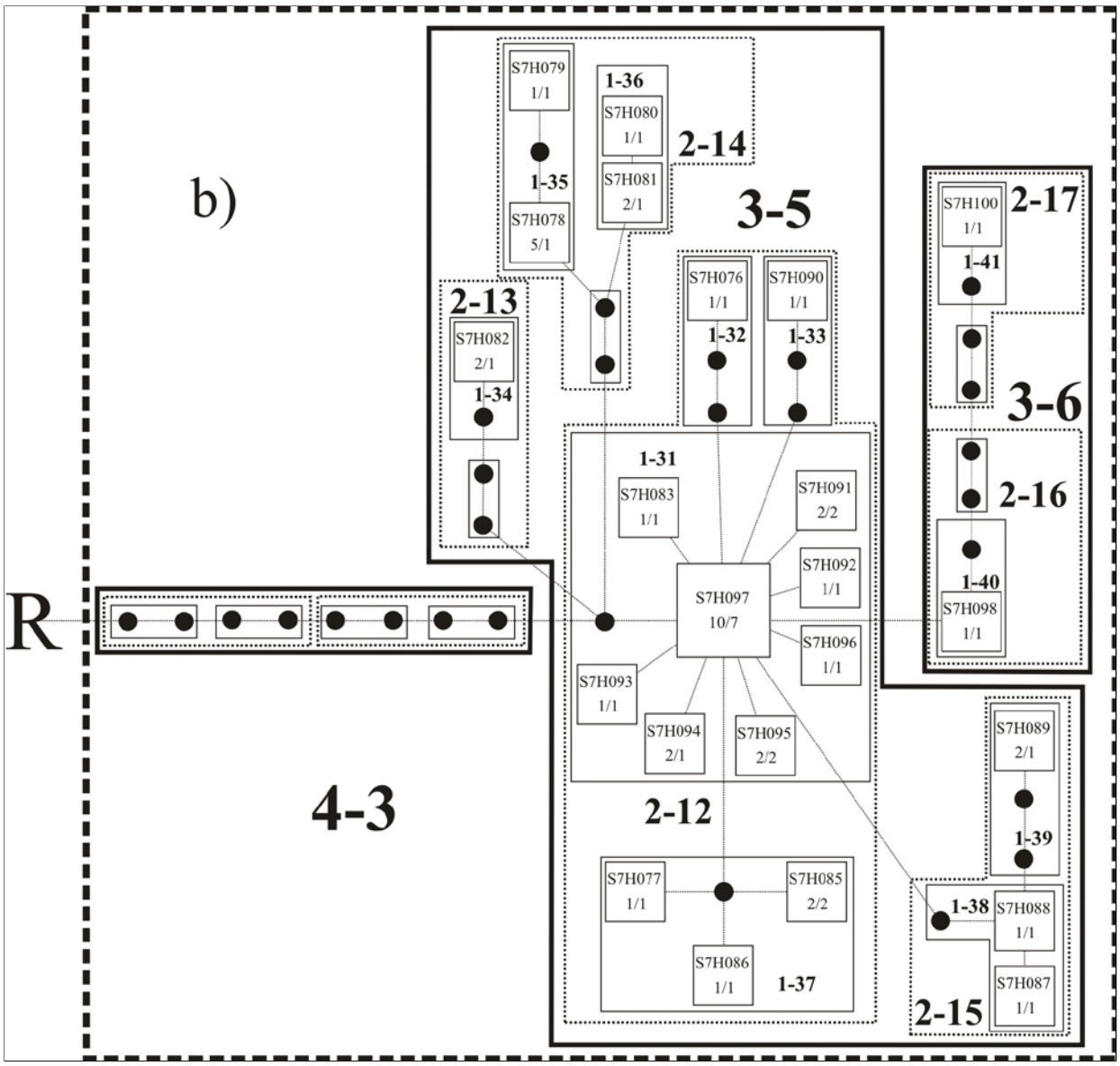
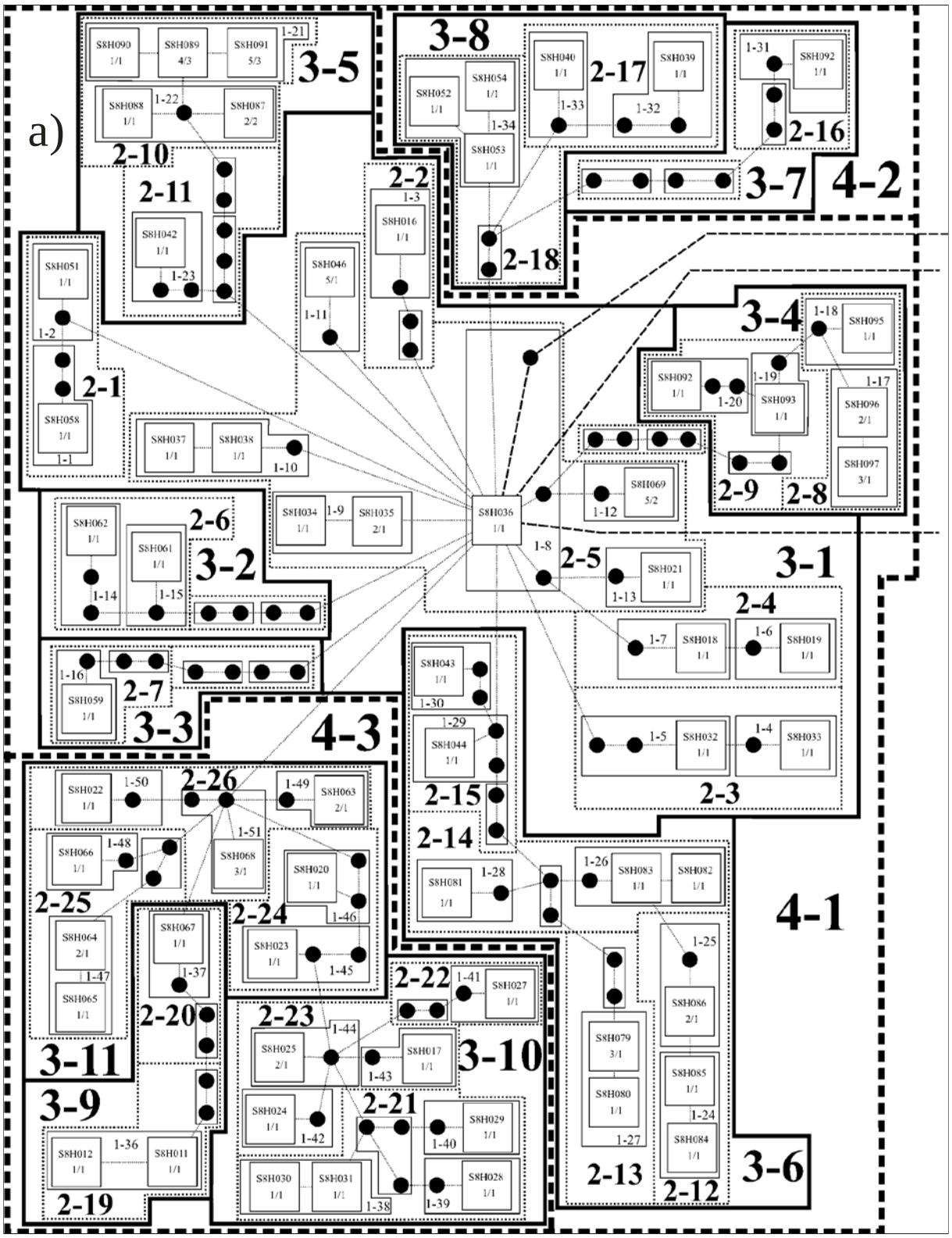

b)
R
4-3
2-12
2-13
2-14
3-5
3-6
2-15
2-16
2-17
S7H079 1/1
1-35
S7H078 5/1
1-36
S7H080 1/1
S7H081 2/1
S7H082 2/1
1-34
S7H076 1/1
1-32
S7H090 1/1
1-33
1-31
S7H083 1/1
S7H091 2/2
S7H092 1/1
S7H097 10/7
S7H096 1/1
S7H093 1/1
S7H094 2/1
S7H095 2/2
S7H077 1/1
S7H085 2/2
S7H086 1/1
1-37
S7H100 1/1
1-41
1-40
S7H098 1/1
S7H089 2/1
1-39
1-38
S7H088 1/1
S7H087 1/1

