Spring 5-9-2010

Assessing the Phylogenetic Utility of DNA Barcoding Using the New Zealand Cicada Genus *Kikihia*

Megan Ribak
University of Connecticut - Storrs, megan.ribak@gmail.com

APPROVAL PAGE

HONORS THESIS

ASSESSING THE PHYLOGENETIC UTILITY OF DNA BARCODING USING THE NEW ZEALAND CICADA GENUS *KIKIHIA*

Presented by

Megan Pierce Ribak

Honors Thesis Advisor_____
Chris Simon

Honors Academic Advisor_____
Janine Caira

Associate Honors Advisor_____
Elizabeth Jockusch

Department of Ecology & Evolutionary Biology
The University of Connecticut
2010

Abstract

DNA Barcoding (Hebert et al. 2003) has the potential to revolutionize the process of identifying and cataloguing biodiversity; however, significant controversy surrounds some of the proposed applications. In the seven years since DNA barcoding was introduced, the Web of Science records more than 600 studies that have weighed the pros and cons of this procedure. Unfortunately, the scientific community has been unable to come to any consensus on what threshold to use to differentiate species or even whether the barcoding region provides enough information to serve as an accurate species identification tool. The purpose of my thesis is to analyze mitochondrial DNA (mtDNA) barcoding's potential to identify known species and provide a well-resolved phylogeny for the New Zealand cicada genus *Kikihia*. In order to do this, I created a phylogenetic tree for species in the genus *Kikihia* based solely on the barcoding region and compared it to a phylogeny previously created by Marshall et al. (2008) that benefits from information from other mtDNA and nuclear genes as well as species-specific song data. I determined how well the barcoding region delimits species that have been recognized based on morphology and song. In addition, I looked at the effect of sampling on the success of barcoding studies. I analyzed subsets of a larger, more densely sampled dataset for the *Kikihia* Muta Group to determine which aspects of my sampling strategy led to the most accurate identifications. Since DNA barcoding would by definition have problems in diagnosing hybrid individuals, I studied two species (*K.* "murihikua" and *K. angusta*) that are known to hybridize. Individuals that were not obvious hybrids (determined by morphology) were selected for the case study. Phylogenetic analysis of the barcoding region revealed insights into the reasons these two species could not be successfully differentiated using barcoding alone.

Acknowledgments

I would like to thank my family and friends for supporting me during the past four years. I also want to thank Dr. Chris Simon for welcoming me into her lab freshman year and for mentoring me in my scientific pursuits since then. Thank you to all of the members of the Simon lab for their camaraderie and aid in completing this project: Chris Owen, Beth Wade, Dave Marshall, Kathy Hill, Geert Goemans, Rakhee Thombre, Emily Ellis, and Pat Gero. I appreciate Dr. Jockusch's willingness to read my thesis. Dr. Caira has been a wonderful advisor for the past three years and I appreciate all of her guidance.

Computational resources were provided by the Bioinformatics Facility of the Biotechnology/Bioservices Center, University of Connecticut, Storrs, Connecticut, and by the CIPRES portal at the San Diego Supercomputer Center. This project was funded by NSF DEB–04-22386 (Phylogeography of NZ Cicadas), NSF DEB-0619012 (REU for Phylogeography of NZ Cicadas), NSF DEB 07 20664 (Systematics and Biogeography of Australian Cicadettini and their relatives worldwide), and NSF DEB 05-29679 (PEET: Partnerships for training new experts in Auchenorrhyncha taxonomy).

Table of Contents

List of Figures

List of Tables

Introduction

1 The Barcoding Project

The Barcoding Project (Hebert et al. 2003) has the potential to revolutionize the process of species identifications and lighten the workload for the diminishing population of taxonomists. However, the scientific community has not been able to come to a consensus on whether the barcoding region provides enough information to be an accurate species identification tool and what threshold to use to differentiate between species. Furthermore, significant controversy surrounds some of the proposed applications of barcoding.

2 The Barcoding Region

DNA barcoding involves sequencing a 650 base pair fragment of the mitochondrial gene COI (cytochrome c oxidase I). There is some controversy over the reason for and applications of barcoding. Various authors have proposed different purposes for DNA barcoding, but the most prevalent concept of barcoding is the creation of a library of sequences that can be used to identify previously described taxa (Meusnier et al. 2008; Rubinoff 2006b).

Because barcoding relies on mitochondrial DNA (mtDNA) it has inherent advantages and disadvantages. The mitochondrial genome is known for having relatively well conserved regions that are excellent for primer creation. Mitochondrial DNA, unlike nuclear DNA, has no introns, rarely experiences recombination, and is maternally inherited in a haploid manner (Hebert et al. 2003; Rubinoff 2006a; Ballard & Rand 2005). Phylogenetics often utilizes mitochondrial DNA because it is useful when studying species-level relationships and recently diverged taxa (Hebert et al. 2003; Rubinoff 2006a). However, the quick rate of

evolution that makes mitochondrial DNA so useful for recent divergences becomes problematic when divergences dating to the Mesozoic or earlier are examined (Mitchell 2008). This rapid rate of evolution of mitochondrial DNA can lead to homoplasy since frequent base pair changes might result in convergently similar sequences in two unrelated taxa (Rubinoff 2006a). Another potential problem is that mitochondrial DNA does not always display such a simple inheritance pattern because heteroplasmy, hybridization, paternal leakage, and incomplete lineage sorting often complicate matters (Fontaine et al. 2007; Rubinoff 2006a; Ballard & Whitlock 2004). Also, since mitochondrial DNA is maternally inherited, it sometimes predicts completely different phylogenetic relationships from nuclear DNA (Rubinoff 2006a; Ballard & Rand 2005).

The barcoding region is a gene segment within a protein-coding region of the mitochondrial genome. Protein-coding regions of DNA have specific constraints that can be useful or detrimental to this application. First of all, since a change in nucleotides will often have an effect on the amino acids and hence the protein that is produced, sequences can only experience limited changes. Fortunately, the third positions of codons are not under strong selection to remain constant because of the redundancy of the amino acid coding system. Therefore, one-third of the nucleotide sites have a higher potential to change once species diverge. Another advantage to using protein-coding regions instead of genes encoding RNA is the relative rarity of indels (Hebert et al. 2003). In protein coding genes indels are partially constrained by the necessity of avoiding frame shifts.

There are advantages of using the gene COI for barcoding. The mitochondrial genome is often associated with well-conserved primers and the COI gene is particularly rich in highly conserved primers. Hebert et al. (2003) report that the primers have been

functional with "representatives of most, if not all, animal phyla." This gene has the added advantage of both being rapid enough (at silent sites) to differentiate between phylogeographic groups within a species and slow enough (at amino acid replacement sites) to determine deeper phylogenetic relationships (Hebert et al. 2003).

The COI region chosen for barcoding does have some disadvantages. Some recently diverged species could lack fixed differences in the barcoding sequence which would result in uninformative data and be problematic for any applications of barcoding (Mitchell 2008; Rock et al. 2008).

3 Applications of Barcoding

Utilizing barcodes for routine species identifications is the most widely accepted of the potential applications. Suggestions have also been made to use DNA barcodes for species descriptions, phylogenetic analysis and conservation efforts although these applications are highly controversial (Rubinoff 2006a, 2006b; Mitchell 2008).

3.1 Assigning specimens to known species

The original application of DNA barcoding was species identification. Hebert et al.'s (2003) plans for barcoding include making a database of all COI barcoding sequences so that future specimens may be identified. This database would become part of a global bioidentification system (GBS) designed to help solve many of the problems associated with morphological taxonomy and help reduce misidentifications (see the final section of this chapter that evaluates the success of barcoding efforts to date) (Hebert et al. 2003). Barcoding could also be used as a quality control system to ensure that study specimens are identified correctly (Mitchell 2008).

One of the biggest problems with using the barcoding region to assign organisms to species is that it only includes a small amount of the available information. There has always been heated debate over what characters need to be considered when defining a species, even when only morphological information was available. Will and Rubinoff (2004) suggest that species boundaries based on morphology are preferable to those based on the barcoding region because morphological traits reflect information coded by multiple genes. DNA barcoding may take too little information into account. Will & Rubinoff (2004) point out that the barcodes can successfully identify a specimen only when its barcode sequence is an exact match of an identified specimen that is already in the database. If the sequence is not identical to one already included in the profile then a researcher will have no sure way of identifying the specimen using barcoding. Genetic diversity within a species is a problem when only a single representative of each species is included in the profile (Rubinoff 2006a). A researcher would not be able to confirm the identity of a specimen without the aid of a taxonomist, which would nullify the purpose of the barcoding.

One problem with barcoding is that differentiating between species is only possible if arbitrary rules are employed. Researchers must assume that intraspecific variation is significantly less than interspecific variation within the barcoding sequence region (Meyer and Paulay 2005; Langhoff et at. 2009; Ward 2009; Lukhtanov et al. 2009). One proposed method to differentiate between species is to set a standard threshold for the percentage of variation that is tolerated for specimen identification. Setting such a boundary is not a solution because the proposed divergence between species (3% for invertebrates and 2% for mammals and birds) is not successful in delineating between all species (Rubinoff 2006a). Although these boundaries are in the right ballpark for most species, the lack of success in

other species is expected because species vary in age. Another proposed method for drawing a distinct line between species is the 10x concept. Advocates of this system suggest that if a pair of sequences diverges by at least 10 times the average divergence of a species group then the specimens will be of different species. This is one suggested way of defining the barcoding gap. This method of differentiating between species will not function with less divergent cryptic species (Ward 2009). Langhoff et al. (2009), Ward (2009), Lukhtanov et al. (2009), and Meyer and Paulay (2005) examined variation between and within species and did not always find a distinct gap between species. Meyer and Paulay (2005) and Ward (2009) found that neither the 3% set divergence nor the 10x concept performed without making mistakes in delineating species. The same is true of variation within and among well-sampled New Zealand cicada species in the genus *Kikihia* (Marshall et al. 2008; Marshall et al. in prep). There is no clear consensus on a rule to determine whether sequences belong to conspecifics.

3.2 Discovering new species

Some controversy has surrounded proposals to use DNA barcodes to discover new species. Rubinoff (2006b) discouraged "the sole use of mtDNA to identify (discover) new species and understand global biodiversity." Mitchell (2008) agreed and promoted the use of barcodes to discover new species as long as subsequent morphological and molecular analyses are performed. He cites a study where a species of moth was originally discovered by barcoding and subsequently verified. The use of DNA barcodes to highlight areas of traditional taxonomy that should be reevaluated has been gaining popularity in recent years (Kerr et al. 2009; Packer et al. 2009). Mitchell (2008) also suggests that barcoding sequences

could be used as placeholders when new species are discovered but not yet formally described. This application would facilitate consistent identification of undescribed taxa.

3.3 Biodiversity and conservation

Another of the proposed uses for DNA barcoding is to quickly catalogue all of the biodiversity on the planet before it disappears and to help determine where to focus conservation efforts (Rubinoff 2006a, 2006b; Langhoff et al. 2009). Rubinoff (2006a, 2006b) is strongly opposed to using barcoding as the primary resource for biodiversity analysis because barcodes do not provide enough information to make decisions about the potential endangerment of the species. He also feels that the public will be less likely to support conservation efforts for species determined by DNA barcodes. Finally, species in need of conservation could lack genetic differentiation at the barcode site (Rubinoff 2006a). Some of Rubinoff's expectations for a conservation-based application are set up for failure because they require more information than DNA barcoding can offer.

3.4 Identifying pests and other financial uses

DNA barcoding can be useful in many situations that would be financially beneficial to the public. One proposed application for DNA barcoding is to be able to reliably identify snake venom so that researchers attempting to design antivenoms can be positive that they have venom from the correct snake. Apparently, snake taxonomy is rapidly changing and samples of venom sent to laboratories are frequently composed of samples from more than one species or from improperly identified species of snakes. When designing a product that has the potential to save lives, it is important for the scientists to have the correct corresponding venom (Pook & McEwing 2005).

It is often difficult to identify the larval stage of insects that are responsible for destroying crops. Barcodes could be used to determine which pest is plaguing a farmer since the organisms' DNA remains constant throughout its lifetime and the adult stages are usually more easily identified. Once the type of pest is quickly identified, the farmer could proceed with treatment more rapidly and lose fewer crops (Mitchell 2008). Phenotypic differences in the lifestages of an organism are also a problem within fisheries. These important food production operations could benefit greatly from DNA barcoding (Rock et al. 2008). The rapid identifications provided by DNA barcoding could also be beneficial for managing invasive species. This technology would be especially useful at commercial ports and national borders, where a speedy identification of taxa could result in swift action that could prevent the spread of the invasive species (Mitchell 2008). These applications of barcoding were available before the mitochondrial barcoding region was selected so the question is whether the COI segment chosen is the best gene segment for the job.

3.5 Phylogenetic analysis

Many researchers have come to the conclusion that phylogenies based solely on the barcoding region are inferior to phylogenies that include more sources of data. This is true of any study based on a small amount of sequence data from a single gene. However, the process of species identification through barcoding requires the assignment of taxa to clusters on a tree based on neighbor-joining phylogenetic analysis (Hebert et al. 2003; Meier et al. 2006; Pagès et al. 2009). Many of the problems with barcoding phylogenies have already been discussed (in the section on the barcoding region), but there is some criticism that focuses specifically on the tree diagrams in the Hebert et al. (2003) study.

Will and Rubinoff (2004) criticize Hebert et al.'s (2003) tree diagrams because the diagrams do not agree with any existing hypothesized phylogenies. They acknowledge that Hebert et al. avoid using the term "phylogeny" in favor of "profile." However, Hebert et al. still used phylogenetic terminology in describing taxa as monophyletic and forming a "cohesive group." Another problem with Hebert et al.'s (2003) study is their use of phenetics with some phylogenetic methods, meaning their results are based solely on similarities. This leads to confusion since the trees appear similar to those that commonly reflect evolutionary relationships (Will & Rubinoff 2004). There are many considerations that need to be built into successful phylogenetic tree construction (reviewed in Simon et al. 2006) and Hebert et al. ignores most of them. Barcoding is often misapplied because people use it to determine evolutionary relationships. Rubinoff states that "barcoding is not meant to and does not provide evolutionary information about taxa; rather, it is intended only as a means of 'yes' or 'no' identification based on predetermined units" (Rubinoff 2006a). He makes an effort to stress that barcoding does not accurately reflect evolutionary relationships and that needs to be kept in mind if the system is going to be used (Rubinoff 2006a).

In spite of all the problems, Hebert et al. (2003) maintain that their barcoding study was mostly successful at identifying species. According to the authors their study's few inaccurate identifications would be remedied by using a larger database. Mitchell (2008) also provides plausible explanations for studies with unusually high failure rates of species identification. Researchers' tendency to study more difficult taxonomic problems is responsible for the low success rate in these studies. Some of these studies used sequences from Genbank which has many incorrect species identifications that would result in apparent misidentifications in the barcoding studies.

4 Barcoding, Taxonomy, and Integration

Taxonomy, the science of assigning names to species and higher taxa, is crucial to other fields of science. Unfortunately, the number of taxonomists is in decline and there are already insufficient numbers of specialists in this field to handle the existing workload (Rubinoff 2006a; Hebert et al. 2003; Packer et al. 2009). Hebert et al. (2003) estimate that "since few taxonomists can critically identify more than 0.01% of the estimated 10-15 million species, a community of 15,000 taxonomists will be required, in perpetuity, to identify life if our reliance on morphological diagnosis is to be sustained." In addition to the large number of taxonomists, a significant time investment is needed – estimates from 940 years to several millennia have been proposed– to describe all existing species if the traditional methods are applied (Meier et al. 2006; Packer et al. 2009). Mitchell (2008) agrees that the shortage of taxonomists is difficult to remedy because taxonomic techniques are time consuming and highly specialized. Clearly the sheer magnitude of the problem puts a limit on the capabilities of taxonomy. Researchers have been searching for an alternative system that would take some of the strain off of taxonomists so that they can focus on other areas of systematics instead of performing species identifications (Will & Rubinoff 2004; Packer et al. 2009).

Barcoding can utilize the expertise of current as well as former taxonomists, since previously identified museum specimens should be used to produce barcodes whenever possible. This is particularly useful because there is an unfortunate decline in the popularity of taxonomy and some groups of organisms no longer have expert taxonomists. The barcode database will help preserve taxonomic information in a novel format and allow laboratories without morphology experts to identify relevant species (Hebert et al. 2003). Taxonomists

will benefit because with fewer samples to identify they will have more time to pursue research on their specialty taxa (Mitchell 2008; Packer et al. 2009).

In addition to the large specialized workforce needed to perform species identifications, morphological taxonomy has other serious challenges. Morphology is not consistent due to phenotypic plasticity and genetic variability within species. In many species, there are also significant differences in morphology between the genders and different life stages. Cryptic species, by definition, are often impossible to identify by morphological characteristics alone. All of these obstacles in species identification often lead to incorrect identification. Another reason for incorrect species identification is that some researchers attempt to use keys without the appropriate level of expertise. Furthermore, some of the taxonomic keys in use are flawed and such keys are rarely revised since to do so is a major undertaking. The current method of species identification is also in need of improvement because it is costly both financially and in terms of time (Hebert et al. 2003; Mitchell 2008; Pagès et al. 2009; Packer et al. 2009).

Some proposed alternatives to the traditional method of morphological taxonomy involve DNA technology. Researchers have attempted to take advantage of the variability present in the genome to identify species (Hebert et al. 2003, Pons et al. 2006, Marshall et al. in prep). There is a precedent of using mitochondrial DNA sequences in addition to morphology to resolve difficult species identifications (Mitchell 2008; Pagès et al. 2009). Microgenomic identification systems are already in place among scientists studying viruses, bacteria, and protists (e.g., Lewis and Lewis 2005; Zettler et al. 2002; Abriouel et al. 2008; Iliff et al. 2008). These groups are nearly impossible to identify using only morphology so a

genetic approach has been accepted for taxon identification (Hebert et al. 2003). Mitochondrial DNA barcodes are very similar to these genetic techniques.

Two aspects of molecular technology, information storage on Genbank or similar databases and the construction of phylogenetic trees, have flaws with potential for improvement. If an author accidentally submits an incorrect sequence to the database, it will not be removed unless the original author updates the sequence. Genomic databases would be much more successful if they had curators to correct such inaccuracies. Phylogenetic analyses require sequence data from multiple unlinked genes (Mitchell 2008). It is both time consuming and expensive to obtain the amount of data needed for a phylogenetic analysis. If accurate relationships could be determined from phylogenetic trees made from smaller numbers of genes, more relationships could be determined to better understand biodiversity.

Many researchers are urging the scientific community to consider an integrated approach (which includes nuclear genes as well as the barcoding region) in order to utilize the benefits of mitochondrial DNA and avoid some of the pitfalls associated with barcoding (Rubinoff 2006a, 2006b). Mitchell (2008) believes that the problems barcoding created by hybridization could be resolved if nuclear genes were also analyzed. Many scientists have been using an integrated approach all along: they rely on morphology as well as genetic data from both nuclear and mitochondrial genes to identify species and determine their relationships to one another.

5 Success Rates of Barcoding

Various studies and analyses of those studies have been performed to determine the success of DNA barcoding for species identification. Meusnier et al. (2008) report barcoding

success levels over 97% in studies involving birds, mammals, fishes, and arthropods. Hebert et al. (2003) created a profile of one hundred species from seven diverse animal phyla and then attempted to identify newly analyzed taxa using this profile. This experiment resulted in a 96% success rate of correctly assigning the taxa to the appropriate phylum (it is important to note that this is not species identification). Furthermore, each species had a different COI sequence for the barcoding region. This process was repeated with a different data set including eight orders of insects and 50 newly analyzed taxa were correctly assigned to each order. Hebert et al. (2003) repeated this experiment once more making a profile for two hundred species that have recently diverged from one another. The 150 newly analyzed individuals were all assigned the correct species identification. Hebert et al. (2003) stated that the majority of the individuals fell into monophyletic groups that reflect their accepted taxonomic relationships. They believe that the barcoding study was exceptionally successful in terms of species identification: "'test' taxa were always either genetically identical to or most closely associated with their conspecific in the profile" (Hebert et al. 2003). Furthermore, they came to the conclusion that COI was capable of handling this application, even when it comes to deeper divergences (Hebert et al. 2003). The majority of barcoding studies have success rates similar to those determined by Hebert et al. (2003). However, there are some studies that have shown a higher rate than 5% failure to determine species assignment by barcoding alone (Mitchell 2008).

