# IMPROVING SPECIMEN IDENTIFICATION: INFORMATIVE DNA USING A STATISTICAL BAYESIAN METHOD

# IMPROVING SPECIMEN IDENTIFICATION: INFORMATIVE DNA USING A STATISTICAL BAYESIAN METHOD

BY MELANIE LOU, B.Sc. Hons., M.Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Doctor of Philsophy

McMaster University

DOCTOR OF PHILOSOPHY (2012) McMaster University
(Biology) Hamilton, Ontario

TITLE: Improving specimen identification: Informative DNA using a statistical Bayesian method

AUTHOR: Melanie Lou, B.Sc. Hons. (York University), M.Sc. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: [xi], 99

## ABSTRACT

This thesis investigates the effect of different levels of information content in sequence data from the mitochondrial gene, cytochrome c oxidase, subunit I (COI), through the observed number of segregating sites, and a new Bayesian statistical method on the accuracy of specimen assignments or delimitations to its species of origin. There are four major parts of this thesis.

The first chapter examines the taxonomic accuracy of GenBank sequences for specimen delimitation. In addition to sequence accuracy (i.e., errors), it is a potential error that may influence accurate specimen diagnoses. Using 5,179 barcode sequences from 590 species and 8,586 GenBank (non-barcode) sequences from 2,900 species, across twelve insect orders, we compared the performance of specimen assignment between NCBI sequences labelled with and without the "BARCODE" keyword (indicates both sequence accuracy and taxonomy are supported) respectively. We expect the two groups of sequences should have identical proportions of unusually divergent (within-species sequences that may potentially be different species) and unusually similar (between-species sequences that may potentially be the same species) sequences. In contrast, non-barcode records had a high proportion of unusually divergent sequences, suggesting an error in sequence accuracy or taxonomy or both. However, there was no evidence of unusually similar sequences, suggesting the correct sequence accuracy or taxonomy or the 3% divergence threshold used for delimition in this group may be inappropriate. This study highlights that caution and a firm understanding of the data should be exercised when using GenBank data for species diagnoses.

The second chapter examines the assignment performance of a Bayesian statistical assignment method, based on segregating sites, in sequence data that lack a clear "barcoding gap" (a region defined by the maximum intraspecific distance and minimum interspecific distance). To our knowledge, it is the first tree-less statistical approach that makes use of segregating sites for species assignments and is also very fast (10,000 simulations in roughly 3 seconds with a 1.6GHz processor running Linux). Sequences from the genus *Drosophila* were used because its taxonomy is well supported but some pairs of sibling species lack a barcoding gap. Using 616 *Drosophila* CO1 sequences from 19 species and simulated sequences, the method performed well in the absence of a barcoding gap. And only when the degree of incomplete lineage sorting (despite species divergence, a lineage from one species may group more closely with a lineage of a less related species) is high does the method falter, but even then the probability of assigning the unknown to its species of origin is still high relative to a less closely related species.

The third chapter focuses on the information content and sampling of reference sequences from a geographically widespread species with migration. The assignment of an unknown specimen to a species that is sufficiently sampled (adequate representation of intraspecific variation) should improve. Using tiger moth (genus *Grammia*) 179 sequences from 13 species and simulated sequences, the addition of at least one reference sequence from a different deme or region of a species distribution returns a greater proportion of results that correctly assign an unknown specimen to its species of origin (e.g., inclusion of one dispersed, simulated, sequence resulted in an 18% increase, from 26% to 44%, in correct assignments). Thus correct delimition depends on adequate representation of conspecific (within-species) variation, particularly with species characterized by population subdivision and gene flow, highlighting the importance of proper sampling protocols to construct complete reference libraries.

Often there is a disconnect between the assumptions made by a model and the true evolutionary signals of the data it is applied to. The final chapter seeks to improve the segregating sites algorithm by incorporating terms to describe the role that other biological phenomena, namely population subdivision, gene flow, and unequal base composition from transition bias may have in shaping genetic diversity. A more comprehensive model should improve estimates of a population genetic parameter, $\theta$, used to measure the level of variation. The modified probability distributions (of observing a number of segregating sites in a number of sequences) are similar but more accurate at resolving the true distribution of genetic variability relative to those calculated under the original theory. The results reinforces that subdivided populations with migration and heterogeneous base composition and substitution rates for transitions and transversions shape patterns of variability and should be considered in models used to describe genetic signals of groups undergoing speciation.

## ACKNOWLEDGEMENTS

I thank my parents, Young Chang and Laura Lou, and my sister, Angela Lou, for their support and patience.

N. Kirischian for her continual friendship and advice.

A. Reyes for his fortitude and love.

My supervisor, Dr. G. B. Golding, for his academic guidance, support, and wisdom.

Friends, both living locally and abroad, who have supported, challenged or joined me on this journey. In particular, Drs. M. Abou-Chakra, A. Ahuja, and M. Huntley.

Past and present members of the Golding and Evans lab for continual support and companionship.

My committee members, Drs. J. Stone and R. Morton, for their constructive criticism.

Gratitude goes to B. Reuter and A. Tracey for handling logistics and teaching assistantships, respectively.

And I appreciate direct and indirect funds that have supported my studies, namely NSERC, Genome Canada, Barcode of Life and McMaster University.

## DECLARATION OF ACADEMIC ACHIEVEMENT

Each chapter of this thesis has been written as a separate manuscript. Data collection, programming, analysis and manuscript preparation for each chapter was primarily an individual effort, with contributions in data preparation and editing from G. Brian Golding.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# Part I

# INTRODUCTION

Species diversity is the main currency of nearly all disciplines of biology. Thus, a good foundation and understanding of species diversity begins with accurate species identifications. DNA barcoding is an initiative for species identification that is based on the surveillance of sequence diversity in a 648 bp region of the mitochondrial gene coding for cytochrome c oxidase, subunit I (COI), a gene that plays an essential role in energy production (Capaldi, 1990; Tsukihara *et al.*, 1996). In this approach, the DNA barcode of an unknown sample or specimen is screened against a reference sequence library and a species assignment is made when the query sequence can be assigned to a species in the reference library. The improvement of sequencing technologies has resulted in an unprecedented amount of genetic information and to fully 'analyze' the data takes a larger amount of time due to a variety of factors that include insufficient models and methods, fewer taxonomic specialists, sequencing error, and more complex data sets (e.g., of poorly known, underrepresented groups, and degraded DNA). Since the success of biological research, conservation, forensic, bio-security, economic and consumer policy efforts depend on correct species identification, to avoid analysis gridlock, attention must be directed to overcoming challenges that hamper accurate species identifications. In previous studies, the effect of some factors on barcoding accuracy, such as the number (one or more) and type (mitochondrial or nuclear) of markers and sampling size (samples per species), have been investigated. A comparision of the reliability of existing GenBank data, relative to taxonomist-verified barcode data, for species identification had not been investigated. And it is often not the number of loci or sequences that is important but the number of informative sites (Simon *et al.*, 2006). There are few, if any, studies that use the pattern of segregating sites as a measure to delimit species. The theory of segregating sites stems from a combination of theories by Kimura (1969) and Watterson (1975) stating that new mutations can only occur at sites not previously mutated (that is, no two mutations ever occur at the same site), given an infinite number of sites, and they are also not subject to recombination, respectively. There are many species identification approaches (and new ones being developed) and performance among them have been explored (Ross, Murugan and Li, 2008; Austerlitz *et al.*, 2009; Little, 2011; Parks, MacDonald and Beiko, 2011; Zhang *et al.*, 2011). The following are descriptions of several categories of methods. Molecular operational taxonomic units or (MOTUs) and evolutionary significant units (ESUs) estimate diversity but fail to connect delineated units with known species (Kizirian and Donnelly, 2004; Blaxter *et al.*, 2005). In ecological niche modelling, environmental variables are identified and associated with the known distribution of a species (Raxworthy *et al.*, 2007). In character-based methods, a unique combination of diagnostic characters are used to define a species (autopomorphic species concept (ASC)-K.C. and Wheeler, 1990; population aggregation analysis (PAA)-Davis and Nixon, 1992; cladistic haplotype analysis (CHA)-Brower, 1999; Characteris-

tic Attribute Organization System (CAOS)-Sarkar, Planet and Desalle, 2008). But the constant change occuring within species (microevolution; Funk and Omland, 2003), reliance on a reference tree (Little, 2011), and lack or subtlety of informative molecular characters (Hudson and Coyne, 2002) may limit their use. By far, the three categories of methods most embraced by the barcoding community are distance-, tree-, or coalescent-based. The use of genetic distances and a threshold (a region, dubbed the "barcoding gap", that is defined by the maximum level of intra- versus minimum level of inter-specific variation and has taken on various values, notably 2%, 3%, 10x intraspecfic variation, and 1% Hebert *et al.*, 2003, 2004; Ratnasingham and Hebert, 2007) and variations of it (fuzzy-set-approach Zhang *et al.*, 2012; support vector machine (SVM)-Seo, 2010) are inadequate because they fail to consider species specific evolutionary rates (Hickerson, Meyer and Moritz, 2006; Meier, Zhang and Ali, 2008; Lim, Balke and Meier, 2012), and a "barcoding gap" is not necessarily a prerequisite for correct species assignment (Ross, Murugan and Li, 2008; Virgilio *et al.*, 2010; Hendrich *et al.*, 2010). In phylogenetic- or tree-based methods, the query belongs to the clade that it groups with (Statistic Assignment Package (SAP)-Munch *et al.*, 2008; pplacer-Matsen, Kodner and Armbrust, 2010). Relative to a tree-based method, a coalescent method is more complex because it models demographic information (population genetics) in conjunction with backward-in-time evolutionary relationships (phylogenetic). An example is the general mixed Yule-coalescent (GMYC) model that distinguishes population-level processes within lineages from processes associated with speciation and extinction (Pons *et al.*, 2006). However, the criterion of reciprocal monophyly (sequences of individuals forming their own clades to the exclusion of others) of tree- and coalescent-based methods is arbitrary since a lack of monophyly does not preclude speciation (Ross, Murugan and Li, 2008). Furthermore, most methods do not provide a measure of statistical confidence or probability of the assignment. To address this shortcoming, Abdo and Golding (2007) introduced a coalescent method that operates within a Bayesian framework. In general, a Bayesian method is ideal because it may attach probabilities to hypotheses (i.e., provides exact versus approximate inferences) by considering the given data (i.e., likelihood) with other relevant information (i.e., prior or past knowledge). However, a Bayesian method can suffer computational problems if the data set is very large (Zhang *et al.*, 2011). To overcome computational demands, the coalescent or 'tree' step in Abdo and Golding (2007) was replaced with a population genetic parameter, Watterson (1975)'s $\theta$, that can be calculated using the number segregating sites. This slight modification produced a faster algorithm and is, to our knowledge, the first Bayesian method, based on the theory of segregating sites (Lou and Golding, 2010). Also lacking are models and methods that account for the information content of the data and how it may affect identifications. There have been a number of 'integrative' studies advo-

cating the inclusion of sequences representative of the intraspecific (within-species) variation of geographically widespread species to improve barcoding accuracy rates (Cameron, Rubinoff and Will, 2006; Padial *et al.*, 2010; DeWalt, 2011; Goldstein and DeSalle, 2011). While Bergsten *et al.* (2012) did investigate geographical sampling on barcoding, a threshold of 1% was used and no Bayesian approaches were evaluated. In addition to population substructure, heterogeneous substitution rates for transitions and transversions shape genetic variation, especially in mitochondrial DNA (mtDNA). It has been proposed that the infinite sites model (ISM) should also be extended to account for different substitution rates for transitions and transversions (David *et al.*, 2012).

While studies have shown the importance of a comprehensive library and firm understanding and use of the sequence data and identification methods for barcoding success, there still exists a lack of correspondence between data and model assumptions made by barcoding methods. That is, data and methods do not sufficiently model the evolutionary signals that describe dynamic species boundaries. Ultimately, the main contribution of this thesis, to the body of knowledge that is species delimitation, is to improve the use and modelling of data by integrating different sources of relevant information.

The main purpose of this study was to investigate how to improve the species assignment framework. This was attempted by investigating the effect of different levels of informativeness (resolving power) of data and a new assignment method on the accuracy of assignments. Specifically, the research questions driving this thesis are: Can GenBank sequences be used to make accurate assignments? Can using a descriptor of variation that models the evolution of the sequences (population mutation rate, $\theta$) and including more informative sites (from sampling sequences across the species geographical range) improve the number of correct assignments? How does a method based on $\theta$ or the number of segregating sites perform relative to comparable assignment methods? Does the modelling of biological forces influencing sequence diversity improve probability estimates of the observed level of variation? Since we are investigating the performance of the data and assignment methods, the units of analysis are the number of correct assignments of an unknown query to its correct species or how well a proposed model represents the level of genetic variation. Results will be generated through a series of assignments using simulated and empirical data. The simulated sequence data will model the evolutionary dynamics expected of animal mtDNA and evolutionary relationships of recently and deeply diverged species. Empirical data will be drawn from the Class Insecta because of easy access to a vast number of sequences that can describe easy and hard assignment scenarios, even though the species relationships are well

defined. To perform the assignments within a statistical framework, attention was focused on using and developing Bayesian methods.

The rest of the thesis is structured, in four chapters, as follows: The first chapter investigates if GenBank data, using a species delimitation threshold (3%), can generate accurate species assignments. The second chapter introduces a new Bayesian statistical assignment method, based on a widely used population genetic parameter ($\theta$), to describe the level of observed variation in a set of sequences, and involves a comparison of how it performs relative to a similar Bayesian method. This is followed by a chapter to investigate if including informative sequences (to capture sufficient intraspecific variation) from across the geographical range of species aids the assignment of an unknown query back to this species. Finally, the last chapter explores if an improvement in probability estimates of the observed genetic variation in a set of sequences can be achieved with a model of evolution that considers sequences sampled from subdivided populations and heterogeneous base and substitution rates for transitions (purine to purine or pyrimidine to pyrimidine interchanges) and tranversions (purine to pyrimidine-and vice versa- interchanges).

# Chapter 1

# Species identifications within GenBank are not accurate enough for barcoding

## 1.1 ABSTRACT

Many studies have used DNA barcoding for specimen identification. Some of these studies have relied on sequences taken from GenBank even though problems with sequence accuracy are well known. We tested the accuracy of GenBank records for another potential source of error: species identification. Insect sequences were used to examine the magnitude of this problem as this group is species rich and is well characterized. In the absence of errors, a comparison of sequences with and without the "BARCODE" designation should have identical characteristics. Non-barcode records were found to have an unusually high proportion of divergent conspecific sequences and expected proportion of similar congeneric sequences. These results suggest the records within GenBank may have little to no errors or that the standard 3% percent sequence divergence cannot be used to distinguish species. The latter explanation is more likely as it has been shown to fail to correctly diagnose insect species for 45% of the cases.

## 1.2 INTRODUCTION

The Barcode of Life project has the goal to find a single standard piece of DNA that can be used to identify species. The success of this project depends on several factors including the support of professional taxonomists, an agreement on a standardized segment of DNA, the accuracy of the DNA sequencing, the accuracy of the identification for the specimens that supply reference barcode sequences, and the statistical accuracy of matching a query sequence to reference taxa.

A 648 bp region near the 5′ end of mitochondrial cytochrome c oxidase, subunit I (COI) has been successfully used for specimen identification (Hebert *et al.*, 2003) and resolving sequence diversity in fungi (Seifert *et al.*, 2007), gastropods (Remigio and Hebert, 2003), amphipod crustaceans (Witt, Threloff and Hebert, 2006), bats (Clare *et al.*, 2007), birds (Hebert *et al.*, 2004), fishes (Ward *et al.*, 2005), and Lepidoptera (Hebert *et al.*, 2004; Hajibabaei *et al.*, 2006).

To help ensure the accuracy of specimen identification, the Consortium for the Barcode of Life has set up a collection of standards that should be met for barcode records. These standards include methods of DNA preparation, methods of sequencing, availability of the sequence trace files to check accuracy, multiple samples for each species and the availability of voucher specimens to double check species identifications and to match sequence information to the corresponding taxonomic information (`http://www.barcodeoflife.org/content/resources/standards-and-guidelines`).

Recognizing the importance of these criteria and standards, NCBI has set up the use of a reserved keyword to mark data records that adhere to these standards. The "`BARCODE`" keyword indicates that the sequence data in the corresponding record met the highest standards.

Several studies (Meier *et al.*, 2006; Elias *et al.*, 2007; Wiemers and Fiedler, 2007) have used entries from GenBank to test and to explore the usefulness of the barcoding concept. It is well known that the accuracy of DNA sequences deposited in GenBank is often less than could be desired (Harris, 2003; Valkiunas *et al.*, 2008). This is understandable since GenBank is a data repository and is not a currated sequence database. However, while the accuracy of the sequencing has been tested and found to be lacking, there are other aspects of the data in GenBank that have not been rigorously examined. Foremost among these is the accuracy of species identifications from which GenBank's data originate. Most of the sequence information comes from model organisms such as *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*, and species identification in these cases is unlikely to be too far from

the correct answer. However, there are many other taxa that have had portions of their DNA sequenced that are much more difficult to identify.

A problem presented by these data is that for most of the sequences in GenBank, species identifications cannot be validated because no voucher specimens are provided. To work around this problem, we have analyzed sequences from insect species retrieved from GenBank with the goal to examine how similar or how dissimilar sequences from a single species might be. Insects were chosen as a well defined and yet diverse group of animals from which many barcode sequences have been collected (Hebert, Ratnasingham and deWaard, 2003). Records with the "BARCODE" keyword were compared to records within GenBank without this keyword. If the accuracy, both for sequence correctness as well as species identification, are similar between GenBank entries with and without this keyword, then current barcode studies can make uncritical use of the vast warehouse of data previously stored in GenBank.

## 1.3 MATERIALS AND METHODS

COI sequences were collected from NCBI in October/November 2007. Among the insects, a total of 6,505 barcode sequences from 747 species and 34,384 non-barcode sequences from 11,385 species were obtained. All sequences were aligned using MUSCLE (Edgar, 2004); scaffold sequences from five different orders (Coleoptera, Diptera, Hemiptera, Hymenoptera, and Lepidoptera) were included in the alignment to prevent incorrect alignments caused by sequences with limited overlap. To obtain a representation of the barcode region in the COI sequences, the region was extracted from the aligned sequences using the NCBI barcode sequence (accession ID: EF180877) as a reference sequence; this entry was chosen because it was the longest barcode sequence available. It was assumed that the region extracted does not include gaps. Only sequences containing at least 85% of the original number of residues were used. The reduced data set consisted of 5,179 barcode sequences from 590 species and 8,586 non-barcode sequences from 2,900 species (Table 1.1). Another data set consisting of common species between barcode and non-barcode data was generated (Table 1.2). For each sequence, taxonomic classification was recorded according to the four following levels: order, family, genus, and species.

Alignment of sequences within a species was done using the corresponding amino acid sequence via MUSCLE (Edgar, 2004) and then translated back to DNA using TRANALIGN. A scaffold sequence was included in all alignments to prevent incorrect alignments caused by sequences with limited overlap. Kimura two-

Table 1.1: Distribution of barcode and non-barcode sequences in 12 insect orders

| | Barcode | | Non-barcode | |
|---|---|---|---|---|
| Order | Sequences | Species | Sequences | Species |
| Coleoptera | 2 | 2 | 1017 | 446 |
| Diptera | 2698 | 123 | 1962 | 412 |
| Ephemeroptera | 0 | 0 | 229 | 80 |
| Hemiptera | 0 | 0 | 326 | 71 |
| Hymenoptera | 78 | 19 | 1294 | 540 |
| Lepidoptera | 2401 | 446 | 2566 | 969 |
| Odonata | 0 | 0 | 95 | 19 |
| Orthoptera | 0 | 0 | 423 | 71 |
| Phthiraptera | 0 | 0 | 122 | 13 |
| Strepsiptera | 0 | 0 | 7 | 6 |
| Thysanoptera | 0 | 0 | 195 | 42 |
| Trichoptera | 0 | 0 | 350 | 231 |

parameter distances (K2P; Kimura, 1980) between members within a species were determined. Any species file containing a distance greater than or equal to 3% is recorded as a species with a divergent member.

To analyze distances between congeneric species, a randomly chosen sequence from each species was blasted against the NCBI database. Sequences were collected from the first fifty BLAST results and aligned. Alignment and K2P distances, between the random query and BLAST results, were determined using the same metholodgy described above. If the K2P distance was less than 3% or if no distance was generated and the query and blast result did not possess the same taxonomic species designations, the blast result is recorded.

## 1.4 RESULTS

Considering all species, the number of species with divergent members (one or more members that differ by more than 3% K2P distance) are shown in table 1.3. Some of the "BARCODE" records contain divergent sequences. This is most notable within Hymenoptera with 12.5%, but these sequences only form a small percentage (less than 1%) of the data set. In contrast, Diptera sequences comprises 75% of

Table 1.2: Distribution of barcode and non-barcode sequences in 2 insect orders, common to both data sets

| | Barcode | | Non-barcode | |
|---|---|---|---|---|
| Order | Sequences | Species | Sequences | Species |
| Diptera | 2158 | 57 | 187 | 57 |
| Lepidoptera | 1425 | 214 | 789 | 214 |

the data set and only 3.8% Dipteran species have divergent members. With regard to the non-barcode data set, the percentage of divergent species is much higher. Consistently over 50% of the species without the "BARCODE" designation contain divergent sequences except for Hemiptera, Lepidoptera, and Trichoptera. Not all species had more than one sequence per species and hence it is not possible for their single sequence to differ. To determine the opportunity for species to be detected in this manner, the last column in table 1.3 shows the average number of sequences per species with two or more sequences.

With reference to common species between barcode and non-barcode data, the results were consistent with those of the all-species data set showing a higher percentage of divergent species among non-barcode records in comparison to barcode records (Table 1.4).

In addition to finding members of a single species with divergent sequences, we also looked for species with sequences that were similar to the sequences of other species. With reference to the all-species data set, Coleopteran and Lepidopteran barcode records have a smaller percentage of abnormally similar species in comparison to their non-barcode neighbours (Table 1.5). However, this result is not entirely unexpected since there is a greater chance that one may find more abnormally similiar species in these orders as indicated in the last column of the table. In contrast, there are more barcode records with similar congeneric species than non-barcode records in the following two orders: Diptera and Hymenoptera.

With equal chance of finding "similar" distinct species in the common-species data set, we find similar percentages in barcode and non-barcode records (Table 1.6).

Table 1.3: Species with one or more members that differ by more than 3% K2P distance in 12 insect orders

| Group | Divergent species | Species with $\geq$ 2 seq | Percent | Avg sequences per species with $\geq$ 2 seq |
|---|---|---|---|---|
| **Barcode data (5,179 seq)** | | | | |
| Diptera | 4 | 105 | 3.8 | 25.52 |
| Hymenoptera | 1 | 8 | 12.5 | 8.38 |
| Lepidoptera | 0 | 359 | 0 | 6.45 |
| **Non-barcode data (8,586 seq)** | | | | |
| Coleoptera | 63 | 85 | 74.1 | 7.72 |
| Diptera | 96 | 164 | 58.5 | 10.45 |
| Ephemeroptera | 18 | 33 | 54.5 | 5.52 |
| Hemiptera | 10 | 27 | 37.0 | 10.44 |
| Hymenoptera | 79 | 127 | 62.2 | 6.94 |
| Lepidoptera | 173 | 428 | 40.4 | 4.73 |
| Odonata | 2 | 3 | 66.7 | 26.33 |
| Orthoptera | 15 | 20 | 75.0 | 18.6 |
| Phthiraptera | 3 | 5 | 60.0 | 22.8 |
| Strepsiptera | 1 | 1 | 100.0 | 2.0 |
| Thysanoptera | 9 | 17 | 52.9 | 10.0 |
| Trichoptera | 20 | 35 | 37.1 | 4.4 |

## 1.5 DISCUSSION

The data in table 1.3 and table 1.4 clearly indicate that more non-barcode sequence records have insect sequences that differ from conspecific sequences by more than 3%. This is not necessarily unexpected as, other than empirical evidence, there is no reason to expect sequence divergences within a species to be limited. However, if the cause of the divergence were simply taxonomic divergence, then the barcode and GenBank records should provide identical results. They do not. GenBank records are more likely to be divergent by more than 3%. One possible cause of this is that the barcode sequences are sampled from only a few members of each species and hence the opportunity for divergence is not as great simply because there are fewer comparisons made within each species. The results for Diptera and Hymenoptera in table 1.3 and Diptera and Lepidoptera in table 1.4 indicate that on

Table 1.4: Species with one or more members that differ by more than 3% K2P distance in 2 insect orders

| Group | Divergent species | Species with $\geq 2$ seq | Percent | Avg sequences per species with $\geq 2$ seq |
|---|---|---|---|---|
| **Barcode data (3,583 seq)** | | | | |
| Diptera | 2 | 57 | 3.5 | 39.91 |
| Lepidoptera | 0 | 173 | 0 | 8.00 |
| **Non-barcode data (976 seq)** | | | | |
| Diptera | 14 | 36 | 38.9 | 4.61 |
| Lepidoptera | 42 | 137 | 30.7 | 5.19 |

average the opposite is true, there are generally more sequences per species in the barcode records and hence there is a greater opportunity for divergent sequences to be discovered. Another possible cause of this difference is that the barcode records are generally more modern records and as techniques have improved, sequencing errors have decreased. An examination of the date of entry for each record did not, however, reveal any apparent relationships between the number of divergent records and the date of entry for the record. Hence, the most likely explanation for this data is a larger level of sequence error in GenBank records. As stated above this is not an unusual or unexpected result.

