More taxa or more characters: which dataset is better?

Stefan Barone

Follow this and additional works at: http://academicworks.cuny.edu/bb_etds

Recommended Citation

Barone, Stefan, "More taxa or more characters: which dataset is better?" (2011). CUNY Academic Works.
http://academicworks.cuny.edu/bb_etds/3
MORE TAXA OR MORE CHARACTERS-
WHICH DATA SET IS BETTER?

STEFAN BARONE

Submitted to the Committee on Undergraduate Honors of Baruch Honors College
of the City University of New York
in partial fulfillment of the requirement for the degree of
Bachelor of Arts in ad hoc Biology with Honors

Submitted Fall Semester 2011

Approved by the Department of Natural Sciences:

Mentor and Chairperson
Dr. Valerie Schawaroch

Committee Member
Dr. Keith Ramig

Committee Member
Dr. Edyta Greer
TABLE OF CONTENTS continued

Total evidence - 36 taxa with complete gene sampling .........................................................48
  Parsimony informative characters per analysis .................................................................48
Trees ..................................................................................................................................48
Tree statistics ....................................................................................................................48
Recovery of clades .............................................................................................................48
  Tree node statistics ........................................................................................................50

DISCUSSION ......................................................................................................................55
  Future work .....................................................................................................................58

ACKNOWLEDGEMENTS .................................................................................................59

REFERENCES .....................................................................................................................62

APPENDIX A: WET LAB PROCEDURES ...........................................................................64
  Culturing fruit flies .........................................................................................................65
  DNA isolation ..................................................................................................................66
  PCR ..................................................................................................................................66
  Agarose gel electrophoresis .............................................................................................67
  DNA sequencing .............................................................................................................67
  Addendum: Advice to the student researcher .................................................................67

APPENDIX B: PROTOCOLS ...........................................................................................69
  PROTOCOL 1: DNA isolation template and directions .....................................................69
  PROTOCOL 2: PCR directions and recipe ........................................................................72
  PROTOCOL 3: Gel template and directions .....................................................................74
  PROTOCOL 4: Performing alignments with MegAlign ...................................................77
  PROTOCOL 5: Directions for tree analysis using PAUP* ...............................................79
  PROTOCOL 6: Directions for bootstrap analysis ............................................................80
  PROTOCOL 7: Directions for Bremer analysis ..................................................................81
SUMMARY

The fruit fly, *Drosophila melanogaster*, has been the model organism for our understanding of many biological processes, including genetics and development. Science initially tries to understand the biochemical and physical processes of a single system. The next logical step is to see how broadly applicable these processes are across species and if any differences among species reflect differences in the processes. This comparison of biochemical and physical processes across species is referred to as comparative biology.

All comparative biological investigations rely upon an understanding of the relationships among the species being compared. Ironically, the relationships among the model organism, *Drosophila melanogaster* and its close relatives in the *Drosophila melanogaster* species group are still unclear and contested among scientists. One goal of this project is to create a family tree describing the relationship among *Drosophila melanogaster* and its relatives.

Phylogenies are family trees of relationships among species. These trees summarize the evolutionary history of the organisms. The data needed to create these evolutionary trees come in two dimensions. One aspect is the number of species (taxa) to be sampled. The other aspect is the amount of information (characters) needed for each species. A second goal of this project is to make phylogenetic investigations more efficient by determining the relative contribution of different types of data (species/taxa versus characters).

The taxonomy of the *Drosophila melanogaster* species group is based on morphological characters (physical features) such as, male genitalia, sex combs, and body coloration. There are not enough of these characters to fully resolve the relationships among the 180 species within the *Drosophila melanogaster* species group; therefore, scientists have been using the genetic letter code of DNA (DNA sequence) for which there is a virtually infinite source of characters.
Two previous studies that created phylogenies for the *Drosophila melanogaster* species group using DNA sequence are Schawaroch (2002) and DaLage, *et al.* (2007). These two studies disagree in the relationships among the species. Some of the species sampled between these two studies were the same. However, each study used different DNA sequence. The purpose of this study was to combine both data sets and create a phylogeny for (1) all of the species sampled (so there was missing data for non-overlapping species) and (2) for just the overlapping species (so no missing data). By comparing all analyses, it was hoped to address the main question at hand, “Is it better to add more taxa [species] to a data set, or more characters”. The definition of “better” was defined as the analysis providing more information on the relationships of the taxa. The study also sought to establish a robust hypothesis of the relationships of three well-known major clades. After comparing all the analyses and their corresponding statistics, it was determined that it is better to add more characters with a complete data set than more taxa. This will have implications for current investigations where whole genomes are being sequenced for several model organisms.
ABSTRACT