As with any other system, DNA barcoding has its share of flaws which are often more informative than the successes. Hebert et al. (2003) admit that some misidentifications did occur during their study. Errors were thought to be due to hybridization, introgression, polyploidization, incomplete lineage sorting, *Wolbachia* infections (in invertebrates) and

"numts" (paralogous copies of mitochondrial genes that are inserted into nuclear DNA) (Hebert et al. 2003; Mitchell 2008; Ward 2009). Other studies cite problems with discovering new species using the distance-based methods that are typical of barcoding (Mitchell 2008). DNA barcoding also encounters problems common to any type of molecular analysis: degradation may make it impossible to amplify a sequence and primers can never be truly universal due to the potential to develop mutations in the primer binding regions (Meusnier 2008; Mitchell 2008). Mitochondrial DNA often provides different relationships between taxa than nuclear DNA. A group of researchers analyzed relevant literature and discovered that there were differences among relationships assigned by nuclear and mitochondrial DNA in more than 20% of the studies they surveyed. Due to the nature of the study, this estimate may be conservative which would mean that the barcoding region could provide information that is inconsistent with relationships based on nuclear genes over 20% of the time (Rubinoff 2006a). Some researchers view this abundance of flaws as a reason for evolutionary biologists to reject the widespread use of DNA barcoding. Others believe that barcoding is the best system we have at present and that almost all systems have some problems associated with them. It seems to be a matter of opinion whether the obstacles that potentially prevent correct identifications are more significant than the benefits to be gained from this novel application of mitochondrial DNA sequencing.

6 Case Studies

There are countless case studies that attempt to determine the success of barcoding. The studies discussed below are only a few of those available, yet these studies each provide different insight into the successes and pitfalls of the controversial practice of barcoding.

Rock et al. (2008) attempted to test barcoding using fish from the Scotia Sea of Antarctica. A total of 124 sequences from 34 putative species of Antarctic fish (assuming morphological identifications are correct) were examined in this study. The study showed that barcodes were successful in identifying most species. For the majority of the taxa, the neighbor joining trees created in this study accurately reflected the accepted relationships. There were some cases where barcoding did not work; however, the authors did not provide a specific success rate. Some of the reasons provided for the few failures of barcodes were hybridization and insignificant differences in COI sequences between recently diverged taxa. The authors found barcoding to be a very useful tool because these Antarctic fish are often difficult to identify using morphology alone. The DNA barcodes often confirmed that uncertain morphological identifications were correct. Occasionally the barcodes suggested that the original morphological identifications were inaccurate and the authors believe that the barcodes provide the correct identification (especially because there were notes made about uncertainties on the field data sheets). Another gene (cytochrome b) was also used to discriminate between species in some situations where COI was not informative enough, so this study is an example of integrated barcoding (Rock et al. 2008).

Culicoides (biting midges, Diptera: Ceratopogonidae) are another group of organisms that are nearly impossible to identify to the species level using only morphology (Pagès et al. 2009). This subgenus does not have a firmly established taxonomy and has been reorganized on many occasions. All their specimens were identified to species level using wing morphology (no hybrids were included in this study). The 95 sequences the authors obtained from five species all showed low intra-specific and high inter-specific diversity which is required for successful barcoding. They included other sequences from GenBank to enlarge

their dataset for phylogenetic analysis. Pagès et al. (2009) performed NJ (neighbor joining) and ML (maximum likelihood) analysis using the Jukes-Cantor model. The authors determined that DNA barcoding was capable of identifying all the species from their focal subgenus, the species formed monophyletic groups, and all the species-level relationships determined were supported by high bootstrap values although cryptic species were present. Pagès et al. (2009) report that barcoding was very successful in the subgenus *Culicoides.*

Hebert et al. (2004) tested the effectiveness of barcoding using 260 species of North American birds. In this study, only half of the species were represented by two or more sequences. Paired sequences from the same taxon were always identical to or grouped most closely with their conspecifics. The branching of the tree generally reflected the current taxonomy of the birds at higher levels as well as at the species level. The study was also successful because the sequence differences between species were always greater than those within species. This created deep divergences between species and shallow divergences within species in the neighbor joining tree. There were only four cases where the study was not as successful. The authors believe that the four polytypic species might represent hidden species and not a failure of barcoding to identify species (Hebert et al. 2004). However, some scientists hypothesize that many species are not monophyletic, so it may not be unusual to have deep divergences within a species (Meier et al. 2006). Hebert et al. (2004) admitted that COI barcoding has difficulty with hybridization and very recently diverged species because sequences will be too similar or identical. They believe that DNA barcodes will occasionally recognize correct classifications that morphological taxonomy has mistaken for years (as with their example of the Snow Goose). The authors also stress the importance of obtaining barcodes from various individuals of the species to avoid problems caused by

paraphyly or polyphyly (which may be due to imperfect taxonomy, hybridization, and incomplete lineage sorting). Overall, they think that their study was successful and proved that barcoding could sort North American birds into the appropriate species (Hebert et al. 2004).

Since birds have been so heavily studied and have a relatively well established taxonomy there are many barcoding studies that focus on avian species. Kerr et al. (2009) analyzed a data set containing 559 different species of Palearctic and North American birds. The authors made an effort to include as much geographic variation in their data set as possible in addition to including multiple specimens as representatives for each species (487 out of 559 species had multiple individuals). Kerr et al. (2009) employed NJ clusters, a threshold program designed for nematodes (MOTU), and a character-based system (CAOS). The authors propose that recent speciation, hybridization, and introgression are possible causes for the lack of differentiated barcodes in some sister species. They found a 90% success rate for species identification using traditional NJ methodology. Kerr et al. (2009) believe that the first two methods they analyzed were preferable to the character-based method of interpreting DNA barcodes for species identification. The authors believe that DNA barcoding is a useful tool for species identification when there is adequate sampling. However, DNA barcoding cannot reliably identify recently diverged species in birds (Kerr et al. 2009).

Baker et al. (2009) also performed a study of DNA barcoding using birds. They attempted to address some of the concerns of opponents of DNA barcoding. Baker et al. (2009) compared NJ analysis of the barcoding region to phylogenies that utilized multiple genes. The found that DNA barcoding could still differentiate between closely related sister-

species of birds. The authors also studied a group of rockhopper penguins that has been a subject of taxonomic debate about whether it is comprised of two separate species or three subspecies. They propose that splitting these clades into three separate species might be the most appropriate based on the results of DNA barcoding. DNA barcoding will not always agree with the traditional taxonomy. In such unresolved situations it is possible that the DNA barcoding is not working or that the taxonomy does need to be reevaluated. Baker et al. (2009) think that DNA barcoding is incredibly advantageous in situations similar to the rockhopper penguins if taxonomists take the opportunity to investigate such discrepancies. The authors note that there is not a barcoding gap so thresholds cannot be foolproof (especially in recently diverged species). However, they have often been successful in identifying avian species. Baker et al. (2009) hypothesize that DNA barcoding can potentially differentiate between species after 100,000 to 150,000 generations. The authors believe that barcoding does have its share of flaws, but they have been overemphasized since DNA barcoding has completed many successful species identifications in birds. Baker et al. (2009) conclude by advocating for an integrated approach to barcoding in certain applications.

Ward analyzed the success of DNA barcoding in birds and fishes using publicly available sequence data (from BOLD, the Barcode of Life Database) in 2009. He compared the genetic divergence at three taxonomic levels: species, genus, and family. The species level analysis is most applicable to evaluating the potential success of DNA barcoding. It can determine whether it is possible to differentiate between species and accurately identify specimens. Ward found that the barcoding gap did not exist in this data set. There were 42 species of bird (out of 657) and 23 species of fish (out of 1088) that could not be

differentiated by barcoding alone. The arbitrary boundary between species that was proposed for mammals and birds (a genetic divergence of 2%) was supported by this study. The probability of a pair of sequences with over 2% divergence belonging to the same species was very low for both birds and fishes. Ward (2009) concluded that overall his study was a success for barcoding and that the cases where barcoding could not differentiate between species were due to recent speciation, incorrect taxonomy, or hybridization.

Meier et al. (2006) tested DNA barcoding on 1333 sequences from 449 Diptera species. Only 127 of the species used in this study had multiple sequences in the data set. This study had a very low success rate when compared with other barcoding studies. The highest identification rate experienced in this study was 68% (there were multiple success rates in this study because the authors analyzed the data in a variety of ways). There were very few misidentifications but many of the species could not be identified using this methodology alone. The authors suggest that successful barcoding studies are overly optimistic due to poor study design, primarily the fact that other studies often do not sample geographic variation within a species. Meier et al. (2006) also accuse other authors of not including many closely related sister species in their studies. Another proposed problem with barcoding is the use of neighbor joining trees because they fail to supply the researcher with more than one tree when there may be tie trees (trees that have the same fit to the data) and the algorithm used to build the trees is not a very efficient search strategy when the number of taxa is high and the amount of sequence is low. Finally, NJ trees are subject to the same model-fitting/systematic error problems as other algorithms when not applied properly. Other problems Meier et al. (2006) encountered were that species frequently had multiple barcodes and the same sequences were found in different species. The lack of a distinct

barcoding sequence for an individual species was more common in those species represented by multiple sequences in the data set. The 22 cases of species having overlapping barcodes are a subject of concern because an exact sequence match cannot guarantee that the organism is assigned to the correct species. The authors hypothesize that many more species would be accurately identified if there was a database that contained multiple sequences for every species on the planet. Although the Barcoding Life Consortium is attempting to establish such a database, it is impossible so barcoding will never have complete success. In spite of all of these problems, Meier et al. (2006) believe that barcoding should not be immediately dismissed because it can be useful in some taxa and in certain situations. Barcoding can be used to assign organisms to species groups (which is all the resolution needed in certain applications), determine that morphologically different genders or lifestages belong to the same species, confirm that products are made from endangered species, and discover cryptic species (Meier et al. 2006).

Langhoff et al. focused their 2009 study on two genera (*Ctenopseustis* and *Planotortrix*) of New Zealand leafroller moths. The phylogeny of these genera has been studied before using genetic data, morphology, and pheromone blends (which play an important role in lepidopteran mating). The selection of these genera for an evaluation of barcoding is ideal since this taxonomic problem has been previously studied and lepidopteran taxa are known for rapid speciation which often results in cryptic species. Langhoff et al.'s methodology differed slightly from other barcoding studies in that they only sequenced a 468 base pair fragment of the COI and they employed other phylogenetic methods than neighbor joining analysis to create their tree. Seven of the twelve species included were not monophyletic (although an additional species had only one representative in the study). The

results of this study showed that there was not a barcoding gap, which made it impossible to distinguish between multiple species in this manner. The authors explain that incomplete lineage sorting, introgression, horizontal gene transfer, error in specimen identification, or incorrect taxonomy could be responsible for the lack of success of barcoding in this particular study. They suggest that analyzing the genes involved in pheromone production (females) and reception (males) could be more successful in this group of organisms. However the need to utilize another segment of DNA would be a failure for DNA barcoding (Langhoff et al. 2009).

Lukhtanov et al. (2009) recognize that many barcoding studies include data sets with poor geographic sampling. They propose to improve geographic sampling in their study by including allopatric species. The data set included 353 species, 285 (81%) of which were represented by multiple specimens. Lukhtanov et al. report that 90.1% (318) of the species clustered in the NJ analysis and were identifiable by DNA barcodes. However, despite this success rate, 34 species (9.6%) did not have unique barcode sequences so they would not be identifiable using this method. They determined that there was no significant difference in the percentage of shared barcodes among sympatric (16.4%) and allopatric (18.6%) species pairs according to the current taxonomy. The authors believe that the current taxonomy should be revised and that the allopatric species pairs lacking unique barcodes should be considered single species. If the taxonomy were revised as they suggested, there would be a significant difference in the percentage of undifferentiated barcodes between allopatric and sympatric species pairs. This would mean that sympatric species pairs are far more likely to lack unique barcode sequences than allopatric species pairs. Lukhtanov et al. (2009) also examined how geographic variation would affect the barcoding gap. The intraspecific

variation significantly increased with increased geographic sampling. This increase in intraspecific variation caused a significant decrease in presence of a gap between species as well as an increase in the number of paraphyletic species. This trend does not have an affect on the success of DNA barcoding because the species still cluster together and are separated by deep sequence divergences which are supported by high bootstrap values. The authors still advise that more information than sequence data from the barcoding region should be considered when dealing with paraphyletic taxa. Based on these results, Lukhtanov et al. (2009) conclude that improving the geographic sampling did not impede identification through DNA barcoding.

Meyer and Paulay (2005) were also concerned about the lack of sampling in DNA barcoding studies. They believed that neither interspecific nor intraspecific variation had been adequately represented because many studies included only one or two individuals per species, a limited geographic sample, or did not include some sister-species. Meyer and Paulay (2005) found an 80% identification success rate using DNA barcoding with cowries. The authors suspect that the 20% failure rate (including ambiguous and incorrect identifications) is due to the fact that some species are not monophyletic. Comprehensive sampling cannot compensate for this obstacle of barcoding. The authors also examined intraspecific and interspecific variation: there was no barcoding gap present as variation within and among species overlapped. The authors believe that no simple guideline for species delineation will be free from all errors in identification. Meyer and Paulay (2005) conclude that DNA barcoding is most successful in groups whose taxonomy has been thoroughly studied since such applications will avoid error from inaccuracies in taxonomy.

7 Synthesis and Concluding Remarks

DNA barcoding has been a hot topic in biology since Hebert et al. published their 2003 study on the topic. Systems of identification similar to DNA barcoding have been in use for some groups of organisms that are often difficult to identify solely by morphology (Mitchell 2008; Pagès et al. 2009; Lewis and Lewis 2005; Zettler et al. 2002; Abriouel et al. 2008; Iliff et al. 2008; Hebert et al. 2003). For such groups of organisms, it would be beneficial to have a standard segment (such as the barcoding region) to use for routine identifications. DNA barcodes have potential to be useful for species identifications without the aid of a taxonomist in certain situations. Barcoding has been proposed as a quality control measure to confirm the identity of specimens (Mitchell 2008; Pook & McEwing 2005). The proposed commercial uses of barcoding, such as pest identification, invasive species detection, and fishery management, are also worth pursuing (Mitchell 2008; Rock et al. 2008).

There are many pitfalls to the use of barcoding for species identification so the scientific community must be cautious in accepting it. Some biological phenomena that potentially interfere with barcoding are heteroplasmy, hybridization, paternal leakage, introgression, polyploidization, recent speciation, incomplete lineage sorting, error in specimen identification, incorrect taxonomy, *Wolbachia* infections and "numts" and all of the above phenomena are known to occur to different degrees depending on the dataset (Hebert et al. 2003, 2004; Mitchell 2008; Ward 2009; Fontaine et al. 2007; Rubinoff 2006a; Ballard & Whitlock 2004; Rock et al. 2008; Langhoff et al. 2009). Another potential problem is that the barcoding region may not be informative enough to identify species. In certain studies, it has become apparent that not all species have differences in their barcoding regions (Mitchell

2008; Rock et al. 2008; Hebert et al. 2004; Meier et al. 2006). Problems also exist in the implementation of the barcoding method. Neighbor joining analysis is used in the process of species identification. Although NJ is a phylogenetic method, it is not the optimal method in terms of search strategy and use of information in the data (Swofford et al. 1996). In addition, most researchers would agree that the barcoding region does not contain enough information for well-supported phylogenetic analysis (Hebert et al. 2003; Meier et al. 2006; Pagès et al. 2009). Finally, the largest problem with barcoding is determining a set of regulations for the boundaries used during species identification. Researchers have examined intraspecific and interspecific variation in an effort to determine the effectiveness of species assignments by barcoding (Meyer and Paulay 2005; Langhoff et at. 2009; Ward 2009; Lukhtanov et al. 2009; Rubinoff 2006a).

DNA barcoding has its share of flaws and the majority of its supporters recognize these flaws. However, traditional taxonomy based on morphology also has its drawbacks (Packer et al. 2009; Hebert et al. 2003; Mitchell 2008; Meier et al. 2006; Pagès et al. 2009; Rubinoff 2006a). It is naïve to hope for a system of identification that can identify all species without making any errors. DNA barcoding has much to offer the scientific community as long as researchers keep its limitations in mind.

Materials and Methods

1 Relative Advantages of Nuclear and Mitochondrial Genes

Mitochondrial DNA (mtDNA) is favored for many molecular studies because it is maternally inherited and small in size. Also, the sequences of mtDNA diverge relatively

quickly and the gene order and composition is relatively uniform (Hoy 1994, Simon et al. 1994). Although mtDNA is sometimes favored because it is maternally inherited, this feature may be problematic if the maternal gene tree is incorrect due to lineage sorting. However, mtDNA is less likely to be affected by lineage sorting than nuclear DNA because it is inherited from only one parent (Simon et al. 2006).

One important advantage of using mitochondrial genes for study is that there are more copies of them so they are easier to amplify. Mitochondrial genes are useful for studying species that diverged recently because they have a high rate of substitution. However, if the divergence event is not recent, nuclear genes work better for phylogenetic analysis (Lin & Danforth 2003). The quick evolution of mtDNA provides a problem in that it may result in many multiple substitutions, which can be misleading for phylogenetic analysis (Simon et al. 2006). Nuclear genes have both exons and introns, which evolve at different rates. This can be advantageous for some phylogenetic analyses. Also, nuclear genes generally work better in different phylogenetic applications. Often, both nuclear and mitochondrial genes are included in a study to reap the benefits that the different genes offer (Lin & Danforth 2003) and to provide independent gene tree estimates of the species tree (Simon et al. 2006).

2 DNA Extraction

The process of DNA extraction requires separation of nuclear and/or organelle DNA from protein, carbohydrate and lipid materials contained in the cell. In the past, various grinding, precipitation and washing methods were used many of which contained toxic materials and time-consuming steps. The last ten years have seen the development of many

commercial kits for faster safer PCR extraction. The following protocol was developed from the NucleoSpin® Tissue Kits User Manual published by Clontech. Tissue equivalent to two or three *Drosophilia* flies is combined with 180µL lysis buffer T1 and 25µL proteinase K and incubated for 12 to 18 hours at 56ºC to free the DNA from any membranes within the cell. Proteinase K is an enzyme that digests proteins. Next, 200µL of buffer B3 is added and the tubes are inverted five to ten times. The lysed samples are incubated at 70ºC for ten minutes. This step has been removed from the protocol of more recent extraction kits. They are briefly centrifuged and then 200µL chilled 100% ethanol is added to the tubes to precipitate the DNA. The solution is transferred into the NucleoSpin columns and centrifuged for one minute at 11,000 x g to bind the DNA to the column. The DNA is then washed with 500µL wash buffer BW and 500µL wash buffer B5. To get the DNA off of the membrane, 100µL of warm elution buffer (warmed in 70ºC water bath) is added to the NucleoSpin column and allowed to rest for one minute before centrifuging the liquid through. This process is repeated for a second elution. DNA can be quantified using a spectrophotometer.

3 PCR

PCR, or the polymerase chain reaction, serves to amplify a specific region of DNA. One of the greatest benefits of PCR is that it may be performed using only small amounts of tissue so the procedure will often work regardless of sample size. PCR has many useful applications since it can be performed with a small initial sample. In PCR cycles, the short products defined by the two primers (the ones that are desired) replicate exponentially and the long products defined by one primer and copied off the original molecule (which are not

desired) increase linearly (Simon et al. 1991). Thus the long products are a relative minor component of the final reaction.

Primer design is an important part of any PCR procedure. Primers are usually 18-25 bases in length. They are complementary to the ends of the region of DNA to be copied and mark the location where strand synthesis begins. The nucleotide sequence of the desired piece of DNA is not always known since the mitochondrial genome is not known for many species. If a close relative of the desired taxon has been sequenced, it is possible to choose primers based on the sequence of the relative. Otherwise, it is necessary to use a highly conserved region of DNA for a primer. Many papers are published with suggestions of primers that can be used on a wide variety of taxa (Kocher et al. 1989, Simon et al. 1994, Palumbi 1996, Simon et al. 2006). When amplifying protein coding genes, degenerate primers provide another option when the sequence of the target taxon is unknown. Degenerate primers are composed of different primers that match a known conserved amino acid sequence (Simon et al. 1991). The codes for amino acids are highly degenerate, meaning multiple codons will code for the exact same amino acid.

Nested PCR is a technique in which two sets of primers are used to amplify a given region of DNA. First one PCR is run with one set of primers and then primers that sit within the region amplified by the first PCR are used for the second reaction. Sometimes primers will unexpectedly bind to more than one part of the DNA molecule. A nested PCR can be used to make sure that only the desired region of DNA is amplified. This method of amplifying DNA is often used when the genome of an organism has not been sequenced.

The PCR reaction begins with an initial denaturation step (at 92-96ºC) which improves the effectiveness of the reaction. At the beginning of the cycle, the template DNA

strands dissociate due to the heat during denaturation period. The temperature is then lowered so the primers can anneal at around 50ºC. In the last step of the cycle, the temperature is raised to 72-74ºC so that the enzyme Taq can add nucleotides at its optimum temperature (Simon et al. 1991). Taq is a DNA polymerase from the bacterium *Thermus aquaticus*. Since Taq is a heat-resistant DNA polymerase, it is incredibly well-suited for use in PCR (Hoy 1994). There are two important factors to keep in mind when determining temperatures for the PCR cycle. During the annealing phase of the cycle, mispriming can occur if the temperature is too low. This is a result of the fact that lower temperatures result in less specific annealing. Another thing to keep in mind is that Taq is an enzyme and works best at a certain temperature (Simon et al. 1991). The optimum temperature range for Taq is specific and depends on the manufacturer's recommendations.