Figure C.11: Nested haplotype network of species 8
The entire cladogram formed a 5-step clade with six 4-step clades: a) clades 4-1, 4-2 and 4-3, b) clades 4-4, 4-5 and 4-6. The lines running off the page connect to the next portion of the network on the following page. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower).


a)
3-5
3-8
3-7
4-2
2-10
2-11
2-2
2-17
2-16
2-18
3-4
2-1
3-2
2-6
3-1
2-5
2-4
2-7
3-3
4-3
2-15
2-14
2-3
4-1
2-26
2-25
2-24
2-22
2-23
2-20
3-11
3-10
3-9
2-21
2-13
2-12
3-6
2-19
S8H090 1/1
S8H089 4/3
S8H091 5/3
S8H088 1/1
S8H087 2/2
S8H042 1/1
S8H046 5/1
S8H052 1/1
S8H054 1/1
S8H053 1/1
S8H040 1/1
S8H039 1/1
S8H092 1/1
S8H016 1/1
S8H051 1/1
S8H058 1/1
S8H037 1/1
S8H038 1/1
S8H095 1/1
S8H092 1/1
S8H093 1/1
S8H096 2/1
S8H097 3/1
S8H069 5/2
S8H034 1/1
S8H035 2/1
S8H036 1/1
S8H062 1/1
S8H061 1/1
S8H021 1/1
S8H018 1/1
S8H019 1/1
S8H059 1/1
S8H043 1/1
S8H044 1/1
S8H032 1/1
S8H033 1/1
S8H022 1/1
S8H063 2/1
S8H066 1/1
S8H068 3/1
S8H020 1/1
S8H081 1/1
S8H083 1/1
S8H082 1/1
S8H064 2/1
S8H065 1/1
S8H067 1/1
S8H023 1/1
S8H027 1/1
S8H025 2/1
S8H017 1/1
S8H024 1/1
S8H029 1/1
S8H030 1/1
S8H031 1/1
S8H028 1/1
S8H079 3/1
S8H080 1/1
S8H086 2/1
S8H085 1/1
S8H084 1/1
S8H012 1/1
S8H011 1/1

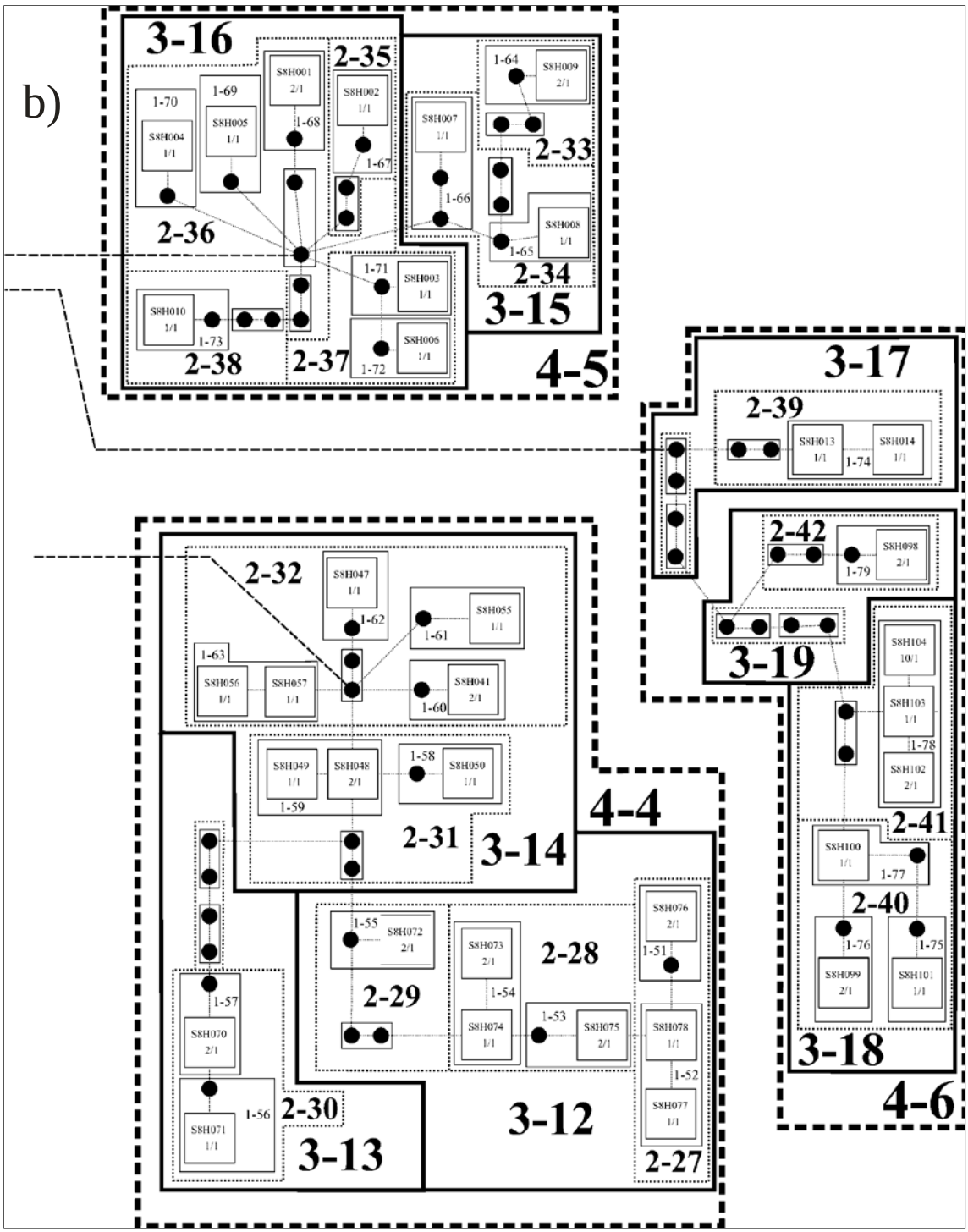
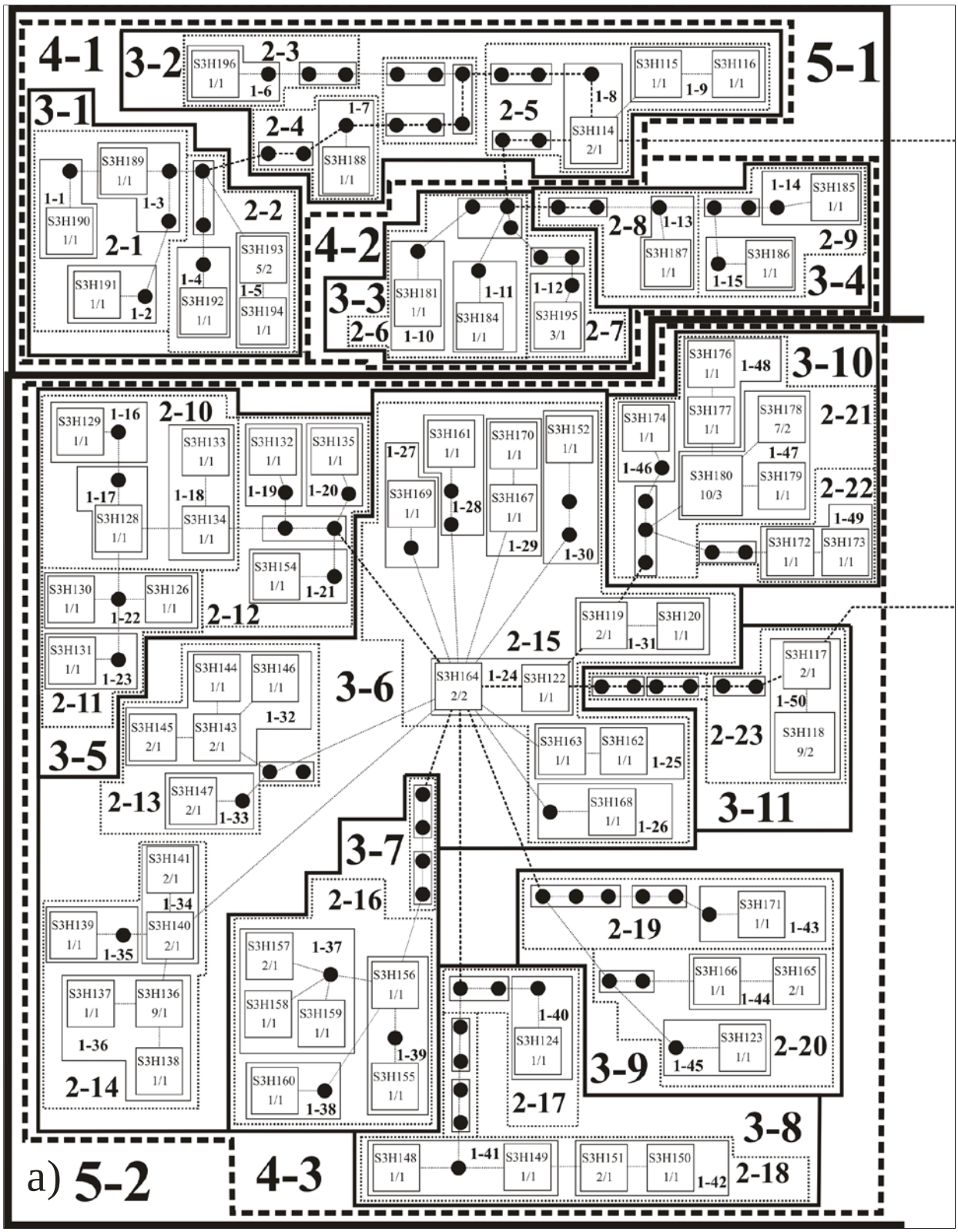
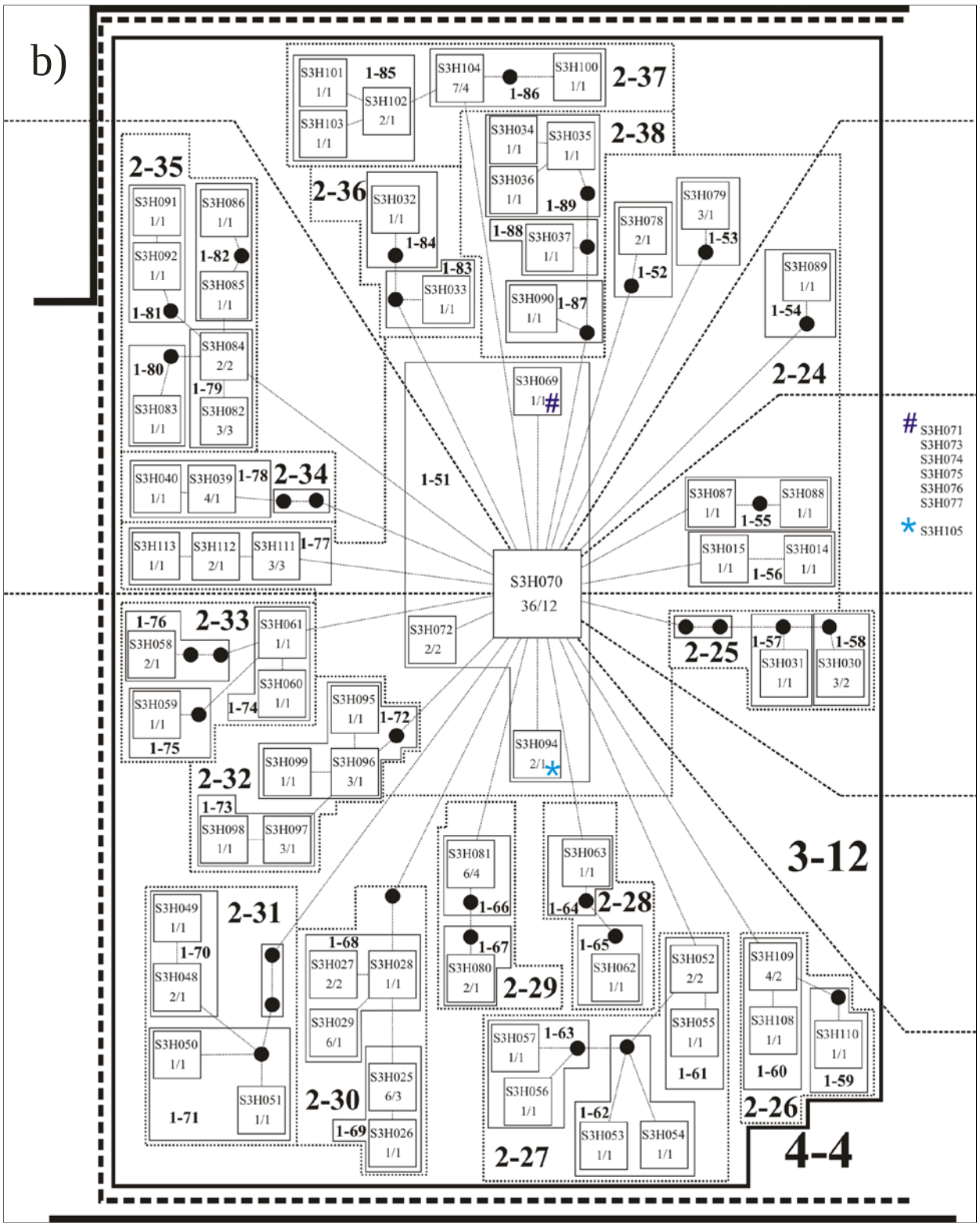