Unfortunately, the results were not consistent when it came to identifying abnormally similar congeneric species. Even though more GenBank records were found to contain sequences that are unusually similar to sequences from other species (particularly Coleoptera and Lepidopera) this was an expected result as there are more species per genera in these two orders in GenBank data hence the chances of finding similar congeneric species is greater. On the other hand, there was a greater chance of finding similar congeneric species in barcode records for Diptera and Hymenoptera and this turned out to be the case. Either there is no sequence error in GenBank sequences, as initially claimed, and there are species in barcode records abnormally similar to other distinct species or the results are the product of an arbitrary criterion cut-off value that has been known to have associated difficulties (Moritz and Cicero, 2004; Meyer and Paulay, 2005; Hickerson, Meyer and Moritz, 2006) and cannot be applied equally across orders or even across species (Cognato, 2006). The latter explanation is more likely than the former as Cognato (2006) found that a standard percent sequence divergence has failed to correctly diagnose

Table 1.5: Species within 3% K2P distance to other species in 12 insect orders

| Group | No. sim spp (SS) | No. spp. | Percent | No. genera | Spp./Genera |
|---|---|---|---|---|---|
| **Barcode data (5,179 seq)** | | | | | |
| Coleoptera | 0 | 2 | 0 | 2 | 1.00 |
| Diptera | 75 | 123 | 60.1 | 17 | 7.24 |
| Hymenoptera | 6 | 19 | 31.6 | 2 | 9.50 |
| Lepidoptera | 87 | 446 | 19.5 | 170 | 2.62 |
| All | 168 | 590 | 28.5 | 191 | 3.10 |
| **Non-barcode data (8,586 seq)** | | | | | |
| Coleoptera | 20 | 446 | 4.5 | 176 | 2.53 |
| Diptera | 162 | 412 | 39.3 | 92 | 4.48 |
| Ephemeroptera | 6 | 80 | 7.5 | 27 | 2.96 |
| Hemiptera | 9 | 71 | 12.7 | 41 | 1.73 |
| Hymenoptera | 87 | 540 | 16.1 | 236 | 2.29 |
| Lepidoptera | 488 | 969 | 50.4 | 323 | 3.00 |
| Odonata | 0 | 19 | 0 | 5 | 3.80 |
| Orthoptera | 22 | 71 | 31.0 | 37 | 1.92 |
| Phthiraptera | 2 | 13 | 15.4 | 5 | 2.60 |
| Strepsiptera | 2 | 6 | 33.3 | 4 | 1.50 |
| Thysanoptera | 2 | 42 | 4.8 | 16 | 2.63 |
| Trichoptera | 103 | 231 | 44.6 | 114 | 2.03 |
| All | 903 | 2,829 | 31.9 | 1,039 | 2.72 |

insect species for 45% of the cases. The use of specific thresholds or statistical approaches of specimen assignment may be more appropriate.

Table 1.6: Species within 3% K2P distance to other species in 2 insect orders

| Group | No. sim spp (SS) | No. spp. | Percent | No. genera | Spp./Genera |
|---|---|---|---|---|---|
| **Barcode data (3,583 seq)** | | | | | |
| Diptera | 38 | 57 | 66.7 | 14 | 4.07 |
| Lepidoptera | 59 | 214 | 27.6 | 96 | 2.23 |
| All | 97 | 271 | 35.8 | 110 | 2.46 |
| **Non-barcode data (976 seq)** | | | | | |
| Diptera | 40 | 57 | 70.2 | 14 | 4.07 |
| Lepidoptera | 61 | 214 | 28.5 | 96 | 2.23 |
| All | 101 | 271 | 37.3 | 110 | 2.46 |

# Chapter 2

# Assigning sequences to species in the absence of a interspecific "barcoding" gap

Lou, M. and Golding, G.B. (2010) *Molecular Phylogenetics and Evolution*. 56: 187-194.

## 2.1 ABSTRACT

Barcoding is an initiative to define a standard fragment of DNA to be used to assign unknown sequences to existing known species groups that have been pre-identified externally (by a taxonomist). Several methods have been described that attempt to place this assignment into a Bayesian statistical framework. Here we describe an algorithm that makes use of segregating sites and we examine how well these methods perform in the absence of an interspecific 'barcoding gap'. When a barcoding gap exists, that is when the data are clearly delimited, most methods perform well. Here we have used data from the *Drosophila* genus because this genus includes sibling species and the species relationships within this genus while complex are, arguably, better understood than in any other group. The results show that Bayesian methods perform well even in the absence of a barcoding gap. The sequences from *Drosophila* are correctly identified and only when the degree of incomplete lineage sorting is extreme in simulations or within the *Drosophila* species do they fail in their identifications and even then, the "correct" species has a high posterior proba-

bility.

## 2.2 INTRODUCTION

DNA barcoding involves the use of a short DNA sequence as a means to taxonomically identify a specimen (Hebert, Ratnasingham and deWaard, 2003; Hebert *et al.*, 2003; Remigio and Hebert, 2003). The key to this concept is to standardize the segment of DNA used for barcoding and then to construct a database of this sequence from as many taxonomically identified species as possible. Storing these data in a searchable database permits new or unknown specimens to be identified via a comparison of their sequence with sequences from characterized species. The recognized utility of this methodology has resulted in a global, synchronized effort with more than 100 member organizations (including museums, zoos, botanical gardens and universities) involved in setting a global standard in taxonomy and in creating a database of DNA barcode sequences.

Although the usefulness of this approach is well established (see for example Hebert, Ratnasingham and deWaard, 2003; Hebert *et al.*, 2004; Hajibabaei *et al.*, 2006), some taxonomic groups, such as cowries (Meyer and Paulay, 2005) and tiger moths (Schmidt and Sperling, 2008), have shown an unacceptably high error rate for identification by DNA barcodes. Part of the reason for this discrepancy is due to similar levels of intra- and interspecific divergence. Under these conditions there may be a small amount of divergence between species relative to the amount of divergence within species. The difference between intra- and interspecific divergences is known as the barcoding gap. Cognato (2006) found substantial overlap between levels of intra- and interspecific variation within several orders of insects resulting in the failure to correctly diagnose insect species for 45% of the cases. Within Diptera, there are congeneric sequences whose distance is within 1% (Meier, 2008). Similarly, the Lepidopteran family Lycaenidae showed an 18% overlap between intra- and interspecific COI divergence (Wiemers and Fiedler, 2007). An overlap may occur for a number of reasons. It may occur when there is a wide variation in rates of molecular evolution among lineages (Sparks and Smith, 2006; Huang *et al.*, 2008). The COI from some animals, such as coral (Huang *et al.*, 2008), evolves too slowly to be useful for barcoding. Incomplete lineage sorting (paraphyly or polyphyly; Moritz and Cicero, 2004; Pollard *et al.*, 2006; Wiemers and Fiedler, 2007; Aliabadian *et al.*, 2009) and poor taxonomy may also explain the lack of a barcoding gap. An inference must be made as to which species (or other taxonomic group) the sequence belongs. It is often difficult to discern whether or

not differences between the query sequences and sequences within the database are due to intraspecific differences or if they are an indication of interspecific differences. The effectiveness of barcoding is associated with a clear distinction between levels of divergence with the level of interspecific divergence greater than intraspecific divergence. Indeed it has been shown that the simplest of methods performs well under these circumstances (Ross, Murugan and Li, 2008; Austerlitz *et al.*, 2009). Although it is not impossible to identify a species in the absence of a barcoding gap, this deficiency makes it much more difficult.

However, these methods lack ways to measure the confidence with which an assignment is made. Hence, there is a need for statistical methods to determine the most appropriate assignment and the degree of confidence with which this assignment can be made, particularly when a barcode gap might be small or nonexistent. Frezal and Leblois (2008) note that population genetics theory is required to account for the level of uncertainty that is contributed by these processes. Here, only Bayesian methods will be examined because these provide the necessary statistical strength to distinguish between well supported assignments versus poor assignments and to provide a strong statistical framework.

There are two Bayesian methods that have been proposed to date. The first is a method that uses the coalescent (Abdo and Golding, 2007). This method calculates the likelihood of coalescents for sequences known to originate from a particular species and then calculates the change in the likelihood when the query sequence is considered a member of this species. The assignment of an unknown individual sequence is to the group, $i$, that minimizes the posterior risk, $R_i$. The posterior risk of group $i$ reflects the posterior probability that the sequence belongs to a coalescent with sequences from species $i$ and the 'loss' of making the decision that the query sequence originated from species $i$. Here, loss is defined as the difference between the sequence of the unknown individual and the consensus sequence of the assumed correct group $k$. The mathematical details for calculating the posterior risk, loss and posterior probability are given in Abdo and Golding (2007).

A coalescent method can be time consuming for data sets with a large number of sequences since it must generate enough coalescent trees to adequately sample all possible coalescent events. Therefore, the coalescent method is amended in this paper by replacing the coalescent-based Markov Chain Monte Carlo (MCMC) algorithm with one that makes use of the number of segregating sites from the sequences of a single species. A segregating sites method uses only sites at which there is a nucleotide change. The theory behind segregating sites allows closed form solutions to be used in place of time consuming MCMCs. It is therefore very rapid. It does, however, involve a loss of information and compresses the entire collection of

sequence data into a single number. For barcoding sequences, which can generally be assumed to be closely related sequences, the loss of information is usually minor.

Another Bayesian method is the SAP (statistical assignment package) algorithm that incorporates taxonomic information from NCBI and uses this information to impose topology constraints on the trees sampled from a MCMC. The probability of assignment is the number of sampled trees showing the unknown sequence branching with a sequence from species $i$ (Munch *et al.*, 2008, a,b). This approach assumes that the branching pattern, as delimited by the taxonomy, is realistic and accurate. It also does not take into account the variability that might be expected around this branching pattern due to unsampled intraspecific differences and it assumes that the species are monophyletic. However, several studies have shown that the expectation of monophyly for recently diverged species is not realistic (Knowles and Carstens, 2007; Hickerson, Meyer and Moritz, 2006; Hudson and Coyne, 2002). It is noted by Nielsen and Matz (2006) that false species assignments can be caused by incomplete lineage sorting and by random mutation processes that can mimic incomplete lineage sorting.

The comparison of population genetic methodologies to phylogenetic methods done here suggests that the posterior probability of species identification is, in general, much smaller for the former. This suggests that these methods are more conservative than phylogenetic methods. The underlying cause of these differences in posterior probabilities are shown to be because these methods estimate the probabilities of different quantities.

## 2.3 MATERIALS AND METHODS

### 2.3.1 Evaluating assignment with segregating sites

Following Abdo and Golding (2007), we evaluate the probability of assigning an unknown sequence to a taxonomic grouping in a Bayesian context. For some unknown DNA sequence, $x$, the goal is to assign the species from which this sequence was taken to the correct taxonomic group, $k$. Hence, we wish to find:

$$Pr(x \in k|x, D, \theta)$$

where $D$ is a database of known sequences with $n$ distinct taxonomic groups and $\theta$ is a known collection of evolutionary parameters. The assignment of sequence $x$ must be made to one of the taxonomic groups.

Table 2.1: *Drosophila* COI sequences tested (Monophyly is taken from the diagram in figure 2.1).

| Group | Species | Monophyletic | Sequences |
|---|---|---|---|
| melanogaster | *D. mauritiana* | no | 3 |
| melanogaster | *D. melanogaster* | – | 10 |
| melanogaster | *D. simulans* | no | 27 |
| quadrisetata | *D. barutani* | – | 6 |
| quadrisetata | *D. beppui* | – | 3 |
| quinaria | *D. falleni* | – | 15 |
| quinaria | *D. innubila* | – | 29 |
| quinaria | *D. recens* | no | 136 |
| quinaria | *D. subquinaria* | no | 136 |
| repleta | *D. arizonae* | – | 17 |
| repleta | *D. mettleri* | – | 24 |
| repleta | *D. mojavensis* | – | 47 |
| repleta | *D. navojoa* | – | 4 |
| repleta | *D. nigrospiracula* | – | 10 |
| virilis | *D. montana* | – | 42 |
| virilis | *D. virilis* | – | 11 |
| – | *D. angor* | no | 13 |
| – | *D. daruma* | – | 4 |
| – | *D. pachea* | – | 79 |
| Total | | | 616 |

It is assumed that different groups that are potential targets of the assignment are fully pre-specified. Each group is assumed to form a panmictic population that follows a Wright-Fisher, neutral model of evolution that does not allow recombination, selection, or migration. Hence, the evolutionary process within each group is governed by one parameter, which is the expected number of mutational events between sequences. This quantity is dependant upon a population measure, $\theta = 4N_e\mu$, and is in turn, reflected in the number of segregating sites between sequences.

Using Bayes rule, assuming that the presampled individuals are assigned correctly by external taxonomists, assuming independence of the evolutionary history between groups and assuming uniform priors, this can be calculated as:

$$Pr(x \in k|x, D, \theta) = \frac{Pr(x, D_k|x \in k, \theta_k)/Pr(D_k|\theta_k)}{\sum_j Pr(x, D_j|x \in j, \theta_j)/Pr(D_j|\theta_j)}$$

(see Abdo and Golding, 2007, for a derivation).

A risk function can be evaluated using this probability and traditionally, an assignment decision is based on the assignment with minimum risk. The risk function can be defined as:

$$R_i = \sum_k L(k, i) Pr(x \in k|x, D, \theta)$$

where $R_i$ is the risk of making the assignment to species $i$ and $L(k, i)$ is the loss associated with an assignment to species $i$ when the correct assignment should be to species $k$ and $Pr(x \in k|x, D, \theta)$ is the posterior probability of membership of the unknown sequence $x$ to taxonomic group $k$.

In Abdo and Golding (2007) a method to evaluate $Pr(x, D_k|x \in k, \theta_k)$ using the coalescent and an MCMC is implemented. However, it is also possible to evaluate $Pr(x, D_k|x \in k, \theta_k)$ using the theory of segregating sites. If the sequence data $\{x, D_k\}$ has $s$ segregating sites, then the probability of the data given $\theta_k$ can be approximated by the probability corresponding to the number of segregating sites, $s$. Hence,

$$Pr(x, D_k|x \in k, \theta_k) \sim Pr(S = s|n, \theta_k)$$

where $s$ is the number of sequences in $\{x, D_k\}$. The basic recursive definition for

the probability that a sample of $n$ sequences will have $s$ segregating sites is:

$$Pr(S = s|n, \theta_k) = \frac{n-1}{n-1+\theta_k} Pr(S = s|n-1, \theta_k) + \frac{\theta_k}{n-1+\theta_k} Pr(S = s-1|n, \theta_k)$$

This recursion makes the assumption that an infinite sites model holds, that the populations are equilibrium single random mating populations of size $N_e$ with mutation to new alleles at a rate $\mu$.

This recursion has been solved by Tavare (1984) to yield a closed form solution of:

$$Pr(S = s|n, \theta_k) = \frac{n-1}{\theta_k} \sum_{i=1}^{n-1} (-1)^{i-1} \binom{n-2}{i-1} (\frac{\theta_k}{i+\theta_k})^{s+1}$$

Our implementation of this formula was found to occasionally be numerically unstable. Therefore, if the closed form solution did not satisfy the recursion with numerical accuracy, we then did an evaluation of the complete recursion.

Attention is focused here on the posterior probability rather than risk (multiple loss functions can be used to quantify risk as described in Abdo and Golding, 2007) to make the results from the segregating sites algorithm comparable with those from the SAP algorithm. To test the assignment of unknown queries using the segregating sites algorithm, we conducted a simulation to test the performance of the algorithm in the absence of a 'barcoding gap'. The simulations use a multi-species coalescent (Degnan and Rosenberg, 2009) to model ten species with a pectinate species tree (Figure 2.2). Each of the ten species has five lineages. The 'unknown' query sequence is simulated as the sixth sequence from the first species. Sequences of length 600bp were simulated using the coalescent tree. The entire length of the sequences were allowed to accumulate substitutions at a constant rate, defined by the parameter $\theta$. This value of $\theta$ is the total mutation rate for all sites in the length of the sequence. At every time interval, defined by $T$, those sequences that had not yet coalesced to a common ancestor were added to the sequences from other "species". The time intervals $T$ were scaled according to $2N_e$ generations and represent the time back to speciation events. However, the coalescents may extend beyond multiple speciation events depending on the size of $T$. In these simulations $T$, ranges from $0.5$ to $3.0$. When $T = 3.0$, the level of interspecific divergence is greater

than the level intraspecific divergence and this represents the ideal situation where a barcoding gap exists and each species is usually monophyletic and is distinct from every other species; in this scenario, we expect a high proportion of correct assignments. When $T = 0.5$, there is a lack of a barcoding gap which may lead to incomplete lineage sorting; we expect a lower proportion of correct assignments. The simulations were repeated 10,000 times and the results are given in table 2.2.

An advantage of the segregating sites algorithm is its speed. The method of segregating sites obviously involves a loss of information in moving from a full coalescent evaluation to an evaluation of a single number, the number of segregating sites. However, it gains a great deal of speed compared to a coalescent method. The analysis of 10,000 simulation runs took only seconds. In addition, for actual data collected from nature, the sequences are from highly conserved genes. Such sequences are anticipated to be very similar and the opportunity for multiple mutations to arise at a single site is small. The results described below document the efficacy of this method.

### 2.3.2 The SAP algorithm

SAP version 1.0.6 was downloaded and installed locally (Munch *et al.*, 2008a). An in-house database constructed from sequences from the *Drosophila* genus were used for searches conducted with a local version of `BLAST` v. 2.2.17. The local database was annotated using the taxonomic information from NCBI. The set of sequence homologues were aligned using a local copy of `ClustalW` v. 2.0 (Thompson, Higgins and Gibson, 1994).

### 2.3.3 *Drosophila* sequences

The *Drosophila* species provide a good data set to test the ability of algorithms to assign sequences to species in the absence of a barcoding gap. Many species are sibling species with small interspecific differences and some have no barcoding gap at all with identical sequences shared among species.

A *Drosophila* data set consisting of 1542 CO1 sequences from 314 species was collected from NCBI and/or Flybase (Tweedie *et al.*, 2009) February 2009. Alignment of sequences within a species was done using the corresponding amino acid sequence via `MUSCLE` (Edgar, 2004) and then translated back to DNA using `TRANALIGN`. Sequences with large indels ($> 10$ amino acids) were removed. The

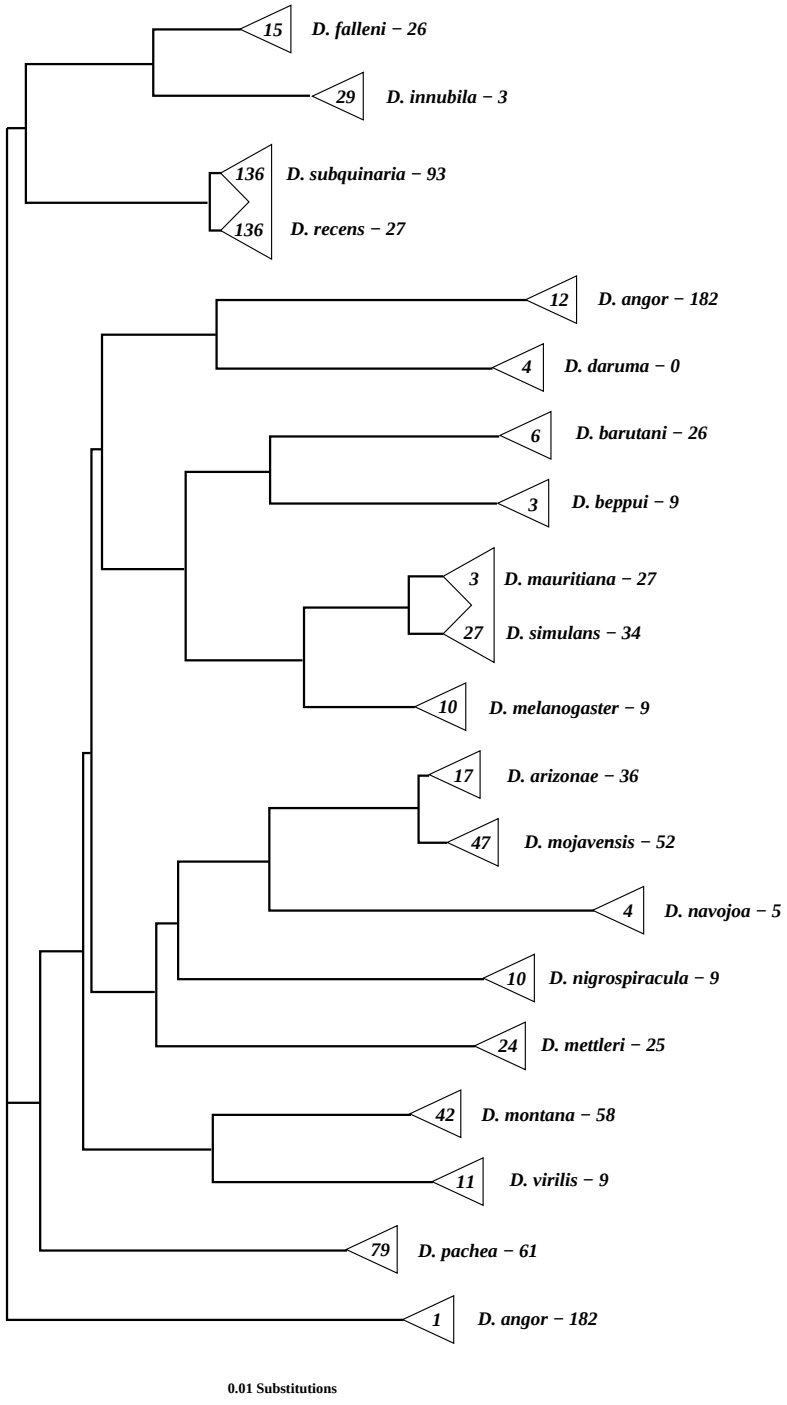



Figure 2.1: A diagram of the relationships of the *Drosophila* COI sequences. The numbers in the triangle give the number of sequences used from each species and the number after the species name is the number of segregating sites within these sequences.

sequences were trimmed to the barcode region (663bp). Sequences were deleted entirely if they contained less than 650bp. Species with two or fewer sequences were removed. Sequences were ensured to originate from distinct strains, from independent wild isolates or from different laboratories, as listed in the GenBank annotation. If there were multiple copies from the same source, the longest sequence from a single strain, isolate, or laboratory was used (refer to supplementary material for a listing of strains and isolates of *Drosophila* species used in the study with, where available, references to literature containing information on where the strain or isolate originates). The remaining data set comprised of 616 sequences from 19 species. A summary of the sequences is shown in table 2.1 (the species and group designations were taken from NCBI; groups are listed only if there are multiple members present). Other commonly known *Drosophila* species have insufficient numbers of sequences or insufficient information that they represent distinct samples to be included by these criteria.

A diagram of the topological relationships between *Drosophila* species is shown in figure 2.1. This diagram is patterned after a phylogeny constructed from Kimura 2-parameter distances (Kimura, 1980) using the Neighbor Joining method (Saitou and Nei, 1987) and with the phylogeny from Flybase (`http://flybase.org/`) with the exceptions of species *D. angor*, *D. barutani*, *D. beppui*, and *D. daruma* (Wang *et al.*, 2006) which are not listed in Flybase. Based on fossil, biogeographic, and molecular clock data, subgenera *Drosophila* (*D. melanogaster*, *D. simulans*, and *D. mauritiana*) and *Sophophora* are estimated to have diverged approximately 62.9 $\pm$ 12.4 million years (MYA) (Powell, 1997; Tamura, Subramanian and Kumar, 2004). Thus, there should be enough interspecific divergence to prevent the assignment of unknown sequences to the incorrect subgenus.