4 A PCR Cycle

The Ex *Taq* PCR Kit from BD Biosciences was used for PCR reactions in my experimental system. The following protocol was adapted from the directions included in the kit's user manual. All reagents are kept on ice (except for the Taq which is taken out of the freezer just before it is added). A master mix is used in this procedure because it allows the components to be mixed and divided uniformly for each reaction. The master mix includes 1.25µL of 10X Ex Taq Buffer, 1.25µL of dNTP's, 0.625µL of forward 10µM primer, 0.625µL of reverse 10µM primer, 0.5µL of $MgCl_2$, 7.2µL of sterile ddH_2O, and 0.05µL of Ex Taq for each reaction. When making the master mix, an extra reaction is included in the calculations to allow for a negative control. The Taq is added at the very end before the master mix is divided among the different tubes so that the enzyme does not have time to

heat up and start working before it is put in the PCR machine. After the master mix is thoroughly mixed, 23µL of master mix is added to each labeled strip tube along with 2µL of genomic DNA. The profile begins with an initial denaturation of 94ºC for 2 minutes, followed by 40 cycles of a 45 second denaturation at 94ºC, a 45 second annealing period at 45ºC, and a 72ºC extension for 1 minute and 30 seconds. The final extension is at a temperature of 72ºC for 10 minutes. The primers used for the barcoding region are COI Barcode Forward (5'GGTCAACAAATCATAAAGATATTGG 3') and COI Barcode Reverse (5'TAAACTTCAGGGTGACCAAAAAATCA3'). These primers amplify a segment of the COI gene approximately 650 bp in length.

5 PCR Optimization

It is important to maintain consistency in the concentrations of the components of the PCR reaction to obtain good results. For example, primer concentration is crucial to the success of PCR because an excessive amount of primers in the reaction solution will result in non-specific priming (with multiple products). Another problem caused by a high primer concentration is that the primers may bind to themselves, resulting in primer-dimers. High concentrations of dNTPs can also result in mispriming. Polymerase chain reaction is most successful when the four dNTPs are present in equal concentrations. Taq concentration can also significantly affect the results of a PCR. Non-specific products are often a result of too much Taq in the reaction solution. Taq is expensive so it is also economically beneficial to use lower concentrations of the enzyme (Simon et al. 1991). Commercially available PCR kits are carefully optimized to work consistently on a large number of different taxa.

6 Gel Electrophoresis

Gel electrophoresis is a procedure which pushes DNA through a gel matrix using an electrical current. DNA is negatively charged and will migrate to the positive end of the gel when an electric current is applied. Shorter bands can move through the gel more quickly so electrophoresis separates DNA into bands based on size. Either agarose or polyacrylamide gels may be used for electrophoresis. The kind of gel used depends on the length (agarose gels are used for longer DNA fragments and polyacrylamide are used for shorter fragments). Ultra-violet light is used to visualize the bands of stained DNA after electrophoresis.

The gels used in this research for visualizing PCR products were one percent agarose gels. Vernier SYBR Safe™ is used to stain the DNA in place of the traditional ethidium bromide (a known teratogen). One bright band for each PCR is ideal when running out reactions on a gel. The single band [of the correct size] signifies success because all of the DNA fragments present are of the same length for a given set of primers.

7 PCR Cleanup

The purpose of PCR cleanup is to remove salts, extra nucleotides and primers before sequencing (Simon et al. 1991). ExoSAP-IT® (USB) was used to clean the PCRs. One part of ExoSAP-IT® was combined with five parts water and vortexed. In a sterile strip tube 1µL of the 1:5 ExoSAP-IT® mixture and 2.5µL PCR were added. The strip tubes were placed in a thermocycler for fifteen minutes at 37°C then fifteen minutes at 80°C. ExoSAP- IT® does not remove any of the salts or inactivated proteins to clean up the reactions. The EXOnuclease digests any single stranded DNA (such as primers) and Shrimp Alkaline

Phosphatase removes phosphates from dNTPs so they are unreactive and can no longer be incorporated into DNA.

8 Sequencing

There are two methods of manual sequencing: the Sanger method and the chemical sequencing method. Maxam and Gilbert designed the chemical sequencing method in 1977. There are three main steps in this method: first bases are modified, the modified bases are then removed from their sugars, and finally the strand is broken where the sugar molecule is exposed (Slightom et al. 1991).

The Sanger method is more frequently referred to as the dideoxy or chain-terminating method of sequencing. DNA synthesis is a key component of the Sanger method so protocols that are based on this method will require a DNA polymerase (such as Taq), deoxyribonucleic acids (dNTPs), a primer, and dideoxyribonucleic acids (ddNTPs). In the past, the dNTPs were frequently radiolabeled with ^{32}P or ^{35}S with ^{35}S providing clearer results. Most sequencing reactions are currently done with fluorescently labeled ddNTPs. The ddNTPs are missing a hydroxyl group at the 3' position of the deoxyribose ring, which prevents additional bases from being added and terminates the extension of the DNA fragment (Hoy 1994).

The primer is annealed to the template in the preincubation step of the dideoxy or chain-terminating reaction. Four distinct reactions are performed (one per base) by combining DNA polymerase, primers, dNTPs, ddNTPs, and the template DNA. There are fewer ddNTPs than dNTPs in the solution so the integration of a ddNTP is random. The DNA molecule is terminated whenever a ddNTP is incorporated into the chain. This results

in a mixture of DNA molecules of different sizes that share the same sequence on one end (due to the primer). Sequencing reactions are done with only one primer at a time and thus only in one direction (forward or reverse). The labeled DNA fragments are run out on a gel to determine the sequence of the bases of the DNA. Since the DNA fragments have been terminated at different bases along the chain, they will run out on a gel and form bands of different lengths. With radiolabeled primers, four different lanes are used for the four reactions; with fluorescent primers, only one lane is needed because each tag is a different color. The radioactive marker on the bases allows each band to be identified as a base in the sequence of the segment of DNA. The smaller pieces of DNA move more quickly creating bands closer to the end of the gel. The larger pieces cannot make their way through the gel as quickly and stay closer to the beginning of the gel. The manual Sanger method uses an X-ray film to determine the location of the bands of DNA (Hoy 1994).

An ABI (Applied Biosystems) Automated Sequencer was used in these experiments. Automated sequencers operate similarly to the manual method of sequencing developed by Sanger (Ferl et al. 1991). The machine performs electrophoresis inside capillary tubes and records the bases using a laser beam that detects fluorescently-labeled ddNTPs instead of having to perform these tasks manually. Also, all the reactions may be done at once now because the method of marking the bases is different. The ABI machine used in these experiments has a 16 capillary array so that 16 reactions may be sequenced at the same time. Radioactive marking is a health hazard and is rarely used in sequencing because of recent developments in fluorescent dyes and automatic sequencing. When sequencing, both forward and reverse reactions should be used to reach a consensus (although this is not always necessary with shorter segments such as the barcoding region). Usually, the

sequencer does not provide clear results at the very beginning and end of the strand of DNA. When both a forward and reverse primer are used (with overlap), it is possible to obtain longer sequences than by using a primer in one direction. The most important benefit of using a forward and reverse primer is that it is possible to compare the two sequences when the sequencer has difficulty calling a base. The sequencing profile used for this experiment was an initial denaturation of 96ºC for 2 minutes, followed by 25 cycles of a 30 second denaturation at 96ºC, a 15 second annealing period at 50ºC, and a 60ºC extension for 2 minutes and 30 seconds. The final extension is at a temperature of 60ºC for 5 minutes. The sequencing primers were identical to the amplification primers.

9 Sequences Used in this Study

A total of 248 specimens from the New Zealand cicada genus *Kikihia* were sequenced for this project (see Tables 1-3 for a list of all specimens and collection locations). Species identifications were based on song (which plays a crucial role in mating) and morphology. Five mitochondrial data sets were created for this thesis: 1) a set of 50 *Kikihia* sequences (COI nonbarcode, COII, A6, A8) taken from a previous study conducted in our laboratory (Marshall et al. 2008), called Multigene-50; 2) a set of 49 *Kikihia* sequences (COI barcode) newly sequenced from the same specimens used in Marshall et al. (2008), called Barcode-49; 3) a set of 149 Muta and Westlandia sequences (COI and COII) from a previous study in our laboratory (Marshall et al. in review) called Muta-multigene-149; 4) a set of 149 specimens (COI barcode) sampled largely from the *Kikihia* Muta and *Kikihia* Westlandica species groups, called Muta-barcode-149; and 5) a set of 69 sequences (COI barcode) collected from

two species suspected of hybridization *K.* "murihikua" and *K. angusta* called Hybrid-69. These are described below.

9.1 Multigene-50 Dataset

Marshall et al. (2008) determined a well resolved phylogeny of the New Zealand cicada genus *Kikihia* based on separate analyses of 2152 bp of mitochondrial DNA (COI, COII, ATPase6, ATPase8 combined) and 1545 bp of the nuclear gene EF-1α. The authors recognized four monophyletic species groups based on this analysis: the Westlandica group (including *K.* "westlandica north," *K.* "westlandica south," *K.* "tasmani," *K. angusta*, *K.* "murihikua," *K.* "flemingi," and *K. subalpina*), the Cutora group (*K. cutora, K. cutora exulis, K. cutora cumberi, K. convicta, K. laneorum, K. dugdalei,* and *K. ochrina*), the Rosea group (*K.* "balaena," *K.* "rosea acoustica," and *K.* "peninsularis"), and the Muta group (*K.* "aotea east," *K.* "aotea west," *K. longula*, *K.* "nelsonensis," *K. muta*, *K.* "tuta," and *K. paxillulae*). These four species groups were found to have diverged more or less simultaneously as an unresolved polytomy. Two of the other three *Kikihia* species (*K. cauta* and *K. scutellaris*), known as the "shade singers," were found to have diverged from the rest of the genus first. The last species, *K. horologium*, also diverges at the unresolved polytomy, and it was not included as a member of any of the subgroups determined by Marshall et al. (2008). There were some differences between the results obtained from the mitochondrial and nuclear DNA in some of the more closely related taxa. The placement of *K.* "acoustica," *K. angusta*, *K.* "tasmani," and *K.* "nelsonensis" varied slightly between the phylograms created from nuclear and mitochondrial data. For this study, the original Marshall et al. (2008) mitochondrial dataset was slightly trimmed because some specimens used in that study were not available for use in this project. A total of 50 sequences were included in this dataset. One *K.* "aotea

(western)" specimen, 01.TO.RCG.01, has two sequence ambiguities which were coded. This specimen was still included in both the Multigene-50 and Muta-barcode-149 datasets.

9.2 Barcode-49 Dataset

The DNA barcoding section of the COI gene was sequenced for the same specimens used in the Marshall et al. (2008) trimmed subset. A total of 49 sequences were included and there were two specimens per species for the majority of the specimens in the dataset.

9.3 Muta-barcode-149 Dataset

The dataset that will be referred to as the "Muta Group" dataset is composed of some of the species from both the Muta group (which includes *K.* "aotea east," *K.* "aotea west," *K. longula*, *K.* "nelsonensis," *K. muta*, *K.* "tuta," and *K. paxillulae*) and the Westlandica group (including *K.* "westlandica north," *K.* "westlandica south," *K.* "tasmani," *K. angusta*, *K.* "murihikua," *K.* "flemingi," and *K. subalpina*). The species that had multiple representatives in this dataset were *K.* "aotea east," *K.* "aotea west," *K.* "nelsonensis," *K. muta*, *K.* "tuta," *K. paxillulae*, *K.* "westlandica north," *K.* "westlandica south," and *K.* "tasmani." *K. paxillulae* and *K.* "tasmani" only had two specimens in this subset of the study. One representative of each of the other *Kikihia* species was included for context while analyzing the Muta-barcode-149 and the Hybrid-69. There is only *K. paxillulae* specimen included in the Muta-barcode-149 dataset. An additional *K. paxillulae* specimen was one of the specimens included in the group of representative species within multiple analyses in this project. This second *K. paxillulae* specimen was included in all of the random subsamples.

9.4 Hybrid-69 Dataset

This dataset included over 30 specimens from *K.* "murihikua" and *K. angusta* from populations spread out over their ranges dataset (a total of 69 sequences). The specimens

included in this dataset were not obvious hybrids with the exception of the four *K. rosea* x *K.* "murihikua" hybrids known from song.

10 Phylogenetic Analysis

10.1 Multigene-50 and Barcode-49 data analysis

The first step of analysis was to run the *Kikihia* Multigene-50 and the *Kikihia* Barcode-49 datasets in Modeltest version 3.7 (Posada and Crandall 1998) to see which model of evolution best fit the data. The model GTR+I+ Γ was selected using the Akaike Information Criterion (AIC: Akaike 1973) and it was used for all of the analyses in this project. The sampling for this dataset was evenly distributed throughout the genus so it is appropriate to use for all of the analyses. The Multigene-50 dataset was reanalyzed in the same manner as the barcoding dataset so that any differences in the results would be from the data and not the methodology.

Maximum likelihood (ML) phylograms with bootstrap support values were created using RAxML version 7.2.6 as configured on the CIPRES portal (Miller et al. 2009), which uses the default settings for this version of the program. RAxML estimates tree topology under a GTR+CAT model, which uses a simplified treatment of among site rate variation, and then optimizes the final branch lengths under GTR+I+Γ. Bootstrap percentages were obtained using 100 pseudoreplicates in each case. The higher the bootstrap value the greater the support for that particular node. ML bootstrap values of 70 and higher are generally regarded as potentially good groupings of taxa. Values in the 90's are most reliable. All bootstrap values less than 50 were removed from the figures in this paper. All trees were mid-point rooted.

10.2 Muta-barcode-149 data analysis

The Muta-barcode-149 dataset was best fit to the HKY+I+ Γ model. This model is less complex than GTR+I+Γ and it is likely that underparameterization (selecting a model that is too simple) causes more problems in phylogenetic analysis than overparameterization (Lemmon and Moriarty 2004), so we elected to use the GTR+I+Γ model.

10.3 Subsampling of Muta-barcode-149 dataset

In order to observe the effect that sample size has on the success of barcoding, random subsamples of the Muta-barcode-149 dataset were taken. We felt that random sampling of the Muta-barcode-149 dataset would allow us insight into various scenarios which can limit specimen collection for any project (weather conditions, travel limitations, accessibility of habitat, knowledge of the full range of a species, etc.). Ten random samples containing 35 specimens (about a quarter of the total specimens within the dataset) were taken. To make random samples, the Muta-barcode-149 specimens were organized in a table in alphabetical order according to specimen code. Each specimen was assigned a number (1-149) based on its placement in the table. MINITAB 14 Student (Minitab Inc.) was used to generate lists of 35 random numbers (in the range of 1-149), and the corresponding specimens were then selected and assembled into new dataset files for analysis. Single representatives of most of the remaining *Kikihia* species were included in the subsample analyses to provide context, just as in the main study. *K. cauta* and *K. scutellaris*, the two most distant *Kikihia* species, were removed after initial analysis because they did not fall as outgroups as observed in the Marshall et al. (2008) study using more data. A possible explanation for this phenomenon is that these species diverge too deeply in the *Kikihia* tree for their relationships to be accurately reconstructed using the smaller datasets.

10.4 Hybrid-69 data analysis

Maximum likelihood (ML) phylograms with bootstrap support values were created using RAxML and the final branch lengths were optimized under GTR+I+Γ.

Results

1 *Kikihia* Multigene-50 and Barcode-49.

The Multigene-49 mitochondrial phylogeny was created from 2152 bp of mitochondrial DNA (COI, COII, ATPase6, ATPase8 combined) with 509 parsimony informative sites. The barcoding phylogeny was based on 651 bp (COI) containing 159 parsimony informative sites. The Multigene-49 dataset had roughly three times more parsimony informative sites than the barcoding region alone.

All of the groups present in the Marshall et al. (2008) mitochondrial phylogeny are still present in the Multigene-50 version modified for this paper. In addition, all of the groups are strongly supported on both trees (ML bootstraps of 100) with the exception of the Rosea group (ML bootstrap of 69 in Marshall et al. 2008; 73 in the Multigene-50 tree). In the original findings of Marshall et al. (2008), the Westlandica group is not as well supported (ML bootstrap 86 versus 100 in the Multigene-50 tree).

An examination of Figure 5 shows that there are differences in the barcode phylogeny versus the phylogeny with multiple mitochondrial genes. In the Multigene-50 phylogram, the Westlandica, Cutora, Rosea, and Muta groups branch off at the same time just as they do in the original publication. In the Barcode-49 phylogram, this polytomy does not exist. The

most serious failing of the Barcode-49 phylogram is the lack of a monophyletic Westlandica group.

2 Muta-barcode-149

It is important to first examine the phylogram of the modified Marshall et al. (in review) dataset which contains sequence from the 3' end of COI and COII (hereafter called the Muta-multigene dataset). Assuming that this phylogram is more reliable because it is based on more information (1467 base pairs of mtDNA with 311 parsimony informative sites compared to the 135 parsimony informative sites from the 654 bp of barcode) it is important to first establish whether the species (which have been determined by song and morphology) are forming monophyletic clades. If this is not the case, DNA barcoding is already set up to fail since the factors used to determine species identification would not be compatible with the process of DNA barcode-based species identification.

In the Muta-multigene phylogram, all of the *K.* "aotea west" specimens form a paraphyletic group. This clade also contains the *K. muta* specimen 01.WI.PUT.02. The majority of the *K.* "aotea east" specimens also form a monophyly with the representative *K. longula* specimen. The four other *K.* "aotea east" specimens (02.HB.OCB.01, 01.HB.ES*K*.01, 02.GB.NUH.01, and 02.HB.GGR.02) form a monophyletic group that is sister to the clade containing the majority of the *K. muta* specimens. The remaining six *K. muta* specimens (03.KA.OKI.01, 03.MB.HNR.01, 01.KA.BDS.05, 02.KA.WBS.04, 03.KA.WK*K*.04, and 03.SC.PFR.01) are found in a clade that also contains one *K. paxilullae* (01.KA.BDS.04), two *K.* "nelsonensis" (01.MB.TWI.10 and 02.NN.KPL.01), and all of the *K.* "tuta." The remaining *K.* "nelsonensis" fall into three separate clades, two of which are

composed solely of this species and one of which also includes a *K.* "westlandica north" (02.NN.FAR.02). The other *K. paxillulae* (97.KA.PPR.81) is sister to all of the specimens discussed up until this point. The remaining *K.* "westlandica north" specimens are paraphyletic within a clade that is sister to *K.* "westlandica south." The two *K.* "tasmani" specimens fall within this clade. All of the *K.* "westlandica south" specimens form a single monophyletic group.

Overall, the Muta-barcode phylogram is compatible with the Muta-multigene phylogeny. However, a general trend that appears when comparing the two trees is that the barcode phylogram has weaker support at the majority of nodes. There is also a false sister relationship between two of the *K.* "nelsonensis" clades and the miscellaneous clade containing all of the *K.* "tuta" specimens.

2.1 Subsampling the Muta-barcode-149 dataset

The results provided by the random sampling of the Muta-barcode-149 dataset are contained in Table 5. Barcoding species identification success was defined by all specimens of the species forming a single monophyletic clade. If a single member of the species was found outside of this clade or any member of another species was within the clade, it was not considered a success for barcoding because it would result in incorrect species identifications.

3 *K.* "murihikua" and *K. angusta* Hybrid-69 dataset

There were 66 parsimony informative sites within the 653 bp of sequence used in this subset of the project. The *K. angusta* sequences appeared in various groups within the phylogram. There is a large monophyletic grouping of *K. angusta* sequences but the node is

not well supported (ML bootstrap was under 50) so it can be thought of as two separate groups. Five additional *K. angusta* are located in two groups dispersed among the *K.* "murihikua" specimens. *Kikihia* "murihikua" forms many small groups and there is not even a suggestion of a larger monophyletic clade for this species. Two of the four *K. rosea* x *K.* "murihikua" hybrids were most closely related to two *K.* "murihikua" specimens with strong support (ML boostrap value was 100). The other two *K. rosea* x *K.* "murihikua" hybrids were most closely related to *K. rosea* and *K.* "acoustica" (also with a ML bootstrap of 100).

Discussion

1 *Kikihia* Multigene-50 and Barcode-49

The barcoding region only offers a third of the parsimony informative sites that the multiple mitochondrial genes of Marshall et al. (2008) offer. This means that there is much less information available to create a phylogeny. In the Barcode-49 phylogeny, there is a false sister relationship between the Muta group and the Cutora group (with ML bootstrap of 76) which is most likely due to random error. This random error occurs because the barcoding region does not provide enough information. The Barcode-49 phylogeny has some clades with very strong support (ML bootstrap is 96) for the Muta group and the Cutora group is also strongly supported (ML bootstrap of 84). The Rosea group has moderate support in both of the phylograms. In the Barcode-49 phylogram, the "Shade singers" have a much weaker support than the rest of the groups (ML boostrap 58) and the Westlandica group is not monophyletic and its paraphyletic sub-components are not well supported. It is important to note that these are deeper level relationships within the genus.