b)
3-16
2-35
S8H001 2/1
S8H002 1/1
1-70
S8H004 1/1
1-69
S8H005 1/1
1-68
1-67
2-36
S8H007 1/1
1-64
S8H009 2/1
2-33
1-66
S8H008 1/1
1-65
2-34
3-15
1-71
S8H003 1/1
S8H010 1/1
1-73
2-38
2-37
1-72
S8H006 1/1
4-5
3-17
2-39
S8H013 1/1
1-74
S8H014 1/1
2-42
S8H098 2/1
1-79
3-19
S8H104 10/1
S8H103 1/1
1-78
S8H102 2/1
2-41
S8H100 1/1
1-77
2-40
1-76
1-75
S8H099 2/1
S8H101 1/1
3-18
4-6
2-32
S8H047 1/1
1-62
S8H055 1/1
1-61
1-63
S8H056 1/1
S8H057 1/1
S8H041 2/1
1-60
S8H049 1/1
S8H048 2/1
1-58
S8H050 1/1
1-59
2-31
3-14
4-4
1-55
S8H072 2/1
S8H073 2/1
2-28
S8H076 2/1
1-51
2-29
1-54
S8H074 1/1
1-53
S8H075 2/1
S8H078 1/1
1-52
S8H077 1/1
2-27
1-57
S8H070 2/1
S8H071 1/1
1-56
2-30
3-13
3-12

Figure C.12: Nested COI haplotype network of species 3
The entire cladogram formed a 6-step clade with two 5-step clades spread across three pages: a) clades 4-3 along with 5-1, b) first half of clade 4-4 and c) second half of clade 4-4. The lines running off the page connect to the next portion of the network on the following page. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The haplotypes denoted by (#) and (*) are directly attached to haplotype S3H070. R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.


4-1
3-2
3-1
5-1
4-2
3-3
3-4
3-10
2-10
2-21
2-22
2-12
2-15
3-6
2-11
3-5
2-13
3-7
2-16
3-11
2-23
2-19
2-20
3-9
2-17
3-8
2-18
2-14
a) 5-2
4-3
S3H196 1/1
S3H115 1/1
S3H116 1/1
S3H114 2/1
S3H189 1/1
S3H190 1/1
S3H188 1/1
S3H185 1/1
S3H187 1/1
S3H186 1/1
S3H191 1/1
S3H193 5/2
S3H192 1/1
S3H194 1/1
S3H181 1/1
S3H184 1/1
S3H195 3/1
S3H176 1/1
S3H174 1/1
S3H177 1/1
S3H178 7/2
S3H180 10/3
S3H179 1/1
S3H172 1/1
S3H173 1/1
S3H129 1/1
S3H133 1/1
S3H132 1/1
S3H135 1/1
S3H128 1/1
S3H134 1/1
S3H154 1/1
S3H130 1/1
S3H126 1/1
S3H131 1/1
S3H161 1/1
S3H170 1/1
S3H152 1/1
S3H169 1/1
S3H167 1/1
S3H119 2/1
S3H120 1/1
S3H164 2/2
S3H122 1/1
S3H117 2/1
S3H118 9/2
S3H163 1/1
S3H162 1/1
S3H168 1/1
S3H144 1/1
S3H146 1/1
S3H145 2/1
S3H143 2/1
S3H147 2/1
S3H141 2/1
S3H139 1/1
S3H140 2/1
S3H137 1/1
S3H136 9/1
S3H138 1/1
S3H157 2/1
S3H158 1/1
S3H159 1/1
S3H156 1/1
S3H155 1/1
S3H160 1/1
S3H171 1/1
S3H166 1/1
S3H165 2/1
S3H123 1/1
S3H124 1/1
S3H148 1/1
S3H149 1/1
S3H151 2/1
S3H150 1/1
1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11
1-12
1-13
1-14
1-15
1-16
1-17
1-18
1-19
1-20
1-21
1-22
1-23
1-24
1-25
1-26
1-27
1-28
1-29
1-30
1-31
1-32
1-33
1-34
1-35
1-36
1-37
1-38
1-39
1-40
1-41
1-42
1-43
1-44
1-45
1-46
1-47
1-48
1-49
1-50