Some of these species are considered sibling species and are difficult to distinguish by anyone other than trained experts. Nevertheless, the species and their relationships are well known (Kelly and Noor, 1996; Powell, 1997). The ability of some taxa to create semi-sterile, usually uni-directional, hybrids has been well documented (Noor, 1995). The species pairs *D. arizonae* & *D. mojavensis*, *D. mauritiana* & *D. simulans*, and *D. recens* & *D. subquinaria* are considered sibling species. In the case of *D. mauritiana* and *D. simulans*, there is a haplotype identified as originating from *D. mauritiana* that is identical to that in *D. simulans* (Satta and Takahata, 1990; Ballard, 2000a,b). The divergence date of these species is estimated as 0.93 $\pm$ 0.49 MYA (Tamura, Subramanian and Kumar, 2004) and so this phenomenon may be due to incomplete lineage sorting or introgression. Similarly, 2 haplotypes (with 1 and 2 representative sequences respectively) out of 109 COI haplotypes from *D. subquinaria* are identical to 2 haplotypes (containing 16

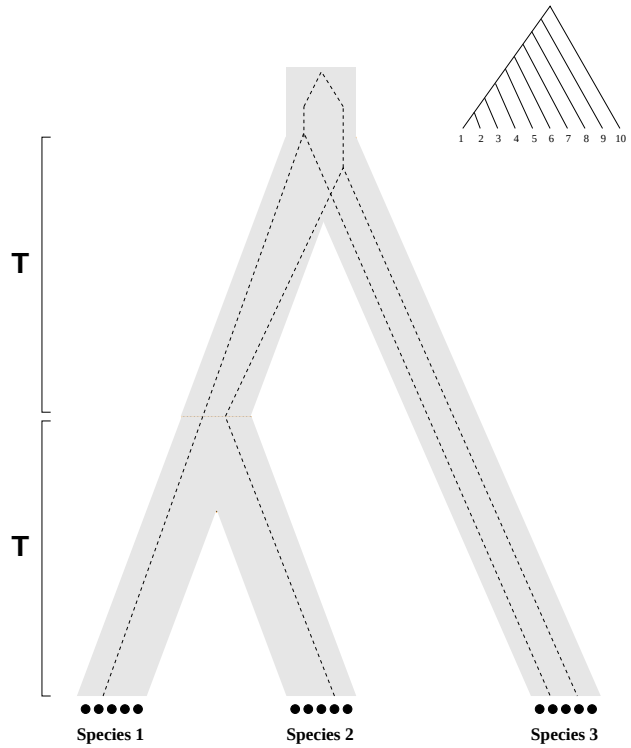



Figure 2.2: The simulation scheme used. Species are added to the tree sequentially up to a total of ten. Three species are expanded here. The length of time separating the divergence of each species can be short, $T << N$, allowing incomplete lineage sorting to occur (as illustrated here lineages within species #2 are more closely related to lineages within species #3 than they are to species #1 despite the implied species relationships).

and 66 sequences respectively) out of 36 COI haplotypes from *D. recens*. These are the result of *Wolbachia*-mediated introgression (Shoemaker *et al.*, 2004; Jaenike *et al.*, 2006). Although *D. arizonae* and *D. mojavensis* are sibling species with an estimated divergence time of 1.91 to 2.97 MYA, their sequences are similar but they do not share any haplotypes (Reed, Nyboer and Markow, 2007).

# 2.4 RESULTS

## 2.4.1 Simulation properties of a segregating sites algorithm

Simulations were conducted to test how well the segregating sites algorithm will assign queries when there is a known degree of similarity between the correct species and its closest relative(s). In this case, each species is progressively more and more distant from the first species (Figure 2.2). The first species is the origin of the query sequence, and the branch length back to the common ancestor encompassing the next species ranges from $T$ = 0.5 to 3.0. With the simulation, the degree to which the histories of the individual species are distinct can be measured by examining the degree to which lineage sorting is complete. The results of this simulation are shown in table 2.2.

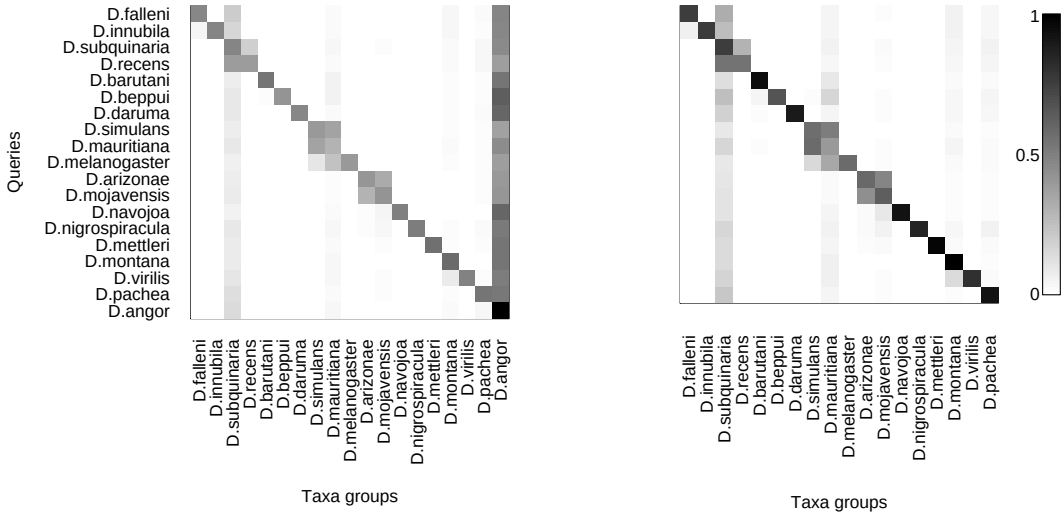
The first row for each simulation run in table 2.2 gives an indication of the extent of incomplete lineage sorting. When the interspecific distance between species is very short ($T = 0.5$) lineage sorting is seldom complete within species 1. Only 15% of the 10,000 simulations have a distinct monophyletic lineage for the five sequences in species 1 while 28% have lineages that confuse species 1 and 2. Nevertheless, the segregating sites method correctly assigns 44% of the queries to species 1. Given the short divergence time and the comparatively small opportunity for distinct substitutions to occur, it is not surprising that the average posterior probabilities for these assignments are low. Because of the similarity between these species, the degree of confidence in these assignments is low.

In general, assignments to species further and further away from species 1 occur in rapidly declining numbers and with declining posterior probabilities. In addition, the estimated value of $\theta$ increases. Thus, the assignments are made to more distantly related species when the number of mutations is, by chance, larger and further blurs the species level distinctions.

As $T$ increases, the proportion of incomplete lineage sorting declines and the assignments become more accurate. In every circumstance, however, the proportion

Table 2.2
Simulation results based on the assignment of 10,000 queries. The query sequence always originates from Taxon #1. The first row indicates how many coalescents for Taxon #1 included sequences from other species (indicated by the column). The second row gives the number of times each species had the highest posterior probability.

| | Taxa | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $T = 3.0, \theta = 2.0$ | | | | | | | | | | |
| No. of taxon 1 coalescents | | | | | | | | | | |
| including other taxa | 9281 | 680 | 39 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| No.Assigned to each taxa | 9388 | 564 | 32 | 14 | 2 | 0 | 0 | 0 | 0 | 0 |
| Avg. Posterior | 0.729 | 0.516 | 0.499 | 0.496 | 0.508 | 0 | 0 | 0 | 0 | 0 |
| Avg. $\hat{\theta}$ | 1.645 | 2.425 | 3.425 | 4.460 | 4.560 | 0 | 0 | 0 | 0 | 0 |
| $T = 2.0, \theta = 2.0$ | | | | | | | | | | |
| No. of taxon 1 coalescents | | | | | | | | | | |
| including other taxa | 7961 | 1744 | 249 | 43 | 2 | 1 | 0 | 0 | 0 | 0 |
| No.Assigned to each taxa | 8543 | 1245 | 173 | 27 | 7 | 3 | 2 | 0 | 0 | 0 |
| Avg. Posterior | 0.601 | 0.444 | 0.404 | 0.477 | 0.439 | 0.347 | 0.551 | 0 | 0 | 0 |
| Avg. $\hat{\theta}$ | 1.705 | 2.145 | 2.748 | 3.493 | 4.594 | 5.280 | 6.720 | 0 | 0 | 0 |
| $T = 1.0, \theta = 2.0$ | | | | | | | | | | |
| No. of taxon 1 coalescents | | | | | | | | | | |
| including other taxa | 4497 | 3497 | 1286 | 471 | 147 | 57 | 28 | 8 | 6 | 3 |
| No.Assigned to each taxa | 6468 | 2238 | 836 | 311 | 95 | 33 | 9 | 7 | 2 | 1 |
| Avg. Posterior | 0.394 | 0.322 | 0.286 | 0.265 | 0.254 | 0.244 | 0.277 | 0.306 | 0.303 | 0.203 |
| Avg. $\hat{\theta}$ | 1.834 | 2.136 | 2.303 | 2.323 | 2.752 | 3.423 | 4.329 | 5.011 | 3.140 | 2.880 |
| $T = 0.5, \theta = 2.0$ | | | | | | | | | | |
| No. of taxon 1 coalescents | | | | | | | | | | |
| including other taxa | 1522 | 2846 | 2176 | 1350 | 823 | 502 | 304 | 201 | 109 | 167 |
| No.Assigned to each taxa | 4431 | 2154 | 1367 | 849 | 520 | 308 | 162 | 114 | 61 | 34 |
| Avg. Posterior | 0.263 | 0.229 | 0.209 | 0.195 | 0.186 | 0.179 | 0.170 | 0.176 | 0.161 | 0.161 |
| Avg. $\hat{\theta}$ | 2.044 | 2.212 | 2.211 | 2.203 | 2.213 | 2.247 | 2.189 | 2.343 | 2.374 | 1.726 |




(a) (b)

Figure 2.3: Average posterior probability of assigning a query to each species group in the local database using the segregating sites algorithm; (a) with *D. angor* and (b) without *D. angor*. A grayscale ramp from white to black represents the average posterior probability assignment from 0.0 to 1.0 respectively. The origin of the query sequence is shown on the y-axis and the taxon for assignment is shown on the x-axis. Shadings off the main diagonal indicate posterior probabilities to incorrect taxon.

of correctly assigned query sequences is larger than the proportion of species #1 that have incomplete lineage sorting. Thus, the correct assignment of sequences can occur even without a barcoding gap but the confidence in that assignment can be variable.

### 2.4.2 The assignment of *Drosophila* sequences

Each *Drosophila* sequence was removed in turn and then assigned to a member species by the algorithms discussed here. The results for the segregating sites algorithm are shown in figure 2.3. The figure gives the average posterior probability that a query sequence (on the y-axis) is assigned to any one of the taxa (on the x-axis). The assignments of *Drosophila* sequences via the segregating sites algorithm (Figure 2.3a) consistently suggest that *D. angor* has a strong posterior probability for each and every one of the query sequences. Indeed, in many cases the posterior probability of an assignment to this group can be larger than that for the correct

taxon. For example the average posterior probability for eleven *D. virilis* sequences is 0.4061 that they originated from a coalescent of the *D. angor* sequences and only 0.3908 that they originated from the coalescent formed by the remaining *D. virilis* sequences.

The *D. angor* sequences are an odd collection. The phylogeny shown in figure 2.1 suggests that these sequences branch polyphyletically throughout the tree. These thirteen sequences form roughly five groups. The first group of four sequences are identical among themselves but differ from the others by 60 to 115 substitutions (within a length of 663 bp; a rather large level of intraspecific divergence). The second group of six sequences differ within the group by 2 to 46 substitutions. The third, fourth and fifth groups are each a single sequence that differs from every other *D. angor* sequence by 75-121, 81-119, 110-121 substitutions. That two sampled sequences from a single species should differ by fully a sixth of their nucleotides in a highly conserved sequence is unusual.

The effect of this on the assignments is to suggest that *D. angor* has a huge (and unrealistic) value of $\theta$ and that the coalescent formed by the *D. angor* sequences can encompass any query. The potential addition of a query sequence to the *D. angor* group does not significantly alter the likelihood of the observed number of segregating sites. This is because only a comparatively few number of additional segregating sites are added with an already very large $\theta$. But since another entire sequence is added, the sample size has increased and, since the query is in the middle of this coalescent, the addition of another sequence with less variation actually improves the likelihood of the observation. This appears to be the cause of the high posterior probabilities of assignment to *D. angor* independent of the query sequence. To a lesser extent, this phenomenon also occurs with *D. subquinaria* since this taxon also has a large amount of sequence variation. To eliminate this effect the *D. angor* sequences were removed and the analysis redone as shown in figure 2.3b.

With the elimination of *D. angor*, most of the query sequences show the highest posterior probability to the taxon from which they originated. Missassignments occur most noticeably in three locations. The missassignment of *D. recens* to *D. subquinaria* (and to a lesser extent, the reverse), a symmetrical confusion between *D. arizonae* and *D. mojavensis*, and missassignments among *D. simulans* and *D. mauritiana*. The missassignments of *D. recens* sequences to the *D. subquinaria* species is because many of these sequences (82 from 2 distinct haplotypes) are identical to sequences labelled as originating from *D. subquinaria* (Shoemaker *et al.*, 2004; Jaenike *et al.*, 2006). The lack of resolution between the *D. arizonae* and *D. mojavensis* species is due to their sibling species status and recent divergence time (Reed, Nyboer and Markow, 2007). The distinction between *D. simulans* and

*D. mauritiana* is even less clear due both to their shared haplotypes and recent divergence (Tamura, Subramanian and Kumar, 2004).

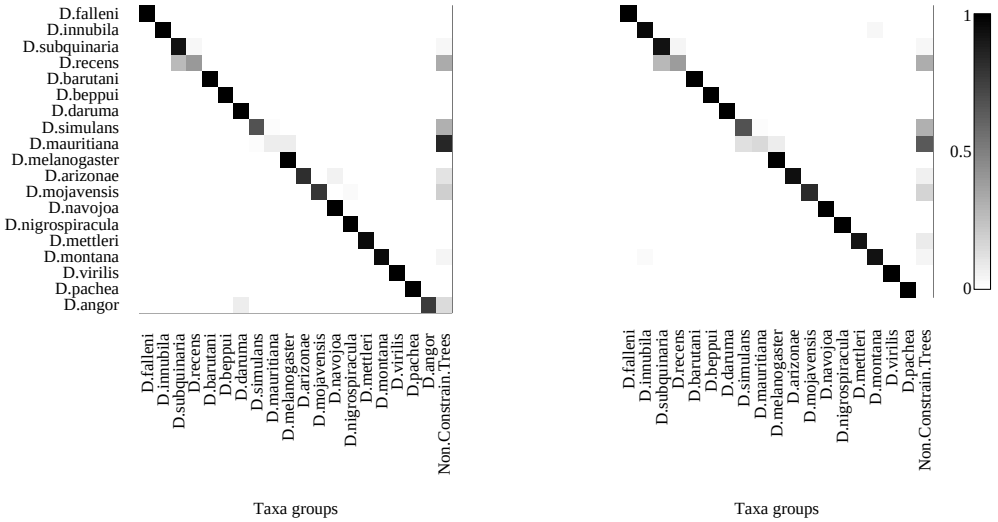
The assignments by the SAP algorithm of query sequences to *Drosophila* species are shown in figure 2.4. This algorithm also had difficulty with the same group of taxa that the segregating sites algorithm had difficulty with. For the most part, these difficulties are not as apparent in the figure since a portion of the sampled trees from the MCMC do not match the given taxonomy from NCBI, termed here non-constrained trees. These trees, that do not match the NCBI annotated taxonomy, are classified separately. These trees represent an ambiguous component of the assignment.

The segregating sites algorithm spent roughly 3 seconds per assignment for the whole 616 sequence data set on a computer with a 1.6GHz processor, running Linux. SAP spent roughly 8 minutes per assignment on the same system. A single assignment of a single query to the 42 sequences of *D. montana* using a coalescent assigner takes many hours to run and even then it is doubtful that it has reached stationarity. A single assignment to the 136 sequences of *D. recens* would take orders of magnitude longer. To complete the data set would require this to be repeated for each of the 616 queries. Hence it is not possible to provide comparable results for the coalescent assigner.

## 2.5 DISCUSSION

Barcoding involves the assignment of a sequence to a pre-existing taxonomic group. This is done using information drawn from a short DNA sequence, COI in many cases (*rbcl* and *matK* in the case of plants; Hollingsworth *et al.*, 2009). The relationships of the sequences among the taxa contains information regarding their likelihood of being samples from a particular species. Unfortunately, when a collection of samples is first made, it is often difficult to determine their taxonomic species of origin. This is particularly the case if the group is little studied and has many sibling species. The *Drosophila* species have many sibling groups but have the advantage that the true species relationships are generally well known.

With the advent of better sequencing technologies, it is expected that the number of alternative species to which an assignment must be made will increase, consequently making the task of assignment more difficult. Thus, the performance of barcoding assignment methods, both in speed and accuracy, given increasing amounts of information, is important.




(a) (b)

Figure 2.4: Average posterior probability of assigning a query to each species using the SAP algorithm; (a) with *D. angor* and (b) without *D. angor*. A grayscale ramp from white to black represents a posterior probability assignment from 0.0 to 1.0 respectively. The origin of the query sequence is shown on the y-axis and the taxon for assignment is shown on the x-axis. Shadings off the main diagonal indicate posterior probabilities to incorrect taxon.

In general, a method to calculate the probability that an unknown sequence originated from a particular species, $x$, is desired. The segregating sites algorithm does not calculate this probability; rather it estimates the probability that the query sequence could originate from a coalescent implied from the knowledge of the current database. The segregating sites algorithm considers all of the sequences from each species. The data from the *D. angor* sequences illustrates this subtle difference. Similarly, the SAP program also does not determine the desired probability. Rather it estimates the probability that a sequence consistently branches next to a single member of species $x$ given the current database. The sequences from *D. angor* do not generally alter these assignments.

The segregating sites algorithm, however, consistently suggest that for each query sequence, there is a significant probability that this sequence might have arisen from *D. angor*. The reason for this is that the given database and the given species identifications are assumed to be correct and, as such, given the huge amount of sequence divergence within the 'hypervariable' species *D. angor*, there is a very real possibility that any of these sequences might have originated from *D. angor*.

Assuming that the given data is indeed accurate, this seems to be the correct answer. The taxonomic assignment of sequences to the species within the database (*D. angor*, for example) are assumed to be correct. This assumption is made at the species level for the segregating sites algorithm. It is similarly made for SAP at deeper taxonomic levels.

If the classification of the sequences of *D. angor* into a single species is correct then the segregating sites algorithm provides correct posterior probabilities. The further consideration of a risk measurement based on distances (which can be incorporated into a segregating sites algorithm) will however warn against over interpretation of the posterior probabilities. The presence of such a hypervariable species is also highlighted by the algorithm's results and suggests a possible alternate interpretation; that the species might be a candidate for further taxonomic scrutiny.

Even if an unknown query sequence is a perfect match to a sequence in a knowledge database, it does not imply that a perfect species identification has been achieved. Other species identifications might have a high or even a higher posterior probability. Therefore, given that a perfect match has been found in the database, this alone does not justify the conclusion that the species of origin has been identified.

The model-based methods analyzed here capitalize on understanding the process governing the system under study and result in more informative and powerful tools to analyze sequence data generated from such systems. In applying any statistical method it is important to understand the boundaries and limitations of its application. The application of the segregating sites algorithm and the SAP algorithm to *Drosophila* data illustrates well that they calculate posterior probabilities of somewhat different quantities. Which method is preferred and should be applied depends on which quantity is desired. The SAP algorithm measures where a sequence branches while the segregating sites algorithm measures if a sequence can 'fit' into an existing species.

The results presented indicate that both Bayesian methods work well to correctly identify species even in the absence of a 'barcode gap'. When uncertainty exists in the assignment, the methods correctly reflect and report this uncertainty. The degree of uncertainty in these methods is directly reflected in the accuracy of the taxonomic reconstructions.

The segregating sites algorithm is available at http://info.mcmaster.ca/TheAssigner/.

## 2.6 ACKNOWLEDGMENTS

This research was funded by an NSERC discovery grant, an NSERC Barcode network grant and an CRC award to GBG. ML is funded by an NSERC Barcode network grant, grants from McMaster University and scholarships.

## 2.7 SUPPLEMENTAL DATA

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2010.01.002.

# Chapter 3

# The effect of sampling population substructure on species identification with DNA barcodes using a Bayesian statistical approach

Lou, M. and Golding, G.B. (2012) *Molecular Phylogenetics and Evolution*. 65: 765-773.

## 3.1 ABSTRACT

Barcoding is an initiative to define a standard fragment of DNA to be used to assign sequences of unknown origin to existing known species whose sequences are recorded in databases. This is a difficult task when species are closely related and individuals of these species might have more than one origin. Using a previously introduced Bayesian statistical tree-less assignment algorithm based on segregating sites, we examine how it functions in the presence of hidden population subdivision with closely related species. Not surprisingly, adding samples to the database from a greater proportion of the species range leads to a consistently higher number of accurate results. Without such samples, query sequences that originate from outside of the sampled range are easily misinterpreted as coming from other species. However, we show that even the addition of a single sample from a different subpopulation is sufficient to greatly increase the probability of placement of unknown

queries into the correct species group. This study highlights the importance of broad sampling, even with five reference samples per species, in the creation of a reference database.

## 3.2 INTRODUCTION

DNA barcoding has become a popular method for species identification and delimitation due to advances in the speed and cost of sequencing and the difficulty in delineating unknown specimens using traditional criteria. In addition to proper biodiversity assessment, barcoding has important implications in various areas such as: effective monitoring of invasive and pest species, identifying disease vectors and protecting consumers from market substitutions (Ball and Armstrong, 2006; Lowenstein *et al.*, 2010; Wong *et al.*, 2011).

Since its initial introduction (Hebert, Ratnasingham and deWaard, 2003), the initiative has evolved from using a distance-based threshold to using a variety of different evolutionary signals to resolve species boundaries (Hebert, Ratnasingham and deWaard, 2003; Hebert *et al.*, 2004; Ratnasingham and Hebert, 2007; Davis and Nixon, 1992; Abdo and Golding, 2007; Munch *et al.*, 2008; Sarkar, Planet and Desalle, 2008; Lou and Golding, 2010). Futhermore, an increasing number of studies advocate the use of traditional lines of evidence (whether behavioural, ecological, geographical, morphological or reproductive) in combination with sequence data to provide further support by showing a correspondence between the two. This combined use of barcoding data with other forms of information has resulted in several well-supported studies that may not have been as reliable if the delimitations had relied solely on sequence data (Hebert *et al.*, 2004; DeSalle, Egan and Siddall, 2005; Siddall and Budinoff, 2005).

The use of additional information may become essential when problems occur from reference sequence data with low information content. One of the benefits of using a mitochondrial marker is that we expect it to better reflect species boundaries because the expected time to obtaining clear, distinct species groups (i.e., reciprocal monophyly) is short because of its small effective population size (Neigel and Avise, 1986). However, population subdivision with limited gene flow can increase the time to coalescence and, consequently, the time required to achieve reciprocal monophyly (Wakeley, 2000; Hudson and Coyne, 2002). With a lengthened time to reciprocal monophyly, lineages between less closely related species may coalesce before lineages within a species (a phenomenon known as incomplete lineage sorting; Neigel and Avise, 1986), thus blurring species boundaries and imped-

ing accurate species diagnoses and delimitations. This is particularly problematic with recently diverged species, a group already prone to incomplete lineage sorting, where the effects of subdivision and migration are more pronounced (Wakeley, 2000). Wong *et al.* (2011) suggested that some incorrect delimitations reflected the failure to consider the geographic divergence of catfish. Similarly, Papadopoulou *et al.* (2008) have shown that different rates of gene flow greatly affect divergences and is one of the reasons that can cause DNA barcoding failures. While the effects of subdivision are explored here, it should be noted that mtDNA is a not perfect marker and may occasionally also show non-neutral evolution, non-clonal inheritence and variation in mutation rates (Galtier *et al.*, 2009).

One way to acknowledge hidden population subdivision is to sample sequences from across a broad geographical range. Any within-species variation is likely to be widely distributed among several geographical localities or demes and sampling this variation is crucial to being able to correctly calculate the probability of origin and distinguish between close sister species. Many barcoding difficulties may, in part, be due to the failure to choose an appropriate sampling scheme (Meyer and Paulay, 2005; Meier *et al.*, 2006; Wiemers and Fiedler, 2007; Wong *et al.*, 2011). Inherent within-species variation may be spread across local, geographical populations of individuals of one species and, by employing a broad sampling scheme, the addition of these dispersed individuals should aid barcoding identification, provided that the sampled sequences sufficiently reflect variation within the species.

The effect of sampling on identification and delimitation has been investigated in distance, tree, and general mixed Yule-coalescent (GMYC) methods (Meyer and Paulay, 2005; Ross, Murugan and Li, 2008; Monaghan *et al.*, 2009; Hendrich *et al.*, 2010; Virgilio *et al.*, 2010; Zhang *et al.*, 2010; Bergsten *et al.*, 2012). Further complexities have also been taken into account, for example, Bergsten *et al.* (2012) have investigated sampling strategies ranging from a local to global scale and Zhang *et al.* (2010) have investigated sampling from two different models of population structure: a linear stepping-stone and an equilibrium island model with unequal sample sizes in three subpopulations. However, no study, to date, has been conducted using a Bayesian statistical method capable of providing an assessment of identification confidence. While Bergsten *et al.* (2012) used a threshold value to calculate the proportion of ambiguous assignments (i.e., the number of queries assigned to more than one reference species) as a measure of method uncertainty, it is not as statistically accurate as a Bayesian method where the probability of assignment describes the assignment to a particular species given that it could also assign to other species possibilities. Setting the classification within a statistical framework to generate posterior probabilities is preferred since difficulties in the

classification of sequences from very recently diverged sibling species are expected via any methodology. We previously introduced the segregating sites algorithm, a fast, Bayesian tree-less method that is able to calculate the probability that the sequence might originate (`PrOr`) from any one of the candidate species (Lou and Golding, 2010). Due to its speed and the large body of theory behind it, the segregating sites algorithm is further explored in this paper to investigate the efficacy of this algorithm when species have recently diverged and exist in subdivided groups.