The focus of barcoding is species-level identification. Whether the species within the genus are well supported monophyletic groups is much more pertinent to whether barcoding is successfully meeting its aims (Hebert et al. 2003, 2004; Mitchell 2008; Rock et al. 2008; Pagès et al. 2009; Kerr et al. 2009; Ward 2009; Lukhtanov et al. 2009; Meyer and Paulay 2005). For successful barcoding species identifications, species should be distinguishable by their placement within the phylogeny. Species should form distinct monophyletic clades. The Marshall et al. (2008) tree as well as the barcoding tree would have had difficulty differentiating between the following species pairs: *K. longula* and *K.* "aotea east," *K.* "acoustica" and *K. rosea*, *K.* "tasmani" and *K.* "westlandica north," and *K.* "murihikua" and *K. angusta*. *K.* "nelsonensis" would also pose problems for species identification in both of the phylograms. The barcoding phylogram shows another three species that would be difficult to identify using these data alone: *K. cutora cutora*, *K.* "flemingi" and *K. subalpina*. These species identifications would be challenging for two reasons: either the species is not monophyletic or the species do not have enough differences to tell them apart (such that the distance is comparable to other instances between conspecific taxa within *Kikihia*). The lack of a barcode gap which occurs in this study adds evidence to previous studies that state the barcoding gap is not universal (Langhoff et al. 2009; Ward 2009; Meyer and Paulay 2005; Lukhtanov et al. 2009; and Rubinoff 2006a). The *Kikihia* genus is not well suited to mtDNA barcoding.

2 Muta Group

DNA barcode dataset recovered many of the clades from the Marshall et al. (in review) phylogeny. Overall, the support is not as strong but the drops in ML bootstraps are

not usually extreme. The barcode phylogram does have some difficulty with the sister relationships between many of the clades.

A comparison of the Muta-multigene-149 tree (Figure 6) to the Muta-barcode-149 tree (Figure 7) shows significant disagreement. Clades labeled A-F on the Muta-multigene-149 tree show completely different sister-group relationships on the Muta-barcode-149 tree including some clades that are falsely related on the barcode tree. These can be represented as ((((A1, (A2, A3)), ((B1, B2), C), D), E), F) on the Muta-multigene-149 tree versus ((A1, (A2, A3)), (((B1, B2), D), ((F, E), C))) on the Muta-barcode-149 tree. If all nodes with less than 50% bootstrap support are collapsed, then the Muta-multigene-149 tree looses only the group B + C to become ((((A1, (A2, A3)), (B1, B2), C, D), E), F) but the Muta-barcode-149 tree becomes (((A1, A2, A3), (B1, B2), D), (C, (E, F))). The group A2 actually looses a few taxa and most of this clade is unresolved. The group (C, (E, F) is a false grouping on the barcode tree compared to the Muta-multigene-149 tree.

The sometimes complex relationships between members of the same species result in failures for species identification using DNA barcoding. Marshall et al. (in review) found that there was no species gap within the mitochondrial data to establish a threshold between intraspecific and interspecific variation in the *Kikihia* Muta group. The findings of this study and those of Marshall et al. (in review) add support to previous studies conclusions that the barcoding gap does not always exist (Langhoff et al. 2009; Ward 2009; Meyer and Paulay 2005; Lukhtanov et al. 2009; and Rubinoff 2006a). Marshall et al. (in review) also cited introgression and incomplete lineage sorting following recent speciation events to be particularly problematic for mtDNA based identifications within this genus. Hybridization, introgression, and incomplete lineage sorting have all been noted as problematic for DNA

barcoding species identifications (Hebert et al. 2003, 2004; Mitchell 2008; Ward 2009; Fontaine et al. 2007; Rubinoff 2006a; Ballard & Whitlock 2004; Rock et al. 2008; Langhoff et al. 2009). A few notable examples of these problematic biological phenomena were discussed in detail are *K.* "tuta" (some specimens with *K.* "tuta" mtDNA have *K. muta, K. paxillulae,* and *K.* "nelsonensis" songs), *K.* "aotea east" (which may not be a completely different species from eastern *K. muta* and *K. longula*), and *K.* "tasmani" (which may have captured the mtDNA of *K.* "westlandica north"). These complex relationships make defining species through mitochondrial DNA particularly challenging for the genus *Kikihia*. This knowledge of the genus (along with the fact that *K. muta* 01.WI.PUT.02 is a hybrid) provides a clearer understanding of why DNA barcoding is unsuited to these species. The only species that can be reliably and consistently identified using DNA barcoding is *K.* "westlandica south" since it forms a monophyletic clade with a long stem (so it is easy to separate from its sister species).

The fact that subsampling fewer species can actually result in higher success rates is problematic for barcoding and suggests random factors are at play. Subsampling could result in higher success rates for barcoding because some of the specimens and/or species that would cause misidentifications may not be selected for analysis. The two subsamples with the highest success rates could identify four out of the six or seven species present (see Table 6). Considering there were a total of ten subsamples this is not very impressive. This study did not provide a very high success rate and the *Kikihia* Muta group's results do not support the positive results reported for some other studies (Meusnier et al. 2008; Hebert et al. 2003, 2004; Kerr et al. 2009).

A sample size of 35 was usually sufficient to recover most of the different clades (see Table 7) in the Muta-multigene phylogeny. All of the individual species were usually recovered in the subsampling datasets (see Table 6). Since there were only two *K.* "tasmani" and one *K. paxillulae* (in addition to the one specimen of this species included within the representatives from the rest of the genus) it is not surprising when subsampling failed to recover multiple specimens from these species. *K.* "tuta" and *K.* "westlandica south" both were completely excluded from one sample and had only one representative in three other samples.

3 *K.* "murihikua" and *K. angusta* Hybrid-69 dataset

Neither *K.* "murihikua" nor *K. angusta* form well supported monophyletic clades. It would be impossible to identify these species using DNA barcoding alone. These findings support the hypothesis that DNA barcoding cannot identify species with known hybridization and introgression. Hybridization is occurring in both directions between *K. rosea* and *K.* "murihikua" since two of the hybrids were most closely related to *K.* "murihikua" and the other two were most closely related to *K. rosea* and *K.* "acoustica." These four hybrid individuals were all found at the same Dunedin-LMR locality. This site was located along a hillside road with fragments of native bush, a lot of disturbed vegetation, planted pines, and grasses for vegetation in the area. Based on my data and that from Marshall et al. (2009; in review), hybridization is clearly occurring within *Kikihia* which makes this genus a bad candidate for species identifications based on mtDNA barcodes. One worrying aspect of the hybrid samples is that many were not identifiable as hybrids by morphology, so the lack of

intermediates does not guarantee that a group of species is a good candidate for DNA barcoding.

For species to be successfully identified using DNA barcoding, there must be a gap that shows the intraspecific variation is less than interspecific variation (Meyer and Paulay 2005; Langhoff et al. 2009; Ward 2009; Lukhtanov et al. 2009). Even if the four hybrid *K. rosea* x *K.* "murihikua" specimens and five *K. angusta* specimens outside of the main monophyletic clade were removed from this analysis, it would not yield successful identifications. There is not a distinct gap separating the main *K. angusta* clade from the multiple clades of *K.* "murihikua." The results of this study corroborate the findings of many other studies that there is not always a clear gap between species (Meyer and Paulay 2005; Langhoff et al. 2009; Ward 2009; Lukhtanov et al. 2009; Rubinoff 2006a).

Conclusions

The DNA barcode had fewer basepairs of sequence (651) and fewer parsimony informative sites (159) than the Marshall et al. (2008) dataset that included multiple mitochondrial genes (2152bp, 509 parsimony informative sites). Since there is less information available, it is understandable that the mtDNA barcode would have more problems than sequence data from multiple mitochondrial genes. DNA barcoding has difficulty determining deeper level relationships within the New Zealand cicada genus *Kikihia*. It was not able to identify the *Kikihia* Westlandica group as monophyletic and there was less support for the other monophyletic groups. Both the Multigene-50 (and Marshall et al. 2008) and Barcode-49 phylogenies would pose nine identification challenges. Three additional species would be

impossible to identify using the barcode phylogeny alone. There would be a total of 11 species that could not be definitively identified within the 29 species of the genus *Kikihia*. The Muta-multiple and Muta-barcode datasets show that DNA barcoding cannot successfully handle the complex species relationships presented by the genus *Kikihia.* The Muta-barcode phylogeny had weaker support at most nodes and reported false sister relationships due to random error. Hybridization, introgression, incomplete lineage sorting, and recent speciation greatly complicate species identification. A clear example of this was provided in the *K.* "murihikua" and *K. angusta* case study. DNA barcoding could not successfully identify these two species due to introgression. As further evidenced by the *K. rosea* x *K.* "murihikua" specimens, hybridization is definitely a concern within this cicada genus.

Literature Cited

Abriouel, Hikmate, Antonio Martín-Platero, Mercedes Maqueda, Eva Valdivia and Manuel Martínez-Bueno. (2008) Biodiversity of the microbial community in a Spanish farmhouse cheese as revealed by culture-dependent and culture-independent methods. *International Journal of Food Microbiology* 127(3): 200-208.

Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary, pp. 267–281.

Baker, Allan J., Erika Sendra Tavares, and Rebecca F. Elbourne. (2009) Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Molecular Ecology Resources* 9(Suppl. 1): 257-268.

Ballard, J. William O. and David M. Rand (2005) The Population Biology of Mitochondrial DNA and Its Phylogenetic Implications. *Annual Review of Ecology, Evolution, and Systematics* 36: 621-642.

Ballard, J. William O. and Michael C. Whitlock (2004) The incomplete natural history of mitochondria. *Molecular Ecology* 13: 729-744.

Crosby, T. *K.*, J. S. Dugdale, and J. C. Watt (1998) Area codes for recording specimen localities in the New Zealand subregion. *New Zealand Journal of Zoology* 25: 175-183.

Ferl, Robert J., et al. (1991) The Application of Automated DNA Sequence Analysis to Phylogenetic Studies. *Phylogenetic Analysis of DNA Sequences*. Ed. Michael M. Miyamoto and Joel Cracraft. New York: Oxford UP. 45-58.

Fontaine, Kathryn M., John R. Cooley, and Chris Simon. (2007) Evidence for Paternal Leakage in Hybrid Periodical Cicadas (Hemiptera: *Magicicada* spp.). *PLoS ONE* 2(9): e892.

Herbert, Paul D. N., Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society* 270: 313-321.

Hebert, Paul D. N., Mark Y. Stoeckle, Tyler S. Zemlak, and Charles M. Francis. (2004) Identification of Birds through DNA Barcodes. *PLoS Biology* 2(10): 1657-1663.

Hoy, Marjorie A. *Insect Molecular Genetics: An Introduction to Principles and Applications*. Boston: Academic Press, 1994.

Iliff, Peter, Robert Ntozini, Kusum Nathoo, Ellen Piwoz, Lawrence Moulton, The ZVITAMBO Study Group, and Jean Humphrey. (2008) Making a working clinical

diagnosis of HIV infection in infants in Zimbabwe. *Tropical Medicine and International Health* 13(12): 1459-1469.

Kerr, Kevin CR, Sharon M Birks, Mikhail V Kalyakin, Yaroslav A Red'kin, Eugeny A Koblik, and Paul DN Hebert. (2009) Filling the gap- COI barcode resolution in eastern Palearctic birds. *Frontiers in Zoology* 6(29).

Langhoff, Pia, Astrid Authier, Thomas R. Buckley, John S. Dugale, Allen Rodrigo, and Richard D. Newcomb. (2009) DNA barcoding of the endemic New Zealand leafroller moth genera, *Ctenopseustis* and *Planotortrix*. *Molecular Ecology Resources* 9: 691-698.

Lemmon, A. R. and E. C. Moriarty (2004). The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology* 53(2): 265-277.

Lewis, Louise A. and Paul O. Lewis. (2005) Unearthing the Molecular Phylodiversity of Desert Soil Green Algae (Chlorophyta). *Systematic Biology* 54(6): 936-947.

Lin, Chung-Ping, Bryan N. Danforth. (Mar 2004) How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined data sets. *Molecular Phylogenetics and Evolution* 30(3): 686-702.

Lukhtanov, Vladimir, Andrei Sourakov, Evgeny V. Zakharov, and Paul D. N. Hebert. (2009) DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources* 9(Suppl. 1): 1302-1310.

Marshall, David C., Karen Slon, John R. Cooley, Kathy B. R. Hill, and Chris Simon. (2008) Steady Plio-Pleistocene diversification and a 2-million-year sympatry threshold in a

New Zealand cicada radiation. *Molecular Phylogenetics and Evolution* 48: 1054-1066.

Marshall, David C., Kathy B. R. Hill, John R. Cooley, and Chris Simon. (In review) Hybridization, Mitochondrial DNA Taxonomy, and Prediction of the Early Stages of Reproductive Isolation: Lessons from New Zealand Cicadas of the Genus *Kikihia*. *Systematic Biology*

Meier, Rudolph, Kwong Shiyang, Gaurav Vaidya, and Peter *K.* L. Ng. (2006) DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success. *Systematic Biology* 55(5): 715-728.

Meusnier, Isabelle, Gregory AC Singer, Jean-François Landry, Donal A Hickey**,** Paul DN Hebert, and Mehrdad Hajibabaei. (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* 9: 214.

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

Miller, MA, MT Holder, R Vos, PE Midford, T Liebowitz, L Chan, P Hoover, and T Warnow. The CIPRES Portals. CIPRES. 2009-08-04. URL: http://www.phylo.org/sub_sections/portal. Accessed: 2009-08-04. (Archived by WebCite(r) at http://www.webcitation.org/5imQlJeQa)

Mitchell, Andrew. (2008) DNA barcoding demystified. *Australian Journal of Entomology* 47: 169-173.

Packer, Laurence, Jason Gibbs, Cory Sheffield, and Robert Hanner. (2009) DNA barcoding and the mediocrity of morphology. *Molecular Ecology Resources* 9: 42-50.

Pagès, N., F. Muñoz-Muñoz, S. Talavera, V. Sarto, C. Lorca, and J.I. Núñez. (2009) Identification of cryptic species of *Culicoides* (Diptera: Ceratopogonidae) in the subgenus *Culicoides* and development of species-specific PCR assays based on barcode regions. *Veterinary Parasitology* 165: 298-310.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic Biology 55:595-609

Pook, C.E. and R. McEwing. (2005) Mitochondrial DNA sequences from dried snake venom: a DNA barcoding approach to the identification of venom samples. *Toxicon* 46(7): 711-715.

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

Rock, J., F. O. Costa, D. I. Walker, A. W. North, W. F. Hutchinson, and G. R. Carvalho. (2008) DNA barcodes of fish of the Scotia Sea, Antarctica indicate priority groups for taxonomic and systematics focus. *Antarctic Science* 20(3): 253-262.

Rubinoff, Daniel. (2006a) Utility of Mitochondrial DNA Barcodes in Species Conservation. *Conservation Biology* 20(4): 1026-1033.

Rubinoff, Daniel. (2006b) DNA Barcoding Evolves into the Familiar. *Conservation Biology* 20(5): 1548-1549.

Simon, Chris, Adrian Franke, Andrew Martin. (1991) The Polymerase Chain Reaction: DNA Extraction and Amplification. *Molecular Techniques in Taxonomy*. Ed. Godfrey M. Hewitt, Andrew W. B. Johnston, and J. Peter W. Young. New York: Springer-Verlag.

Simon, Chris, Thomas R. Buckley, Francesco Frati, James B. Stewart, and Andrew T. Beckenbach. (2006) Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* 37: 545-579.

Simon, Chris, et al. (1994) Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers. *Annals of the Entomological Society of America* 87(6): 651-701.

Simon, Chris, et al. (2006) Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* 37: 545-79.

Slightom, Jerry L., David R. Siemieniak, and Leang C. Sieu. (1991) DNA Sequencing: Strategy and Methods to Directly Sequence Large DNA Molecules. *Phylogenetic Analysis of DNA Sequences*. Ed. Michael M. Miyamoto and Joel Cracraft. New York: Oxford UP 18-44.

Ward, Robert D. (2009) DNA barcode divergence among species and genera of birds and fishes. *Molecular Ecology Resources* 9: 1077-1085.

Will, Kipling W. and Daniel Rubinoff. (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20: 47-55.

Zettler, Linda A. Amaral, Felipe Gómez, Erik Zettler, Brendan G. Keenan, Ricardo Amils, and Mitchell L. Sogin. (2002) Eukaryotic Diversity in Spain's River of Fire. *Nature* 417: 137.

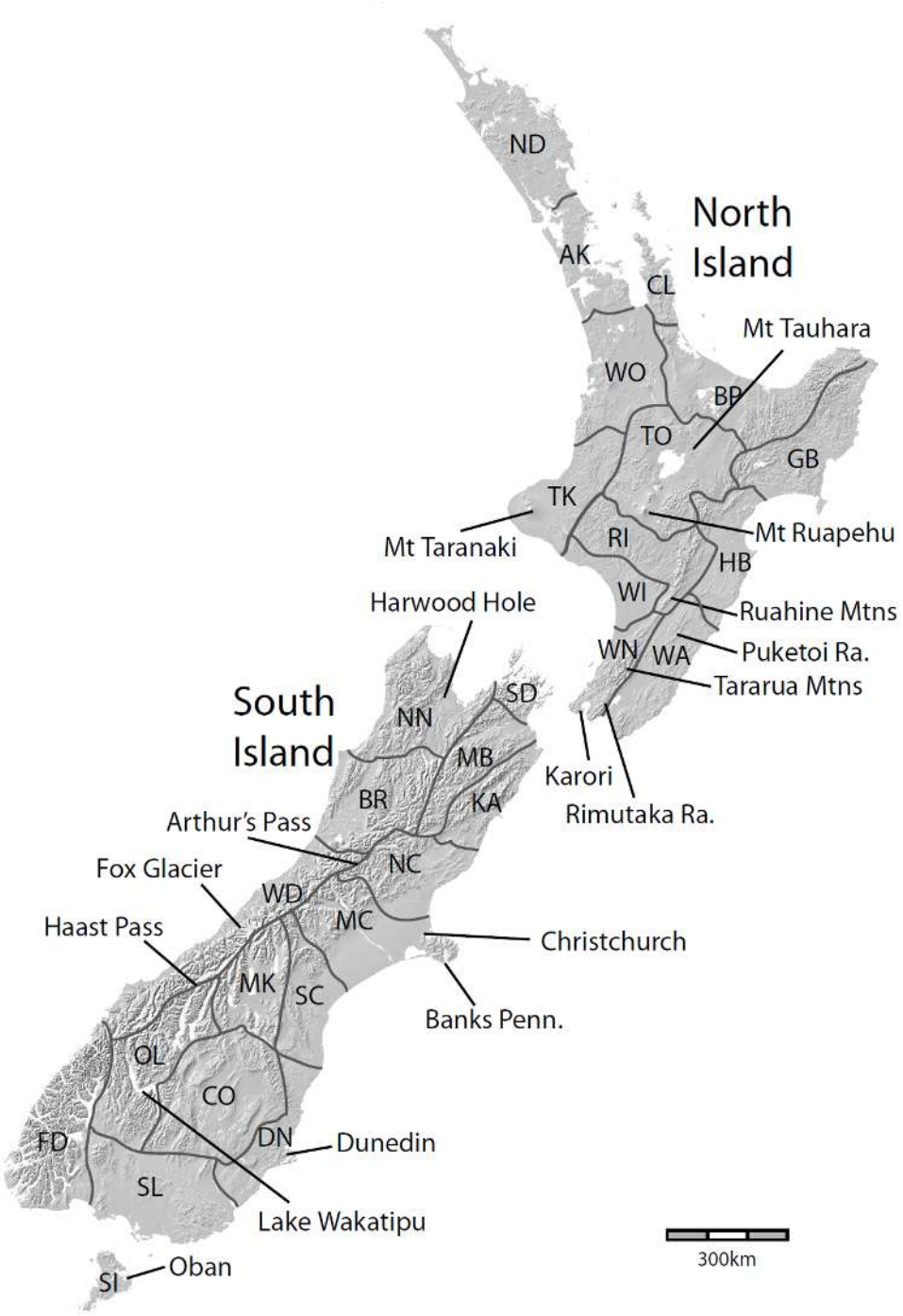

Figure 1. A map of NZ district codes designated by Crosby et al. 1998; map redrawn by Kathy Hill. The codes for the districts of the North Island are AK (Auckland), BP (Bay of Plenty), CL (Corornandel), GB (Gisborne), HB (Hawkes Bay), ND (Northland), Rl (Rangitikei), TK (Taranaki), TO (Taupo), WA (Wairarapa), Wl (Wanganui), WN (Wellington), and WO (Waikato). The South Island district codes are BR (Buller), CO (Central Otago), DN (Dunedin), FD (Fiordland), KA (Kaikoura), MB (Mariborough), MC (Mid Canterbury), MK (Mackenzie), NC (North Canterbury), NN (Nelson), OL (Otago Lakes), SC (South Canterbury), SD (Mariborough Sounds), SI (Stewart Island), SL (Southland), and WD (Westland).