b)
S3H101 1/1
1-85
S3H102 2/1
S3H103 1/1
S3H104 7/4
1-86
S3H100 1/1
2-37
S3H034 1/1
S3H035 1/1
2-38
S3H036 1/1
1-89
2-35
2-36
S3H032 1/1
1-84
S3H078 2/1
1-52
S3H079 3/1
1-53
S3H091 1/1
S3H086 1/1
1-82
S3H092 1/1
S3H085 1/1
1-88
S3H037 1/1
1-83
S3H033 1/1
S3H090 1/1
1-87
S3H089 1/1
1-54
1-81
S3H084 2/2
1-80
1-79
S3H083 1/1
S3H082 3/3
2-24
S3H069 1/1 #
S3H071 S3H073 S3H074 S3H075 S3H076 S3H077
* S3H105
S3H040 1/1
S3H039 4/1
1-78
2-34
1-51
S3H087 1/1
1-55
S3H088 1/1
S3H113 1/1
S3H112 2/1
S3H111 3/3
1-77
S3H015 1/1
1-56
S3H014 1/1
S3H070 36/12
1-76
2-33
S3H058 2/1
S3H061 1/1
S3H072 2/2
2-25
1-57
1-58
S3H031 1/1
S3H030 3/2
S3H059 1/1
1-75
S3H060 1/1
1-74
S3H095 1/1
1-72
S3H099 1/1
S3H096 3/1
2-32
S3H094 2/1 *
1-73
S3H098 1/1
S3H097 3/1
3-12
S3H081 6/4
1-66
S3H063 1/1
1-64
2-28
S3H049 1/1
2-31
1-70
S3H048 2/1
1-68
S3H027 2/2
S3H028 1/1
1-67
S3H080 2/1
2-29
1-65
S3H062 1/1
S3H052 2/2
S3H055 1/1
1-61
S3H109 4/2
S3H108 1/1
1-60
S3H110 1/1
1-59
S3H029 6/1
S3H050 1/1
S3H057 1/1
1-63
S3H056 1/1
S3H025 6/3
2-30
1-71
S3H051 1/1
1-69
S3H026 1/1
1-62
S3H053 1/1
S3H054 1/1
2-26
2-27
4-4

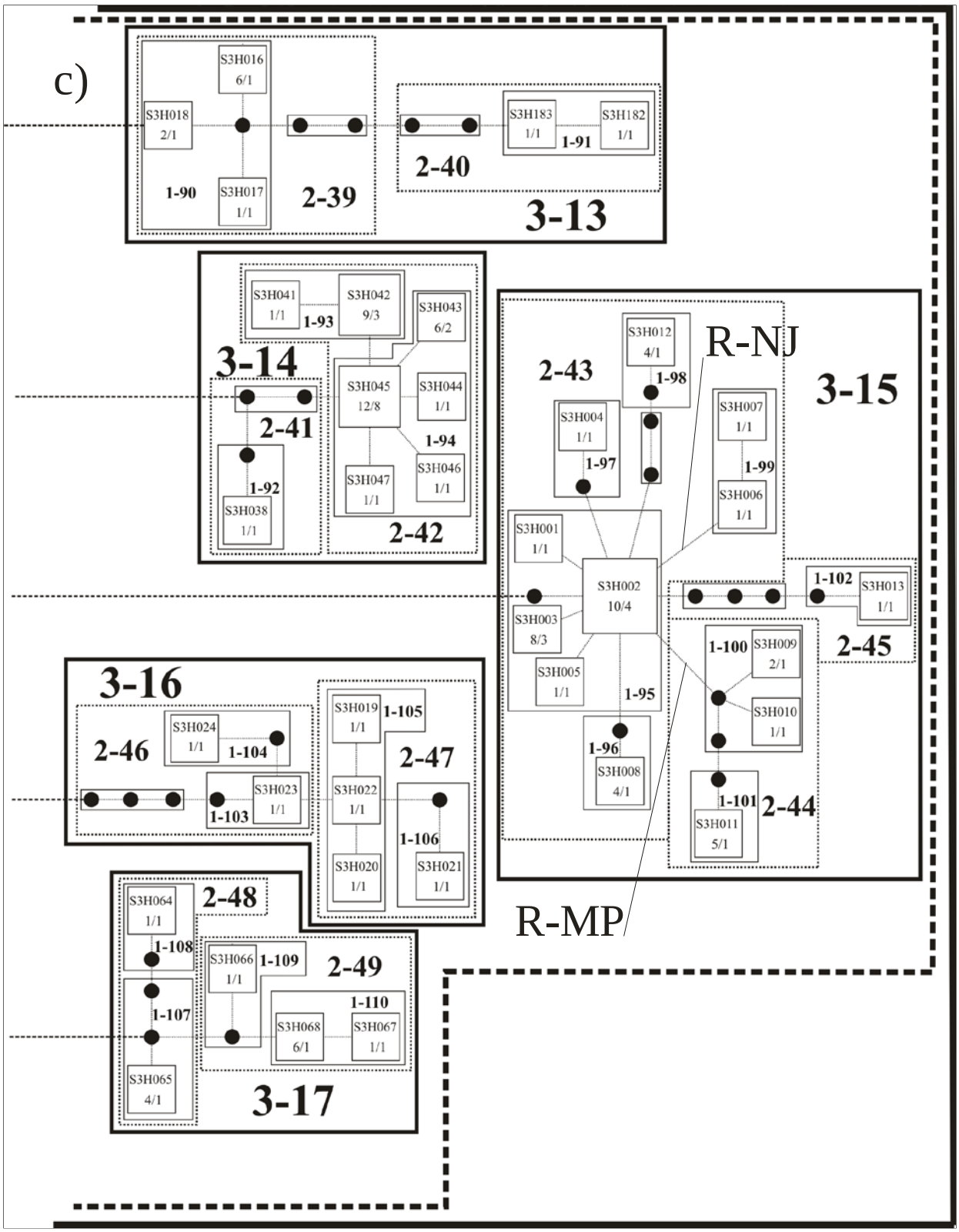

c)
S3H016
6/1
S3H018
2/1
S3H017
1/1
1-90
2-39
S3H183
1/1
1-91
S3H182
1/1
2-40
3-13
S3H041
1/1
1-93
S3H042
9/3
S3H043
6/2
3-14
S3H045
12/8
S3H044
1/1
1-94
S3H046
1/1
S3H047
1/1
2-41
1-92
S3H038
1/1
2-42
S3H012
4/1
1-98
R-NJ
2-43
3-15
S3H004
1/1
1-97
S3H007
1/1
1-99
S3H006
1/1
S3H001
1/1
S3H002
10/4
S3H003
8/3
S3H005
1/1
1-95
1-102
S3H013
1/1
2-45
1-100
S3H009
2/1
S3H010
1/1
1-96
S3H008
4/1
1-101
S3H011
5/1
2-44
3-16
S3H024
1/1
1-104
2-46
S3H023
1/1
1-103
S3H019
1/1
1-105
2-47
S3H022
1/1
S3H020
1/1
1-106
S3H021
1/1
R-MP
S3H064
1/1
1-108
2-48
1-107
S3H065
4/1
S3H066
1/1
1-109
2-49
1-110
S3H068
6/1
S3H067
1/1
3-17

Figure: C.13: Nested COI haplotype network of species 4

The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The statistical parsimony criterion split the network into 3 major lineages: a) lineage A, b) lineage B, and c) lineage C. R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.