Here the identification performance of DNA barcodes with broader samples is analyzed using our Bayesian statistical method, the segregating sites algorithm (Lou and Golding, 2010), in a population stucture model based on isolation by distance. To investigate the efficacy of barcoding in species with population substructure, we simulated sequences based on three parameters: a sampling scheme of reference sequences (to represent differences in the number and location of dispersed samples among demes), rates of migration between demes, and times to divergence between species. For various combinations of these parameters, we examined the probability that a query sequence originates from each species as calculated by the segregating sites algorithm. As an application, the same testing procedure was carried out with cytochrome c oxidase, subunit 1 (*CO1*) sequences of the genus *Grammia* (Lepidoptera: Noctuidae). The tiger moth species of this genus provides a good case study where classical morpho- and ecological traits do not agree with species groupings based on mitochondrial DNA (mtDNA). As 54% of the sampled species share haplotypes with at least one other species, under the barcoding gap criterion that no overlap between intra- and interspecies divergences be present, this would result in incorrect diagnoses for 32% of the species (Schmidt and Sperling, 2008). Both our simulated results and the empirical findings show that including at least one dispersed sample can aid sequence identification, even with recently diverged species and that including more dispersed samples further improves these results.

Our results highlight the importance of considering population subdivision and gene flow to the barcoding workflow, particularly for species known to have wide distribution ranges, and to sample broadly whenever possible to ensure that representative samples that contribute to describing the species boundary are included. Minimally, the results show that a single extra sample from another locality goes a long way to ensure accuracy.

## 3.3 METHODS AND DATA

### 3.3.1 Population spatial substructure

The simulation is based on an isolation by distance population model where every individual is restricted in its local movement to neighbouring demes (two-dimensional movement within a $d \times d$ square lattice where $d \times d$ represents the number of demes). Therefore, individuals are much more closely related to nearby individuals than to distant individuals. Let $\theta = 4N_e\mu$ be the population mutation rate ($\mu$ is the mutation rate per locus per generation), $M = 4N_em$ be the symmetric migration rate between demes ($m$ is the proportion of the population that migrates between two demes per generation) and $N_e$ is the effective population size. All demes are assumed to be of constant and equal size. The taxonomy of the reference sequence data is assumed to be correct.

### 3.3.2 Coalescent model with population substructure

Let each species exist within its own lattice and let each deme within the lattice contain any number of sampled sequences from the species. In generating a coalescent history of the lineages, the occurrence of a coalescent or migration event depends on where the lineages exist on the lattice. The probability of a coalescent event is more likely if many lineages are found within the same deme; otherwise a migration event is more likely. Coalescent theory with a consideration for population structure is well developed. The times until a coalescent or migration event are exponentially distributed with means:

$$I_{coal} = d\sum_{i=1}^{d} \frac{k_i(k_i-1)}{2}$$

and

$$I_{migr} = \frac{Mdk}{2}$$

respectively (where $k_i$ represents the number of lineages in deme $i$ and $k = \sum_{i=1}^{d} k_i$ is the total number of lineages in all demes; Hein, Schierup and Wiuf 2005).

The sum of the above two, $I_{coal} + I_{migr}$, represents the total rate until the occurrence of an event and the probability that the next event is a coalescent event or migration event is:

$$\frac{I_{coal}}{I_{migr} + I_{coal}} = \frac{\sum_{i=1}^{d} k_i^2 - k}{k(M-1) + \sum_{i=1}^{d} k_i^2}$$

and

$$\frac{I_{migr}}{I_{migr} + I_{coal}} = \frac{kM}{k(M-1) + \sum_{i=1}^{d} k_i^2}$$

respectively. For further details, refer to Hein, Schierup and Wiuf (2005).

### 3.3.3 Simulation

We simulated a multi-species coalescent (Degnan and Rosenberg, 2009), based on a total of ten species. Each species has five sampled sequences. Five is the recommended minimum by the Consortium for the Barcode of Life (CBOL) (Hajibabaei *et al.*, 2007) and via simulation study (Ross, Murugan and Li, 2008). The first species has one additional sampled sequence, which is used as the unknown query sequence. Each of the remaining nine species are progressively more and more distant from the first species (in a pectinate or asymmetric pattern). Other patterns were simulated with qualitatively similar results. These simulations permit incomplete lineage sorting but do not address introgression. Two lineages can coalesce only if their sequences exist within the same deme. Going back in time, the lineages will coalesce at a rate determined by the population size and migration rates. At a predetermined time, $T$, speciation is assumed to occur. At this time, lineages of either species, whether coalesced to a single ancestor or not, are randomly placed on this new lattice and thereafter treated as a single species. This process is repeated until a full coalescent history of all ten species is obtained.

Once the full coalescent is constructed, random substitutions are placed on the

branches of the coalescent, according to the rate $\theta$, and the resulting sequence data at the leaves are taken as the simulated data.

Given 51 simulated reference sequence data, the query sequence is removed from the reference data set and it, along with the remaining simulated sequences, are tested by the segregating sites algorithm (Lou and Golding, 2010) to determine the probabilities of origin (`PrOr`) for the query from each of the ten species. We have previously shown that the segregating sites algorithm can reliably assign unknown specimens even in the absence of a barcoding gap (a separation between intra- and interspecific variation).

We hypothesize that the probability that the query sequence originated from the first species should be greater when at least one or more dispersed sequences are included in the analysis. A sequence from the correct species but located in a spatially distinct deme adds important intraspecific variation that would not be obtained if all the reference samples originate from a single deme. At a minimum, the number of simulations where the `PrOr` is highest for the first species should be at least equal to the number where the first species is monophyletic. This should represent a minimum expectation.

### 3.3.4 Simulated data

The sampling scheme of reference sequences on the lattice, the number of demes, time to coalescence, and rates of migration are allowed to vary. We set the DNA sequence length equal to 600 bp, $\theta$ to 2.0, and modelled ten species, each represented by 5 lineages. The sequence length chosen is approximately the length of the 648 bp barcoding region (Hebert *et al.*, 2004) and the level of sequence variation ($\theta$) was chosen to be sufficient so that it mimics a marker like *CO1* that is able to discriminate at the species level and yet remain relatively conserved given its indispensable role in energy production (Capaldi, 1990). Simulations show that the number of alleles sampled per locus does not have a significant effect on the time to coalescents that exhibit reciprocal monophyly (Hudson and Coyne, 2002; Knowles and Carstens, 2007).

Table 3.1 shows the various sampling schemes for the reference species; all configurations are placed in a lattice containing 4 demes ($2 \times 2$) or 9 demes ($3 \times 3$). The scheme `all` represents a sampling situation where all the reference sequences are from one deme or base region. The schemes `1other` and `2other` represent situations where one or two dispersed samples, respectively, are included in the reference species. We were also interested in the effect of a larger lattice or sampling

Table 3.1: Lattice sampling schemes analyzed. Each `r` represents a reference sequence belonging to the species from which the query sequence, `Q`, originates. By default, $d \times d$ = 4 while the suffix 'L' sets $d \times d$ = 9 (see `all_L, 1other_L, 2other_L`).

| Sampling scheme | Lattice layout | Description |
|---|---|---|
| all | r r<br>r r r<br>Q | `all` reference sequences from one, base, sampling region and query, `Q`, in region furthest from the base |
| 1other | r<br>r r r<br>Q | One dispersed reference sequence adjacent to the base region |
| 2other | r<br>r r<br>r<br>r Q | Two dispersed reference sequences in independent regions, adjacent to the base region |
| all_L | r r<br>r r r<br><br>Q | `all` reference sequences from base region; $d \times d$ = 9 |
| 1other_L | r<br>r r<br>r r<br><br>Q r | One dispersed reference sequence from a region furthest from the base; $d \times d$ = 9 |
| 2other_L | r<br>r r<br>r<br>r<br>Q r | Two dispersed reference sequences in independent regions: one is in the center deme of the lattice and the other is in from a region furthest from the base; $d \times d$ = 9 |
| withQ | r<br>r r r<br>r Q | One dispersed reference sequence in the same region as the query |
| Qcloser_1other | r<br>r r r<br>Q | One dispersed reference and query sequence in independent regions, adjacent to the base region |
| Qcloser_2other | r<br>r r<br>r<br>Q r | Two dispersed reference and query sequences in independent regions, adjacent to the base region; the query is closer to the base region |

area (`all_L`, `1other_L` and `2other_L` where $d \times d$ = 9). We also investigated the effect of query placement, relative to the base region (`Qcloser_1other` and `Qcloser_2other`) and, lastly, a configuration where a reference and query sample originate from the same deme (`withQ`).

The time to speciation (backward-in-time), $T$, was set to 10 and 3 (scaled in units of $2N_e d * d$ generations). The symmetric rate of migration, $M$, ranged from 0.1 to 1000 ($M$ up to 10 shown here), to model different rates of movement among demes within a lattice. When the time to speciation is long and the migration rate is high, the lineages within each species should coalesce with each other first and the level of variation within a species should be less than between species; this represents the ideal situation where each population is a distinct and monophyletic species (Avise, 1989), and we expect most of these simulations to have `PrOr` largest for the first species. When the time to speciation is short or the migration rate is low, there will be more incomplete lineage sorting and this would result in a lower proportion of the simulations where the `PrOr` is largest for the first species. Each combination of parameters are based on 10,000 simulation runs.

Since these simulations are conducted within a statistical framework, we have the advantage of not only identifying correct assignments but also those that occur with high confidence. Thus, to be conservative, we additionally considered analyses of simulations where the `PrOr` is $\geq$ 80%. Due to difficulties with obtaining simulations that satisfied this criterion, these results are based on 100 simulation runs. The difficulty arises because conspecific lineages will take a long time to coalesce if they are spread among many demes that seldom migrate when the migration rate is low, thereby increasing the chance of paraphyletic coalescents (Wakeley, 2000).

### 3.3.5 Empirical data: *Grammia*

Species of the *Grammia* genus have a large geographic range, exhibit interspecific hybridization and incomplete lineage sorting, making them an ideal data set to explore the use of dispersed samples on assignment fidelity. Of several *Grammia* species for which sequence information is available, we chose *Grammia nevadensis* as our focal species because of the paraphyly of its lineages with those from most of the species in the Western clade. It has been widely sampled from 16 locations spanning several provinces of Canada and northwestern US states (Schmidt and Sperling, 2008; Schmidt, 2009).

All 225 *Grammia* sequences from 33 species (Schmidt and Sperling, 2008; Schmidt, 2009) were downloaded from NCBI. Species represented by at least 5

Table 3.2: Summary of *CO1* data for 12 *Grammia* species (Schmidt and Sperling, 2008; Schmidt, 2009) and for *Holarctia obliterata* which served as an outgroup.

| Species | Monophyletic | Sequences |
|---|---|---|
| *Grammia arge* | yes | 5 |
| *Grammia celia* | no | 5 |
| *Grammia figurata* | no | 11 |
| *Grammia nevadensis* | no | 18 |
| *Grammia ornata* | no | 9 |
| *Grammia parthenice* | no | 13 |
| *Grammia phyllira* | yes | 7 |
| *Grammia quenseli* | no | 10 |
| *Grammia virgo* | no | 9 |
| *Grammia virguncula* | no | 37 |
| *Grammia williamsii* | no | 44 |
| *Grammia williamsii tooele* | no | 6 |
| *Holarctia obliterata* | yes | 5 |
| **Total** | - | 179 |

individuals were kept for further analysis. This criterion limited our reference data set to 179 *Grammia* sequences from 13 species (Table 3.2). For sampling scheme `all`, *G. nevadensis* contained sequences only from British Columbia, and the query was chosen to be from Utah. Dispersed sequences for schemes `1other-5other` are sampled from two provinces in Canada (Alberta, Saskatchewan) and three states in the US (Washington, Oregon, Colorado). The inclusion of one or more dispersed sequence(s) was compensated by a reduction of sequences from British Columbia to maintain a total of five reference sequences for the species.

## 3.4 RESULTS

### 3.4.1 Simulation

Using the segregating sites algorithm, an assignment is considered correct when the `PrOr` is highest for the first species. A multi-species coalescent consisting of distinct and monophyletic species should possess sufficient divergence within and among species to permit the correct assignment of the query to the first species. So we expect a higher proportion of correct assignments when a monophyletic coalescent is recovered for the first species relative to a coalescent that is paraphyletic and includes sequences from other species (i.e., when the time to speciation is short and when the migration rate is low). In other words, the proportion of correct assign-

ments should be at least equal to the proportion of monophyletic trees for the first species.

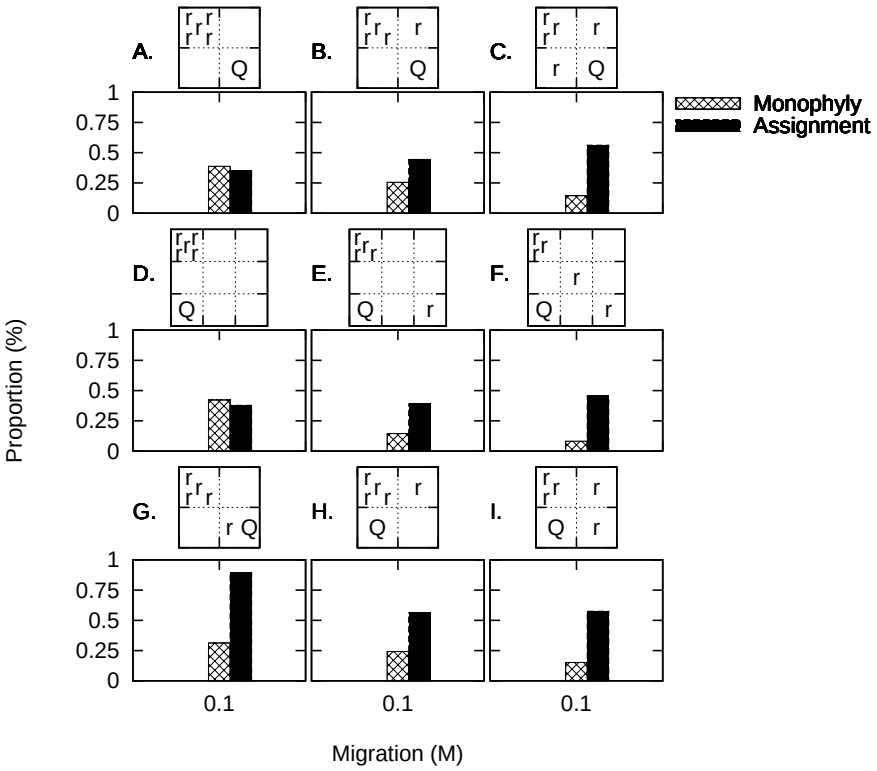



Figure 3.1: Inclusion of dispersed samples aids correct identification with recently diverged species. Each histogram is based on 10,000 simulations when $T$ = 3.0 and $M$ = 0.1. Each subfigure, A-I, has a specific placement of reference (r) and query (Q) sequences for the correct species (Table 3.1). Monophyly represents the proportion of monophyletic coalescents for the first species (double-hatched bars). Assignment represents the proportion of correct assignments where the query assigned to the first species (solid bars).

**Effect of population subdivision.** Here we focus on the results of simulations where the sampling is largely restricted to one deme (`all`) and the species boundaries are not yet clearly distinct ($T = 3.0$). When the migration rate is low ($M = 0.1$; Figure 3.1A), the proportion of correct assignments, 35%, is less than the proportion of simulations with the first species monophyletic, 39%. This indicates that

at this level of divergence and migration the assignment of an unknown query sequence is difficult and/or misleading when the reference sequences of the species are sampled from just one location. In a larger sampling area, the distance between the base region and the query is larger and we would expect a decrease in the proportion of correct assignments reflecting the lengthened time required for coalescence. While the result is approximately the same (38% correct assignments versus 42% monophyletic coalescents; Figure 3.1D), additional simulations with an even larger lattice ($d \times d$ = 16) confirmed the prediction (data not shown).

When one dispersed sequence was included in the reference dataset (`1other`, `1other_L` and `Qcloser_1other`; Table 3.1), the number of simulations where `PrOr` was largest for the first species increased whether sampled in a small (44% vs 26%; Figure 3.1B) or large sampling area (39% vs 14%; Figure 3.1E) or when the query is sampled closer to the base region (56% vs 24%; Figure 3.1H).

When two dispersed sequences were included in the reference dataset (`2other`, `2other_L` and `Qcloser_2other`; Table 3.1), the proportion of simulations with `PrOr` largest for the first species increased further: in a small (56% vs 14%; Figure 3.1C) or large sampling area (46% vs 8%; Figure 3.1F) and when the query is closer to the base region (57% vs 15%; Figure 3.1I).

The sampling scenario that returned the highest proportion of correct assignments (90% vs 31%) is when the query and a reference sample are sampled from the same deme (Figure 3.1G).

**Effect of migration.** To examine the effect of migration, simulations were repeated with 10-fold increases in the rate of migration. A high migration rate allows lineages to move with greater ease among the demes of a lattice. Consequently, conspecific lineages will coalesce sooner and in turn increase the chances of monophyly.

When the rate of migration was $\geq 1$, close to 100% of the simulations were monophyletic for the first species and the `PrOr` was, correctly, largest for the first species (Figures 3.2 and 3.3). If the rates of migration are large and there is sufficient variation to distinguish species then employing a comprehensive sampling scheme is not necessary.

**Conservative assessments.** When we restrict our analyses to simulations where the `PrOr` is strongly supported (`PrOr` $\geq$ 80%), there is a large increase in the number of correct assignments when discrete species are considered (from 46%

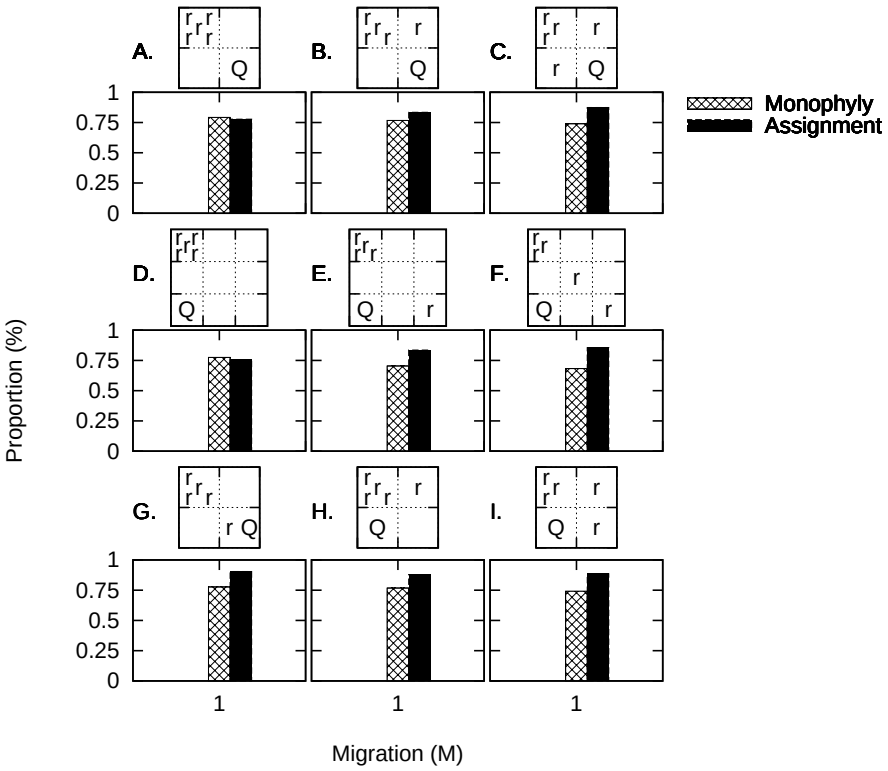



Figure 3.2: Inclusion of dispersed samples aids correct identification with recently diverged species. Each histogram is based on 10,000 simulations when $T$ = 3.0 and $M$ = 1. See figure 3.1 for simulation and legend details.

(A) to 83% (C), Figure 3.4 (L)ower; $T$ = 10) but the effect disappears for newly divergent species (Figure 3.4 (L)ower; $T$ = 3).

Sampling schemes with two dispersed sequences often did not return any simulation runs where the `PrOr` was larger than 80%. (Figure 3.4 C,F, and I of (L)ower; $T$ = 3). When the migration rate is low and the time to speciation is short, any additional variation at some point cannot compensate for the increased levels of paraphyly. However, for every sampling scheme, the number of simulations where the `PrOr` is largest for the first species increase relative to the amount of monophyly. It is simply that the degree of certainty for these has been reduced.

**Performance.** We wanted to investigate how the probability changes as the com-

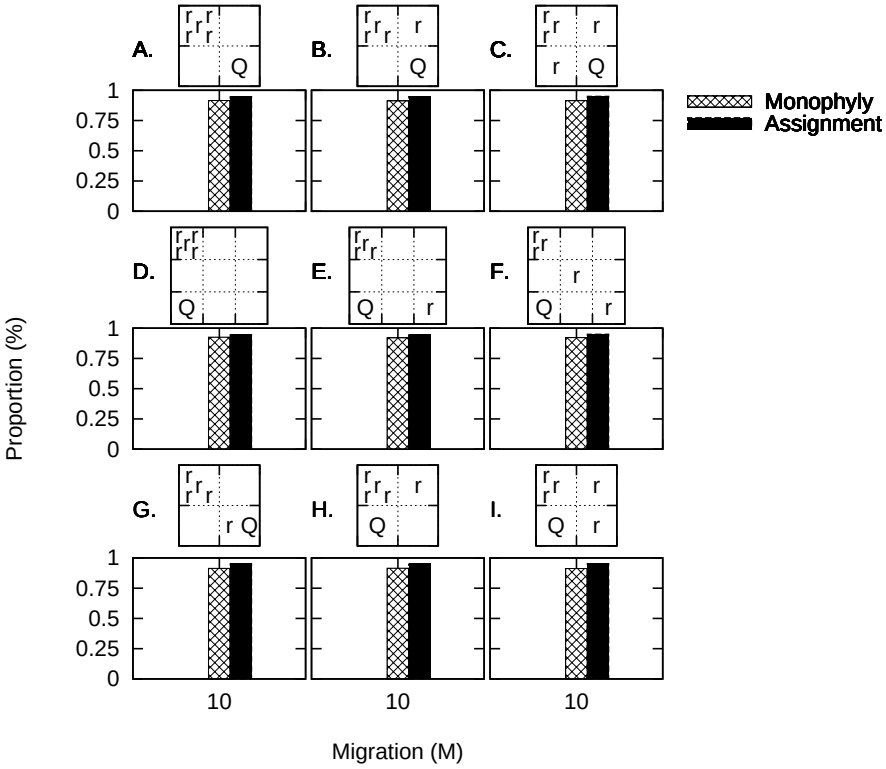



Figure 3.3: Inclusion of dispersed samples aids correct identification with recently diverged species. Each histogram is based on 10,000 simulations when $T$ = 3.0 and $M$ = 10. See figure 3.1 for simulation and legend details.

position of the first species is changed to more accurately reflect the variation of species with a wide distribution range when both gene flow and the time to speciation are low ($M$ = 0.1 and $T$ = 3.0 respectively). To do this, we began with a simulation in which all sequences from the first species are restricted to one deme (`all`; Table 3.1) and then repetitively change the sequence composition to increasingly reflect a species with a wider distribution (that is, all the sequences are randomly dispersed on the lattice, each individual in its own deme). Each simulation was repeated 10,000 times. Performance is measured as the ratio of the number of simulations observed where `PrOr` is largest to the first species relative to the number where the first species is monophyletic.

As expected, as more dispersed sequences are included, there is a decrease in monophyletic coalescents with a corresponding increase in the number where `PrOr`

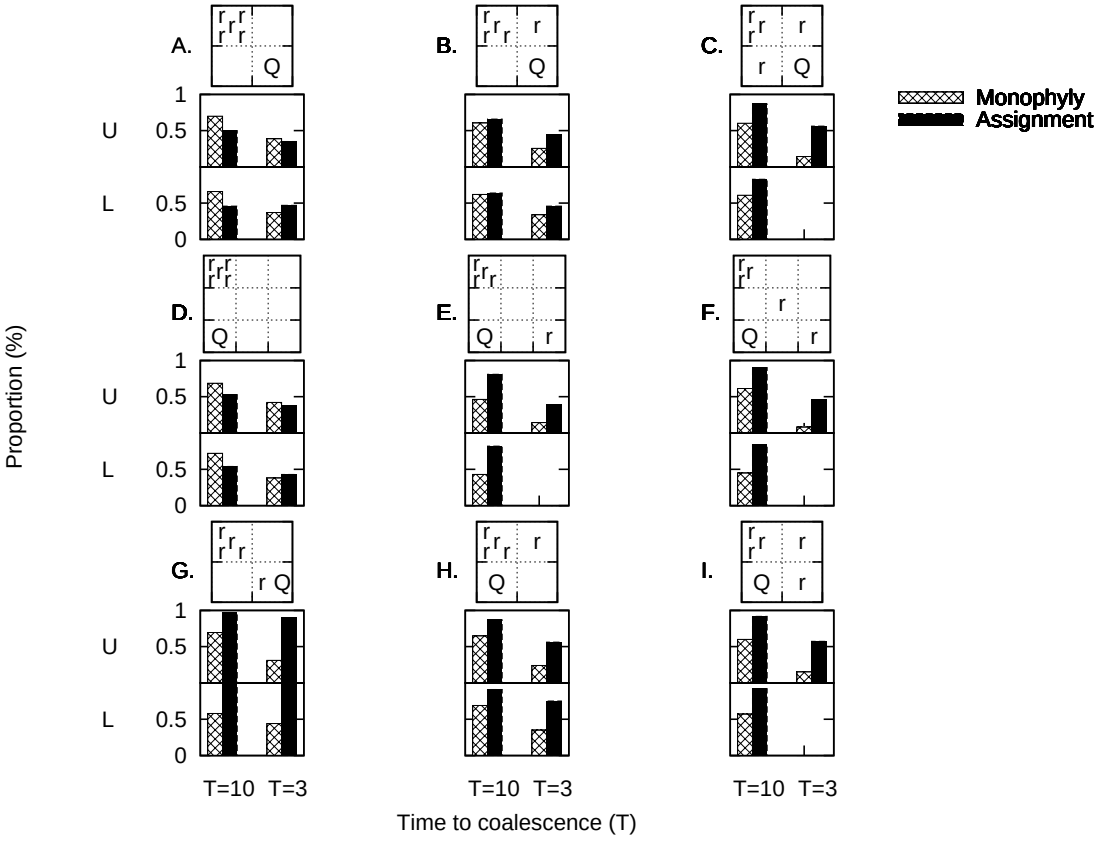



Figure 3.4: Simulations based on $M = 0.1$ and $T$ is scaled in units of $2N_e d * d$ generations. The upper row of histograms (U) is based on 10,000 simulations and the lower row of histograms (L) is based on 100 simulations where the `PrOr` is $\geq$ 80%. Monophyly represents the proportion of monophyletic coalescents for the first species (double-hatched bars). Assignment represents the proportion of correct assignments where the query assigned to the first species (solid bars). The inclusion of dispersed samples aids correct identification with high confidence. This is not confirmed for recently diverged species (C,E,F,I) because of a lack of high confidence assignments due to a higher level of paraphyly.

is largest for the first species. This strongly supports the use of dispersed sequences to form the reference datasets. However, when the correct species is entirely composed of dispersed sequences, the performance decreases from 15 to 13 correct assignments/monophyletic tree (Figure 3.5, number of dispersed samples = 5) suggesting that the number of correct assignments returned cannot be expected to do

much better when the correct species consists entirely of dispersed sequences because of the greater level of paraphyly.