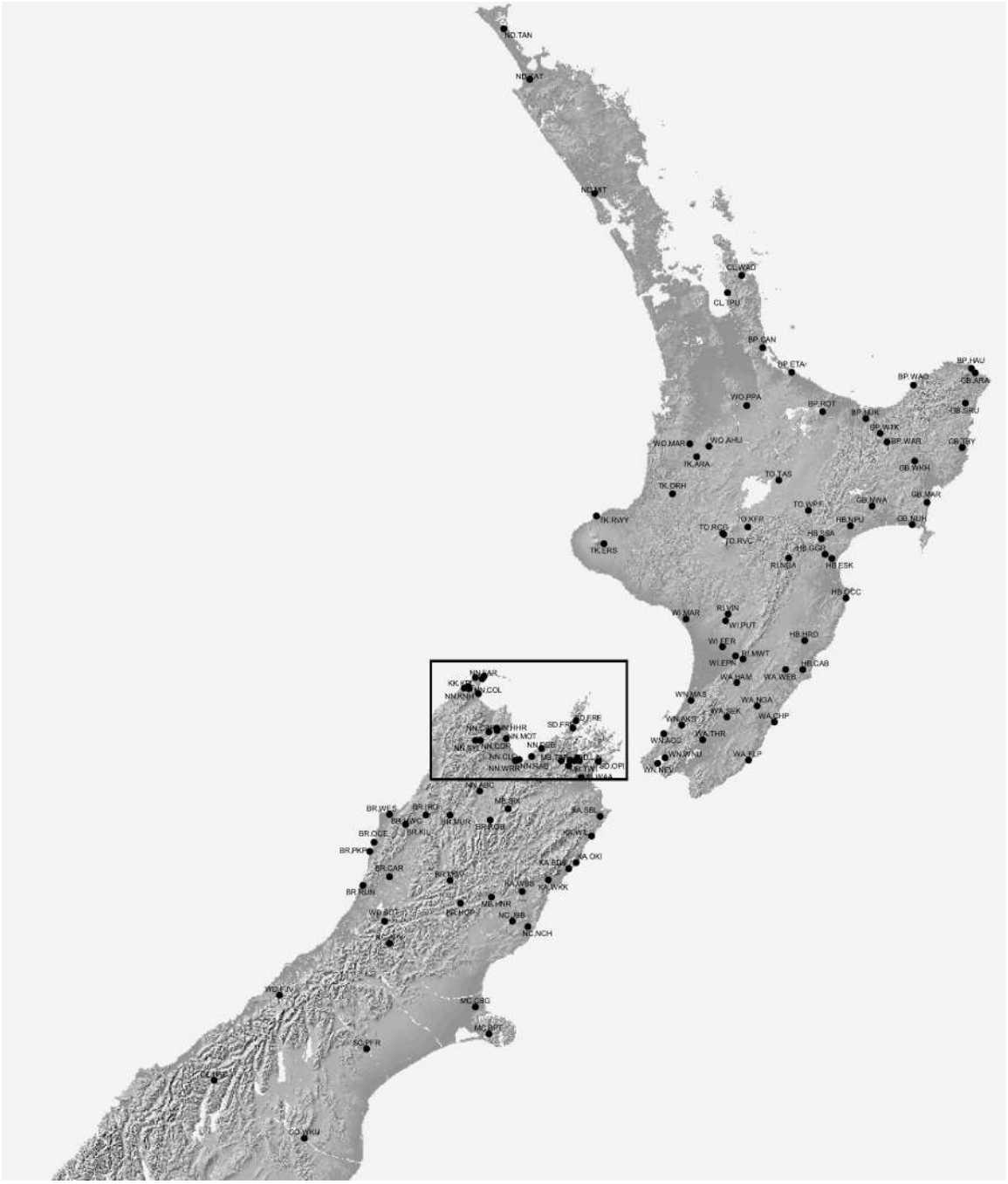

Figure 2. Map of New Zealand with "Muta group" dataset collection localities. A close-up of the highlighted region is available in Figure 3. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

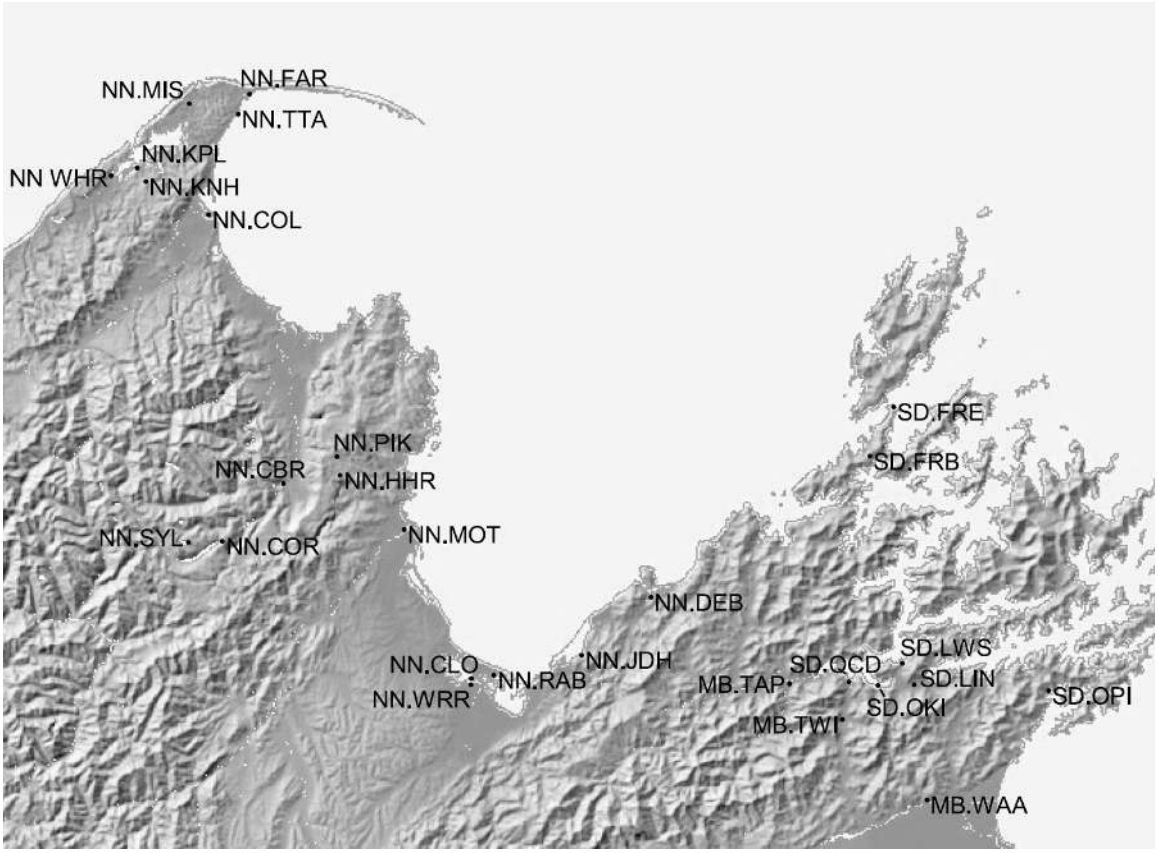

Figure 3. A close-up of "Muta group" dataset collection localities from the northern South Island of New Zealand. The region within this close-up is shown on the map of the entire country in Figure 2. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

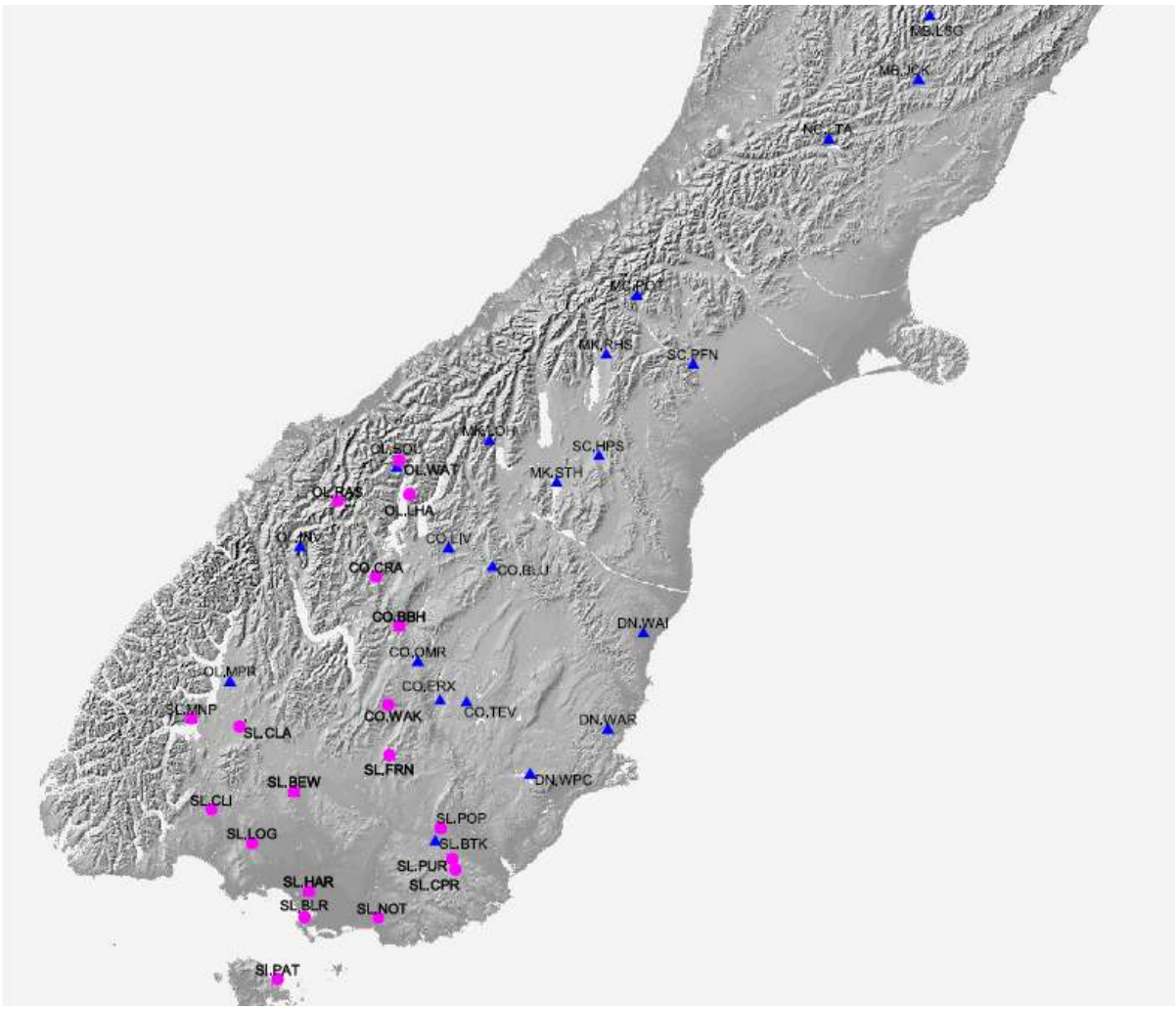

Figure 4. A map of collection localities for *K. angusta* and *K.* "murihikua." *K.* "murihikua" collection locations are in pink and *K. angusta* specimen localities are colored blue.

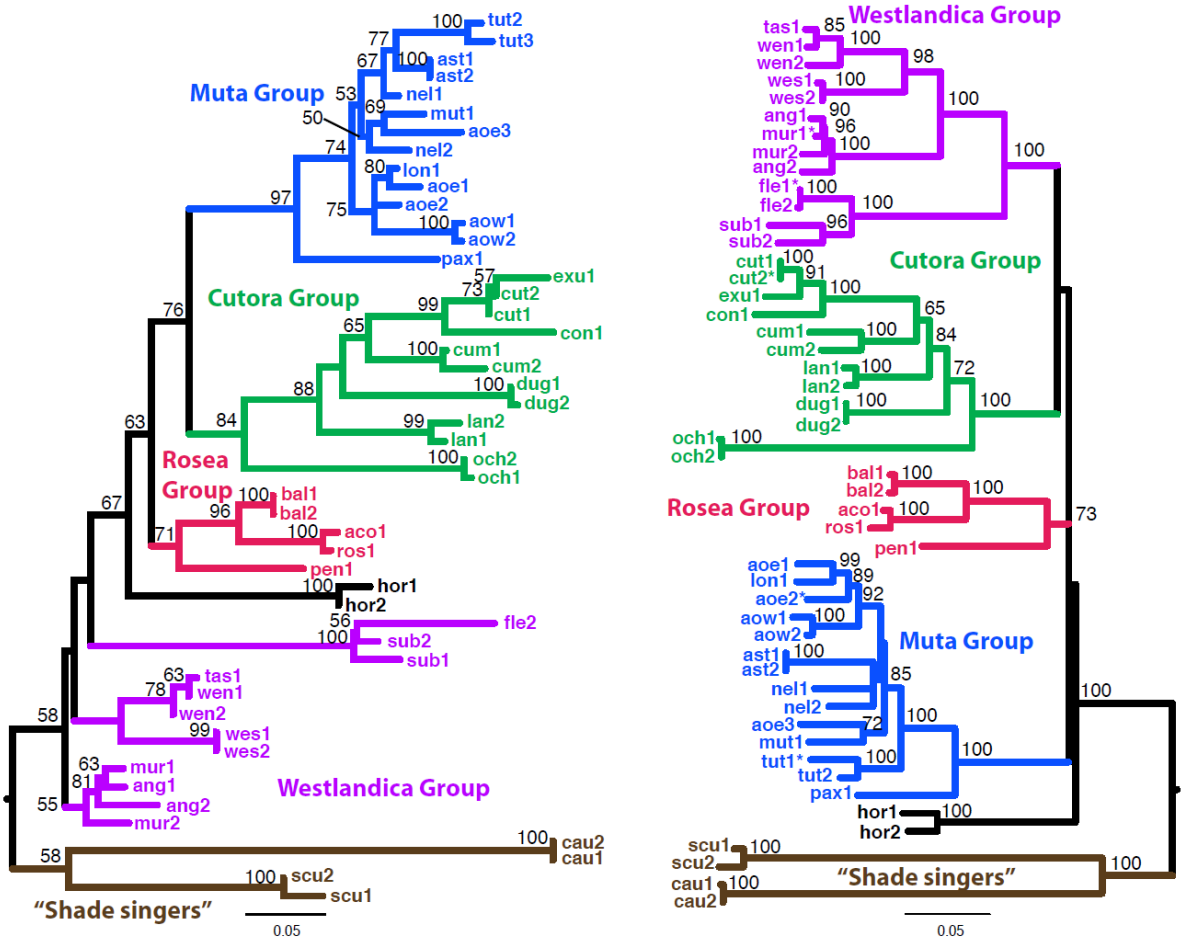

Figure 5. ML (GTR+I+Γ) phylograms of mtDNA Barcode-49 (left) and trimmed Marshall et al. (2008) mitochondrial dataset, Multigene-50 (right). ML bootstrap values less than 50 are excluded from the phylograms.

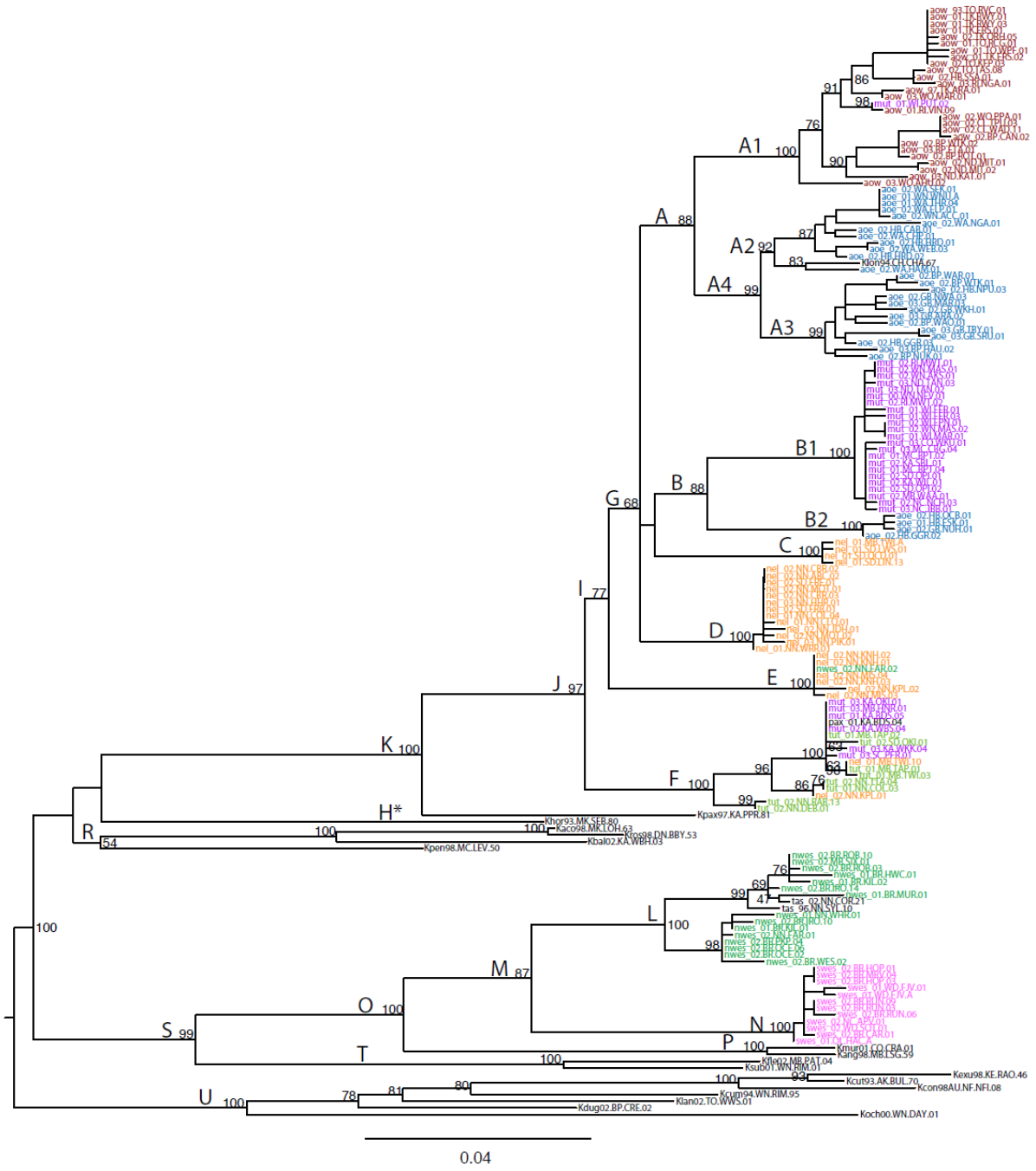

Figure 6. ML (GTR+I+Γ) phylogram of the trimmed Marshall et al. (in review) mitochondrial sequence Muta-multigene-149 dataset. ML bootstrap values less than 50 are excluded from the phylogram. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