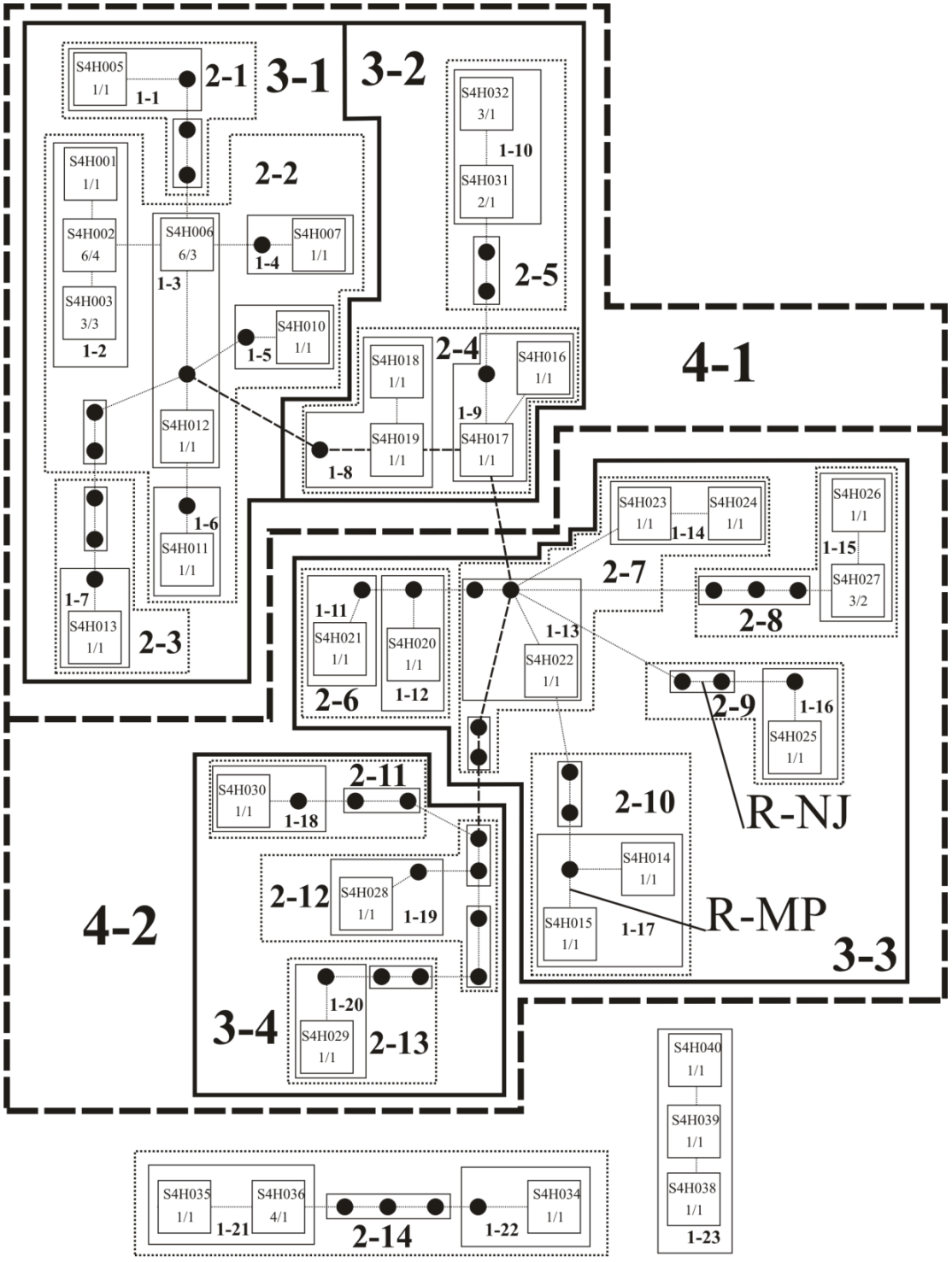


Figure: C.13: Nested COI haplotype network of species 5 lineage A

The entire cladogram formed a 5-step clade. The numbers associated with each non-inferred node gives the species and haplotype (upper), and sample size and number of sites (lower). R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.

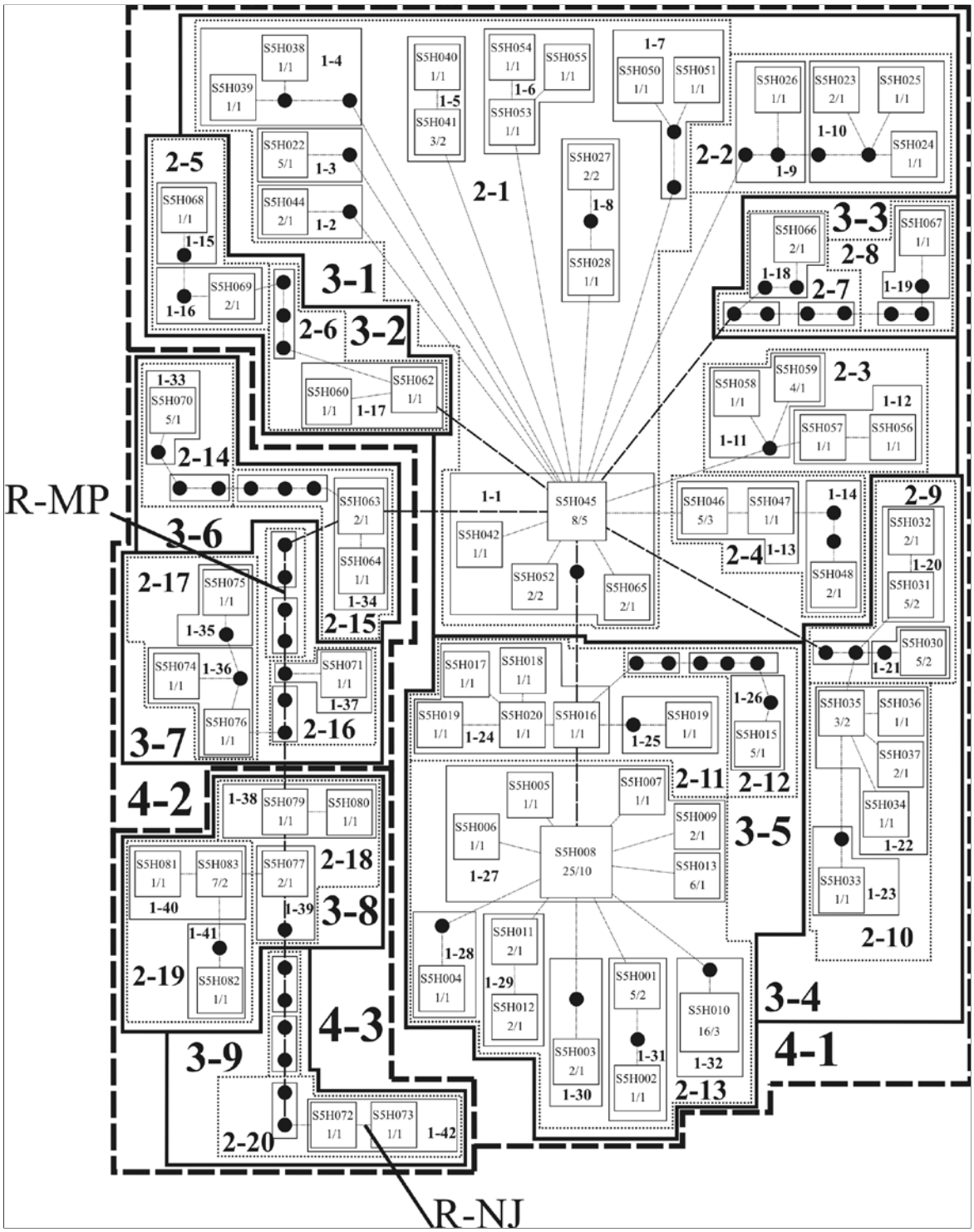


Figure: C.14: Nested COI haplotype network of species 5 lineage B
The entire cladogram formed a 4-step clade. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower).