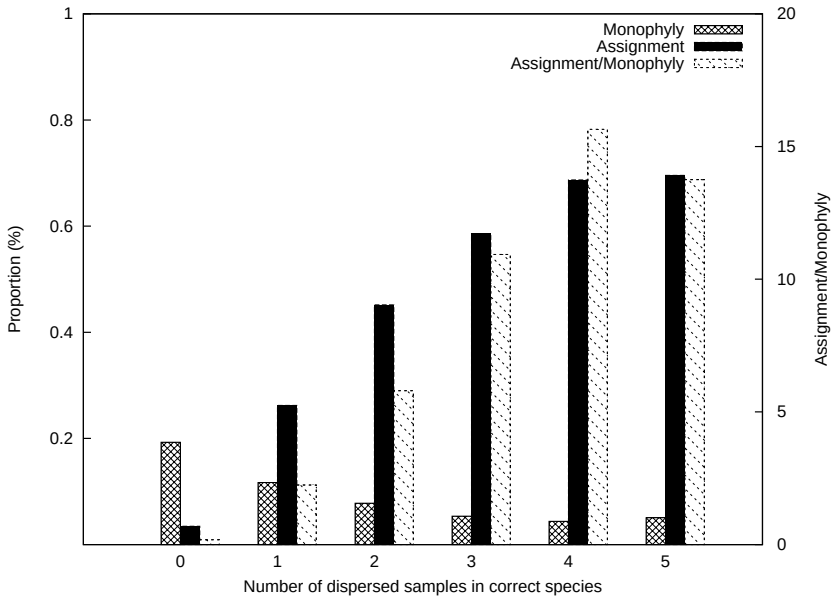



Figure 3.5: Increasing performance in assignment when the correct species is composed of more dispersed sequences. Each histogram is based on 10,000 simulations when $M$ = 0.1 and $T$ = 3.0. When the correct species is entirely composed of dispersed sequences, performance decreases because there is a greater level of paraphyly.

### 3.4.2 *Grammia* (Tiger moth) example

Our analysis of all possible assignments concerning the composition of *G. nevadensis* (Figure 3.6) shows that the probability of correctly assigning the query increases when at least two dispersed sequences are included (Table 3.3, columns '% Max' and 'Max P(CA)').

Low probabilities are largely attributed to the extensive paraphyly among western *Grammia* species and, to a lesser extent, the nature of the segregating sites algorithm which calculates high probabilities of assignment to distantly related taxa if they have extensive sequence variation. As expected, the Utah *G. nevadensis* query sequence had high `PrOr` to species found in the Western *Grammia* haplogroup. However, the query consistently had the highest `PrOr` to *Grammia williamsii*. There is extensive sequence variation among the 50 *G. williamsi* specimens that

broadly span the US, with some sequences in both Western and Eastern haplotype clades (Schmidt, 2009). Among the 23 haplotypes, some are unique to subspecies *G. williamsii tooele*, some share haplotypes with a few Eastern clade species, and some have introgressed with other species (Schmidt and Sperling, 2008; Schmidt, 2009). Because of its hyper-variability, *G. williamsi* acts as a single, morphological species that is capable of generating a coalescent that includes any query. In all cases, however, the statistical risk is always lowest for *G. nevadensis* (Table 3.4). Minimal statistical risk (a metric included in the segregating sites algorithm) to *G. nevadensis* suggests that the 'loss' of assigning the query to *G. nevadensis*, given that it could assign to other species, is small and that it is the species of origin (see Abdo and Golding, 2007, for a further details).

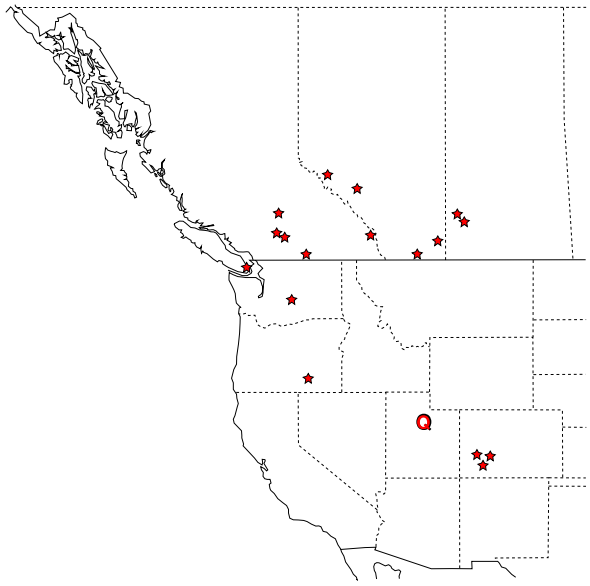



Figure 3.6: Geographical locations of *G. nevadensis* samples (stars) and query (Q). Samples are from (with number of sequences in parentheses) British Columbia (6), Alberta (4), Saskatchewan (2), Washington (1), Oregon (1), Colorado (3).

## 3.5 DISSCUSSION

Barcoding with the mitochondrial *CO1* gene sequence has been successful in many groups of animals (Hebert *et al.*, 2004) but has proved less successful in some other groups (Meyer and Paulay, 2005; Monaghan *et al.*, 2005). The problem is the lack of correspondence between sequence-delimited groups and taxonomically recog-

Table 3.3: Including dispersed sequences for the correct species increases the number of correct assignments. See footnote for details.

| Sampling scheme | Total | P(CA) | % Max | Max P(CA) | Mis-assigned to | Avg P |
|---|---|---|---|---|---|---|
| `all` | 1 | 0.111 | 0 | 0.000 | *G. williamsii* | 0.174 |
| `1other` | 66 | 0.015 | 0 | 0.000 | *G. williamsii* | 0.192 |
| `2other` | 825 | 0.142 | 59 | 0.178 | *G. williamsii* | 0.178 |
| `3other` | 3300 | 0.158 | 50 | 0.173 | *G. williamsii* | 0.168 |
| `4other` | 4950 | 0.161 | 52 | 0.170 | *G. williamsii* | 0.166 |
| `5other` | 2772 | 0.162 | 63 | 0.168 | *G. williamsii* | 0.166 |

Values are based on assignment among 10 species.
In `all`, the composition of the correct species, *G. nevadensis*, contains 6 samples strictly from British Columbia. The composition of *G. nevadensis* is modified to contain one or more dispersed samples (`1other-5other`). The dispersed samples are sampled from two provinces in Canada (Alberta, Saskatchewan) and three states in the US (Washington, Oregon, Colorado). The query is from Utah. See figure 3.6 for the geographical locations of the sampled sequences and the query.
`Total` indicates the total number of possible combinations for assignment. Correct assignment (CA) indicates assignments to the first species (*G. nevadensis*). P(CA) is the average `PrOr` from *G. nevadensis*. % Max is the proportion of assignments when P(CA) is the largest for *G. nevadensis*. Max P(CA) is the average P(CA) when it is the largest for *G. nevadensis*. If the query is incorrectly assigned, the incorrect species (Mis-assigned to) and the average `PrOr` from this species (Avg P) is given.

nized species. This lack of agreement is attributable to a variety of phenomena such as incomplete lineage sorting (Hudson and Coyne, 2002), allopatric speciation (Coyne and Orr, 2004), gene- and species-tree discordance (Funk and Omland, 2003) and the criteria used to determine species boundaries (Mayr, 1942). Other practical problems include incomplete reference databases with insufficient within-species sampling, which is required for accurate species authentication (Meyer and Paulay, 2005; Siddall and Budinoff, 2005).

When gene flow is high and when species divergence times are large, methods to classify sequences to species groups should be relatively straightforward. However, when gene flow is low and species divergence times are small the ability to correctly classify sequences will then be impaired. In these situations, methods that determine the posterior probability that the sequence originates from each species become critical.

We would expect that the number of simulations with the highest probability of

Table 3.4: For all assignments, correct or incorrect, the statistical risk of assigning to the correct species is always the lowest. This suggests that the query originates from the correct species. See footnote for details.

| Sampling scheme | Total | Risk(CA) | % Min | Min risk(CA) |
|---|---|---|---|---|
| `all` | 1 | 0.007 | 100 | 0.007 |
| `1other` | 66 | 0.011 | 100 | 0.011 |
| `2other` | 825 | 0.011 | 100 | 0.011 |
| `3other` | 3300 | 0.012 | 100 | 0.012 |
| `4other` | 4950 | 0.013 | 100 | 0.013 |
| `5other` | 2772 | 0.013 | 100 | 0.013 |

See table 3.3 for simulation details.
`Total` indicates the total number of possible combinations for assignment. Correct assignment (CA) indicates assignment to the first species (*G. nevadensis*). Risk(CA) is the average statistical risk of assignment to *G. nevadensis*. % Min is the proportion of assignments when Risk(CA) is the lowest for *G. nevadensis*. Min risk(CA) is the average Risk(CA) of assignment when it is the lowest for *G. nevadensis*.

origin (`PrOr`) to the first species should be at least equal to or larger than the number with a monophyletic relationship among the reference sequences. However, our results suggest that this is not always true. Whenever, significant population subdivision exists and reference sequences have not been collected from different demes, a query sequence from a different deme will appear sufficiently different from the reference sequences to prevent correct identification (Figure 3.1A). On the other hand, adding just a single reference sequence from a divergent deme can reverse this, and the `PrOr` will be higher than the proportion of monophyletic reference species (Figure 3.1B). This is a result of the increased estimate of conspecific variation as represented by increased $\theta$ values. Continuing to add more samples from divergent demes further improves the relative ratios (Figure 3.5).

The tiger moth species of the *Grammia* genus are an example of a group with geographically widespread populations connected by gene flow. Despite extensive non-monophyly among these species, the `PrOr` from *G. nevadensis* increased when the database contained dispersed samples spanning the geographic locales between the base sampling region (British Columbia) and the origin of the query (Utah) (Figure 3.6).

Both the simulation and empirical results suggest that the success of mtDNA barcodes depend on sufficient reference sequences that are representative of the

within-species variation and when it is undersampled, from substructured genetic variation (population subdivision) or newly divergent species or both, inaccurate species identifications and delimitations may result. This reflects the requirement from traditional taxonomy to ensure that sufficient variation is sampled in order to determine if characters are taxonomically useful (DeWalt, 2011; Trewick, 2007; Wong *et al.*, 2011). Thus, mtDNA sequence is a valuable tool but only with a comprehensive database consisting of complete conspecific reference sequences, especially from species with wide geographical distributions or that have recently diverged, and our study attests to the need for methods to consider adequate representation of the natural variation within the species.

Futhermore, accurate species delimitations have important implications in the development of proper guidelines and policies used to manage and protect both biodiversity and consumer interests. This includes areas such as, but not limited to, conservation and disease biology and aquaculture.

Our methods could be expanded to allow the coalescent-speciation transitions to vary in space (e.g., along different branches of the tree; Monaghan *et al.*, 2009) and in time (e.g., unsampled lineages in demes that have gone extinct; Lohse, 2009). The examination of peripheral populations is of particular importance for recently speciated groups. We also assumed that lineages migrate in a discrete and symmetric fashion but it would be more accurate to model continuous movement among demes. A recent method by Lemey *et al.* (2010) uses a continuous spatial diffusion model to identify the ancestral geographical history of a sample of sequences but may not be applicable in our simulations since it is not meant to infer population-based spatial histories (Bloomquist, Lemey and Suchard, 2010). Although the current model is limited in these respects, it is sufficient to illustrate how broad sampling of within-species divergence is essential for accurate barcoding identifications, how this variation affects identifications, and that even minimal sampling goes a long way.

While it is important to include singletons (species described by a single sample; Lim, Balke and Meier, 2012) in the biodiversity inventory, a singleton cannot capture any of the variation or complexity of a species (Ross, Murugan and Li, 2008). This variation is critical for any population genetic method, such as segregating sites algorithm, that describes the conspecific variation via a summary statistic ($\theta$) and this calculation requires multiple samples. For this reason, singletons are excluded from the reference data set used here. However, despite their exclusion, if extraneous information is used to estimate this variation then a Bayesian method such as the segregating sites algorithm algorithm should be able to identify queries that originate from singletons in the reference database. One such source of extraneous

information might be to assume that the singleton species has a level of variation ($\theta$) equal to that of sibling species. At the other end of the sampling size spectrum, Bergsten *et al.* (2012) recommended a minimum of 20 samples per species for any sampling strategy. However, the authors also note that the choice of the identification algorithm will determine acceptable sample sizes, identification performance, and error rates. Furthermore, Zhang *et al.* (2010) found that a universal sample size is unrealistic for different species and that it ultimately depends on the evolutionary history of the species. By evaluting the segregating sites algorithm via simulations, we assess its general performance across a range of evolutionary scenarios without particular focus on the *CO1* gene and we find that while more samples will provide better results, a large improvement in the number of correct assignments can be achieved with even a single dispersed sample from a total of five samples per species.

## 3.6 CONCLUSION

Using the segregating sites algorithm and a minimum five samples per species, both simulated and *Grammia* (tiger moth) analyses show that ensuring at least one reference sequence is sampled from a different region or deme of a species distribution returns a greater proportion of results that correctly assign an unknown specimen to its species of origin. Our results highlight the importance of broad sampling to improve the information content of reference samples and that a single dispersed sample can greatly improve the identification of sequences to species.

## 3.7 ACKNOWLEDGEMENTS

We thank Dr. Richard Morton and Dr. Jonathan Dushoff for revision and helpful comments. This work was supported by grants from NSERC and Genome Canada.

# Chapter 4

# An extended theory of segregating sites: effect of subdivided populations and heterogeneous substitution rates

## 4.1 ABSTRACT

Often there is a disconnect between the assumptions made by a model and the true evolutionary signals of the data it is applied to. For instance, current theory assumes that the pattern of segregating sites sufficiently describes the observed level of variation in a set of sequences. However the pattern may be influenced by various phenomena that are unaccounted for, such as population subdivision with gene flow and unequal base composition from transition bias. This study seeks to improve the theory of segregating sites by incorporating terms to account for these biological proccesses. A more comprehensive model should improve probability estimates of the observed level of genetic diversity. The modified probability distributions (of observing a number of segregating sites in a number of sequences) are similar but more accurate at resolving the true distribution of genetic variability relative to those calculated under the original theory. Additionally, the results reinforce the important role subdivided populations with migration and heterogeneous base composition and substitution rates have on shaping polymorphism and should be considered in models used to describe genetic signals of groups undergoing speciation.

## 4.2 INTRODUCTION

The pattern of variation that permits discrimination between species is produced by dynamic processes, which may be influenced by many factors. And each of the many available methods of the barcoding initiative capitalizes on one or more genetic signatures useful for the accurate assignment of specimens to species. The central assumption across all strategies for molecular species recognition is that within-species individuals are more similar to each other than to individuals from other species. However, barcoding failures usually stem from a violation of this basic assumption. For instance, genetic distances and thresholds, also known as the "barcoding gap" (a region defined by the maximum level of intra- versus minimum level of inter-specific variation) and the criterion of reciprocal monophyly (a group of sequences or individuals under one species name or forming their own clade to the exclusion of others) are arbitrary and a lack of either property does not preclude speciation (Hickerson, Meyer and Moritz, 2006; Meier, 2008; Ross, Murugan and Li, 2008; Virgilio *et al.*, 2010). Microevolution (constant change in within-species variation; Funk and Omland, 2003), the reliance on a reference tree (Little, 2011) and lack of informative molecular characters (Hudson and Coyne, 2002) limit the use of character-based methods.

Often, the lack of correspondence between the data and model occurs because the model is too simple: it fails to sufficiently describe complex biological events governing levels of genetic variation. Both Avise (1992) and Soltis *et al.* (2006) found that species exhibit distinct intraspecific (within-species) mitochondrial DNA (mtDNA) patterns associated with geography. A dynamic evolutionary history, consisting of repeated colonizations, extinctions, periods of isolation in refugia, spatial and ecological barriers, can give rise to regional, species-specific genetic patterns (Soltis *et al.*, 2006; Trewick, 2007; Lohse, 2009; Tavares, de Kroon and Baker, 2010; Carr *et al.*, 2011). Thus, failing to sample from each locale in describing a geographically dispersed species may lead to biased barcoding inferences. For instance, failure to include samples from multiple bio-localities substantially underestimated the level of variation found within a species, reducing the "barcoding gap" and correct specimen identifications (Lukhtanov *et al.*, 2009; Zhou *et al.*, 2011). And when conspecific variation is properly sampled, accurate barcoding rates have been achieved (Allcock *et al.*, 2011; Zhou *et al.*, 2009; Tavares, de Kroon and Baker, 2010; Zhou *et al.*, 2010, 2011; Pappalardo *et al.*, 2011). In a previous study, the inclusion of at least one dispersed sequence not only increased the number of correct specimen assignments but also increased both the proportion of segregating sites in a set of sequences and the estimated population mutation rate, $\theta$ (Lou and Golding, 2012). The inferred $\theta$ values were unusually high because the method does not

attribute the added intraspecfic variation as originating from a distinct geographic locale or deme.

Different types of nucleotide substitutions may have different rates, allowing for different patterns of variation to be observed in different species and this may negatively impact barcoding inferences (Yang, 1996; Roe and Sperling, 2007). In mtDNA, the number of polymorphisms that result from transitions will be higher than those that are transversions, but this excess is only evident in very closely related or recently diverged species. For instance, all (100%) 7 substitutions of 112 COI base positions are transition differences between two closely related members of the *Equus* genus, the extant mountain zebra and extinct quagga (Higuchi *et al.*, 1984). With increased sequence divergence, the proportion of observed transition differences is expected to decrease. This has been observed in invertebrate COI (fruit flies, the spruce budworm pest, and ground bettles; Satta, Ishiwa and Chigusa, 1987; Sperling and Hickey, 1994; Martinez-Navarro, Galian and Serrano, 2005) as well as other genetic segments (e.g., COII, NADH dehydrogenase, ribosomal RNA) and organisms (e.g., insects, rat versus both cow and human and between cow and mouse; Brown *et al.*, 1982; Brown and Simpson, 1982; DeSalle *et al.*, 1987; Liu and Beckenbach, 1992). The disappearance of transitions (or accumulation of transversions) may be due to multiple substitutions at the same site, perhaps resulting in eventual saturation (no change in sequence divergence despite increasing time or a poor signal-to-noise ratio) (Brown *et al.*, 1982). It is possible that the observed base difference may not have been achieved in one step and may be the result of two or more base substitutions (an observed transition may mask, previous, multiple tranversions or is the result of transversions that have erased themselves; Holmquist, 1983). Using COII data, Jukes (1987) showed that most of the accumulated transversions occurred at previously unsubstituted sites (also observed in other genes and organisms) and multiple hits and saturation is likely neglible in recently diverged or closely related species. The lack of correspondence between observed patterns and true substitution mechanisms at work suggest that the modelling of heterogeneous base composition, from a high proportion of transitions, is warranted.

Since both population subdivision with gene flow and heterogeneous base composition from transition bias influence the evolutionary history and genetic signatures in data, the theory of segregating sites should be modified to reflect the influence of these forces on the observed level of variation. The modified model should provide a more realistic representation of the evolutionary dynamics of species and improve the metric for accurate species identification.

Here we attempt to develop an extended mathematical model of the theory of

segregating sites under the joint effect of population structure with migration and unequal base composition and substitution rates in mtDNA. The modified model should produce more realistic probability estimates of the genetic variability in a set of sequences (relative to the original model). The model will compare the structure of variation observed in 1 to 3 subdivided populations and when segregating sites are categorized as transitions and transversions.

We will first briefly describe the original theory of segregating sites. This is followed by a description of the model for population subdivision with migration and the modified recursion equations adopting this model. This description format is repeated for modelling transitions and transversions. Finally the modified probabilty distributions of the genetic diversity expected under the influence of each biological phenomenon are presented and discussed.

## 4.3 THEORY

### 4.3.1 A review of the basic theory of segregating sites

The goal is to assign an unknown DNA sequence, $x$, to the correct taxonomic group, $k$. The probability of this assignment (`PrOr`) is

$$Pr(x \in k|x, D, \theta)$$

where $D$ is a database of known sequences with distinct taxonomic groups and $\theta$ (= $4N_e\mu$) is a known collection of population mutation rates. The assignment of sequence $x$ must be made to one of the taxonomic groups.

Each taxonomic group is represented by $n$ sequences. According to the theory of segregating sites, $\theta$ is reflected in the number of segregating sites, $s$, between a set of sequences. Using Bayes rules, `PrOr` is calculated as:

$$Pr(x \in k|x, D, \theta) \sim \frac{Pr(s_k|n_k, x \in k, \theta_k)/Pr(s_k|\theta_k)}{\sum_j Pr(s_j|n_j, x \in j, \theta_j)/Pr(s_j|\theta_j)} \tag{4.1}$$

where the probability of membership of the unknown sequence $x$ to taxonomic group $k$ is

$$Pr(s_k|n_k, x \in k, \theta_k)$$

Following Lou and Golding (2010), the basic recursive definition for the proba-

bility that a sample of $n_k$ sequences will have $s_k$ segregating sites is:

$$Pr(s_k|n_k, x \in k, \theta_k) = \frac{\theta_k}{\theta_k + n_k - 1} Pr(s_k - 1|n_k, x \in k, \theta_k) + \frac{n_k - 1}{\theta_k + n_k - 1} Pr(s_k|n_k - 1, x \in k, \theta_k) \quad (4.2)$$

This recursion assumes an infinite sites model (Kimura, 1969), the samples are equilibrium single random mating populations of size $N_e$ and $\mu$ mutations occur per locus per generation.

### 4.3.2 Modelling population spatial substructure

The methodology will follow Golding (1984, 2002) and Lou and Golding (2010) to calculate equilibrum recursion equations to describe the number of segregating sites between 2 or more populations.

To extend the theory to the number of segregating sites that might be obtained in a sample that originates from a subdivided population, each taxonomic group, $k$, has its sample of sequences, $N_k$, divided into a total of $d$ subpopulations or demes. For each deme $i$, $N_i$ is the number of diploid sequences that undergo random mating internally and $n_i$ is the number of sampled sequences.

Let $m_{st}$ designate the probability of migration, per generation, from deme $s$ to deme $t$. We will assume that migration is reciprocal or symmetric ($\forall s \forall t : m_{st} \equiv m_{ts}$) and that it is irreducible (no isolated subsets of demes).

To maintain constant population sizes, within each deme, over time, it is necessary that the migration parameters satisfy a detailed balance,

$$\sum_t m_{st} N_s = \sum_r m_{rs} N_r$$

such that the number leaving is equal to the number entering the $s^{th}$ deme. Generations are assumed to be discrete and non-overlapping. The model is depicted in figure 4.1.

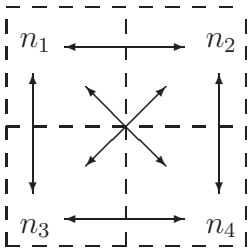



Figure 4.1: A subdivided population represented as a lattice (dashed lines) with $d$ demes (e.g., four demes). Each deme consists of $N_i$ diploid sequences and each sequence is permitted to migrate to and from each deme at an equal rate of $m$ sequences/deme/generation. Let $n_i$ be the number of sampled sequences from deme $i$.

### 4.3.3 Modified theory of segregating sites with population spatial substructure

Equilibrium recursion equations can be derived to describe the number of expected segregating sites between 2 or more populations from one generation to the next.