What is a sufficient amount of taxa and characters to sample in order to generate a robust evolutionary hypothesis? The *Drosophila melanogaster* species group is well suited for addressing this question because it has a varying number of taxa sampled for several genes. Schawaroch (2002) sampled 49 taxa from the ingroup (*melanogaster* species group) and 6 from the outgroup (*obscura* species group) for three gene regions (nuclear gene regions of *Adh*, *hb*, and mitochondrial gene region *CoII*). DaLage, *et al.* (2007) sampled 68 taxa from the ingroup and 6 from the outgroup for one complete gene (nuclear gene *Amyrel*). This investigation re-analyzes data previously published in DaLage, *et al.* (2007) and Schawaroch (2002). In addition, to a re-analysis of these two studies, two new phylogenetic analyses were conducted. The first analysis combines all taxa of the DaLage, *et al.* (2007) and the Schawaroch (2002) studies; however, some of the gene/gene regions have missing data. A second analysis only included the taxa that are shared between the two studies, so that the gene/gene regions are completely sampled for each taxon. All analyses were heuristic searches employing a maximum parsimony criterion in PAUP* (Swofford, 2000). Comparisons were made amongst the trees generated from the four analyses described above for the recovery of clades previously established by morphological investigations. The stability of these clades was examined with measures of bootstrap and decay values. The results of this study not only provided a hypothesis of phylogenetic relationships within the *melanogaster* species group, but insight on how increasing the amount of taxa or characters affects the reconstruction of phylogenies. In addition, this study examined the affects of missing data (?) on phylogenetic reconstruction. Overall, it was concluded that the addition of taxa will have less of an effect and that it is more important to have a large number of characters with a complete dataset (no missing characters).
INTRODUCTION

Evolutionary history displayed as a phylogenetic tree

Life on Earth can be categorized into nested sets based on their features. A system to describe this observation is known as taxonomy which was established by Carolus Linnaeus with the publication of *Systema Naturae* in 1735. How this variation originated and continues is explained by evolutionary theory. The relationships among the species or populations in the context of how they descend from common ancestors are represented as trees called phylogenies (Eldredge and Carcraft, 1980).

The phylogenetic tree contains branches, which diverge from a node (Figure 1). Each branch ends in a group or taxon that represents a population, species, genus, etc. that is either extinct or still in existence today. The node is where the groups split or diverged (Figure 1). In this study, each taxon represents a single species of fruit fly in the genus *Drosophila*.

Figure 1. Simplified example of phylogenetic trees. Phylogenetic trees contain branches (lines in figures) which terminate in a taxon. Nodes are represented as red circles with numbers. In tree A, node 1 and node 2 display bifurcations. The relationships are as described: Taxon A and B are more closely related to each other than either are to Taxon C. Taxon A and B are referred to as sister taxa and Taxon C is basal and sister to that clade (group) formed by Taxon A and B. Tree B demonstrates a tree with no resolution because there is no bifurcation at the node. Multiple divergences (more than 2) from a single node is referred to as a polytomy or "comb".
Reconstructing phylogenies

There are two possible types of data, which can be collected and analyzed to create a phylogenetic tree. The first type of data is morphological, which represents the physical characteristics of the organism. In the past, this was the only source of data available, which was widely useful in classifying fossil organisms in which bones structures were available for analysis. Most of what is known about the evolutionary relationships of *Drosophila* species has been derived from morphological data. Morphological data is limited in the number of characters (e.g., setae (hairs), sex comb length and orientation, genitalic apparatus, etc.). In addition to changes in morphology, changes occur within the DNA sequence, eventually leading to the evolution of separate species and the nested groups of organisms in the taxonomic hierarchy (e.g., species, species subgroup, species group, genus, etc.).

The method used to create phylogenies is based on the evolutionary idea of descent with modification and is called Cladistics. Cladistics is a method of creating phylogenetic structure (tree shape/topology) where the features of the organisms at the node are homologies (shared by a nearest common ancestor) referred to as a case of “special similarity” by Willi Hennig in his book *Phylogenetic Systematics*, published in 1966.