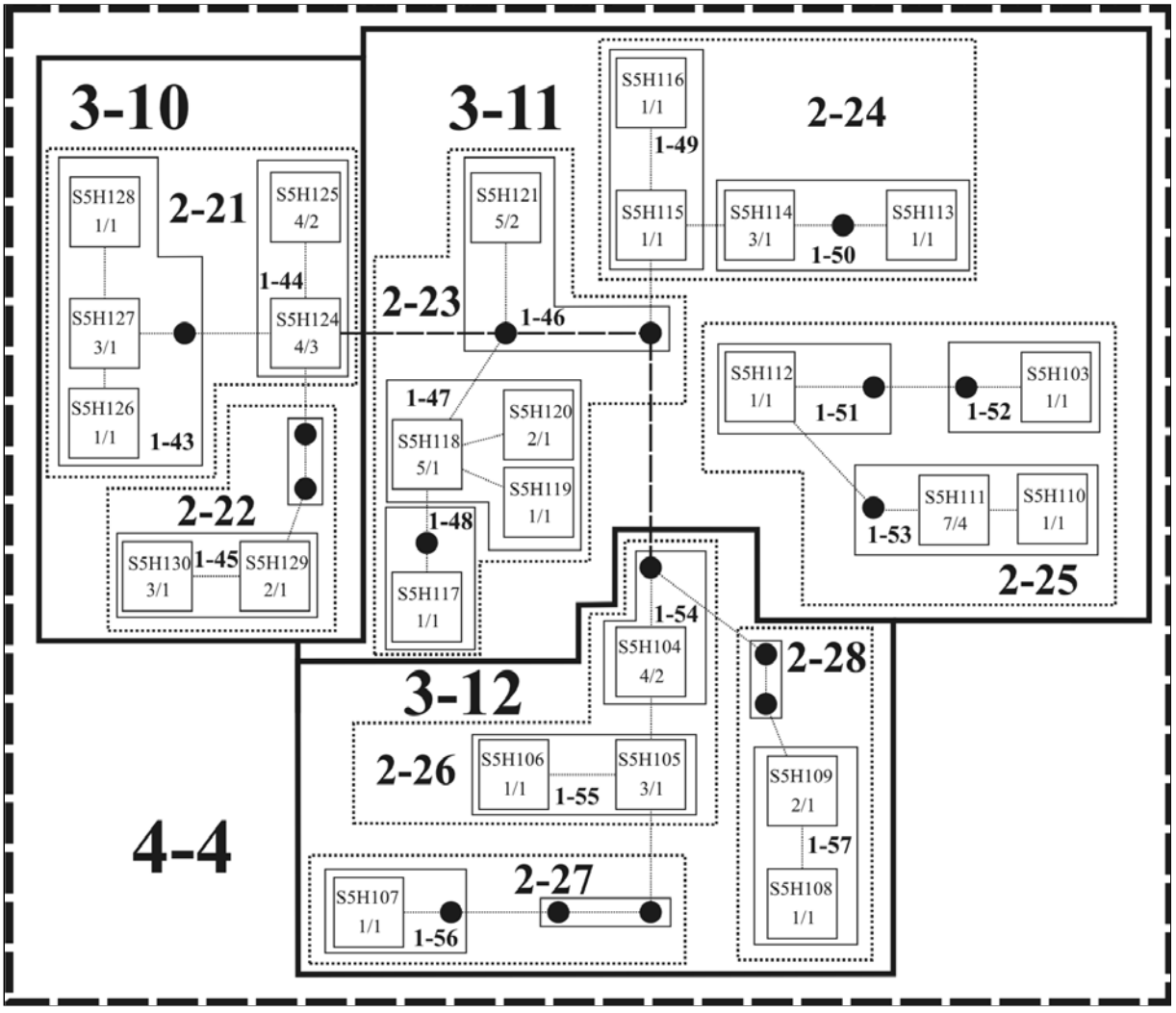


Figure: C.15: Nested COI haplotype network of species 5 lineages C, D and E
The haplotype networks are as follows: a) Lineage C (clade 4-5), b) lineage D (clade 1-64) and c) lineage E (clade 3-15). The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower).