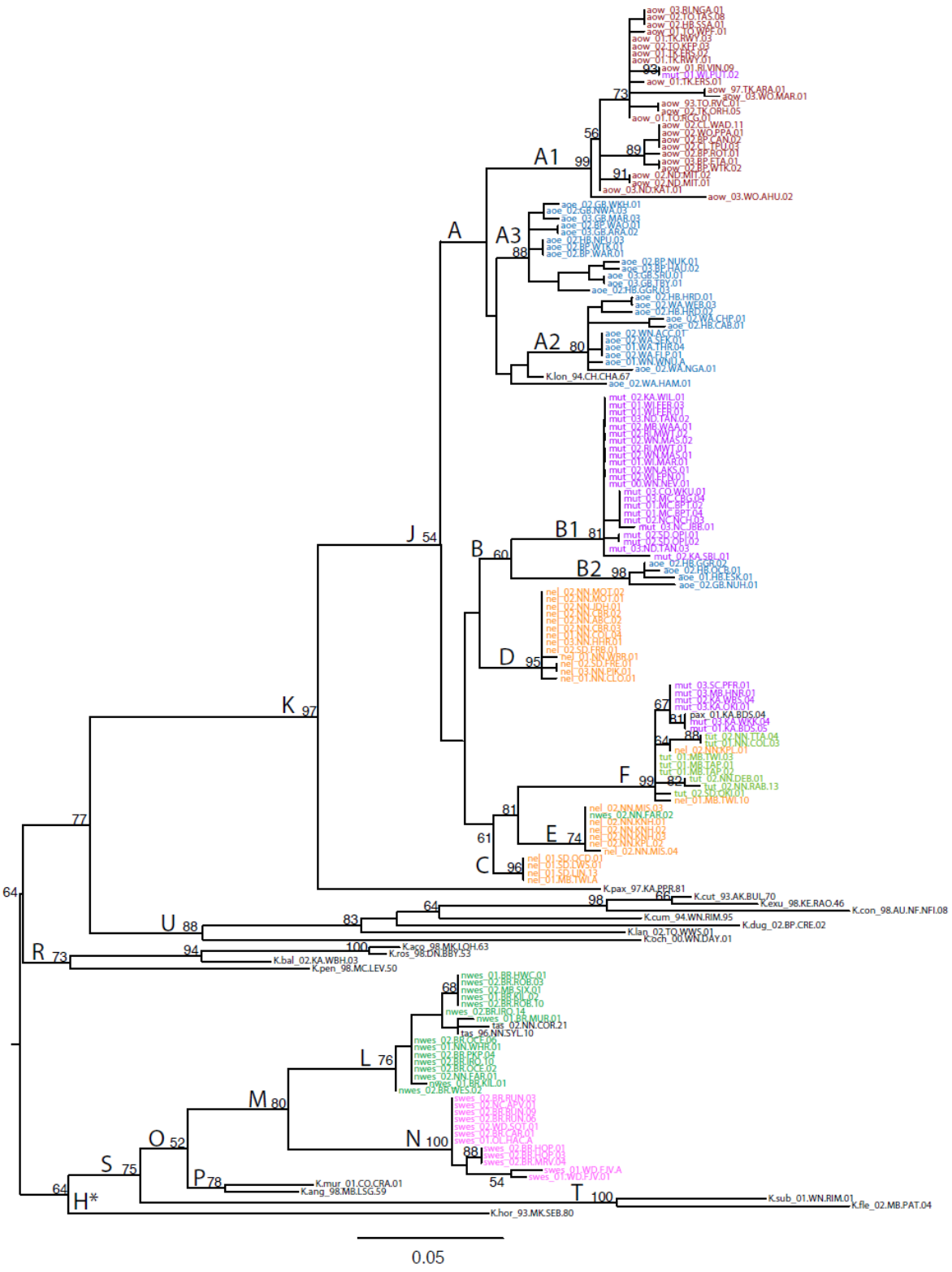

Figure 7. ML (GTR+I+Γ) phylogram of mtDNA barcode for the Muta-barcode-149 dataset. ML bootstrap values less than 50 are excluded from the phylogram. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

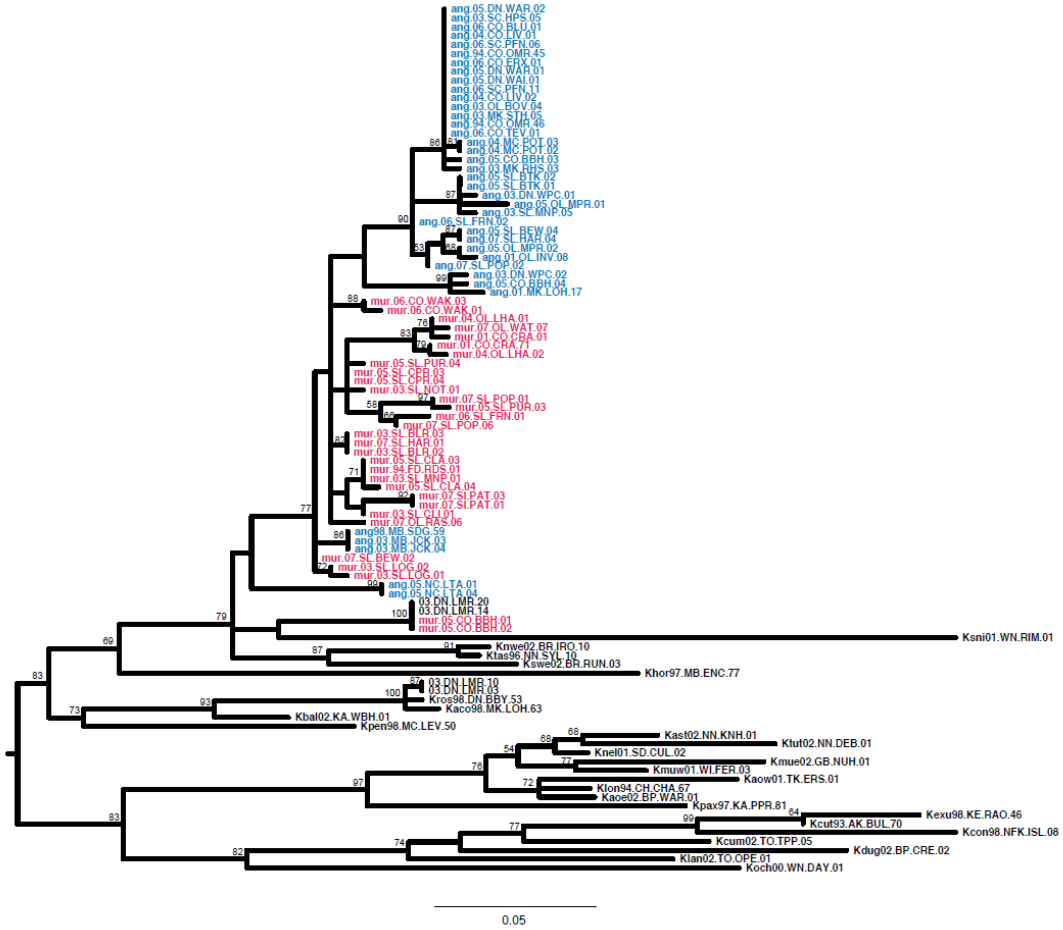

Figure 8. ML (GTR+I+Γ) phylogram of the *K.* "murihikua" and *K. angusta* Hybrid-69 dataset. *K.* "murihikua" are in pink and *K. angusta* specimens are colored blue. ML bootstrap values less than 50 are excluded from the phylogram. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

Table 1. Specimen List of the *Kikihia* Multigene-50 dataset. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

Species	Specimen code	Latitude	Longitude	Location	Tree label
K. "aotea (eastern)"	01.WN.WNU.A	-41.249283	174.921166	Wainuiomata Hill track nr. Tawa Track	aoe1
K. "aotea (eastern)"	02.BP.WAR.01	-38.3049	177.3956	SH2, 7.0 km S. of Wairata	aoe2
K. "aotea (eastern)"	02.GB.NUH.01	-39.044166	177.73765		aoe3
K. "aotea (western)"	01.TK.ERS.01	-39.31255	174.146433	Pembroke Rd., 8 km W. Cardiff Rd.	aow1
K. "aotea (western)"	02.HB.SSA.01	-39.2147	176.6883	SH5 19.4 km NW Glengary Rd., Hawkes B.	aow2
K. "balaena"	02.KA.WBH.01	-42.495033	173.18275	SH70, 3.3 km N. Lyford Lodge, Kaikoura	bal1
K. "balaena"	02.KA.WBH.03	-42.495033	173.18275	SH70, 3.3 km N. Lyford Lodge, Kaikoura	bal2
K. "flemingi"	02.MB.PAT.04	-41.589533	173.29875	W. of L*K.* Chalice Lookout., Mt. Patriarch	fle2
K. "flemingi"	98.OL.TWE.51	-45.0685	168.5458	Twelve Mile Delta, Lake Wakitipu	fle1*
K. "murihikua"	01.CO.CRA.01	-44.903683	168.984916	SH89, 5.2 km SW of Cardrona	mur2
K. "murihikua"	94.FD.RDS.01			SH94, Fiordland	mur1
K. "nelsonensis"	01.NN.WRR.01	-41.295	173.12	SH4, 3.9 km N. Wanganui River Rd.	nel2
K. "nelsonensis"	01.SD.CUL.02	-41.27375	173.787866	Up track S. of Cullens Pt. across SH6	nel1
K. "astragali"**	02.NN.KNH.01	-40.637583	172.5634	Knuckle Hill Summit, NW Nelson	ast2
K. "astragali"**	02.NN.KNH.03	-40.637583	172.5634	Knuckle Hill Summit, NW Nelson	ast1
K. "peninsularis"	98.MC.LEV.50			Rd. to Port Levy, Banks Peninsula	pen1
K. "rosea acoustica"	98.MK.LOH.63	-44.2373	169.82275	0.6 km S. of Lake Ohau lodge	aco1
K. "tasmani"	96.NN.SYL.10	-41.108333	172.633333	NW Nelson Trail, Lake Sylvester	tas1
K. "tuta"	01.NN.WCR.02	-40.5799	172.6274	12 km N. of rd. to Knuckle Hill Trail	tut1*
K. "tuta"	02.NN.DEB.01	-41.1798	173.4294	1.7 km N. Cable Bay Rd on Maori Pa Rd.	tut2
K. "westlandica (north)"	02.BR.IRO.10	-41.786683	172.031	SH6, Iron Bridge, Buller R.	wen2
K. "westlandica (north)"	02.BR.IRO.14	-41.786683	172.031	SH6, Iron Bridge, Buller R.	wen1
K. "westlandica (south)"	02.BR.RUN.03	-42.412716	171.249083	SH6, 0.5 km S. of Runanga at memorial	wes1
K. "westlandica (south)"	02.NC.APV.01	-42.946566	171.563733	Railway Stn., Arthurs Pass Village	wes2
K. angusta	01.OL.INV.07	-44.731183	168.456033	Rees R. Vly Rd. at Invincible Mine Rd	ang2

K. angusta	98.MB.LSG.59	-42.136666	172.9125	Lake Sedgemere, Marlborough	ang1
K. cauta	94.WN.HAW.72	-41.325	174.73	Hawkins Hill, Wellington	cau1
K. cauta	94.WN.RIM.76	-41.1146	175.232066	SH2, Rimutaka Summit	cau2
K. convicta	98.NF.NFI.08	-29.033333	167.95	Norfolk Island	con1
K. cutora cumberi	02.TO.TPP.05	-38.690683	176.067533	Park in Taupo by Lake Taupo	cum1
K. cutora cumberi	94.WN.RIM.95	-41.1146	175.232066	Rimutaka Summit Trail	cum2
K. cutora cutora	93.AK.BUL.70			Bullock Track, Mahurangi West	cut2
K. cutora cutora	00.AK.HAT.01	-36.566666	174.695	Sun Vly. Rd. at Hatfield's Beach, Orewa	cut1
K. cutora exulis	98.KE.RAO.46	-29.248333	178.07	Raoul, Kermadec Islands	exu1
K. dugdalei	00.WN.DAY.02	-41.278333	174.916666	Days Bay, Wellington	dug2
K. dugdalei	02.BP.CRE.02	-38.154166	176.264	SH38 at Rotorua Crematorium and Cem.	dug1
K. horologium	93.MK.SEB.80	-42.25	172.883333	Mt. Sebastopol, Mt. Cook NP	hor2
K. horologium	97.MB.ENC.77	-41.58	173.256666	Above Enchanted Lookout, Lake Chalice	hor1
K. laneorum	02.TO.OPE.01	-38.768666	176.217766	SH5 at Opepe Historic Preserve	lan1
K. laneorum	02.TO.WWS.01	-38.894916	175.495983	16 km W. SH4/SH41 Jct., W. Waituhi Sdl.	lan2
K. longula	94.CH.CHA.67	-41.1067	172.6917	Chatham Islands	lon1
K. muta muta	01.WI.FER.03	-40.229883	175.571616	SH54, 0.7 km E. of Feilding Town Centre	mut1
K. ochrina	00.WN.DAY.01	-41.278333	174.916666	Days Bay, Wellington	och2
K. ochrina	94.WN.NEV.03	-41.301983	174.829216	164 Nevay Rd., Miramar, Wellington	och1
K. paxillulae	97.KA.PPR.81			Puhi-Puhi Res. at Hapuku R., Kaikoura	pax1
K. rosea	98.DN.BBY.53	-45.847166	170.624166	Broad Bay, Dunedin	ros1
K. scutellaris	94.WN.QEP.93			Queen Elizabeth Park, Paekakariki	scu2
K. scutellaris	97.TO.OPE.60	-38.768666	176.217766	SH5 at Opepe Historic Preserve	scu1
K. subalpina subalpina	01.TO.TSR.16	-39.296316	175.735416	Tukino Skifield Rd. ca. 0.3 km W. of SH1	sub2
K. subalpina subalpina	01.WN.RIM.01	-41.1146	175.232066	Rimutaka Summit Trail	sub1

* : not in the *Kikihia* Barcode-49 dataset

** : these specimens are identified as *K.* "astragali" in Marshall et al. (2008) and *K.* "nelsonensis" in Marshall et al. (in review)

Table 2. Specimen List of the *Kikihia* Barcode-49 dataset. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

Species	**Specimen code**	**Latitude**	**Longitude**	**Location**	**Tree label**
K. "aotea (eastern)"	01.WN.WNU.A	-41.249283	174.921166	Wainuiomata Hill track nr. Tawa Track	aoe1
K. "aotea (eastern)"	02.BP.WAR.01	-38.3049	177.3956	SH2, 7.0 km S. of Wairata	aoe2
K. "aotea (eastern)"	02.GB.NUH.01	-39.044166	177.73765		aoe3
K. "aotea (western)"	01.TK.ERS.01	-39.31255	174.146433	Pembroke Rd., 8 km W. Cardiff Rd.	aow1
K. "aotea (western)"	02.HB.SSA.01	-39.2147	176.6883	SH5 19.4 km NW Glengary Rd., Hawkes B.	aow2
K. "balaena"	02.KA.WBH.01	-42.495033	173.18275	SH70, 3.3 km N. Lyford Lodge, Kaikoura	bal1
K. "balaena"	02.KA.WBH.03	-42.495033	173.18275	SH70, 3.3 km N. Lyford Lodge, Kaikoura	bal2
K. "flemingi"	02.MB.PAT.04	-41.589533	173.29875	W. of L*K.* Chalice Lookout., Mt. Patriarch	fle2
K. "murihikua"	01.CO.CRA.01	-44.903683	168.984916	SH89, 5.2 km SW of Cardrona	mur2
K. "murihikua"	94.FD.RDS.01			SH94, Fiordland	mur1
K. "nelsonensis"	01.NN.WRR.01	-41.295	173.12	SH4, 3.9 km N. Wanganui River Rd.	nel2
K. "nelsonensis"	01.SD.CUL.02	-41.27375	173.787866	Up track S. of Cullens Pt. across SH6	nel1
K. "astragali"**	02.NN.KNH.01	-40.637583	172.5634	Knuckle Hill Summit, NW Nelson	ast2
K. "astragali"**	02.NN.KNH.03	-40.637583	172.5634	Knuckle Hill Summit, NW Nelson	ast1
K. "peninsularis"	98.MC.LEV.50			Rd. to Port Levy, Banks Peninsula	pen1
K. "rosea acoustica"	98.MK.LOH.63	-44.2373	169.82275	0.6 km S. of Lake Ohau lodge	aco1
K. "tasmani"	96.NN.SYL.10	-41.108333	172.633333	NW Nelson Trail, Lake Sylvester	tas1
K. "tuta"	01.NN.COL.03	-40.681016	172.670683	Jct. of SH60/Poplar Ln. nr. Collingwood	tut3*
K. "tuta"	02.NN.DEB.01	-41.1798	173.4294	1.7 km N. Cable Bay Rd on Maori Pa Rd.	tut2
K. "westlandica (north)"	02.BR.IRO.10	-41.786683	172.031	SH6, Iron Bridge, Buller R.	wen2
K. "westlandica (north)"	02.BR.IRO.14	-41.786683	172.031	SH6, Iron Bridge, Buller R.	wen1
K. "westlandica (south)"	02.BR.RUN.03	-42.412716	171.249083	SH6, 0.5 km S. of Runanga at memorial	wes1
K. "westlandica (south)"	02.NC.APV.01	-42.946566	171.563733	Railway Stn., Arthurs Pass Village	wes2
K. angusta	01.OL.INV.07	-44.731183	168.456033	Rees R. Vly Rd. at Invincible Mine Rd	ang2
K. angusta	98.MB.LSG.59	-42.136666	172.9125	Lake Sedgemere, Marlborough	ang1

K. cauta	94.WN.HAW.72	-41.325	174.73	Hawkins Hill, Wellington	cau1
K. cauta	94.WN.RIM.76	-41.1146	175.232066	SH2, Rimutaka Summit	cau2
K. convicta	98.NF.NFI.08	-29.033333	167.95	Norfolk Island	con1
K. cutora cumberi	02.TO.TPP.05	-38.690683	176.067533	Park in Taupo by Lake Taupo	cum1
K. cutora cumberi	94.WN.RIM.95	-41.1146	175.232066	Rimutaka Summit Trail	cum2
K. cutora cutora	93.AK.BUL.70			Bullock Track, Mahurangi West	cut2
K. cutora cutora	00.AK.HAT.01	-36.566666	174.695	Sun Vly. Rd. at Hatfield's Beach, Orewa	cut1
K. cutora exulis	98.KE.RAO.46	-29.248333	178.07	Raoul, Kermadec Islands	exu1
K. dugdalei	00.WN.DAY.02	-41.278333	174.916666	Days Bay, Wellington	dug2
K. dugdalei	02.BP.CRE.02	-38.154166	176.264	SH38 at Rotorua Crematorium and Cem.	dug1
K. horologium	93.MK.SEB.80	-42.25	172.883333	Mt. Sebastopol, Mt. Cook NP	hor2
K. horologium	97.MB.ENC.77	-41.58	173.256666	Above Enchanted Lookout, Lake Chalice	hor1
K. laneorum	02.TO.OPE.01	-38.768666	176.217766	SH5 at Opepe Historic Preserve	lan1
K. laneorum	02.TO.WWS.01	-38.894916	175.495983	16 km W. SH4/SH41 Jct., W. Waituhi Sdl.	lan2
K. longula	94.CH.CHA.67	-41.1067	172.6917	Chatham Islands	lon1
K. muta muta	01.WI.FER.03	-40.229883	175.571616	SH54, 0.7 km E. of Feilding Town Centre	mut1
K. ochrina	00.WN.DAY.01	-41.278333	174.916666	Days Bay, Wellington	och2
K. ochrina	94.WN.NEV.03	-41.301983	174.829216	164 Nevay Rd., Miramar, Wellington	och1
K. paxillulae	97.KA.PPR.81			Puhi-Puhi Res. at Hapuku R., Kaikoura	pax1
K. rosea	98.DN.BBY.53	-45.847166	170.624166	Broad Bay, Dunedin	ros1
K. scutellaris	94.WN.QEP.93			Queen Elizabeth Park, Paekakariki	scu2
K. scutellaris	97.TO.OPE.60	-38.768666	176.217766	SH5 at Opepe Historic Preserve	scu1
K. subalpina subalpina	01.TO.TSR.16	-39.296316	175.735416	Tukino Skifield Rd. ca. 0.3 km W. of SH1	sub2
K. subalpina subalpina	01.WN.RIM.01	-41.1146	175.232066	Rimutaka Summit Trail	sub1

* : not in the *Kikihia* Multigene-50 dataset

** : these specimens are identified as *K.* "astragali" in Marshall et al. (2008) and *K.* "nelsonensis" in Marshall et al. (in review)

Table 3. Specimen List of the Muta-barcode-149 dataset. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

Species	Specimen Code	Latitude	Longitude
K. "aotea (eastern)"	01.HB.ESK.01	-39.387916	176.82185
K. "aotea (eastern)"	01.WA.THR.04	-41.08075	175.365583
K. "aotea (eastern)"	01.WN.WNU.A	-41.249283	174.921166
K. "aotea (eastern)"	02.BP.NUK.01	-38.101583	177.139716
K. "aotea (eastern)"	02.BP.WAO.01	-37.7758	177.672166
K. "aotea (eastern)"	02.BP.WAR.01	-38.3049	177.3956
K. "aotea (eastern)"	02.BP.WTK.01	-38.2275	177.314633
K. "aotea (eastern)"	02.GB.NUH.01	-39.044166	177.73765
K. "aotea (eastern)"	02.GB.NWA.03	-38.8941	177.262383
K. "aotea (eastern)"	02.GB.WKH.01	-38.464583	177.730916
K. "aotea (eastern)"	02.HB.CAB.01	-40.407333	176.530866
K. "aotea (eastern)"	02.HB.GGR.02	-39.3501	176.736866
K. "aotea (eastern)"	02.HB.GGR.03	-39.3501	176.736866
K. "aotea (eastern)"	02.HB.HRD.01	-40.146483	176.5394
K. "aotea (eastern)"	02.HB.HRD.02	-40.146483	176.5394
K. "aotea (eastern)"	02.HB.NPU.03	-39.08215	177.017883
K. "aotea (eastern)"	02.HB.OCB.01	-39.742916	177.01065
K. "aotea (eastern)"	02.WA.CHP.01	-40.896066	176.218033
K. "aotea (eastern)"	02.WA.FLP.01	-41.253333	175.919166
K. "aotea (eastern)"	02.WA.HAM.01	-40.550666	175.749833
K. "aotea (eastern)"	02.WA.NGA.01	-40.754166	176.003016
K. "aotea (eastern)"	02.WA.SEK.01	-40.86545	175.6433
K. "aotea (eastern)"	02.WA.WEB.03	-40.417833	176.328066
K. "aotea (eastern)"	02.WN.ACC.01	-41.03295	174.896433
K. "aotea (eastern)"	03.BP.HAU.02	-37.597983	178.320983
K. "aotea (eastern)"	03.GB.ARA.02	-37.6346	178.368916
K. "aotea (eastern)"	03.GB.MAR.03	-38.838233	177.894883
K. "aotea (eastern)"	03.GB.SRU.01	-37.915683	178.271933
K. "aotea (eastern)"	03.GB.TBY.01	-38.318316	178.2718
K. "aotea (western)"	01.RI.VIN.09	-39.926266	175.627583
K. "aotea (western)"	01.TK.ERS.01	-39.31255	174.146433
K. "aotea (western)"	01.TK.ERS.02	-39.31255	174.146433
K. "aotea (western)"	01.TK.RWY.01	-39.058016	174.057966
K. "aotea (western)"	01.TK.RWY.03	-39.058016	174.057966
K. "aotea (western)"	01.TO.RCG.01	-39.191916	175.531683
K. "aotea (western)"	01.TO.WPF.01	-38.957166	176.523916
K. "aotea (western)"	02.BP.CAN.02	-37.4937	175.928183
K. "aotea (western)"	02.BP.ROT.01	-38.058916	176.643866
K. "aotea (western)"	02.BP.WTK.02	-38.2275	177.314633
K. "aotea (western)"	02.CL.TPU.03	-37.003866	175.50825
K. "aotea (western)"	02.CL.WAD.11	-36.8431	175.664133
K. "aotea (western)"	02.HB.SSA.01	-39.2147	176.688266
K. "aotea (western)"	02.ND.MIT.01	-36.1228	173.98985
K. "aotea (western)"	02.ND.MIT.02	-36.1228	173.98985
K. "aotea (western)"	02.TK.ORH.05	-38.847883	174.9336