Terms denoted by a prime represent an equivalent probability in the next generation. The population substructure of sequences is denoted within square brackets; for example, $Pr(s|[2,1])$ denotes a total of three sequences (2 are found in the first subpopulation and 1 is found in second subpopulation).

$$\begin{aligned} Pr(s|\mathbf{n})' = (1 - \sum_i n_i\mu - \sum_i n_i \sum_{j\neq i} m_{ji} - \sum_i n_i(n_i-1)\frac{1}{4N_i}) \cdot Pr(s|\mathbf{n}) \\ + \sum_i n_i\mu \cdot Pr(s-1|\mathbf{n}) \\ + \sum_i n_i \sum_{j\neq i} m_{ji} \cdot Pr(s|...n_i - 1, n_j + 1, ...) \\ + \sum_i n_i(n_i-1)\frac{1}{4N_i} \cdot Pr(s|...n_i - 1, ...) \end{aligned} \quad (4.3)$$

If $N \equiv N_i$, $m \equiv m_{ji}$ then at equilibrium,

$$
\begin{aligned}
Pr(s|\mathbf{n}) = (1 - \sum n_i\mu - \sum n_i m - \sum n_i(n_i-1)\frac{1}{4N}) \cdot Pr(s|\mathbf{n}) \\
+ \sum n_i\mu \cdot Pr(s-1|\mathbf{n}) \\
+ \sum n_i m \cdot \sum_j Pr(s|...n_i - 1, n_j + 1, ...) \\
+ \sum n_i(n_i-1)\frac{1}{4N} \cdot Pr(s|...n_i - 1, ...)
\end{aligned} \tag{4.4}
$$

Combining like terms on the left side,

$$
\begin{aligned}
(\sum n_i\mu + \sum n_i m + \sum n_i(n_i-1)\frac{1}{4N}) \cdot Pr(s|\mathbf{n}) \\
= \sum n_i\mu \cdot Pr(s-1|\mathbf{n}) \\
+ \sum n_i m \cdot \sum_j Pr(s|...n_i - 1, n_j + 1, ...) \\
+ \sum n_i(n_i-1)\frac{1}{4N} \cdot Pr(s|...n_i - 1, ...)
\end{aligned} \tag{4.5}
$$

If $\theta = 4N\mu$ and $M = 4Nm$ then,

$$
\begin{aligned}
(\sum n_i\theta + \sum n_i M + \sum n_i(n_i-1)) \cdot Pr(s|\mathbf{n}) \\
= \sum n_i\theta \cdot Pr(s-1|\mathbf{n}) \\
+ \sum n_i M \cdot \sum_j Pr(s|...n_i - 1, n_j + 1, ...) \\
+ \sum n_i(n_i-1) \cdot Pr(s|...n_i - 1, ...)
\end{aligned} \tag{4.6}
$$

and the probability that a sample of $n$ sequences will have $s$ segregating sites in a subdivided population with migration is:

$$Pr(s|\mathbf{n}) = \frac{\sum n_i\theta}{\sum n_i\theta + \sum n_i M + \sum n_i(n_i-1)} \cdot Pr(s-1|\mathbf{n}) + \frac{\sum n_i M}{\sum n_i\theta + \sum n_i M + \sum n_i(n_i-1)} \cdot \sum_j Pr(s|...n_i-1, n_j+1, ...) + \frac{\sum n_i(n_i-1)}{\sum n_i\theta + \sum n_i M + \sum n_i(n_i-1)} \cdot Pr(s|...n_i-1, ...) \quad (4.7)$$

Note that when $\mathbf{n} = n_1 = n$, with all other $n_i = 0$ and with $m = 0$, equation (4.3) reduces to,

$$Pr(s|n)' = (1 - n\mu - n(n-1)\frac{1}{4N}) \cdot Pr(s|n) + n\mu \cdot Pr(s-1|n) + n(n-1)\frac{1}{4N} \cdot Pr(s|n-1) \quad (4.8)$$

and if $\theta = 4N\mu$, $M = 4Nm$ then at equilibrium,

$$(\theta + n - 1) \cdot Pr(s|n) = \theta \cdot Pr(s-1|n) + (n-1) \cdot Pr(s|n-1) \quad (4.9)$$

or

$$Pr(s|n) = \frac{\theta}{\theta + n - 1} \cdot Pr(s-1|n) + \frac{n-1}{\theta + n - 1} \cdot Pr(s|n-1) \quad (4.10)$$

which is the familiar recursion equation for the number of segregating sites in a single population (4.2).

### 4.3.4 Modelling transitions and tranversions

The number of segregating sites may be categorized into two or more substitutional categories. Here we will consider transitions and transversions. However, more information may be captured by categorizing even further: transitions may be described as interchanges between purines (A and G) or pyrimidines (C and T). And, if not symmetric, purine-purine interchanges may be recorded as 'A-to-G' or 'G-to-A' interchanges. Each category is described by a unique rate of substitution. The most appropriate number and type of categories to use depends on the level of resolution desired.

### 4.3.5 Modified theory of segregating sites considering heterogeneous substitution rates

For a sample of $n$ sequences, let $S = [s_{xy}]$ be a matrix of observed differences (segregating sites) where $s_{xy}$ is a difference that records a change from nucleotide $x$ to nucleotide $y$ and

$$x, y = \{A, G, C, T\}$$

For example, $[s_{xy} \equiv 0]$ indicates 0 segregating sites; $[s_{AG} = 3]$ indicates 3 segregating sites (A $\rightarrow$ G); and $[s_{AG} = 3, s_{AT} = 2]$ indicates 5 segregating sites (3 of A $\rightarrow$ G and 2 of A $\rightarrow$ T).

The proportion of segregating sites is categorized by all possible nucleotide interchanges and each is characterized by a unique substitution rate, $\mu_{xy}$.

$$\begin{aligned} Pr(S|n)' = (1 - n\sum_{x}\sum_{x\neq y}\mu_{xy} - n(n-1)\frac{1}{4N}) \cdot Pr(S|n) \\ + n\sum_{x}\sum_{x\neq y}\mu_{xy} \cdot Pr([s_{xy} - 1]|n) \\ + n(n-1)\frac{1}{4N} \cdot Pr(S|n-1) \end{aligned} \tag{4.11}$$

Let $\mu_S$ be a matrix of the rates of substitution for each category. Then let $\sum_x \sum_{x\neq y} \mu_{xy} \equiv \mu_S$,

$$\begin{aligned} Pr(S|n)' = (1 - n\mu_S - n(n-1)\frac{1}{4N}) \cdot Pr(S|n) \\ + n\mu_S \cdot Pr([s_{xy} - 1]|n) \\ + n(n-1)\frac{1}{4N} \cdot Pr(S|n-1) \end{aligned} \tag{4.12}$$

If $\mu_S \equiv \mu$ (the rates of substitution for each category are equal) then at equilibrium we have (4.2).

More generally, assuming symmetry (between interchanges of the same bases, such as A2G = G2A), let $\alpha$ and $\beta$ represent the substitution rate for transitions, $P$ (purine-purine, pyrimidine-pyrimidine interchanges), and transversions, $Q$ (purine-pyrimidine interchanges), respectively.

$$\begin{aligned} Pr(S|n)' = (1 - n\alpha - n\beta - n(n-1)\frac{1}{4N}) \cdot Pr(S|n) \\ + n\alpha \cdot Pr([s_P - 1]|n) \\ + n\beta \cdot Pr[(s_Q - 1]|n) \\ + n(n-1)\frac{1}{4N} \cdot Pr(S|n-1) \end{aligned} \tag{4.13}$$

Assuming equilibrium and combining like terms on the left side,

$$\begin{aligned} (\alpha + \beta + (n-1)\frac{1}{4N}) \cdot Pr(S|n) = \alpha \cdot Pr([s_P - 1]|n) \\ + \beta \cdot Pr[(s_Q - 1]|n) \\ + (n-1)\frac{1}{4N} \cdot Pr(S|n-1) \end{aligned} \tag{4.14}$$

Let $\theta_\alpha = 4N\alpha$ and $\theta_\beta = 4N\beta$,

$$(\theta_\alpha + \theta_\beta + n - 1) \cdot Pr(S|n) = \theta_\alpha \cdot Pr([s_P - 1]|n) + \theta_\beta \cdot Pr([s_Q - 1]|n) + n - 1 \cdot Pr(S|n-1) \quad (4.15)$$

and the probability that a sample of $n$ sequences will have $s$ segregating sites partitioned into two substitution categories is:

$$Pr(S|n) = \frac{\theta_\alpha}{\theta_\alpha + \theta_\beta + n - 1} \cdot Pr([s_P - 1]|n) + \frac{\theta_\beta}{\theta_\alpha + \theta_\beta + n - 1} \cdot Pr([s_Q - 1]|n) + \frac{n-1}{\theta_\alpha + \theta_\beta + n - 1} \cdot Pr(S|n-1) \quad (4.16)$$

## 4.4 RESULTS

### 4.4.1 Expectations of a modified theory of segregating sites

The equilibrium recursion equations calculate the probability of obtaining a certain number of segregating sites in samples from subdivided populations with migration and samples that exhibit heterogeneous substitution rates.

In the absence of population substructure with migration and heterogeneous substitution rates, using equation (4.2) and assuming $\theta$ = 2.0, the probabilities of 0 and 1 segregating sites in 2 sequences are $Pr(0|2)$ = 0.33 (Figure 4.2A, $\theta$=2.0) and $Pr(1|2)$ = 0.22 (Figure 4.2B, $\theta$=2.0) respectively.

Given population substructure and migration, when the rate of migration is zero, it is expected that the probability of obtaining a number of segregating sites should be the same as the probability seen in a population without substructure. At intermediate rates of migration, that are neither zero nor infinite, the probability of observing a number of segregating sites differs depending on the population structure of the sampled sequences. Among sequences sampled from the same population (e.g., [2,0]), the probability should decrease to reflect the decreasing chance that all the observed variation originates from sequences in one deme in light of migration.

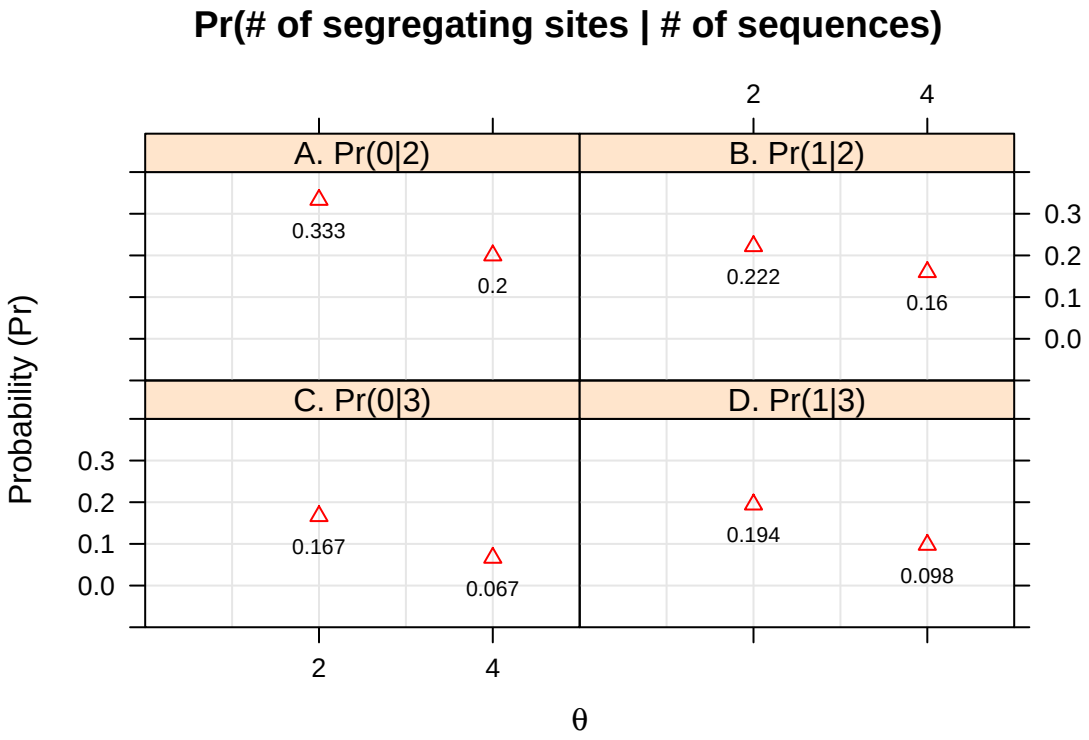



Figure 4.2: Probability of $s$ segregating sites in $n$ sequences when $\theta$ = 2.0 and 4.0. The probability values are calculated using equation (4.2).

Among sequences sampled from 2 or more populations (e.g., [2,1]), the probability of variation should increase to reflect the greater chance that the variation may stem from sequences found in different demes. When the rate of migration is very large or infinite, the probability of variation reflects a metapopulation that acts as a single population. Thus, the probability of variation is expected to reflect a larger number of sampled sequences.

In the case of unequal base substitutions, in the most general case of categorizing the proportion of segregating sites as transitions and tranversions, the modified probability value reflects the chance that the observed variation is due to different combinations of each substitution type; this is accomplished by incorporating different rates for each substitution type. For example, $\theta$ is then expressed as $\theta_\alpha$ + $\theta_\beta$. Then the probabilities of 0 and 1 segregating sites, based on the original and modified recursion equations, are:

$$Pr(0|2) = \frac{1}{\theta + 1} \qquad Pr(1|2) = \frac{\theta}{\theta + 1} \cdot Pr(0|2)$$

$$Pr([0]|2) = \frac{1}{\theta_\alpha + \theta_\beta + 1} \qquad Pr([1]|2) = \frac{\theta_\alpha}{\theta_\alpha + \theta_\beta + 1} \cdot Pr(0|2) + \frac{\theta_\beta}{\theta_\alpha + \theta_\beta + 1} \cdot Pr(0|2)$$

respectively, where $S = [s_{xy}] = [s]$ = [$s_P$, $s_Q$] is a matrix of observed segregating sites partitioned into transitions, $P$, and tranversions, $Q$. While the probability values of the modified and original recursion equations remain the same, the left and right terms, or component probabilities, of $Pr([1]|2)$ should differ depending on the number and substitution rate of each type of substitution. In general, probability of a transition should be greater if the observed number of transitions are greater than the number of transversions and vice versa.

### 4.4.2 Effect of population spatial substructure and migration

We focus on the results where sequences are sampled from a total of 2 subpopulations.

**Sequences in 1 subpopulation**

**Migration is zero.** The probabilities of 0 and 1 segregating sites in 2 sequences in 1 (of 2) subpopulation are $Pr(0|[2,0])$ = 0.33 and $Pr(1|[2,0])$ = 0.22, respectively (Figure 4.3A and B, $M \sim 0$). The samples or individuals remain in their respective

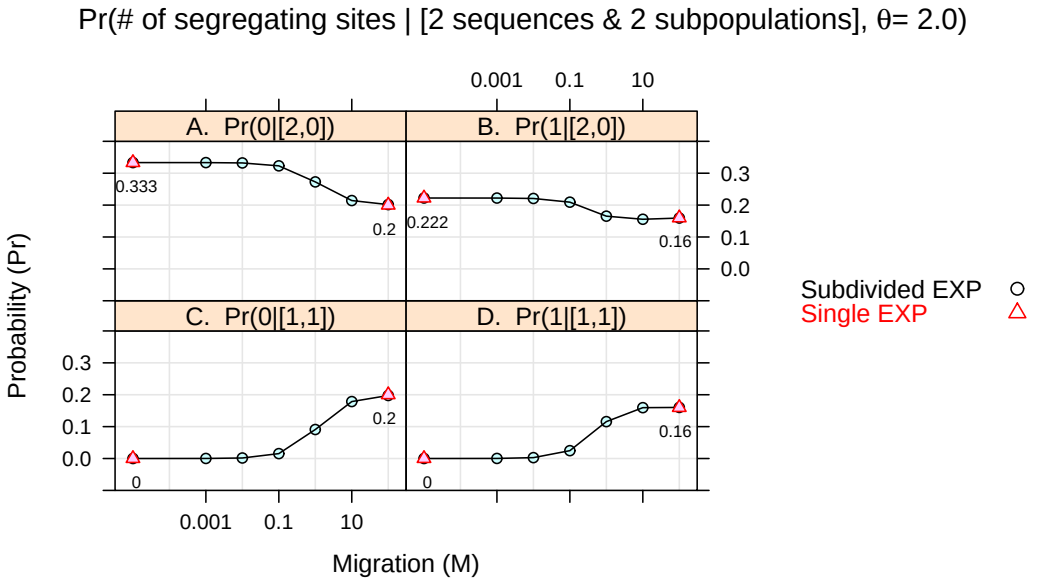



Figure 4.3: Probability of $s$ segregating sites in 2 sequences from 2 subpopulations when $0.0001 \geq M \geq 100$. The probability values are calculated using equation (4.7).

demes and the 2 sequences found in the same subpopulation acts as a single population with no hidden substructure. So all of the variation is entirely attributed to the sequences in that 1 subpopulation. As expected, the results provided by our recursion equations agree with theoretical probability values in the absence of population substructure ($\triangle$ in figure 4.3A and B, $M \sim 0$, correspond to those in figure 4.2A and B, $\theta$ = 2.0).

**Intermediate migration.** When the migration rate is greater than zero but less than being infinitely large (i.e., $0 < M < \infty$) the probability decreases for 2 sequences (Figure 4.3A and B). As the rate of migration increases, samples are permitted to move among subpopulations more easily. Thus, the probability of observing segregating sites in a sample of sequences decreases because the variation may originate from samples in other demes. Thus, the decrease in probability suggests that the number of segregating sites in sequences of 1 subpopulation is less likely if variation is permitted to come from other subpopulations via migration.

**Migration is very large.** When migration is very large (i.e., $M = 100$), the probabilities of observing 0 and 1 segregating sites in 2 sequences in 1 (of 2) subpopulations are $Pr(0|[2,0])$ = 0.20 and $Pr(1|[2,0])$ = 0.16, respectively (Figure 4.3A and B, $M = 100$). Sequences are permitted to move easily between subpopulations,

allowing the metapopulation to act as a single population but it is characterized by more individuals or a higher level of variation (i.e., doubling of $\theta$ to 4.0). So it is not surprising that the results from our modified recursion equations should and do approach the values expected from equations describing a single population ($\triangle$ correspond to those in figure 4.2A and B, $\theta$ = 4.0).

**Sequences in more than 1 subpopulation**

**Migration is zero.** Sampling 1 sequence from each of 2 (of 2) subpopulations, the probabilities of 0 and 1 segregating sites are $Pr(0|[1,1])$ = 0 and $Pr(1|[1,1])$ = 0, respectively (Figure 4.3C and D, $M \sim 0$). At equilibrium, when the sampled sequences are in different demes, the probability of observing segregating sites is 0 because segregating sites cannot be observed among sequences in different demes. Provided that we do not consider an infinite number of sequences, the result is true for any number of sequences (Figure 4.4C and D, $M \sim 0$).

**Intermediate migration.** When sequences are spread out among several demes, an increasing migration rate permits mixture among samples or individuals confined to distinct demes. Since we expect that there is a greater chance that the variation is attributed to samples from distinct demes, with increasing migration rate, there should be a corresponding increase in the probability of observing segregating sites among sequences found in subdivided populations and this is what is observed (Figures 4.3C and D).

**Migration is very large.** Whether sequences are in 1 subpopulation (i.e., [2,0]) or both (i.e., [1,1]), the probabilities are the same (Figure 4.3, $M = 100$) and agree with the theoretical outcome of a metapopulation acting as a single population (Figure 4.2A and B, $\theta$ = 4.0).

**Increasing the number of sequences sampled from subdivided populations**

The probability patterns of 0 and 1 segregating sites observed for 3 sampled sequences from 2 subdivided populations are similar to those observed for 2 sampled sequences but smaller (Figure 4.4). Across increasing migration rates, assuming $\theta = 2.0$, the probabilities are consistently smaller than probabilities based on 2 sequences because observing 0 and 1 segregating sites in more sequences is less likely.

**Increasing the number of subdivided populations**

When the total number of subpopulations increase, the pattern of probability

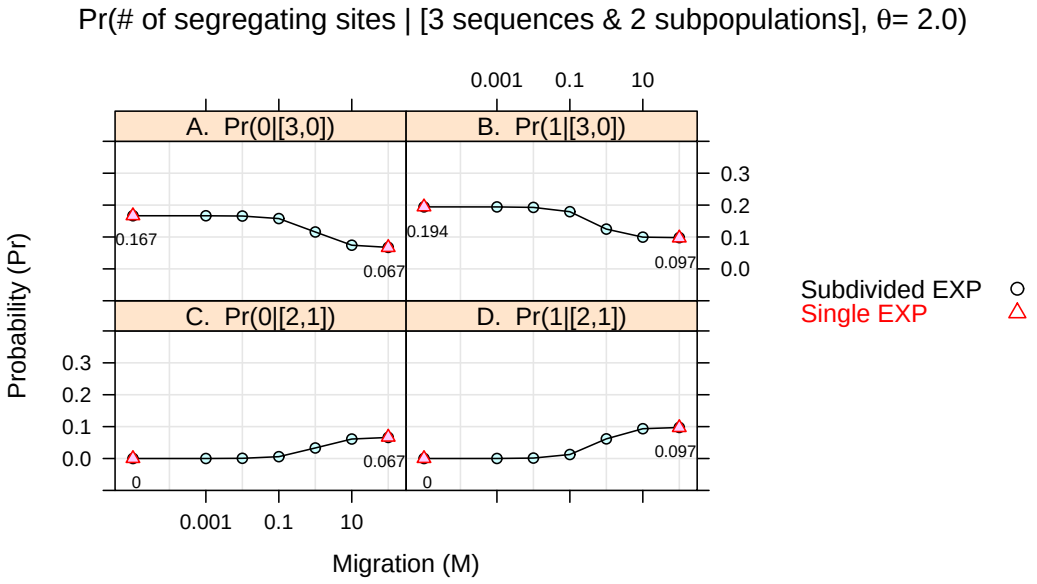



Figure 4.4: Probability of $s$ segregating sites in 3 sequences from 2 subpopulations when $0.0001 \geq M \geq 100$. The probability values are calculated using equation (4.7).

values expected with 2 sampled sequences in 3 subpopulations should be similar to the pattern observed in 2 subpopulations. For example, the trend of probability values for $Pr(0|[2,0,0])$ are similar to those seen for $Pr(0|[2,0])$ except that the slope declines steeply as the migration rate becomes very large (Figure 4.5A versus Figure 4.3A). With the addition of a third subpopulation, when the migration rate is very large, the probability of observing 0 segregating sites is smaller because it is expected that, in 3 subpopulations, the variation will reflect a larger number of sampled sequences (that is, $3\theta$). The same reasoning can be applied to $Pr(0|[1,1,0])$ except that its probability slope increases mildly as the migration rate becomes very large (Figure 4.5B versus Figure 4.3C).

**Probability distribution of the modified theory of segregating sites**

We examine the probability distributions of the number of segregating sites in samples from a single or subdivided population.

To avoid ambiguity, definitions of several terms used to describe a probability distribution are provided:

**tail** region of least frequent occurring values in a distribution

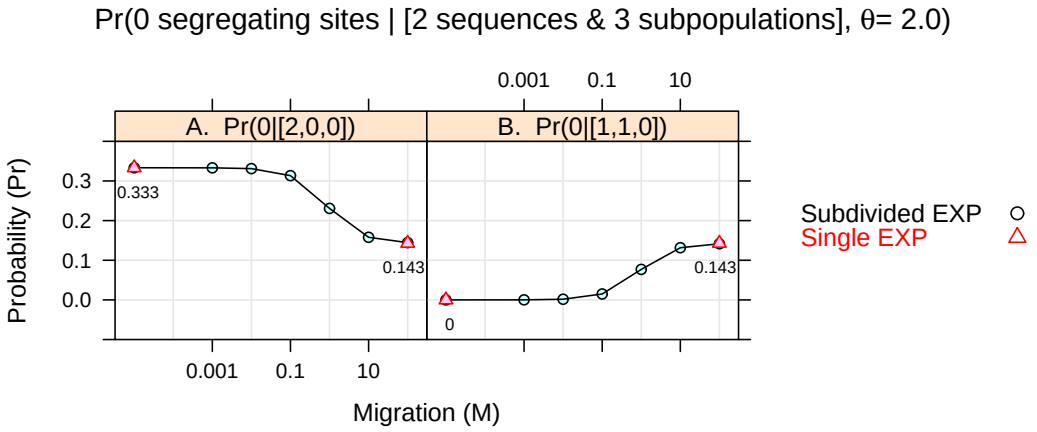



Figure 4.5: Probability of 0 segregating sites in 2 sequences from 3 subpopulations when $0.0001 \geq M \geq 100$. The probability values are calculated using equation (4.7).

**skew** measure of asymmetry of the probability distribution

**negative skew** left tail of the probability distribution is longer than the right (i.e., there are few low values) and the bulk of the mass (values) lie to the right of the distribution. Synonymous terms include left-skewed, left-tailed, and skewed to the left.

**positive skew** right tail of the probability distribution is longer than the left (i.e., there are few high values) and the bulk of the mass lie to the left of the distribution. Synonymous terms include right-skewed, right-tailed, and skewed to the right.

**Sequences in 1 subpopulation.** For 2 sequences, when the migration rate is zero, the mass of the distribution is concentrated on the left (i.e., positively skewed). In other words, the probability decreases as the number of segregating sites increases. The trend is true and the probability values are the same for sequences sampled from a single or subdivided population (Figure 4.6A, C M=0.0001, ◯).