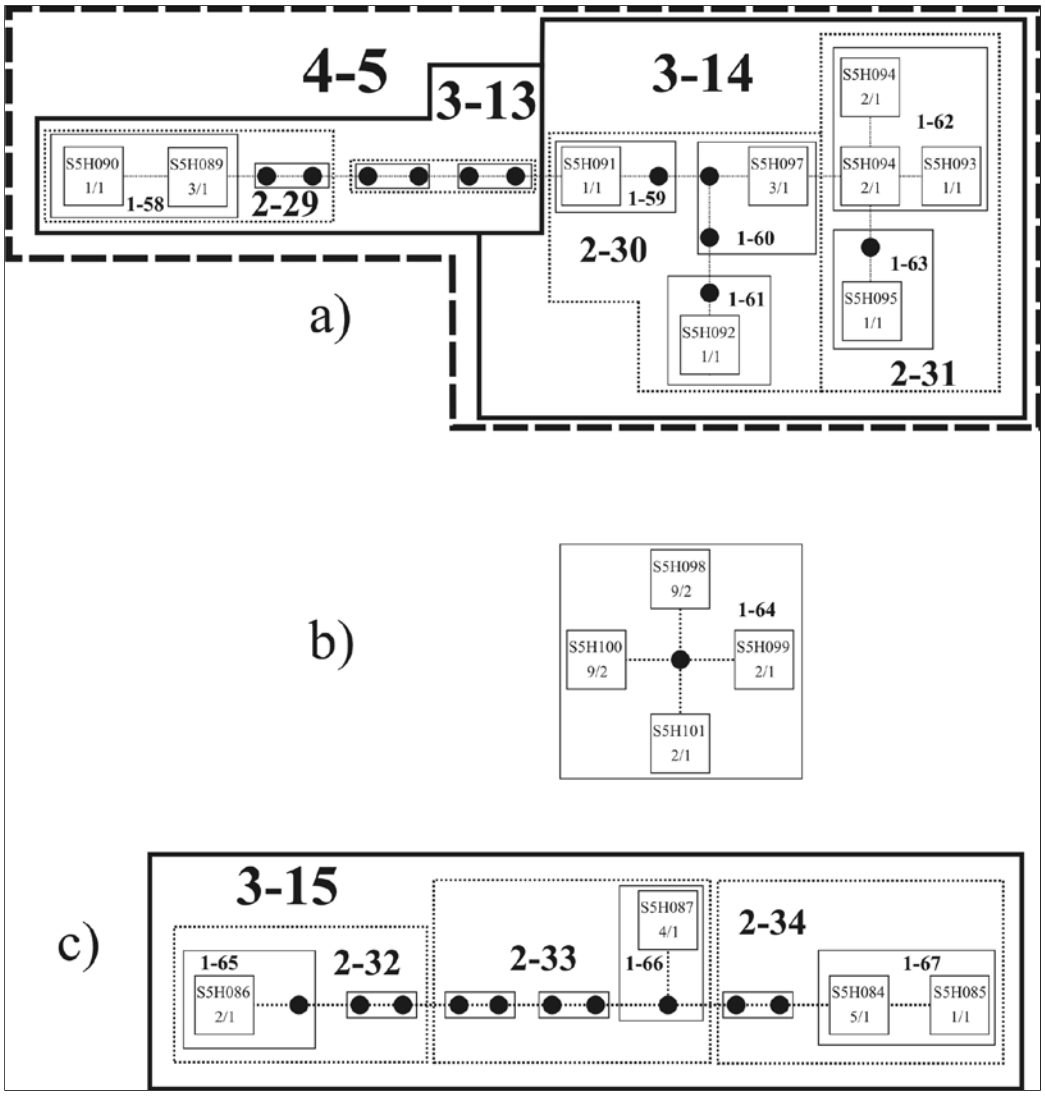
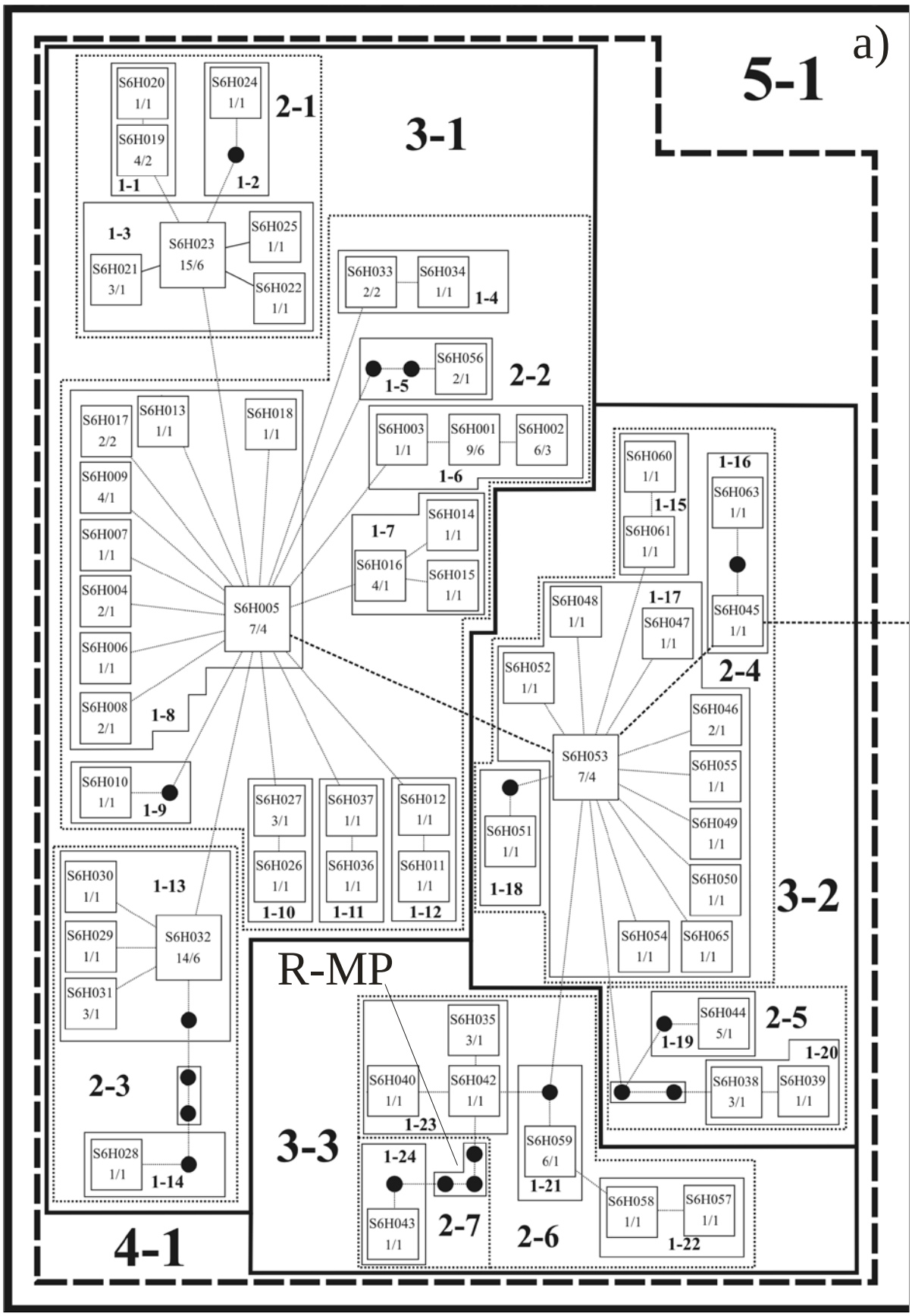
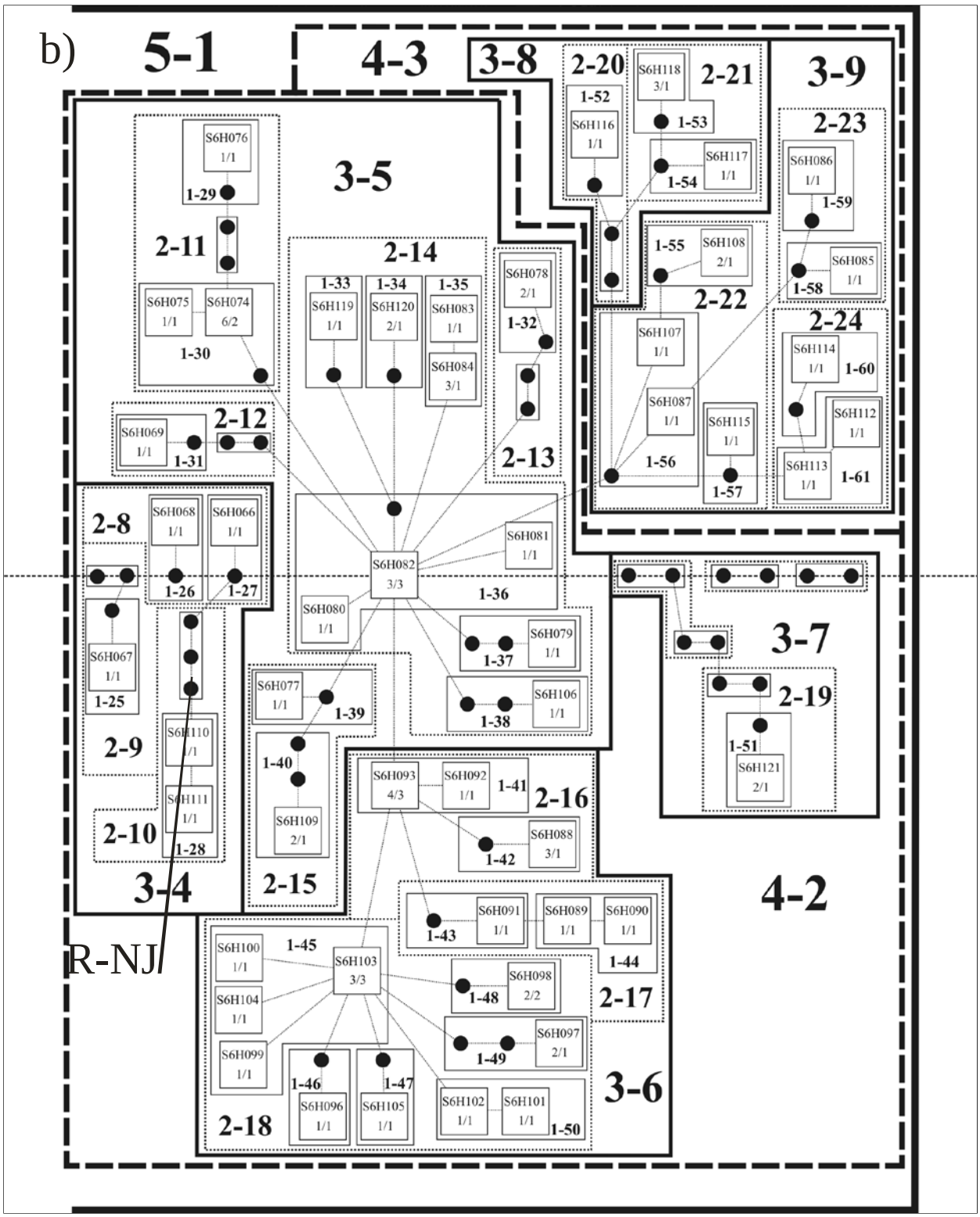


Figure C.16: Nested COI haplotype network of species 6
The entire species 6 cladogram formed a 6-step clade spread across three pages: a) clade 4-1, b) clade 4-2, c) clades 5-2 and 5-3. The lines running off the page connect to the next portion of the network on the following page. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). R-NJ and R-MP denote the roots inferred by neighbor-joining and parsimony, respectively.


a)
5-1
3-1
2-1
S6H020 1/1
S6H019 4/2
1-1
S6H024 1/1
1-2
1-3
S6H021 3/1
S6H023 15/6
S6H025 1/1
S6H022 1/1
S6H033 2/2
S6H034 1/1
1-4
1-5
S6H056 2/1
2-2
S6H017 2/2
S6H013 1/1
S6H018 1/1
S6H009 4/1
S6H007 1/1
S6H004 2/1
S6H006 1/1
S6H008 2/1
1-8
S6H005 7/4
S6H003 1/1
S6H001 9/6
S6H002 6/3
1-6
1-7
S6H014 1/1
S6H016 4/1
S6H015 1/1
S6H010 1/1
1-9
S6H027 3/1
S6H026 1/1
1-10
S6H037 1/1
S6H036 1/1
1-11
S6H012 1/1
S6H011 1/1
1-12
S6H030 1/1
S6H029 1/1
S6H031 3/1
1-13
S6H032 14/6
2-3
S6H028 1/1
1-14
4-1
R-MP
S6H060 1/1
1-15
S6H061 1/1
1-16
S6H063 1/1
S6H045 1/1
2-4
S6H048 1/1
1-17
S6H047 1/1
S6H052 1/1
S6H053 7/4
S6H051 1/1
1-18
S6H046 2/1
S6H055 1/1
S6H049 1/1
S6H050 1/1
S6H054 1/1
S6H065 1/1
3-2
S6H035 3/1
S6H040 1/1
S6H042 1/1
1-23
3-3
1-24
S6H043 1/1
2-7
S6H059 6/1
1-21
2-6
S6H058 1/1
S6H057 1/1
1-22
S6H044 5/1
1-19
2-5
1-20
S6H038 3/1
S6H039 1/1