K. "aotea (western)"	02.TO.KFP.03	-39.132533	175.823866
K. "aotea (western)"	02.TO.TAS.08	-38.6959	176.163066
K. "aotea (western)"	02.WO.PPA.01	-38.0243	175.76325
K. "aotea (western)"	03.BP.ETA.01	-37.709866	176.271416
K. "aotea (western)"	03.ND.KAT.01	-35.090233	173.2551
K. "aotea (western)"	03.RI.NGA.01	-39.400216	176.312383
K. "aotea (western)"	03.WO.AHU.02	-38.4037	175.3453
K. "aotea (western)"	03.WO.MAR.01	-38.389	175.12715
K. "aotea (western)"	93.TO.RVC.01	-39.205	175.545
K. "aotea (western)"	97.TK.ARA.01	-38.505833	175.203333
K. "nelsonensis"	01.MB.TWI.10	-41.337783	173.760616
K. "nelsonensis"	01.MB.TWI.A	-41.337783	173.760616
K. "nelsonensis"	01.NN.CLO.01	-41.286666	173.120833
K. "nelsonensis"	01.NN.COL.04	-40.681016	172.670683
K. "nelsonensis"	01.NN.WRR.01	-41.295	173.12
K. "nelsonensis"	01.SD.LIN.13	-41.291683	173.884233
K. "nelsonensis"	01.SD.LWS.01	-41.26385	173.8636
K. "nelsonensis"	01.SD.QCD.01	-41.288566	173.771216
K. "nelsonensis"	02.NN.ABC.02	-41.569633	172.688166
K. "nelsonensis"	02.NN.CBR.02	-41.031916	172.79755
K. "nelsonensis"	02.NN.CBR.03	-41.031916	172.79755
K. "nelsonensis"	02.NN.JDH.01	-41.255783	173.310616
K. "nelsonensis" **	02.NN.KNH.01	-40.637583	172.5634
K. "nelsonensis"	02.NN.KNH.02	-40.637583	172.5634
K. "nelsonensis" **	02.NN.KNH.03	-40.637583	172.5634
K. "nelsonensis"	02.NN.KPL.01	-40.619133	172.549166
K. "nelsonensis"	02.NN.KPL.02	-40.619133	172.549166
K. "nelsonensis"	02.NN.MIS.03	-40.5358	172.638283
K. "nelsonensis"	02.NN.MIS.04	-40.5358	172.638283
K. "nelsonensis"	02.NN.MOT.01	-41.092016	173.004383
K. "nelsonensis"	02.NN.MOT.02	-41.092016	173.004383
K. "nelsonensis"	02.SD.FRB.01	-40.993933	173.803783
K. "nelsonensis"	02.SD.FRE.01	-40.929283	173.8442
K. "nelsonensis"	03.NN.HHR.01	-41.021683	172.895016
K. "nelsonensis"	03.NN.PIK.01	-40.996983	172.889616
K. "north westlandica"	01.BR.HWC.01	-41.865	171.783333
K. "north westlandica"	01.BR.KIL.01	-41.86555	171.781883
K. "north westlandica"	01.BR.KIL.02	-41.86555	171.781883
K. "north westlandica"	01.BR.MUR.01	-41.786666	172.325
K. "north westlandica"	01.NN.WHR.01	-40.629566	172.50405
K. "north westlandica"	02.BR.IRO.14	-41.786683	172.031
K. "north westlandica"	02.BR.OCE.02	-42.022583	171.392466
K. "north westlandica"	02.BR.OCE.06	-42.022583	171.392466
K. "north westlandica"	02.BR.PKP.04	-42.109566	171.336933
K. "north westlandica"	02.BR.ROB.03	-41.83445	172.810516
K. "north westlandica"	02.BR.ROB.10	-41.83445	172.810516
K. "north westlandica"	02.BR.WES.02	-41.7733	171.592166
K. "north westlandica"	02.MB.SIX.01	-41.73525	173.0289

K. "north westlandica"	02.NN.FAR.01	-40.52345	172.741066
K. "north westlandica"	02.NN.FAR.02	-40.52345	172.741066
K. "south westlandica"	01.OL.HAC.A	-44.14595	169.322283
K. "south westlandica"	01.WD.FJV.01	-43.392083	170.180816
K. "south westlandica"	01.WD.FJV.A	-43.392083	170.180816
K. "south westlandica"	02.BR.CAR.01	-42.340933	171.572383
K. "south westlandica"	02.BR.HOP.01	-42.591066	172.447183
K. "south westlandica"	02.BR.HOP.03	-42.591066	172.447183
K. "south westlandica"	02.BR.MRV.04	-42.380333	172.314583
K. "south westlandica"	02.BR.RUN.03	-42.412716	171.249083
K. "south westlandica"	02.BR.RUN.06	-42.412716	171.249083
K. "south westlandica"	02.BR.RUN.09	-42.412716	171.249083
K. "south westlandica"	02.NC.APV.01	-42.946566	171.563733
K. "south westlandica"	02.WD.SOT.01	-42.74535	171.505616
K. "tasmani"	02.NN.COR.21	-41.107066	172.692066
K. "tasmani"	96.NN.SYL.10	-41.108333	172.633333
K. "tuta"	01.MB.TAP.01	-41.292133	173.669116
K. "tuta"	01.MB.TAP.02	-41.292133	173.669116
K. "tuta"	01.MB.TWI.03	-41.337783	173.760616
K. "tuta"	01.NN.COL.03	-40.681016	172.670683
K. "tuta"	02.NN.DEB.01	-41.1798	173.4294
K. "tuta"	02.NN.RAB.13	-41.282183	173.1594
K. "tuta"	02.NN.TTA.04	-40.549533	172.721566
K. "tuta"	02.SD.OKI.01	-41.292716	173.822216
K. "westlandica"	02.BR.IRO.10	-41.786683	172.031
K. muta muta	00.WN.NEV.01	-41.301983	174.829216
K. muta muta	01.KA.BDS.05	-42.275366	173.771183
K. muta muta	01.MC.BPT.02	-43.780016	172.788166
K. muta muta	01.MC.BPT.04	-43.780016	172.788166
K. muta muta	01.WI.FER.01	-40.229883	175.571616
K. muta muta	01.WI.FER.03	-40.229883	175.571616
K. muta muta	01.WI.MAR.01	-39.980833	175.13
K. muta muta	02.KA.SBL.01	-41.791866	174.14805
K. muta muta	02.KA.WBS.04	-42.486866	173.201816
K. muta muta	02.KA.WIL.01	-41.974583	174.041183
K. muta muta	02.MB.WAA.01	-41.441666	173.908333
K. muta muta	02.NC.NCH.03	-42.80645	173.274283
K. muta muta	02.RI.MWT.01	-40.335916	175.817516
K. muta muta	02.RI.MWT.02	-40.335916	175.817516
K. muta muta	02.SD.OPI.01	-41.297233	174.11575
K. muta muta	02.SD.OPI.02	-41.297233	174.11575
K. muta muta	02.WI.EPN.01	-40.305316	175.7301
K. muta muta	02.WN.AKS.01	-40.948566	175.108183
K. muta muta	02.WN.MAS.01	-40.723333	175.212533
K. muta muta	02.WN.MAS.02	-40.723333	175.212533
K. muta muta	03.CO.WKU.01	-44.7013	170.435116
K. muta muta	03.KA.OKI.01	-42.219716	173.858616
K. muta muta	03.KA.WKK.04	-42.379666	173.521916

K. muta muta	03.MB.HNR.01	-42.535116	172.8211
K. muta muta	03.MC.CBG.04	-43.533466	172.620533
K. muta muta	03.NC.JBB.01	-42.753333	173.079816
K. muta muta	03.ND.TAN.02	-34.629316	172.967266
K. muta muta	03.ND.TAN.03	-34.629316	172.967266
K. muta muta	03.SC.PFR.01	-43.902716	171.252983
K. muta muta	01.WI.PUT.02	-39.990366	175.5968
K. paxillulae	01.KA.BDS.04	-42.275366	173.771183

** : these specimens are identified as *K.* "astragali" in Marshall et al. (2008) and *K.* "nelsonensis" in Marshall et al. (in review)

Table 4. Specimen List of the *K.* "murihikua" and *K. angusta* Hybrid-69 dataset. See the caption of Figure 1 for the district codes designated by Crosby et al. (1998).

Species	**Specimen Code**	**Latitude**	**Longitude**
K. angusta	01.MK.LOH.17	-44.2373	169.82275
K. angusta	01.OL.INV.07	-44.731183	168.456033
K. angusta	01.OL.INV.08	-44.731183	168.456033
K. angusta	03.DN.WPC.01	-45.928183	170.027333
K. angusta	03.DN.WPC.02	-45.928183	170.027333
K. angusta	03.MB.JCK.03	-42.45945	172.837666
K. angusta	03.MB.JCK.04	-42.45945	172.837666
K. angusta	03.MK.RHS.03	-43.823983	170.659466
K. angusta	03.MK.STH.05	-44.46125	170.28605
K. angusta	03.OL.BOU.04	-44.353266	169.168333
K. angusta	03.SC.HPS.05	-44.33375	170.5871
K. angusta	03.SL.MNP.05	-45.56655	167.610683
K. angusta	04.CO.LIV.01	-44.773883	169.508583
K. angusta	04.CO.LIV.02	-44.773883	169.508583
K. angusta	04.MC.POT.02	-43.529666	170.8847
K. angusta	04.MC.POT.03	-43.529666	170.8847
K. angusta	05.CO.BBH.03	-45.155983	169.13155
K. angusta	05.CO.BBH.04	-45.155983	169.13155
K. angusta	05.DN.WAI.01	-45.232116	170.869316
K. angusta	05.DN.WAR.01	-45.715116	170.597283
K. angusta	05.DN.WAR.02	-45.715116	170.597283
K. angusta	05.NC.LTA.01	-42.75585	172.2201
K. angusta	05.NC.LTA.04	-42.75585	172.2201
K. angusta	05.OL.MPR.01	-45.3881	167.90635
K. angusta	05.OL.MPR.02	-45.3881	167.90635
K. angusta	05.SL.BTK.01	-46.248666	169.320183
K. angusta	05.SL.BTK.02	-46.248666	169.320183
K. angusta	06.CO.BLU.01	-44.874033	169.81385
K. angusta	06.CO.ERX.01	-45.53915	169.398333
K. angusta	06.CO.TEV.01	-45.5523	169.590583
K. angusta	06.SC.PFN.06	-43.883366	171.263316
K. angusta	06.SC.PFN.11	-43.883366	171.263316
K. angusta	06.SL.FRN.02	-45.80565	169.018833
K. angusta	07.SL.BEW.04	-45.964133	168.317716
K. angusta	07.SL.HAR.04	-46.471016	168.3821
K. angusta	07.SL.POP.02	-46.184933	169.365283
K. angusta	94.CO.OMR.45	-45.338766	169.2543
K. angusta	94.CO.OMR.46	-45.338766	169.2543
K. angusta	98.MB.LSG.59	-42.136666	172.9125
K. "murihikua"	01.CO.CRA.01	-44.903683	168.984916
K. "murihikua"	01.CO.CRA.71	-44.903683	168.984916
K. "murihikua"	03.SL.BLR.02	-46.600116	168.3401
K. "murihikua"	03.SL.BLR.03	-46.600116	168.3401
K. "murihikua"	03.SL.CLI.01	-46.030666	167.71535
K. "murihikua"	03.SL.LOG.01	-46.21275	167.99105

K. "murihikua"	03.SL.LOG.02	-46.21275	167.99105
K. "murihikua"	03.SL.MNP.01	-45.56655	167.610683
K. "murihikua"	03.SL.NOT.01	-46.62365	168.879316
K. "murihikua"	04.OL.LHA.01	-44.4988	169.2481
K. "murihikua"	04.OL.LHA.02	-44.4988	169.2481
K. "murihikua"	05.CO.BBH.01	-45.155983	169.13155
K. "murihikua"	05.CO.BBH.02	-45.155983	169.13155
K. "murihikua"	05.SL.CLA.03	-45.620283	167.95215
K. "murihikua"	05.SL.CLA.04	-45.620283	167.95215
K. "murihikua"	05.SL.CPR.03	-46.398133	169.458366
K. "murihikua"	05.SL.CPR.04	-46.398133	169.458366
K. "murihikua"	05.SL.PUR.03	-46.3461	169.437516
K. "murihikua"	05.SL.PUR.04	-46.3461	169.437516
K. "murihikua"	06.CO.WAK.01	-45.55365	169.02685
K. "murihikua"	06.CO.WAK.03	-45.55365	169.02685
K. "murihikua"	06.SL.FRN.01	-45.80565	169.018833
K. "murihikua"	07.OL.RAS.06	-44.512083	168.742116
K. "murihikua"	07.OL.WAT.07	-44.322383	169.1876
K. "murihikua"	07.SI.PAT.01	-46.9046	168.11835
K. "murihikua"	07.SI.PAT.03	-46.9046	168.11835
K. "murihikua"	07.SL.BEW.02	-45.964133	168.317716
K. "murihikua"	07.SL.HAR.01	-46.471016	168.3821
K. "murihikua"	07.SL.POP.01	-46.184933	169.365283
K. "murihikua"	07.SL.POP.06 ?	-46.184933	169.365283

Table 5. Parameters for the model GTR+I+ Γ for the various datasets.

Dataset	I value (invar)	alpha shape parameter	rate A <-> C	rate A <-> G	rate A <-> T	rate C <-> G	rate C <-> T	rate G <-> T
Muta-barcode-149	0.671641	1.375707	2.774653	18.376569	1.054509	0.511574	20.795387	1
Muta-multigene-149	0.590641	1.027335	4.026437	31.241643	1.91422	1.459461	32.085984	1
Random A	0.659201	0.955298	4.031776	26.113848	1.681096	0.405275	31.594218	1
Random B	0.67534	1.195223	3.671537	23.238083	1.074448	0.300246	25.481283	1
Random C	0.691963	1.526981	3.431647	21.049628	1.082108	0.707195	24.630754	1
Random D	0.700099	1.822291	2.882851	16.402955	0.782842	0.233978	19.903919	1
Random E	0.689635	1.398084	3.057085	18.945086	1.025265	0.655015	23.54055	1
Random F	0.657139	1.122786	3.106044	17.323167	0.953941	0.255871	22.433357	1
Random G	0.675468	1.273428	4.009819	24.363628	1.324332	0.357734	29.724794	1
Random H	0.693589	1.644011	3.297652	20.180964	1.052595	0.278766	23.22879	1
Random I	0.682292	1.402791	3.181789	17.814274	0.958636	0.284205	22.982783	1
Random J	0.68196	1.466134	3.099047	19.251538	0.923622	0.249657	20.777513	1
Hybrid-69	0.712901	1.835123	3.422027	20.423215	0.886512	0.000017	23.107079	1

Table 6. Results of the "Muta group" random sampling.

Sample (# parsim. inf. sites)	Species	Total # of Specimens	# Monophyletic Clades	# Specimens in Largest Clade	Successful ID with Barcode?
A (119)	*K.* "aotea east"	7	2	5	No
	K. "aotea west"	5	1	2	No
	K. "nelsonensis"	6	2	2	No
	K. muta	10	1	9	No
	K. "tuta"	1	----------	----------	----------
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	3	1	3	Yes
	K. "westlandica south"	2	1	2	Yes
	K. "tasmani"	1	----------	----------	----------
B (117)	*K.* "aotea east"	7	2	3	No
	K. "aotea west"	8	2	6	No
	K. "nelsonensis"	5	1	4	No
	K. muta	6	1	4	No
	K. "tuta"	----------	----------	----------	----------
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	7	1	6	No
	K. "westlandica south"	2	1	2	Yes
	K. "tasmani"	----------	----------	----------	----------
C (118)	*K.* "aotea east"	6	2	3	No
	K. "aotea west"	9	2	2	No
	K. "nelsonensis"	7	2	4	No
	K. muta	7	1	5	No
	K. "tuta"	3	0	----------	No
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	2	1	2	Yes
	K. "westlandica south"	1	----------	----------	----------
	K. "tasmani"	----------	----------	----------	----------
D (118)	*K.* "aotea east"	7	2	5	No
	K. "aotea west"	5	1	5	Yes
	K. "nelsonensis"	6	1	3	No
	K. muta	7	1	6	No
	K. "tuta"	2	0	----------	No
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	5	1	4	No
	K. "westlandica south"	3	1	3	Yes
	K. "tasmani"	----------	----------	----------	----------
	K. "aotea east"	5	2	2	No
	K. "aotea west"	4	1	4	Yes

E (117)	*K.* "nelsonensis"	6	2	4	No
	K. muta	10	1	9	No
	K. "tuta"	4	0	----------	No
	K. paxillulae	2	0	----------	No
	K. "westlandica north"	2	1	2	yes?
	K. "westlandica south"	2	1	2	Yes
	K. "tasmani"	1	----------	----------	----------
F (119)	*K.* "aotea east"	6	2	4	No
	K. "aotea west"	3	1	3	Yes
	K. "nelsonensis"	7	1	5	No
	K. muta	8	2	6	No
	K. "tuta"	1	----------	----------	----------
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	8	1	8	Yes
	K. "westlandica south"	2	1	2	Yes
	K. "tasmani"	----------	----------	----------	----------
G (120)	*K.* "aotea east"	9	3	5	No
	K. "aotea west"	7	1	7	Yes
	K. "nelsonensis"	7	2	3	No
	K. muta	7	1	7	Yes
	K. "tuta"	2	1	2	Yes
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	2	1	2	Yes
	K. "westlandica south"	1	----------	----------	----------
	K. "tasmani"	----------	----------	----------	----------
H (117)	*K.* "aotea east"	6	2	3	No
	K. "aotea west"	5	1	5	Yes
	K. "nelsonensis"	4	0	----------	No
	K. muta	10	2	7	No
	K. "tuta"	3	0	----------	No
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	4	1	3	No
	K. "westlandica south"	1	----------	----------	----------
	K. "tasmani"	2	0	----------	No
I (115)	*K.* "aotea east"	4	1	3	No
	K. "aotea west"	4	1	4	Yes
	K. "nelsonensis"	8	2	4	No
	K. muta	13	2	11	No
	K. "tuta"	2	0	----------	No
	K. paxillulae	1	----------	----------	----------
	K. "westlandica north"	3	1*	3*	yes*

	K. "westlandica south"	----------	----------	----------	----------
	K. "tasmani"	1	----------	----------	----------
J (125)	*K.* “aotea east”	12	3	6	No
	K. "aotea west"	3	1	3	Yes
	K. "nelsonensis"	4	1*	3*	No
	K. muta	5	1	5	Yes
	K. "tuta"	1	----------	----------	----------
	K. paxillulae	2	0	----------	No
	K. "westlandica north"	7	1	7	Yes
	K. "westlandica south"	2	1	2	Yes
	K. "tasmani"	----------	----------	----------	----------

* : not well supported (ML bootstrap value less than 50)

Table 7. Muta group clades supported by greater than 50% ML bootstrap in the random samples. Letters (A1, A2, A3, B1, B2, C, D, E, F, L, and N) are assigned to clades in Figure 6.