When migration is very large, the probability distribution is positively skewed for samples from a single and subdivided population. But, relative to a single population, the probabilities from the subdivided data set are lower when s>=3 (Figure 4.6A, C, M=100, △).

**Sequences in more than 1 subpopulation.** As expected, when migration is near zero, the probability distribution is left-tailed (i.e., the probability is 1 when the

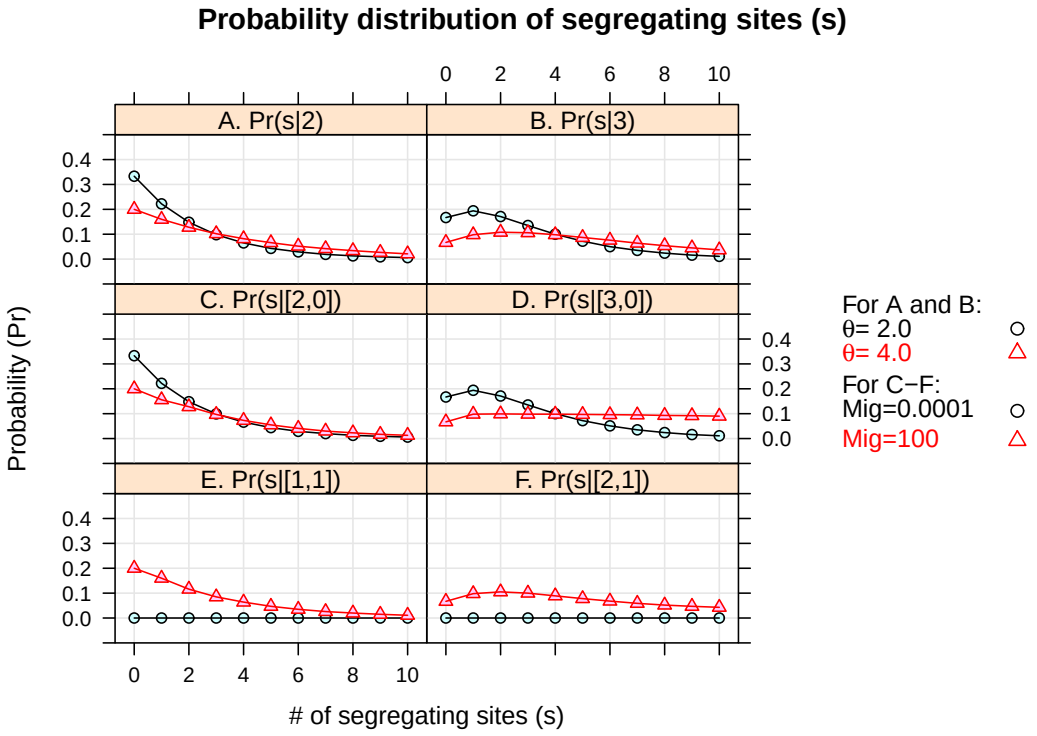



Figure 4.6: Probability distributions of $s$ segregating sites in $n$ sequences without (A and B) and with (C-F) subdivided populations when $0.0001 \geq M \geq 100$. The probability values are calculated using equations (4.2) and (4.7)

number of segregating sites is infinite; Figure 4.6E, M=0.0001, ◯).

When migration is very large, the probability distribution is similar, but with lower values, relative to the positive skew seen with samples in 1 subpopulation (Figure 4.6E, M=100, △).

**More sequences in 1 subpopulation.** For 3 sequences, the probability distribution mass is, similarly, positively skewed but slightly peaks when there are 1 and 2 segregating sites. This trend is the same for sequences from a single or subdivided population (Figure 4.6B, D M=0.0001, ◯).

When migration is very large in a single population, the peak of the positively-skewed distribution occurs when there are 2 segregating sites (Figure 4.6B, s=2, M=100, △). However, with population substructure, the peak is relatively flat (Figure 4.6D, M=100, △). When there are greater than 5 segregating sites, the probabilities from a subdivided population are greater (i.e., $Pr(10|3) = 0.037 < Pr(10|[3,0]) = 0.090$; Figure 4.6B vs D, s=10, M=100, △).

**More sequences in more than 1 subpopulation.** Similar to $Pr(s|[1,1])$, when the migration rate is zero, the tail of the probability is longer on the left and the probability is 1 when the number of segregating sites is infinite (Figure 4.6F, M=0.0001, ◯).

When the migration is very large, the probability distribution is the same as $Pr(s|3)$ and $Pr(s|[3,0])$ when there are 0 and 1 segregating sites, intermediate when 2<=s<=3 (e.g., $Pr(2|3) = 0.106 < Pr(2|[2,1]) = 0.105 < Pr(2|[3,0]) = 0.098$), smaller when 4<=s<=8 (e.g., $Pr(8|3) = 0.054 > Pr(8|[2,1]) = 0.052 < Pr(8|[3,0]) = 0.093$), and intermediate when s>8 (e.g., $Pr(10|3) = 0.037 < Pr(10|[2,1]) = 0.043 < Pr(10|[3,0]) = 0.090$) (Figure 4.6F, M=100, △).

### 4.4.3 Effect of heterogeneous transition and transversion rates in mtDNA

The original recursion equations (4.2) give the following probabilities that a sample of $n$ sequences has 0 to $s$ segregating sites:

$$Pr(s|n) = \frac{\theta}{\theta+1} \cdot Pr(s-1|n) = \frac{\theta^s}{(\theta+1)^{s+1}} \tag{4.17}$$

For $n$ number of sequences, the sum of $Pr(s|n)$, where $s$ ranges from 0 to infinity, is 1.

Under the modified recursion equations (4.16), each probability is the sum of component probabilities and each component probability describes a possible arrangement of the observed segregating sites partitioned into categories. For $n$ sequences and two substitution categories, namely transitions and tranversions, $Pr([1]|n)$ is composed of two component probabilities: the probability that 1 segregating site is either a transition (i.e., $Pr([1,0]|n)$) or a transversion (i.e., $Pr([0,1]|n)$). The number of possible arrangements increases as the number of segregating sites increase. With two substitution categories, each increase in the number of segregating sites introduces an additional component probability. For instance, there are three possible arrangments for 2 segregating sites: both segregating sites are either 2 transitions ([2,0]) or 2 transversions ([0,2]) or 1 transition and 1 transversion ([1,1]).

The following equations describe the breakdown of $Pr([s]|n)$ into its component probabilities for observing up to 5 segregating sites in 2 sequences:

$$
\begin{aligned}
Pr([0]|2) &= Pr(0|2) \\
Pr([1]|2) &= Pr([1,0]|2) + Pr([0,1]|2) \\
Pr([2]|2) &= Pr([2,0]|2) + Pr([0,2]|2) + Pr([1,1]|2) \\
Pr([3]|2) &= Pr([3,0]|2) + Pr([0,3]|2) + Pr([1,2]|2) + Pr([2,1]|2) \\
Pr([4]|2) &= Pr([4,0]|2) + Pr([0,4]|2) + Pr([1,3]|2) + Pr([3,1]|2) + Pr([2,2]|2) \\
Pr([5]|2) &= Pr([5,0]|2) + Pr([0,5]|2) + Pr([1,4]|2) + Pr([4,1]|2) + Pr([2,3]|2) \\
&\quad + Pr([3,2]|2)
\end{aligned}
$$

In general, for $n$ number of sequences, the sum of $Pr([s]|n)$, where $[s]$ ranges from 0 to infinity, is 1.

At equilibrium, the probabilities that a sample of $n$ sequences has 0 to $[s]$ segregating sites are:

$$
Pr([s]|n) = \frac{\theta_\alpha + \theta_\beta}{\theta_\alpha + \theta_\beta + 1} \cdot Pr([s-1]|n) = \frac{(\theta_\alpha + \theta_\beta)^s}{(\theta_\alpha + \theta_\beta + 1)^{s+1}} \tag{4.18}
$$

A comparison of the solved probabilities for the original (4.17) and modified recursion equations (4.18) show that the probability values are the same. But the values for each component probability will depend on the substitution rates for each

category.

For instance, if $\theta_\alpha$ = 1, $\theta_\beta$ = 1, then

$$
\begin{aligned}
Pr([1]|2) &= \frac{\theta_\alpha}{\theta_\alpha + \theta_\beta + 1} \cdot Pr([0]|2) + \frac{\theta_\beta}{\theta_\alpha + \theta_\beta + 1} \cdot Pr([0]|2) \\
Pr([1]|2) &= \frac{1}{1+1+1} \cdot Pr([0]|2) + \frac{1}{1+1+1} \cdot Pr([0]|2) \\
Pr([1]|2) &= \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9} + \frac{1}{9} \\
Pr([1]|2) &= 0.111 + 0.111
\end{aligned}
$$

The first and second terms of $Pr([1]|2)$ are the probabilities that 1 segregating site is either a transition ($Pr([1,0]|2)$) and or transversion ($Pr([0,1]|2)$) in 2 sequences.

If $\theta_\alpha$ = 2.0, $\theta_\beta$ = 0.0, then

$$
\begin{aligned}
Pr([1]|2) &= \frac{2}{0+2+1} \cdot Pr([0]|2) + \frac{0}{0+2+1} \cdot Pr([0]|2) \\
Pr([1]|2) &= \frac{2}{3} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} \\
Pr([1]|2) &= \frac{2}{9} + 0 \\
Pr([1]|2) &= 0.222 + 0
\end{aligned}
$$

If $\theta_\alpha$ = 1.5, $\theta_\beta$ = 0.5, then

$$
\begin{aligned}
Pr([1]|2) &= \frac{1.5}{1.5+0.5+1} \cdot Pr([0]|2) + \frac{0.5}{1.5+0.5+1} \cdot Pr([0]|2) \\
Pr([1]|2) &= \frac{1.5}{3} \cdot \frac{1}{3} + \frac{0.5}{3} \cdot \frac{1}{3} \\
Pr([1]|2) &= \frac{1.5}{9} + \frac{0.5}{9} \\
Pr([1]|2) &= 0.167 + 0.0556
\end{aligned}
$$

The distribution for 2 component probabilities, $Pr([s,0])$ and $Pr([0,s])$, are shown in figure 4.7.

The shape of the distributions are similar to $Pr(s|n)$. For instance, for each

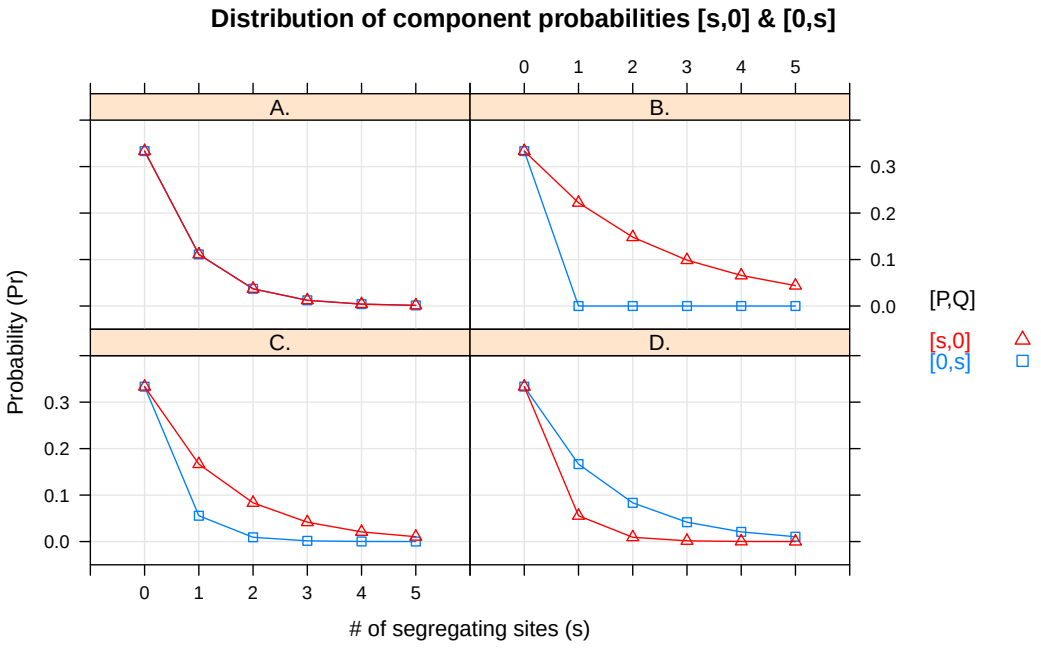



Figure 4.7: Component probabilities of $s$ segregating sites in 2 sequences with heterogeneous substitution rates. Let the substitution rates for transitions ($P$) and transversions ($Q$) be $\theta_\alpha$ and $\theta_\beta$ respectively. A. $\theta_\alpha = 1.0$ and $\theta_\beta = 1.0$, B. $\theta_\alpha = 2.0$ and $\theta_\beta = 0.0$, C. $\theta_\alpha = 1.5$ and $\theta_\beta = 0.5$, and D. $\theta_\alpha = 0.5$ and $\theta_\beta = 1.5$. When $\theta_\alpha = 1.0$ and $\theta_\beta = 1.0$ (A), the probabilities of observing transitions or transversions are equal. If $\theta_\alpha > \theta_\beta$ (B, C), the probability of observing a transition is greater than observing a transversion. If $\theta_\alpha < \theta_\beta$ (D), the probability of observing a transversion is greater than observing a transition.

unique pair of heterogeneous substitution rates, both $Pr([s,0]|2)$ and $Pr([0,2]|2)$ are right-tailed (that is, there are few high probability values as the number of segregating sites increase). A similar trend is observed for $Pr(s|2)$ (Figure 4.6A, M=0.0001, ◯).

The probability of observing 0 segregating sites ($Pr([0,0]|n)$) is independent of whether heterogeneous transition and tranversion rates are considered (e.g., $Pr(0|2)$ = $Pr([0,0]|2)$ = 0.333). With 1 or more segregating sites, the probability of observing a transition or transversion depends on the number and rate of each type of substitution. When $\theta_\alpha = 1.0$ and $\theta_\beta = 1.0$, the probability of 1 or more transitions ($Pr([s,0]|2)$) is identical to the chance of observing 1 or more transversions ($Pr([0,s]|2)$; Figure 4.7A △ vs □ respectively). When $\theta_\alpha > \theta_\beta$ and there are more observed transitions than transversions, the probability of observing a transition is higher than observing a transversion (Figure 4.7B and C). This is especially true when $\theta_\alpha$ = 2.0 and $\theta_\beta$ = 0.0; the chance that the variation originates from 1 or more transversions is 0 (Figure 4.7B □). When the transversion substitution rate is greater than the rate for transitions, the probability that the observed variation originates from transversions is greater than the latter (Figure 4.7D).

Similar, flattened, distribution patterns are observed when 1 or 2 segregating sites are fixed as transitions or transversions (Figures 4.8 and 4.9). The flattened distributions occur because the chance of observing higher levels of variation is smaller with 2 sequences. The flattened effect is stronger with 2 fixed transitions or transversions; the probability of observing variation is virtually 0 with 1 additional segregating site (Figure 4.9C).

## 4.5 DISCUSSION

An important and consistently used population genetic parameter to describe the diversity pattern of a set of sequences is $\theta$ and it is the product of the effective population size and mutation rate. The Watterson (1975) equation describes the relationship between $\theta$ and the expected number of segregating sites (total number of substitutions observed in the data).

Inaccurate $\theta$ estimates may result when the assumptions of the model are violated. For example, violations may arise from sequences sampled from geographically dispersed species or sequences that exhibit heterogeneous base composition from a transition bias.

This study investigated the effect of modelling population substructure with mi-

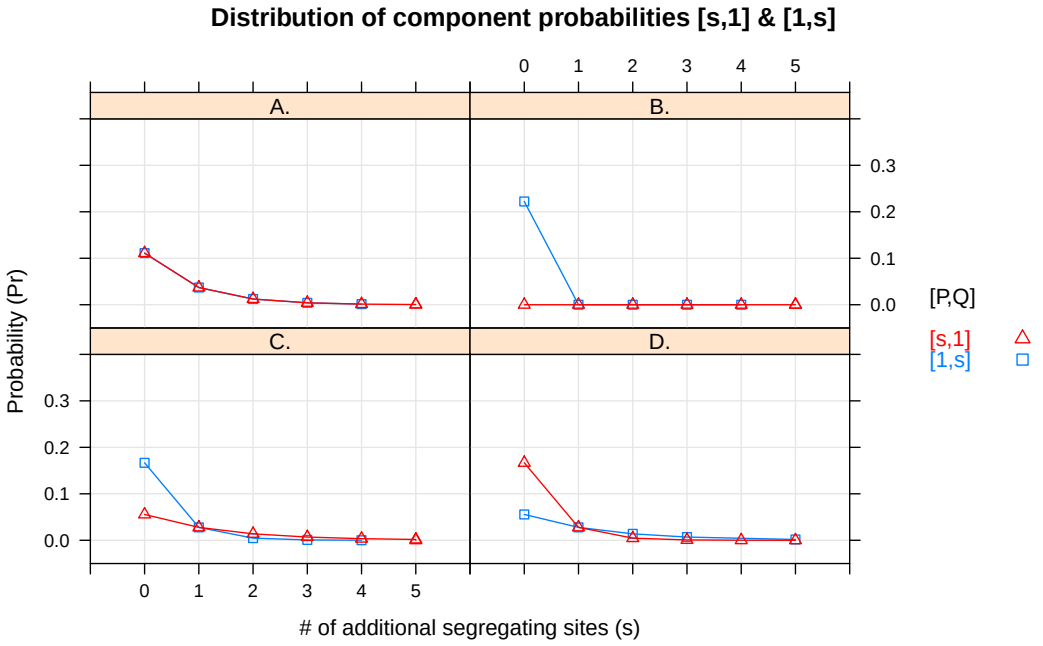



Figure 4.8: Component probabilities of $s$ segregating sites, given 1 transition or transversion, in 2 sequences with heterogeneous substitution rates. Let the substitution rates for transitions ($P$) and transversions ($Q$) be $\theta_\alpha$ and $\theta_\beta$, respectively. A. $\theta_\alpha = 1.0$ and $\theta_\beta = 1.0$, B. $\theta_\alpha = 2.0$ and $\theta_\beta = 0.0$, C. $\theta_\alpha = 1.5$ and $\theta_\beta = 0.5$, and D. $\theta_\alpha = 0.5$ and $\theta_\beta = 1.5$. The distributions are similar, but flattened, relative to those seen in figure 4.7.

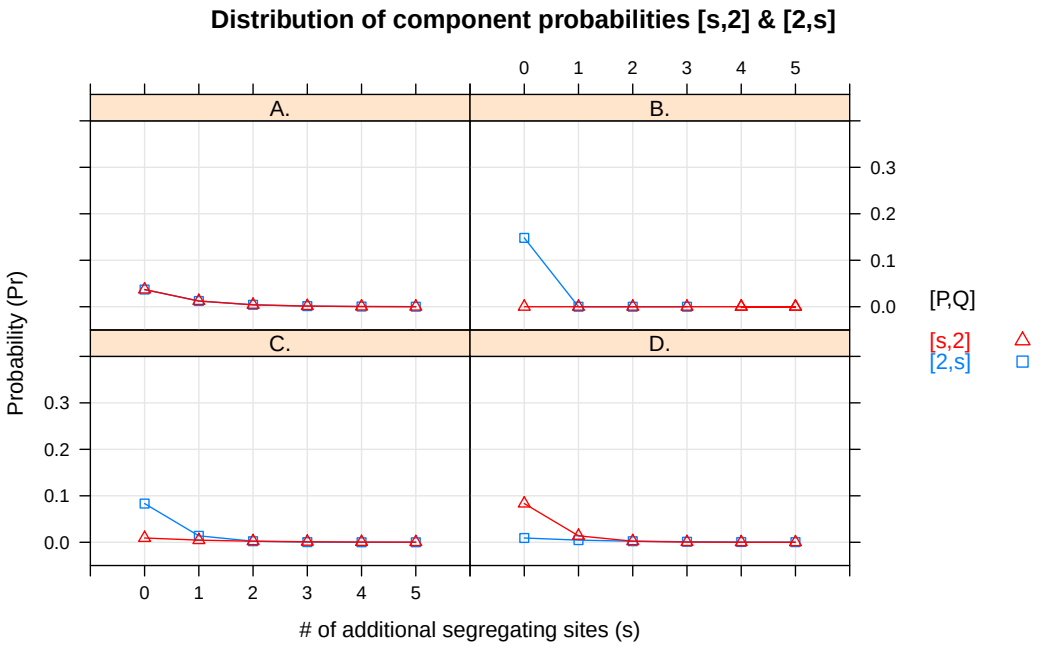



Figure 4.9: Component probabilities of $s$ segregating sites, given 2 transitions or transversions, in 2 sequences with heterogeneous substitution rates. Let the substitution rates for transitions ($P$) and transversions ($Q$) be $\theta_\alpha$ and $\theta_\beta$ respectively. A. $\theta_\alpha = 1.0$ and $\theta_\beta = 1.0$, B. $\theta_\alpha = 2.0$ and $\theta_\beta = 0.0$, C. $\theta_\alpha = 1.5$ and $\theta_\beta = 0.5$, and D. $\theta_\alpha = 0.5$ and $\theta_\beta = 1.5$. The distributions are similar, but flattened, relative to those seen in figures 4.7 and 4.8.

gration and heterogeneous substututions rates on estimating the probability of observing a number of segregating sites in a sample of sequences. To our knowledge, this is the first study that attempts to modify the theory of segregating sites to account for biological phenomena known to affect levels of variation and an estimate of an important measure of genetic diversity, $\theta$, for a sample of sequences or population.

Our results highlight the importance of considering various biological forces as potential sources of variation when describing genetic diversity in a set of sequences or population. Specifically, the recursion equations of the modified theory of segregating sites generate probability values that more accurately reflect the chance of observing a number of segregating sites when processes that contribute to variation, previously assumed to have little effect and were not modelled, are accounted for.

Representative probability values of the observed levels of sequence diversity should be able to improve the assignment of an unknown query to its correct species. An unknown query should assign to a species that it is most similar to or shares the least number of differences with. Given the current theory of segregating sites, assuming informative sequences (that is, the data sufficiently capture intraspecific variation), there should be more observed differences between sequences from different subpopulations thus resulting in an unusually high $\theta$ value, even though the sequences originate from a single species. And an assignment of an unknown query to this group will likely be rejected, even though the query does belong. Under the modified theory, which allows for the possibility of variation from other sources, the estimated value of $\theta$, if modelled correctly, should more closely resemble the true value and a correct assignment might be made. This is in line with our previous study where inflated $\theta$ values were observed when species were composed of one or more conspecific sequences sampled from distinct demes, suggesting that the system was not modelled correctly and the increase in variation is, falsely, from a higher population mutation rate rather than population substructure (Lou and Golding, 2012).

Different substitution pressures may exist between different species. Upon examining counts of each type of change, the pattern of each substitution type should be similar between conspecifics. Therefore, a similar pattern of substitution types and $\theta$ values should exist between an unknown query and its species of origin and allow for accurate specimen identification. Yang (1996) showed that failing to account for among-site rate variation (ASRV) or different substitution rates at different sites will severely underestimate levels of genetic variation and transition-transversion rate ratios. While our study explored the partition of segregating sites as transitions and tranversions, depending on the data and the desired level of res-

olution, a different choice of partitions may be more appropriate. For sequences with low divergences, it may be sufficient to simply count the number of transitions and transversions and similar proportions of each partition type would be expected among conspecifics. For a finer level of resolution, the above strategy may be applied to particular sites since the transition/transversion ratio has been shown to differ among nondegenerate, twofold degenerate, and fourfold degenerate sites in COI (Xia, Hafner and Sudman, 1996; Martinez-Navarro, Galian and Serrano, 2005).

Confidence in our results could be strengthened by conducting a comparison study between the original and modified theories on the performance of specimen identification. Specifically, under the modified model, the proportion of correct assignments should increase and $\theta$ estimates should be closer to the true value. The methodology follows from Lou and Golding (2012).

While our method may be used with all types of genetic data, a limitation of our method an the assumption associated with the use of the infinite sites model. That is, each mutation always occurs in a new position in a long DNA sequence with a low mutation rate (Kimura, 1969). To accommodate this assumption, our modified method should be used with species data that have short times to speciation (i.e., recently diverged or closely related species) or where the occurrence of multiple hits and saturation is very small or neglible (e.g., lack of homoplasy or similarity arising from parallel or convergent evolution). While our method may be less robust when these assumptions are not met, other factors, that may have a greater effect, may be more likely to confound the results first.

In fact, these other factors may serve as significant sources of variation that may be used to further refine the model and provide potential avenues for future research. Our current research investigated different substitution biases (transitions and transversions), each with different rates. It is possible that the rates may change over time. That is, positions may evolve at different rates in different lineages. The term used to describe this phenomenon is among-lineage rate variation (ALRV) or heterotachy (Simon *et al.*, 2006). Though Schwartz and Mueller (2010) have shown that ALRV may have a limited effect on phylogenetic estimation. Other confounding factors may include introgression (an extreme form of hybridization where the genetic content of one species is completely replaced by that in another) and selective sweeps (the favoured mutation and its neighbouring neutral variation become more prevalent, reducing total genetic variation). Introgression has been documented in *Drosophila* mtDNA (Ballard, 2000b) and, along with heterotachy, can cause different species to look prematurely similar. In contrast, selective sweeps may artificially increase reciprocal monophyly and may not reflect the evolutionary relationships among populations (Ballard and Rand, 2005).