When a scientific investigator collects morphological or molecular features of organisms, he/she is making a hypothesis about homology between these features of the taxa. The information is placed in a table called a character matrix which is given to a phylogenetic analysis program such as PAUP* version 4.0b4a (Swofford, 2000) to create a phylogenetic tree (for further explanation see the Methods section of this paper). The summarization of these features is in the building of a phylogenetic tree, which describes the homology among the organisms. The hypotheses of homology are tested when a phylogenetics tree is created.
Homology is when the feature appears at the node where the group of taxa first split/bifurcated (Figure 1). When a hypothesis of homology originates more than once on the tree, then it is referred to as homoplasy. Homoplasy is an erroneous homology statement. The famous example of homoplasy in a morphological feature is a bat wing and a bird wing.

The current study

Background

The understanding of basic biological concepts, such as genetics and development, has been predicated on the foundation of knowledge gained from investigations on fruit flies (drosophilids), most notably *Drosophila melanogaster*. *Drosophila melanogaster* is one of 180 closely related species within the *melanogaster* species group (Ashburner, *et al*., 2005).

Although species within the *melanogaster* species group have been closely studied since the early 19th century (Figure 2, see next page), a better understanding of the evolutionary relationships among *Drosophila melanogaster* and its nearest relatives is needed to advance insights based on comparative biological investigations. A hypothesis of evolutionary relationships among species is called a phylogeny, which is pictorially represented by a tree, or cladogram.

Two recent phylogenetic analyses, Schawaroch (2002) and DaLage, *et al.* (2007), proposed a scheme of relationships for the biologically important *melanogaster* species group. Schawaroch (2002) was directly focused on relationships within the *melanogaster* species group, sampling 43 ingroup taxa (for this study, the words taxa and species will be synonymous) and 6 taxa from the sister group, the *obscura* species group (Table 1 on pages 11-14). Schawaroch sampled 3 gene regions for all 49 taxa: 2 nuclear (*alcohol dehydrogenase* (*Adh*) and *hunchback* (*hb*)) and one
mitochondrial (*cytochrome oxidase II (CoII)*). Although, the DaLage, *et al.* (2007) investigation focused on the relationships within the genus *Drosophila*, this study heavily sampled species within the *melanogaster* species group (68 taxa) and 6 outgroup taxa from the *obscura* species group (Table 1 on pages 11-14). The DaLage, *et al.* (2007) phylogeny was based on DNA sequence of a single nuclear gene called *Amyrel*.