b)
5-1
4-3
3-8
3-9
3-5
2-20
2-21
2-23
2-11
2-14
2-22
2-24
2-12
2-13
2-8
2-9
2-10
3-4
2-15
2-16
3-7
2-19
4-2
2-17
3-6
2-18
R-NJ
S6H076 1/1
1-29
S6H075 1/1
S6H074 6/2
1-30
S6H069 1/1
1-31
1-33
S6H119 1/1
1-34
S6H120 2/1
1-35
S6H083 1/1
S6H084 3/1
S6H078 2/1
1-32
S6H116 1/1
1-52
S6H118 3/1
1-53
S6H117 1/1
1-54
S6H086 1/1
1-59
S6H085 1/1
1-58
1-55
S6H108 2/1
S6H107 1/1
S6H087 1/1
1-56
S6H115 1/1
1-57
S6H114 1/1
1-60
S6H112 1/1
S6H113 1/1
1-61
S6H068 1/1
S6H066 1/1
1-26
1-27
S6H067 1/1
1-25
S6H110 1/1
S6H111 1/1
1-28
S6H082 3/3
S6H081 1/1
1-36
S6H080 1/1
S6H079 1/1
1-37
S6H106 1/1
1-38
S6H077 1/1
1-39
1-40
S6H109 2/1
S6H093 4/3
S6H092 1/1
1-41
S6H088 3/1
1-42
1-43
S6H091 1/1
S6H089 1/1
S6H090 1/1
1-44
1-45
S6H100 1/1
S6H104 1/1
S6H099 1/1
S6H103 3/3
1-48
S6H098 2/2
1-49
S6H097 2/1
1-46
S6H096 1/1
1-47
S6H105 1/1
S6H102 1/1
S6H101 1/1
1-50
S6H121 2/1
1-51

c)

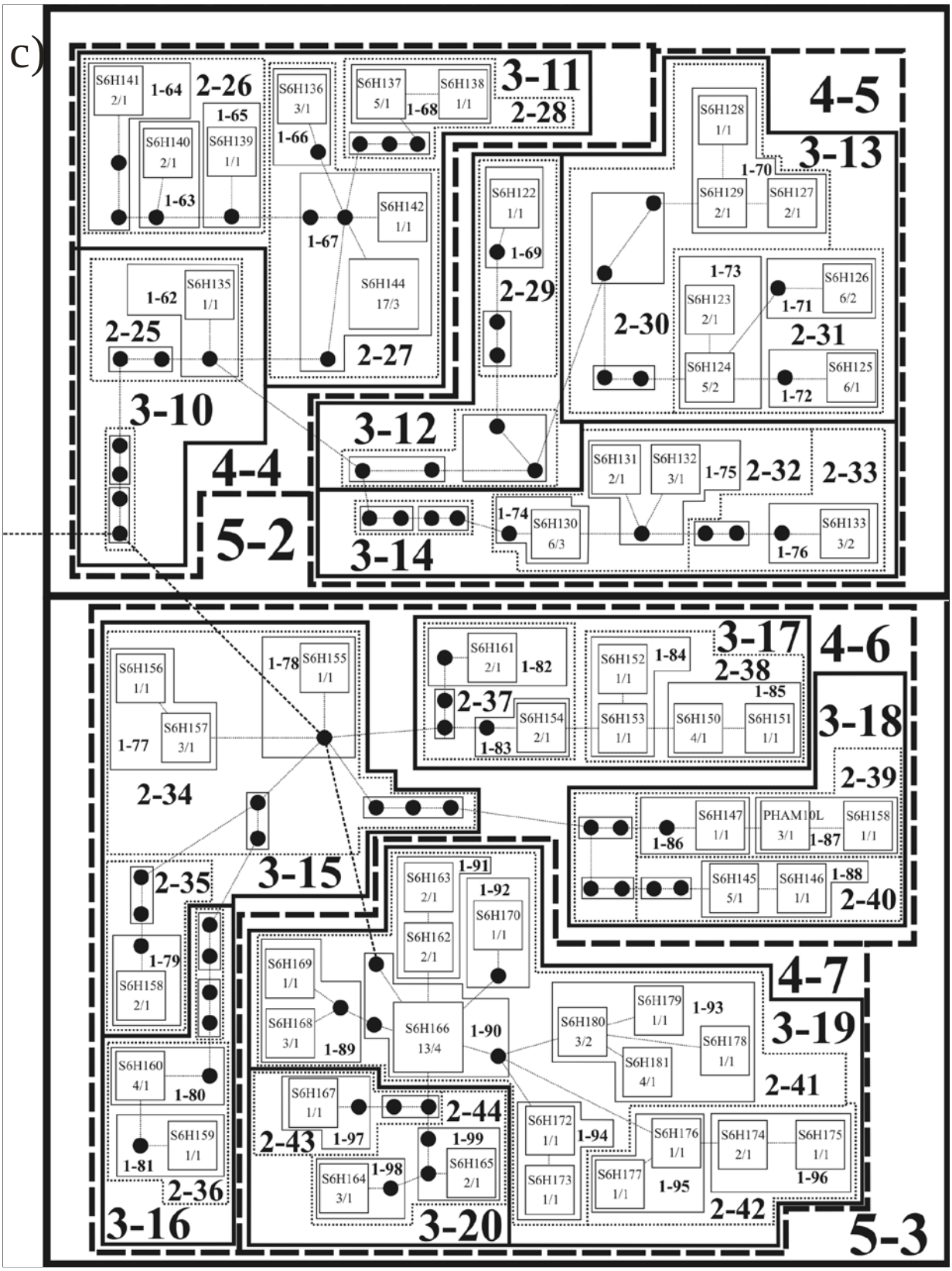

S6H141 2/1
1-64
2-26
1-65
S6H140 2/1
S6H139 1/1
1-63
S6H136 3/1
1-66
S6H137 5/1
1-68
S6H138 1/1
3-11
2-28
S6H142 1/1
1-67
S6H144 17/3
2-27
S6H135 1/1
1-62
2-25
3-10
4-4
5-2
S6H122 1/1
1-69
2-29
3-12
3-14
S6H131 2/1
S6H132 3/1
1-75
2-32
2-33
1-74
S6H130 6/3
S6H133 3/2
1-76
S6H128 1/1
4-5
3-13
1-70
S6H129 2/1
S6H127 2/1
2-30
1-73
S6H123 2/1
S6H126 6/2
1-71
2-31
S6H124 5/2
S6H125 6/1
1-72
S6H156 1/1
1-78
S6H155 1/1
S6H157 3/1
1-77
2-34
S6H161 2/1
1-82
2-37
S6H154 2/1
1-83
3-17
4-6
S6H152 1/1
1-84
2-38
1-85
S6H153 1/1
S6H150 4/1
S6H151 1/1
3-18
2-39
S6H147 1/1
PHAM10L 3/1
1-87
S6H158 1/1
1-86
S6H145 5/1
S6H146 1/1
1-88
2-40
3-15
2-35
1-79
S6H158 2/1
S6H163 2/1
1-91
1-92
S6H170 1/1
S6H162 2/1
S6H169 1/1
S6H168 3/1
1-89
S6H166 13/4
1-90
4-7
3-19
S6H180 3/2
S6H179 1/1
1-93
S6H178 1/1
S6H181 4/1
2-41
S6H160 4/1
1-80
S6H159 1/1
1-81
2-36
3-16
S6H167 1/1
2-44
2-43
1-97
1-99
S6H165 2/1
S6H164 3/1
1-98
3-20
S6H172 1/1
1-94
S6H173 1/1
S6H176 1/1
S6H174 2/1
S6H175 1/1
1-96
S6H177 1/1
1-95
2-42
5-3

Figure: C.17: Nested COI haplotype network of species 9

The species 9 cladogram formed a single 5-step clade. The numbers associated with each non-inferred node gives species and haplotype (upper), and sample size and number of sites (lower). The root is denoted by (R).