In addition to the biological or biotic factors mentioned, there are potential abiotic sources affecting polymorphism levels that models will need to address. These include machine-read errors (i.e., sequencing) and systemic errors (e.g., storage, preparation, and computational processing). Each error will likely increase the number of rare variant sites (i.e., polymorphic positions where one sampled sequence, or singleton, exhibits a unique base relative to the shared nucleotide of others) thus it is especially problematic for approaches based on segregating sites. The impact of error on parameter estimation is not new and increased sequence coverage may lessen error bias but only if it is able to overcome low signal-to-noise ratio, common to low diversity sequences (Clark and Whittam, 1992). If the error rate is unknown, using only shared polymorphisms is an option but this will result in a loss of information because singletons and other random errors are ignored (Knudsen and Miyamoto, 2009). If the error rate is known, it may be incorporated using Phred quality scores (sequence error rate; Ewing and Green, 1998) or an error rate for each nucleotide site (Liu *et al.*, 2010). However, it is important to note that different sequencing platforms, and individual runs on each platform, may have different error distributions. Thus, the theory of segregating sites may benefit from including terms for the accurate modelling of sequence error when it is not negligible (that is, when the error rate is high or the sample size is large or both).

Ultimately, our modified theory of segregating sites helps to untangle the true source of variation, allows for better estimates of the genetic diversity seen in a sample of sequences and confirms our hypothesis that it is important to account for biological phenomena that can affect the accuracy of descriptors used to summarize the level of genetic variation.

# Part II

# CONCLUSION

This thesis describes the usability of a standardized mitochondrial marker and Bayesian methods or models to reflect the evolutionary dynamics of species for robust identifications. Specifically, chapters one and three, on the integrity and usability of existing sequence data for robust identifications, suggest that caution should be exercised when using GenBank (non-barcode) sequences (potential evidence for sequence and taxonomic errors from unusually divergent within-species sequences), and data informativeness (level of within-species variation) can be improved with the addition of a single sample from a different region of a species distribution. In chapter two, a new Bayesian tree-less statistical method, based on segregating sites, provided fast, high probability assignments, even in difficult assignment scenarios characterized by an absence of a "barcoding gap" (overlap in the level of within- and between-species variation). In conjuction with chapter two, chapters three and four show that the pattern of segregating sites (via population genetic parameter, Watterson (1975)'s $\theta$) is an informative measure of genetic diversity. The performance of both the segregating sites algorithm (for assignment) and the modified theory of segregating sites (to predict the true level of genetic diversity in a set of sequences or population) improved when accounting for biological proccesses (i.e., population substructure with migration and unequal base composition and substitution rates) that influence genetic heterogeneity.

Overall, the derived analyses and models reveal and promote the importance of integrating different sources of information to obtain robust species identifications. This is achieved by improving the correspondence between the data and model assumptions by understanding the properties of the marker or data and accommodating these properties in the model or method used.

These findings do not argue that segregating site or Bayesian approaches should replace the use of other methodologies, but rather should be viewed as supplementary tools for reflecting species dynamics and may be used at different phases of the barcoding workflow (i.e., use more-conservative methods when building reference libraries and less-conservative methods once the intra- and interspecific variation is sufficiently sampled) or in conjunction with other lines of non-molecular evidence (e.g., ecological and behavioural).

For simulated and empirical data, the assignment of species status was based on the genetic divergence of a single mitochondrial gene, an issue that has been the subject of much criticism. However, the analyses need not depend on the gene used - it depends on the number of informative sites and this differs among genes and at different evolutionary depths (i.e., times). Thus, the theory and methods may be applied for any set of informative sites.

Broadly, the culmination of the work presented in this thesis aids the task of describing and quantifying diversity and highlights the interconnectively that exists between different evolutionary processes at the genetic and ecosystem level. Though the standard marker and methods described here are simple representations of complex processes at work, they have shown to be effective for species identification and are steps in the right direction.

# Part III

# REFERENCES

# Bibliography

Abdo, Z. and G. B. Golding (2007). A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology. 56*, 44–56.

Aliabadian, M., M. Kaboli, V. Nijman, and M. Vences (2009). Molecular identification of birds: performance of distance-based DNA barcoding in three genes to delimit parapatric species. *PLoS One. 4*, e4119.

Allcock, A., I. Barratt, M. Elaume, K. Linse, M. Norman, P. Smith, D. Steinke, D. Stevens, and J. Strugnell (2011). Cryptic speciation and the circumpolarity debate: A case study on endemic Southern Ocean octopuses using the COI barcode of life. *Deep Sea Research Part II: Topical Studies in Oceanography. 58*, 242–249.

Austerlitz, F., O. David, B. Schaeffer, K. Bleakley, M. Olteanu, R. Leblois, M. Veuille, and C. Laredo (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics. 10*, S10.

Avise, J. (1989). Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution. 43*, 1192–1208.

Avise, J. (1992). Molecular Population Structure and the Biogeographic History of a Regional Fauna: A Case History with Lessons for Conservation Biology. *Oikos. 63*, 62–76.

Ball, S. L. and K. F. Armstrong (2006). DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). *Canadian Journal of Forest Research. 36(2)*, 337–350.

Ballard, J. W. (2000a). Comparative genomics of mitochondrial DNA in members of the Drosophila melanogaster subgroup. *Journal of Molecular Evolution. 51*, 48–63.

Ballard, J. W. (2000b). When one is not enough: introgression of mitochondrial DNA in Drosophila. *Molecular Biolology and Evolution. 17*, 1126–1130.

Ballard, J. W. and D. M. Rand (2005). The population biology of mitochondrial DNA and its phylogenetic implications. *Annual Review of Ecology Evolution and Systematics. 36*, 621–642.

Bergsten, J., D. T. Bilton, T. Fujisawa, M. Elliott, M. T. Monaghan, M. Balke, L. Hendrich, J. Geijer, J. Herrmann, G. N. Foster, I. Ribera, A. N. Nilsson, T. G. Barraclough, and A. P. Vogler (2012). The Effect of Geographical Scale of Sampling on DNA Barcoding. *Systematic Biology. 61*, 851–869.

Blaxter, M., J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of The Royal Society B: Biological Sciences. 360*, 1935–1943.

Bloomquist, E. W., P. Lemey, and M. A. Suchard (2010). Three roads diverged? Routes to phylogeographic inference. *Trends in Ecology and Evolution. 25*, 626–632.

Brower, A. (1999). Delimitation of phylogenetic species with DNA sequences: a critique of Davis and Nixon's population aggregation analysis. *Systematic Biology. 48*, 199–213.

Brown, G. G. and M. V. Simpson (1982). Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proceedings of the National Academy of Sciences. 79*, 3246–3250.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution. 18*, 225–239.

Cameron, S., D. Rubinoff, and K. Will (2006). Who will actually use DNA barcoding and what will it cost? *Systematic Biology. 55*, 844–847.

Capaldi, R. A. (1990). Structure and function of cytochrome c oxidase. *Annual Review of Biochemistry. 59*, 569–596.

Carr, C. M., S. M. Hardy, T. M. Brown, T. A. Macdonald, and P. D. Hebert (2011). A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes. *PLoS One. 6*, e22232.

Clare, E. L., B. K. Lim, M. D. Engstrom, J. L. Eger, and P. D. N. Hebert (2007). DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes. 7*, 184–190.

Clark, A. G. and T. S. Whittam (1992). Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution. 9*, 744–752.

Cognato, A. I. (2006). Standard percent DNA sequence difference for insects does not predict species boundaries. *Journal of Economic Entomology. 99*, 1037–1045.

Coyne, J. A. and H. A. Orr (2004). *Speciation*. Sunderland, Massachusetts: Sinauer & Associates.

David, O., C. Laredo, R. Leblois, B. Schaeffer, and N. Vergne (2012). Coalescent-based DNA barcoding: multilocus analysis and robustness. *Journal of Computational Biology. 19*, 271–278.

Davis, J. and K. Nixon (1992). Populations, Genetic variation, and the Delimitation of Phylogenetic Species. *Systematic Biology. 41*, 421–435.

Degnan, J. H. and N. A. Rosenberg (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution. 24*, 332–340.

DeSalle, R., M. G. Egan, and M. Siddall (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of The Royal Society B: Biological Sciences. 360*, 1905–1916.

DeSalle, R., T. Freedman, E. M. Prager, and A. C. Wilson (1987). Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian Drosophila. *Journal of Molecular Evolution. 26*, 157–164.

DeWalt, R. (2011). DNA barcoding: a taxonomic point of view. *Journal of the North American Benthological Society. 30*, 174–181.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research. 32*, 1792–1797.

Elias, M., R. I. Hill, K. R. Willmott, K. K. Dasmahapatra, A. V. Brower, J. Mallet, and C. D. Jiggins (2007). Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of The Royal Society B: Biological Sciences 274*, 2881–2889.

Ewing, B. and P. Green (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research. 8*, 186–194.

Frezal, L. and R. Leblois (2008). Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics of Infectious Diseases 8*, 727–736.

Funk, D. and K. Omland (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics. 34*, 397–423.

Galtier, N., B. Nabholz, S. Glemin, and G. D. D. Hurst (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology. 18*, 4541–4550.

Golding, G. B. (1984). The sampling distribution of linkage disequilibrium. *Genetics. 108*, 257–274.

Golding, G. B. (2002). Reconstructing the prior probabilities of allelic phylogenies. *Genetics. 161*, 889–896.

Goldstein, P. Z. and R. DeSalle (2011). Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays. 33*, 135–147.

Hajibabaei, M., D. H. Janzen, J. M. Burns, W. Hallwachs, and P. D. N. Hebert (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences. 103*, 968–971.

Hajibabaei, M., G. A. Singer, P. D. Hebert, and D. A. Hickey (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics. 23*, 167–172.

Harris, D. J. (2003). Can you bank on GenBank? *Trends in Ecology & Evolution. 18*, 317–319.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences. 270*, 313–321.

Hebert, P. D. N., E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences. 101*, 14812–14817.

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

Hebert, P. D. N., M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis (2004). Identification of Birds through DNA Barcodes. *PLoS Biology. 2*, 1657–1663.

Hein, J., M. H. Schierup, and C. Wiuf (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford, UK: Oxford University Press Inc.

Hendrich, L., J. Pons, I. Ribera, and M. Balke (2010). Mitochondrial cox1 sequence data reliably uncover patterns of insect diversity but suffer from high lineage-idiosyncratic error rates. *PLoS One. 5*, e14448.

Hickerson, M. J., C. P. Meyer, and C. Moritz (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology. 55*, 729–739.

Higuchi, R., B. Bowman, M. Freiberger, O. A. Ryder, and A. C. Wilson (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature. 312*, 282–284.

Hollingsworth, P. M., L. L. Forrest, J. L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. W. Chase, R. S. Cowan, D. L. Erickson, A. J. Fazekas, S. W. Graham, K. E. James, K.-J. Kim, W. J. Kress, H. Schneider, J. van AlphenStahl, S. C. H. Barrett, C. van den Berg, D. Bogarin, K. S. Burgess, K. M. Cameron, M. Carine, J. Chacn, A. Clark, J. J. Clarkson, F. Conrad, D. S. Devey, C. S. Ford, T. A. J. Hedderson, M. L. Hollingsworth, B. C. Husband, L. J. Kelly, P. R. Kesanakurti, J. S. Kim, Y.-D. Kim, R. Lahaye, H.-L. Lee, D. G. Long, S. Madrin, O. Maurin, I. Meusnier, S. G. Newmaster, C.-W. Park, D. M. Percy, G. Petersen, J. E. Richardson, G. A. Salazar, V. Savolainen, O. Seberg, M. J. Wilkinson, D.-K. Yi, and D. P. Little (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences. 106*, 12794–12797.

Holmquist, R. (1983). Transitions and transversions in evolutionary descent: an approach to understanding. *Journal of Molecular Evolution. 19*, 134–144.

Huang, D., R. Meier, P. A. Todd, and L. M. Chou (2008). Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *Journal of Molecular Evolution. 66*, 167–174.

Hudson, R. R. and J. A. Coyne (2002). Mathematical consequences of the genealogical species concept. *Evolution. 56*, 1557–1565.

Jaenike, J., K. A. Dyer, C. Cornish, and M. S. Minhas (2006). Asymmetrical reinforcement and Wolbachia infection in Drosophila. *PLoS Biology. 4*, e325.

Jukes, T. H. (1987). Transitions, transversions, and the molecular evolutionary clock. *Journal of Molecular Evolution. 26*, 87–98.

K.C., N. and Q. Wheeler (1990). An amplification of the phylogenetic species concept. *Cladistics. 6*, 211–223.

Kelly, J. K. and M. A. Noor (1996). Speciation by reinforcement: a model derived from studies of Drosophila. *Genetics. 143*, 1485–1497.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics. 61*, 893–903.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution. 16*, 111–120.

Kizirian, D. and M. A. Donnelly (2004). The criterion of reciprocal monophyly and classification of nested diversity at the species level. *Molecular Phylogenetics and Evolution. 32*, 1072–1076.

Knowles, L. L. and B. C. Carstens (2007). Delimiting species without monophyletic gene trees. *Systematic Biology. 56*, 887–895.

Knudsen, B. and M. M. Miyamoto (2009). Accurate and fast methods to estimate the population mutation rate from error prone sequences. *BMC Bioinformatics. 10*, 247.

Lemey, P., A. Rambaut, J. J. Welch, and M. A. Suchard (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution. 27*, 1877–1885.

Lim, G. S., M. Balke, and R. Meier (2012). Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods. *Systematic Biology. 61*, 1–5.

Little, D. (2011). DNA Barcode Sequence Identification Incorporating Taxonomic Hierarchy and within Taxon Variability. *PLoS One. 6*, e20552.

Liu, H. and A. T. Beckenbach (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Molecular Phylogenetics and Evolution. 1*, 41–52.

Liu, X., Y. X. Fu, T. J. Maxwell, and E. Boerwinkle (2010). Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. *Genetics. 20*, 101–109.

Lohse, K. (2009). Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). *Systematic Biology. 58*, 439–442.

Lou, M. and G. B. Golding (2010). Assigning sequences to species in the absence of large interspecific differences. *Molecular Phylogenetics and Evolution. 56*, 187–194.

Lou, M. and G. B. Golding (2012). The effect of sampling from subdivided populations on species identification with DNA barcodes using a Bayesian statistical approach. *Molecular Phylogenetics and Evolution. 65*, 765–773.

Lowenstein, J. H., J. Burger, C. W. Jeitner, G. Amato, S. O. Kolokotronis, and M. Gochfeld (2010). DNA barcodes reveal species-specific mercury levels in tuna sushi that pose a health risk to consumers. *Biology Letters. 6*, 692–695.

Lukhtanov, V., A. Sourakov, E. Zakharov, and P. Hebert (2009). DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources. 9*, 1302–1310.

Martinez-Navarro, E. M., J. Galian, and J. Serrano (2005). Phylogeny and molecular evolution of the tribe Harpalini (Coleoptera, Carabidae) inferred from mitochondrial cytochrome-oxidase I. *Molecular Phylogenetics and Evolution.. 35*, 127–146.

Matsen, F. A., R. B. Kodner, and E. V. Armbrust (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics. 11*, 538.

Mayr, E. (1942). *Systematics and the origin of species*. New York: Columbia University Press.

Meier, R. (2008). *DNA sequences in taxonomy, opportunities and challenges. In: Wheeler, QD, The new taxonomy*. Boca Raton: CRC Press.

Meier, R., K. Shiyang, G. Vaidya, and P. K. Ng (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology. 55*, 715–728.

Meier, R., G. Zhang, and F. Ali (2008). The Use of Mean Instead of Smallest Interspecific Distances Exaggerates the Size of the "Barcoding Gap" and Leads to Misidentification. *Systematic Biology. 57*, 809–813.

Meyer, C. P. and G. Paulay (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology. 3*, 2229–2237.

Monaghan, M. T., M. Balke, T. R. Gregory, and A. P. Vogler (2005). DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philosophical Transactions of The Royal Society B: Biological Sciences. 360*, 1925–1933.

Monaghan, M. T., R. Wild, M. Elliot, T. Fujisawa, M. Balke, D. J. Inward, D. C. Lees, R. Ranaivosolo, P. Eggleton, T. G. Barraclough, and A. P. Vogler (2009). Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology. 58*, 298–311.

Moritz, C. and C. Cicero (2004). DNA barcoding: promise and pitfalls. *PLoS Biology. 2*, e354.

Munch, K., W. Boomsma, J. P. Huelsenbeck, E. Willerslev, and R. Nielsen (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology. 57*, 750–757.

Neigel, J. E. and J. C. Avise (1986). *Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation*, pp. 515–534. London: Academic Press.

Nielsen, R. and M. Matz (2006). Statistical approaches for DNA barcoding. *Systematic Biology. 55*, 162–169.

Noor, M. A. (1995). Speciation driven by natural selection in Drosophila. *Nature. 375*, 674–675.

Padial, J. M., A. Miralles, I. De la Riva, and M. Vences (2010). The integrative future of taxonomy. *Frontiers in Zoology. 7*, 16.

Papadopoulou, A., J. Bergsten, T. Fujisawa, M. T. Monaghan, T. G. Barraclough, and A. P. Vogler (2008). Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. *Philosophical Transactions of The Royal Society B: Biological Sciences. 363*, 2987–2996.

Pappalardo, A. M., F. Guarino, S. Reina, A. Messina, and V. De Pinto (2011). Geographically widespread swordfish barcode stock identification: a case study of its application. *PLoS One. 6*, e25516.

Parks, D. H., N. J. MacDonald, and R. G. Beiko (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics. 12*, 328.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen (2006). Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS Genetics. 2*, 1634–1647.

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.

Powell, J. (1997). *Progress and Prospects in Evolutionary Biology*. New York: Oxford University Press, Inc.

Ratnasingham, S. and P. D. Hebert (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes. 7*, 355–364.

Raxworthy, C. J., C. M. Ingram, N. Rabibisoa, and R. G. Pearson (2007). Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (Phelsuma) from Madagascar. *Systematic Biology. 56*, 907–923.

Reed, L. K., M. Nyboer, and T. A. Markow (2007). Evolutionary relationships of Drosophila mojavensis geographic host races and their sister species Drosophila arizonae. *Molecular Ecology. 16*, 1007–1022.

Remigio, E. A. and P. D. N. Hebert (2003). Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Molecular Phylogenetics and Evolution. 29*, 641–647.

Roe, A. D. and F. A. Sperling (2007). Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution. 44*, 325–345.

Ross, H. A., S. Murugan, and W. L. Li (2008). Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology. 57*, 216–230.

Saitou, N. . and M. . Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution. 4*, 406–425.

Sarkar, I. N., P. J. Planet, and R. Desalle (2008). CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources. 8*, 1256–1259.

Satta, Y., H. Ishiwa, and S. I. Chigusa (1987). Analysis of nucleotide substitutions of mitochondrial DNAs in Drosophila melanogaster and its sibling species. *Molecular Biology and Evolution. 4*, 638–650.

Satta, Y. and N. Takahata (1990). Evolution of Drosophila mitochondrial DNA and the history of the melanogaster subgroup. *Proceedings of the National Academy of Sciences. 87*, 9558–9562.

Schmidt, B. C. (2009). Taxonomic revision of the genus *Grammia Rambur* (Lepidoptera: Noctuidae: Arctiinae). *Zoological Journal of the Linnean Society. 156*, 507–597.

Schmidt, B. C. and F. A. H. Sperling (2008). Widespread decoupling of mtDNA variation and species integrity in *Grammia* tiger moths (Lepidoptera: Noctuidae). *Systematic Entomology. 33*, 613–634.

Schwartz, R. S. and R. L. Mueller (2010). Limited effects of among-lineage rate variation on the phylogenetic performance of molecular markers. *Molecular Phylogenetics and Evolution. 54*, 849–856.

Seifert, K. A., R. A. Samson, J. R. Dewaard, J. Houbraken, C. A. Levesque, J. M. Moncalvo, G. Louis-Seize, and P. D. N. Hebert (2007). Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proceedings of the National Academy of Sciences. 104*, 3901–3906.

Seo, T. K. (2010). Classification of nucleotide sequences using support vector machines. *Journal of Molecular Evolution. 71*, 250–267.

Shoemaker, D. D., K. A. Dyer, M. Ahrens, K. McAbee, and J. Jaenike (2004). Decreased diversity but increased substitution rate in host mtDNA as a consequence of Wolbachia endosymbiont infection. *Genetics. 168*, 2049–2058.

Siddall, M. and R. Budinoff (2005). DNA-barcoding evidence for widespread introductions of a leech from the South American *Helobdella triserialis* complex. *Conservation Genetics. 6*, 467–472.

Simon, C., T. Buckley, F. Frati, J. Stewart, and A. 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.

Soltis, D. E., A. B. Morris, J. S. McLachlan, P. S. Manos, and P. S. Soltis (2006). Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology. 15*, 4261–4293.

Sparks, J. S. and W. L. Smith (2006). Sicyopterus lagocephalus: widespread species, species complex, or neither? A critique on the use of molecular data for species identification. *Molecular Phylogenetics and Evolution. 40*, 900–902.

Sperling, F. A. and D. A. Hickey (1994). Mitochondrial DNA sequence variation in the spruce budworm species complex (Choristoneura: Lepidoptera). *Molecular Biology and Evolution. 11*, 656–665.

Tamura, K., S. Subramanian, and S. Kumar (2004). Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Molecular Biology and Evolution. 21*, 36–44.

Tavare, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology. 26*, 119–164.

Tavares, E. S., G. H. de Kroon, and A. J. Baker (2010). Phylogenetic and coalescent analysis of three loci suggest that the Water Rail is divisible into two species, Rallus aquaticus and R. indicus. *BMC Evolutionary Biology. 10*, 226.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research. 22*, 4673–4680.

Trewick, S. (2007). DNA Barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera: Arcididae). *Cladistics. 24*, 240–254.

Tsukihara, T., H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. *Science. 272*, 1136–1144.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, and H. Zhang (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research. 37*, D555–9.

Valkiunas, G., C. T. Atkinson, S. Bensch, R. N. Sehgal, and R. E. Ricklefs (2008). Parasite misidentifications in GenBank: how to minimize their number? *Trends in Parasitology. 24*, 247–248.

Virgilio, M., T. Backeljau, B. Nevado, and M. De Meyer (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics. 11*, 206.

Wakeley, J. (2000). The effects of subdivision on the genetic divergence of populations and species. *Evolution. 54*, 1092–1101.

Wang, B. C., J. Park, H. A. Watabe, J. J. Gao, J. G. Xiangyu, T. Aotsuka, H. W. Chen, and Y. P. Zhang (2006). Molecular phylogeny of the Drosophila virilis section (Diptera: Drosophilidae) based on mitochondrial and nuclear sequences. *Molecular Phylogenetics and Evolution. 40*, 484–500.

Ward, R. D., T. S. Zemlak, B. H. Innes, P. R. Last, and P. D. N. Hebert (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London, Series B Biological Sciences. 360*, 1847–1857.

Watterson, G. A. (1975). On the number of segregating sites in genetical model without recombination. *Theoretical Population Biology. 7*, 256–276.

Wiemers, M. and K. Fiedler (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology. 4*, 1–16.

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

Wong, L. L., E. Peatman, J. Lu, H. Kucuktas, S. He, C. Zhou, U. Na-nakorn, and Z. Liu (2011). DNA barcoding of catfish: species authentication and phylogenetic assessment. *PLoS One. 6*, e17812.

Xia, X., M. S. Hafner, and P. D. Sudman (1996). On transition bias in mitochondrial genes of pocket gophers. *Journal of Molecular Evolution. 43*, 32–40.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution. 11*, 367–372.

Zhang, A. B., L. J. He, R. H. Crozier, C. Muster, and C. D. Zhu (2010). Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution. 54*, 1035–1039.

Zhang, A. B., C. Muster, H. B. Liang, C. D. Zhu, R. Crozier, P. Wan, J. Feng, and R. D. Ward (2012). A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology. 21*, 1848–1863.

Zhang, C., D. X. Zhang, T. Zhu, and Z. Yang (2011). Evaluation of a bayesian coalescent method of species delimitation. *Systematic Biology. 60*, 747–761.

Zhou, X., S. J. Adamowicz, L. M. Jacobus, R. E. Dewalt, and P. D. Hebert (2009). Towards a comprehensive barcode library for arctic life - Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Frontiers in Zoology. 6*, 30.

Zhou, X., L. M. Jacobus, R. E. Dewalt, S. J. Adamowicz, and P. D. Hebert (2010). Ephemeroptera, Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): insights into biodiversity patterns from DNA barcoding. *Journal of North American Benthological Society. 29*, 814–837.

Zhou, X., J. Robinson, C. Geraci, O. Parker, C.R. Flint Jr., D. Etnier, D. Ruiter, R. DeWalt, L. Jacobus, and P. Hebert (2011). Accelerated construction of a regional DNA-barcode reference library: caddisflies (Trichoptera) in the Great Smoky Mountains National Park. *Journal of North American Benthological Society. 30*, 131–162.