<table>
<thead>
<tr>
<th>subgenus</th>
<th>species group</th>
<th>species subgroup</th>
<th># of representative taxa used in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Sophophora</em></td>
<td><em>dentissima</em></td>
<td><em>flavohirta</em> (1 sp.)</td>
<td>0, 1, 1</td>
</tr>
<tr>
<td>Sturtevant,</td>
<td><em>fima</em></td>
<td><em>denticulata</em> (3 sp.)</td>
<td>0, 0, 0</td>
</tr>
<tr>
<td>1942</td>
<td><em>saltans</em></td>
<td><em>elegans</em> (5 sp.)</td>
<td>1, 2, 2</td>
</tr>
<tr>
<td>*Sturtevant,</td>
<td><em>willistoni</em></td>
<td><em>eugracilis</em> (1 sp.)</td>
<td>1, 1, 1</td>
</tr>
<tr>
<td>1942</td>
<td><em>obscura</em></td>
<td><em>ficusphila</em> (6 sp.)</td>
<td>1, 2, 2</td>
</tr>
<tr>
<td>*Sturtevant,</td>
<td><em>melanogaster</em></td>
<td><em>ananassae</em> (25 sp.)</td>
<td>6, 18, 18</td>
</tr>
<tr>
<td>1942</td>
<td><em>dispar</em></td>
<td><em>melanogaster</em> (9 sp.)</td>
<td>3, 9, 9</td>
</tr>
<tr>
<td>Mather, 1955</td>
<td>or <em>populi</em></td>
<td><em>suzukii</em> (17 sp.)</td>
<td>3, 3, 3</td>
</tr>
<tr>
<td>or <em>Thockmorton, 1975</em></td>
<td></td>
<td><em>takahashii</em> (13 sp.)</td>
<td>4, 3, 5</td>
</tr>
<tr>
<td><em>montium</em> (88 sp.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ashburner, <em>et al.</em>, 2005</td>
<td></td>
<td><em>longissima</em> (2 sp.)</td>
<td>0, 0, 0</td>
</tr>
<tr>
<td><em>Toda, 1991</em></td>
<td>or <em>longissima</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Comparison of taxon sampling per species group/subgroup between Schawaroch (2002) and DaLage, *et al.* (2007). This figure is based on Figure 1 of Schawaroch (2000) with additional information from Ashburner, *et al.* (2005). The amount of representative taxa sampled from Schawaroch (2002)- red text, DaLage, *et al.* (2007)- blue text, this study- black text. There are differences in the amount of taxa sampled for each species subgroup. For the *obscura* sp. group- Schawaroch sampled 6 taxa; DaLage, *et al.* sampled 6 taxa (some different). The DaLage, *et al.* (2007) study included a representative from the *flavohirta* sp. subgroup. Schawaroch (2002) did not have any representatives for this sp. subgroup. None of the studies sampled representative species from the *denticulata*, *rhopaloo*, and *longissima* sp. subgroups. *Although I have displayed that there are 37 representative taxa from the *montium* sp. subgroup, it should be noted that this number includes the two taxa of *jambulina* ("jambulinaVS" and "jambulinaDaLage", see caption for Table 1 on pages 11-14).
Table 1. Listing of Taxa Sampled. The species groups and subgroups reported here are based on the traditional taxonomic classification. The taxa list for the DaLage, et al. (2007) study shown here is only the portion of sampled taxa that include species form the *melanogaster* and *obscura* species groups. There is an overlap of 36 taxa between the two studies (names in bold). A “---” indicates that the taxon was not sampled. This study will have four different taxon sampling regimes: (1) re-analysis of DaLage, et al. (2007) with the 74 taxa noted here, (2) re-analysis of Schawaroch (2002) with the 49 taxa noted here, (3) a combined taxon and character sampling of all 87 taxa even if a portion of the DNA character set is missing, and (4) a combined taxon and character sampling for the 36 taxa shared by both data sets. *Note: Schawaroch (2002) concluded, based on locality collected, that the *D. jambulina* may actually be *D. watanabe*. DaLage, et al. (2007) did not identify the collection locality for the *D. jambulina* sampled. Therefore, each will be treated as different species; *D. jambulina* from Schawaroch (2002) will be named “jambulinaVS” and *D. jambulina* from DaLage, et al. (2007) will be named “jambulinaDaLage”.