Random Sample	A1	A2	A3	B1	B2	C	D	E	F	L	N	total absent
A	P	P	P	P	A	P	P	W	P	P	P	1
B	P	W	P	P	W	A	P	P	P	P	P	1
C	P	P	P	P	P	P	P	P	P	P	P	0
D	P	P	P	P	A	W	P	P	P	P	P	1
E	P	P	W	P	P	A	P	P	P	P	P	1
F	P	P	P	P	A	P	P	P	A	P	P	2
G	P	P	P	P	P	P	P	P	P	P	P	0
H	P	P	P	P	W	P	W	P	P	P	P	0
I	P	P	P	P	P	A	P	P	P	P	A	2
J	P	N	P	P	P	W	P	P	P	P	P	1

P : present
A : absent
W : weak support
N : not monophyletic

Appendix I

New Zealand biogeography: The Oligocene Drowning of New Zealand

The scientific community agrees that there was a sizable decrease in the land area of New Zealand above water due to a marine transgression or "drowning" in the late Oligocene (ca. 26 Ma) (Knapp et al. 2007). The only question is whether there was complete or incomplete submergence of the landmass. Various authors have approached the question in a number of different ways from analysis of New Zealand's geology or fossil record to molecular studies of the plants and animals that could have survived the drowning or dispersed from elsewhere afterwards.

1 Flora

1.1 Kauri (Agathis australis)

A large portion of the debate over the extent of the submergence of New Zealand during the Oligocene surrounds one plant species, New Zealand's *Agathis australis*, also known as the Kauri. Support for a vicariant origin (and therefore incomplete submergence of New Zealand) is found in Lambert et al.'s (1993) study on resin. The modern *Agathis australis* resin is very similar to resin found in New Zealand from the Eocene, Oligocene, and Miocene (Knapp et al. 2007). Stöckler et al. (2002) cite a fossil of an extinct species of *Agathis* from the late Early or Late Cretaceous (113-65 Ma) that was found in the Clarence Valley of New Zealand as proof of the continuous presence of the plant on the landmass (Parrish et al. 1998). Although it is a different species, the morphology of the plant was found to be more similar to *Agathis australis* than any other extant species of *Agathis.*

Stöckler et al. (2002) believe that this fossil supports an uninterrupted lineage on New Zealand since the separation of Gondwana. Waters and Craw (2006) feel that this fossil is completely irrelevant to the debate on the extent of the Oligocene drowning because it dates to a time before New Zealand separated from the rest of Gondwana. Therefore, the fossil cannot indicate whether the range of the species was limited solely to New Zealand and cannot suggest a constant lineage on the island any more than it can suggest dispersal from another Gondwanan landmass (Waters and Craw 2006).

The findings of Knapp et al. (2007) suggest that the line of *Agathis australis* diverged from other living species of *Agathis* during the Eocene (54-33 Ma) which was around the time that New Zealand separated from other landmasses and significantly after the separation of New Zealand from Gondwana (80 Ma). This indicates that *Agathis* survived the Oligocene drowning of New Zealand and is evidence against complete submergence. The New Zealand species (*Agathis australis*) diverged very early from the Australian species and another geographically close species (*Agathis macrophylla*) which is located in Tropical Australasia from the Solomon Islands to Fiji. The distant genetic relationship between the New Zealand Kauri and other nearby *Agathis* species makes dispersal an unlikely explanation for the plant's current presence on New Zealand (Knapp et al. 2007). The molecular analysis of Stöckler et al. (2002) also supports a continuous presence of the Kauri as well as an incomplete Oligocene drowning. Their results showed that *Agathis australis* was always the earliest species to diverge from the rest of the genus and therefore the most genetically different from the other species (Stöckler et al. 2002).

Knapp et al. (2007) recognize that this is not the only possibility that their results provide for the *Agathis* lineage. They acknowledge that the date of divergence could result

from the origination of the species *Agathis australis* in Australia, dispersal to New Zealand, and the subsequent extinction of the Australian progenitor. This alternative hypothesis may be supported by the fossil record due to the presence of species from a closely related genus (*Araucaria*) that have gone extinct in Australia. The exact relation of these species to *Agathis australis* is still unknown so further study must be done to determine the validity of the alternative hypothesis (Knapp et al. 2007).

1.2 Southern beeches (Nothofagus)

The southern beeches (*Nothofagus*) of New Zealand are often cited as an ideal example of Gondwanan distribution that would support the incomplete submergence hypothesis (Waters and Craw 2006). Fossil leaves and fruit were found and compared to New Zealand's current fauna to see if there was any resemblance that would suggest an uninterrupted presence on the islands. Specimens resembling *Nothofagus* were found in the fossils, suggesting that this genus survived the Oligocene drowning or that *Nothofagus* was present before the drowning, went extinct, and dispersed back to New Zealand after the land re-emerged. However, these fossils seem much more similar to the *Nothofagus* forests of central eastern Australia; especially since the other fossils resemble other forests in eastern Australia (Pole 1994). The similarity of the fossils to the forest beech trees of Australia would strongly suggest that dispersal could best explain the presence of today's southern beeches on New Zealand.

Molecular analysis has rejected vicariance as an explanation for the presence of the southern beeches on New Zealand. There are two subgenera (*Fuscospora* and *Lophozonia*) that independently colonized New Zealand, thus the New Zealand species of *Nothofagus* do

not form a monophyletic group. The research provided evidence for long-distance dispersal from Australia around 30 Ma (during the Oligocene) (Knapp et al. 2005).

Dispersal is also strongly supported for some species by palynology. Some types of *Nothofagus* pollen were discovered to have appeared in Australia before also being found in New Zealand. Only three species of *Nothofagus* are potential candidates for a vicariant origin based on the pollen study (Pole 1994). With the exception of these three species, *Nothofagus* does not provide strong support for the incomplete submergence hypothesis.

2 Fauna

2.1 Reptiles

New Zealand is home to a vast number of *Oligosoma* skinks. Hickson et al. (2000) tried to determine the time at which the skink population diversified from nucleotide substitution rates of mitochondrial 12rRNA. Their data suggest that *Oligosoma* skinks' diversification probably began around 23 Ma, during the Oligocene or early Miocene. Hickson et al. (2000) do not question the presence of islands remaining above water during the Oligocene drowning. Furthermore, they feel that an island model of speciation would support the diversification patterns and timeline their research suggested for the New Zealand skinks. However, if the nucleotide substitution rate of skinks is significantly slower than the calibration rate (based on bovids and ratite birds) used in their study, it is possible that the skinks diversified while New Zealand was still connected to Gondwana (Hickson et al. 2000).

Hickson et al. (2000) believe that they can rule out a more recent time of diversification for the New Zealand skinks based on the lack of support in their data for a more rapid rate of nucleotide substitution as well as the excessive genetic diversity of the skinks. Based on their genetic evidence they suggest that the most likely hypothesis is that skinks dispersed to New Zealand prior to the Oligocene drowning but after the breakup of Gondwana. Their reasoning is that the poor dispersal ability of skinks would require some terrestrial assistance which could have been provided by an island chain along the Lord Howe or Norfolk rise which, except for a small group of islands surrounding the current remnant of Lord Howe island, has been submerged since the Oligocene or early Miocene. Although they claim that skinks are poor dispersers, they mention other island localities where the presence of skinks must be explained by dispersal. Also, sequence data suggests that a lineage of New Zealand skinks has recently diverged from a New Caledonian species which could imply a more recent dispersal of some skinks to the islands (Hickson et al. 2000). A genetic comparison of New Zealand skinks with Australian skinks would be required to get a better idea of the date of separation of New Zealand species. Smith et al. (2007) performed a study on skinks that included specimens from New Zealand, Lord Howe Island, New Caledonia, New Guinea, and Australia. Their results suggest that dispersal was responsible for the spread of skinks to New Caledonia, Lord Howe Island, and New Zealand. They estimate that the time of divergence for the New Zealand, New Caledonia, and Lord Howe Island skinks from other skinks is between 12.7 and 40.7 Ma (during the Eocene, Oligocene, or Miocene) (Smith et al. 2007). This work strongly supports Waters and Craw's (2006) statement that dispersal should not be dismissed as an explanation for the origin of New Zealand skinks.

The tuatara are widely accepted as having Gondwanan origins and are often cited as evidence for an incomplete submergence of New Zealand during the Oligocene drowning (Worthy et al. 2006). Although this reptile is only found in New Zealand, Waters and Craw feel that it provides no support for either side of the debate because the genetic divergence date is significantly earlier than the New Zealand's separation from Gondwana. The tuatara could have been isolated on another landmass and dispersed to New Zealand at any time before or after the Oligocene drowning. The presence of fossils of the now extinct ancestors of tuatara in many different places also makes this lineage completely irrelevant in the debate over complete or incomplete submergence (Waters and Craw 2006). Waters and Craw are too quick to write off one of the more widely accepted pieces of evidence for an incomplete submergence. The presence of the tuatara solely on New Zealand and an early divergence date would strongly suggest Gondwanan distribution as the most likely hypothesis to explain their current distribution.

2.2 Mammals

Until recently, the only terrestrial mammals in New Zealand were three species of bats that supposedly dispersed to the islands during the mid or late Cenozoic (65-0 Ma). Therefore, they probably arrived after the Oligocene drowning and there were no mammal groups available to provide evidence for or against complete submergence. Landis et al. (2008) commented that "an absence of mammals makes the biota of New Zealand more similar to that of emergent oceanic islands than a continental landmass." This observation implies that all extant organisms arrived via dispersal and which requires a complete submergence of New Zealand during the Oligocene drowning. The recent discovery of an

extinct species of mammal (Worthy et al. 2006) may have a significant impact on the debate over complete or incomplete submergence of the islands.

Three fossilized bones (two mandibular fragments and one femoral fragment) have been found on the South Island of New Zealand dating from 19 to 16 Ma. These bones come from at least one species of nonflying terrestrial mouse-sized mammal. It is also possible that the femur's original owner is an older animal and the mandibles are from a much more recent mammal. This is the first known mammal that lived after the Oligocene drowning. Worthy et al. (2006) believe that this group of terrestrial mammals lived in New Zealand since the separation from Gondwana. They give an estimate of the length of the mammal's presence on New Zealand: from the Late Cretaceous to the Early Miocene. The vicariant origin of this mammal would provide evidence for incomplete submergence during the Oligocene drowning. Other fossils found with the mammal should also be analyzed to see whether other Gondwanan taxa survived the Oligocene drowning as well. This interpretation of their findings would mean that the fauna of New Zealand somehow changed significantly since the late early Miocene (Worthy et al. 2006).

Worthy et al. (2006) argue against the alternative hypothesis of post-Oligocene dispersal to New Zealand from Australia (followed by extinction in Australia). They provide three arguments against this hypothesis, two of which have to do with the poor dispersal abilities of the mammal. First of all, the shape of the femur does not lend itself to swimming and would certainly not have belonged to a mammal capable of flying. Also, Australia has a rich fossil record including a plethora of terrestrial mammals. In all of Australia's fossils, no mammals resembling this recently discovered terrestrial mammal have been found. Although it is possible that there are fossils of this mammal somewhere in Australia, the

extensive collection of known mammal fossils makes it seem unlikely that this species would have completely escaped notice. Furthermore, the distance and lack of these terrestrial mammals on any nearby landmasses makes dispersal an unlikely candidate for their origin. This reasoning strongly suggests Gondwanan origins for the terrestrial mammal which requires some land to remain above water during the Oligocene drowning of New Zealand (Worthy et al. 2006).

2.3 Birds

Cooper and Cooper (1995) analyzed DNA sequences of kiwi, moa, and wren groups to determine when the species diversified. Mitochondrial DNA analysis allowed them to determine that each group diversified from one or a few closely related maternal lines. The data also suggests that the ratite groups diversified in the early Miocene approximately 19-24 Ma and could also include the early Oligocene to the late Miocene. These findings would strongly support a bottleneck in the Oligocene. The ratites support incomplete submergence during the Oligocene drowning because there is evidence of a significant decrease in diversity within each lineage before the radiations occur. An alternative hypothesis that could produce the same results is the independent dispersal of all three groups to New Zealand during the early Miocene. However, the authors feel that a vicariant origin of the ratites is a much more plausible explanation (Cooper and Cooper 1995).

The fossil record provides more straightforward results than the molecular analyses. Other ratite fossils have been found in South America and Antarctica dating from the Paleocene and Eocene. These fossils as well as molecular studies support a flightless vicariant origin for ratites (Paton et al. 2002). Another problem with the dispersal hypothesis

is the lack of nearby relatives that would support dispersal to New Zealand during the Miocene (Cooper and Cooper 1995).

Waters and Craw (2006) once again look upon the supposed Gondwanan origins of the ratites with suspicion. Since the 1995 Cooper and Cooper paper was published, new data lends strong support to the hypothesis that the moa and kiwi groups came to New Zealand through separate dispersal events (Cooper et al. 2001). When using a molecular clock that was calibrated assuming that moas are Gondwanan (82 Ma), DNA analyses of the kiwi showed that it arrived to New Zealand 68 Ma- much more recently than previously assumed. They argue that the moa could have also arrived since the split from Gondwana since the kiwi arrived more recently. If the date of the moa's arrival to New Zealand was changed, it could result in an even more recent arrival of the kiwi since the 68 Ma date was based on the Gondwanan distribution of the moa (Waters and Craw 2006).

3 Geology

The geology of New Zealand during the Oligocene has been described as a low relief region of coastal plains with a maximum altitude of a few hundred meters. There were many rivers across the plains of New Zealand which were subject to cycles of flooding by the sea. Fluctuations of the sea level during the Cenozoic had a significant impact on the extent of flooding because the plains were not very high above sea level. Maps have been created from suggestions for the extent of the flooding that are based on the distribution of sedimentary rock and the results of the Cretaceous Cenozoic Project of the Institute of Geological and Nuclear Sciences (Cooper and Cooper 1995).

One of the most recent additions to the literature of the debate is Landis et al.'s 2008 paper on the Waipounamu Erosion Surface. They argue that there is not enough evidence to support the continuous presence of land during the Oligocene so there is no reason to doubt complete submergence. They claim that the main evidence used to support incomplete submergence during the Oligocene drowning is "(1) the nature and diversity of the *modern* New Zealand flora and fauna, (2) the fossil record and (3) the absence today of middle Cenozoic marine sedimentary rocks from inland portions of North and South island as well as from central Fiordland and Steward Island" (Landis et al. 2008). They argue that sediments from the Cenozoic period support a significant if not complete drowning during the late Oligocene and very early Miocene. One particular geologic feature on which they focus their attention is the Waipounamu Erosion Surface. This erosion surface was previously classified as a peneplain, which is formed due to erosion from being exposed to air. Landis et al. (2008) have found that this erosion surface was actually formed from coastal and shallow marine erosion which supports submergence during the Oligocene. It is possible that other erosion surfaces, like those in Canterbury and Otago, are also due to marine erosion and are parts of the Waipounamu Erosion Surface. This would result in even less land being above water than previously hypothesized. The authors do not provide any reasoning for complete submergence, but they feel it is just as arbitrary to assume incomplete submergence as it is to assume complete submergence (Landis et al. 2008).

4 Discussion

The only New Zealand taxa reviewed in this paper are those that have potential for Gondwanan distribution. The fossil record and genetic analysis of the New Zealand Kauri

mostly support the incomplete submergence hypothesis. However, the southern beeches have all arrived in New Zealand via dispersal (perhaps with the exception of three species). The current analysis of skinks suggests dispersal, but other studies still favor a vicariant origin for this taxon. The tuatara definitely has Gondwanan origins; the only question is whether this reptile dispersed from a nearby landmass or if it remained on New Zealand. The one species of mammal discovered supports incomplete submergence. The ratites were thought to have vicariant origins, although recent studies are starting to question these findings. The small number of taxa that support an incomplete submergence of New Zealand could be explained by the fact that New Zealand was reduced to a few small islands, suggesting that only a few lineages would be able to survive given the limited resources and ecological niches left above water (Knapp et al. 2007). Overall, the evidence presented in this paper suggests an incomplete submergence of New Zealand during the Oligocene drowning.

Many of the people studying and writing about the Oligocene drowning are biased towards either complete or incomplete submergence. Often, the authors do not provide evidence disproving alternate hypotheses or completely ignore them. Other authors cite poor dispersal for a taxon as evidence for a vicariant origin although it has been repeatedly shown that animals previously thought to be incapable of dispersing have done so. Some authors acknowledge the evidence for dispersal within their own papers and still cite poor dispersal ability as proof for vicariance.

The authors are sometimes overwhelmed by the futility of their efforts. There will always be debate over whether New Zealand was completely submerged or whether part of the landmass remained above water. Landis et al. (2008) are frustrated by the fact that they

can never disprove the continuous existence of some land during the Oligocene drowning. Similarly, Knapp et al. (2007) accept the impossibility of disproving dispersal. Although they argue for a vicariant origin, they know that they can never prove it beyond doubt. We know that many taxa dispersed to New Zealand after the Oligocene (San Martin & Ronquist 2004), by this reasoning, it is likely that many taxa dispersed to New Zealand prior to the Oligocene but after New Zealand split from Gondwana. So even pre-Oligocene taxa are not evidence for vicariance. The debate will continue because the evidence does not clearly support one side over the other. There is no doubt that there was a marine transgression in New Zealand during the Oligocene. However, the extent of the submergence is still unknown. More research is needed and the Southern Hemisphere Tricoptera that are the subject of my honors thesis are good candidates for providing relevant information.

References

Cooper, A., C. Lalueza-Fiox, S. Anderson, A. Rambaut, J. Austin, and R. Ward. "Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution." Nature 409: 704-707 (2001).

Cooper, Alan and Roger A. Cooper. "The Oligocene Bottleneck and New Zealand Biota: Genetic Record of a past Environmental Crisis." Proceedings: Biological Sciences 261(1362): 293-302 (1995).

Hickson, Robert E., Kerryn E. Slack, and Peter Lockhart. "Phylogeny recapitulates geography, or why New Zealand has so many species of skinks." Biological Journal of the Linnean Society 70: 415-433 (2000).

Knapp, M., *K.* Stöckler, D. Havell, F. Delsuc, F. Sebastiani, and P. J. Lockhart. "Relaxed molecular clock provides evidence for long-distance dispersal of *Nothofagus* (southern beech)." PLoS Biology 3: 38-43 (2005).

Knapp, Michael, Ragini Mudaliar, David Havell, Steven J. Wagstaff, and Peter J. Lockhart. "The Drowning of New Zealand and the Problem of *Agathis*." Systematic Biology 56(5): 862-870 (2007).

Lambert, J. B., S. C. Johnson, G. O. Poinar, and J. S. Frye. "Recent and fossil resins from New Zealand and Australia." Geoarchaeology 8: 141-155 (1993).

Landis, C. A., H. J. Campbell, J. G. Begg, D. C. Mildenhall, A. M. Paterson, and S. A. Trewic*K.* "The Waipounamu Erosion Surface: questioning the antiquity of the New Zealand land surface and terrestrial fauna and flora." Geological Magazine 145(2): 1-25 (2008).

Parrish, Judith Totman, Ian L. Daniel, Elizabeth M. Kennedy, and Robert A. Spicer. "Paleoclimatic significance of mid-Cretaceous floras from the middle Clarence Valley, New Zealand." Palaios 13: 149-159 (1998).

Paton, Tara, Oliver Haddrath, and Allan J. Baker. "Complete mitochondrial DNA genome sequences show that modern birds are not descended from transitional shorebirds." Proceedings of the Royal Society 269(1493): 839-846 (2002).

Pole, Mike. "The New Zealand Flora- Entirely Long-Distance Dispersal?" Journal of Biogeography 21(6): 625-635 (1994).

Sanmartin, Isabel and Fredrick Ronquist. "Southern Hemisphere Biogeography Inferred by Even-Based Models: Plant versus Animal Patterns." Systematic Biology 53(2): 216-243 (2004).

Smith, Sarah A., Ross A. Sadlier, Aaron M. Bauer, Christopher C. Austin, and Todd Jackman. "Molecular phylogeny of the scincid lizards of New Caledonia and adjacent areas: Evidence for a single origin of the endemic skinks of Tasmantis." Molecular Phylogenetics and Evolution 43: 1151-1166 (2007).

Stöckler, Karen, Ian L. Daniel, and Peter J. Lockhart. "New Zealand Kauri (*Agathis australis* (D.Don) Lindl., Araucariaceae) Survives Oligocene Drowning." Systematic Biology 51(5): 827-832 (2002).

Waters, Jonathan M. and Dave Craw. "Goodbye Gondwana? New Zealand Biogeography, Geology, and the Problem of Circularity." Systematic Biology 55(2): 351-356 (2006).

Worthy, Trevor H., Alan J. D. Tennyson, Michael Archer, Anne M. Musser, Suzanne J. Hand, Craig Jones, Barry J. Douglas, James A. McNamara, and Robin M. D. Bec*K*. "Miocene mammal reveals a Mesozoic ghost lineage on insular New Zealand,

southwest Pacific." Proceedings of the National Academy of Sciences of the United States of America 103(51): 19419-19423 (2006).