### Comparison of Taxon Sampling

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>affinis</td>
<td>affinis</td>
<td><em>obscura</em></td>
<td></td>
</tr>
<tr>
<td>ambigua</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bifasciata</td>
<td>bifasciata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>imaii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>persimilis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pseudoobscura</td>
<td>pseudoobscura</td>
<td></td>
<td>suzukii</td>
</tr>
<tr>
<td>---</td>
<td>subobscura</td>
<td></td>
<td>melanogaster</td>
</tr>
<tr>
<td>tolteca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>kitumensis</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>biarmipes</td>
<td>biarmipes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lucipennis</td>
<td>lucipennis</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>mimetica</td>
<td>mimetica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>erecta</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>orena</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>mauritiana</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>melanogaster</td>
<td>melanogaster</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>santomea</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>sechelli</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------</td>
<td>---------------</td>
<td>------------------</td>
</tr>
<tr>
<td>---</td>
<td>simulans</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>teissieri</td>
<td>teissieri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yakuba</td>
<td>yakuba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>elegans</td>
<td>elegans</td>
<td>elegans</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>subelegans</td>
<td>eugracilis</td>
<td></td>
</tr>
<tr>
<td>eugracilis</td>
<td>eugracilis</td>
<td>ficusphila</td>
<td></td>
</tr>
<tr>
<td>ficusphila</td>
<td>ficusphila</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>levii</td>
<td>flavohirta</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>flavorhirta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>takahashii</td>
<td>takahashii</td>
<td>takahashii</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>psuedotakahashii</td>
<td>melanogaster</td>
<td></td>
</tr>
<tr>
<td>lutescens</td>
<td>lutescens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>paralutea</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prostipennis</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ananassae</td>
<td>ananassae</td>
<td>ananassae</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>atripex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ercepeae</td>
<td>ercepeae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>bipectinata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>lachaisei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>malerkotliana pallens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m. malerkotiana</td>
<td>m. malerkotiana</td>
<td>ananassae</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>merina</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>monieri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>ochrogaster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pallidosa</td>
<td>pallidosa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>papaensis-like</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>parabipectinata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------</td>
<td>---------------</td>
<td>------------------</td>
</tr>
<tr>
<td><em>phaeopleura</em></td>
<td><em>phaeopleura</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>pseudoananassae nigrens</em></td>
<td></td>
<td><em>ananassae</em></td>
</tr>
<tr>
<td>---</td>
<td><em>p. pseudoananassae</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>vallismaia</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>varians</em></td>
<td><em>varians</em></td>
<td></td>
<td><em>asahinai</em></td>
</tr>
<tr>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>auraria</em></td>
<td><em>auraria</em></td>
<td></td>
<td><em>auraria</em></td>
</tr>
<tr>
<td><em>baimaii</em></td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>bakoue</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>barbarae</em></td>
<td><em>barbarae</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. biauraria</em></td>
<td><em>biauraria</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. bicornuta</em></td>
<td><em>bicornuta</em></td>
<td></td>
<td><em>melanogaster</em></td>
</tr>
<tr>
<td><em>D. birchii</em></td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>bocqueti</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>bocqueti-like</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>burlai</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>cauverii</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>chauvacaee</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>davidi</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>diplacantha</em></td>
<td><em>diplacantha</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>dossouii</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>greeni</em></td>
<td><em>greeni</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>jambulina</em></td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td><em>jambulina</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>kanapiae</em></td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>ikkawaii</em></td>
<td><em>ikkawaii</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Comparison of Taxon Sampling contd.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>---</em></td>
<td><em>leontia</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>lini</em></td>
<td><em>lini</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>---</em></td>
<td><em>malagassya</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>mayri</em></td>
<td><em>mayri</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>---</em></td>
<td><em>nagarholensis</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>nikananu</em></td>
<td><em>nikananu</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>orosa</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>parvula</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>punjabiensis</em></td>
<td><em>punjabiensis</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>quadraria</em></td>
<td><em>quadraria</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>rufa</em></td>
<td><em>rufa</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>seguyi</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>serrata</em></td>
<td><em>serrata</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>triauraria</em></td>
<td><em>triauraria</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>tsacasi</em></td>
<td><em>tsacasi</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
<tr>
<td><em>vulcana</em></td>
<td><em>vulcana</em></td>
<td><em>---</em></td>
<td><em>---</em></td>
</tr>
</tbody>
</table>

For the most part, the Schawaroch (2002) and DaLage, *et al.* (2007) studies are in agreement with the existence the species subgroups but differ with respect to the specific relationships among some of the species within each species subgroup. Many phylogenetic and taxonomic investigations (starting with Bock and Wheeler, 1972) have agreed upon the existence of three major clades within the *melanogaster* species group (i.e., *melanogaster* + Asian, sp. subgroups, *montium* sp. subgroup and *ananassae* sp. subgroup clades). However, the relationships among the three clades have been controversial and weakly supported by available data. In fact, the
Schawaroch (2002) and DaLage, et al. (2007) investigations are in conflict with respect to their schemes of relationships amongst these three clades (Figure 3).

**Figure 3. Comparison of hypotheses for relationships among three major clades within the *melanogaster* species group.** Tree A shows the hypothesis of relationships for the three major clades proposed by Schawaroch (2002). The *ananassae* and *montium* subgroups are sister taxa and the *melanogaster* + Asian sp. subgroups clade is basal to that clade. In contrast, Tree B shows DaLage’s, et al. (2007) hypothesis which places the *melanogaster* + Asian sp. subgroups clade and *montium* subgroup as sister taxa with the *ananassae* subgroup basal to that clade. In addition, DaLage, et al., (2007) have proposed elevating the *montium* and *ananassae* species subgroups to species groups. Asian subgroups sampled include representatives from the *elegans, eugracilis, ficsuphil*, *flavohirta, melanogaster, suzukii*, and *takahashii* subgroups.

This conflict may have been caused by differences in project design. The number of gene regions (or characters) sampled and the genomic location of the DNA sequences varied between the two studies. Schawaroch utilized three gene regions (two nuclear and one mitochondrial); whereas, DaLage, et al. (2007) utilized one complete nuclear gene. Although some of the taxa within the *melanogaster* and *obscura* species groups overlapped between the two studies the
total number of taxa varies (Table 1 & Table 2). Schawaroch (2002) sampled 49 representative taxa (43 from the ingroup and 6 from the outgroup); in comparison to the DaLage, et al. (2007) study which contained 74 taxa (68 taxa from the ingroup and 6 from the outgroup).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amount</td>
<td>Percent of Known Taxa</td>
<td>Amount</td>
<td>Percent of Known Taxa</td>
</tr>
<tr>
<td>melanogaster sp. group</td>
<td>180</td>
<td>24%</td>
<td>69</td>
<td>38%</td>
</tr>
<tr>
<td>suzuki sp. subgr.</td>
<td>17</td>
<td>18%</td>
<td>3</td>
<td>18%</td>
</tr>
<tr>
<td>melanogaster sp. subgroup</td>
<td>9</td>
<td>33%</td>
<td>9</td>
<td>100%</td>
</tr>
<tr>
<td>elegans sp. subgroup</td>
<td>5</td>
<td>20%</td>
<td>2</td>
<td>40%</td>
</tr>
<tr>
<td>ficushila sp. subgroup</td>
<td>6</td>
<td>17%</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>eugraculis sp. subgroup</td>
<td>1</td>
<td>100%</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>flavohirta sp. subgroup</td>
<td>1</td>
<td>0%</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>takahashii sp. subgroup</td>
<td>13</td>
<td>31%</td>
<td>3</td>
<td>23%</td>
</tr>
<tr>
<td>ananasae sp. subgroup</td>
<td>25</td>
<td>24%</td>
<td>18</td>
<td>72%</td>
</tr>
<tr>
<td>montium sp. subgroup</td>
<td>88</td>
<td>27%</td>
<td>30</td>
<td>34%</td>
</tr>
</tbody>
</table>

Table 2. Number of representative taxa used throughout studies. The table displays the amount of representative taxa for each sp. subgroup within the melanogaster sp. subgroup utilized by the Schawaroch (2002) study, DaLage, et al. (2007) study, and this investigation. This study increases amount of representative taxa for the melanogaster, takahashii, and montium sp. subgroups. The amount of representative taxa for the montium sp. subgroup used in this study (highlighted in yellow) includes “jambulinaVS” and “jambulinaDaLage”, which may not be two different sp.

The identification of characters in the matrix and handling of characters in the analysis effects the topology of the phylogeny produced. Schawaroch (2002) weighted all base pairs equally. DaLage, et al. (2007) utilized differential weighting based on codon position (i.e., first, second, and third codon positions had the weighting scheme of 4:4:1 respectively) and on nucleotide base change where transitions were down weighted (i.e., transversions:transitions were weighted as 2:1). Both studies appealed to computer programs to generate unbiased alignment in cases where the DNA sequence had indels (insertions and/or deletions of basepairs).
The parameters and programs were cited, however, the DaLage, *et al.* (2007) study made final adjustments to the alignment by eye. Tree-descriptive statistics of length, CI and RI (see *Methods* section for description of each statistic) can be inflated if all characters are used (Kitching, *et al.*, 1998). Hence, Schawaroch only used parsimony informative characters (for further discussion see “Tree reconstruction” of *Methods* section of this study) to generate trees and tree statistics, but DaLage, *et al.* (2007) did not state if all or only parsimony informative characters were used. Both studies reconstructed the phylogenies via maximum parsimony, and DaLage, *et al.* (2007) performed an additional Bayesian analysis (see *Methods* section).

One of the major issues with generations of phylogenies is sampling of taxa and characters. Of course, each plays a role in recovery of nodes and relationships among species (Graybeal, 1998). However, it is still not known which has a more direct influence on the phylogeny (although it is felt that the addition of taxa will affect the phylogeny more rapidly than the addition of characters, and at least 10 taxa should be sampled (Graybeal, 1998)). With this comes the issue and question at hand, “what is a sufficient amount of taxa and characters to sample to construct a robust phylogeny?”

**Purpose**

The Schawaroch (2002) and DaLage, *et al.* (2007) data sets will be used to generate a phylogeny for species within the *Drosophila melanogaster* species group. Phylogenies established here will be compared to previous phylogenies with respect to the clades recovered. This study will establish a stable scheme of relationships among the species that can be used for future comparative biological investigations.
METHODS

Project design

This study ultimately seeks to combine the data sets of Schawaroch (2002) and DaLage, et al. (2007) and to compare the new trees to one another and to trees generated by a re-analysis of the original individual studies (Figure 4). The topologies of these trees will be compared by analyzing clade node recovery with respect to established taxonomy.

Figure 4. Flow chart of phylogenetic hypotheses to be compared. Four phylogenetic analyses will be performed, while keeping search parameters the same among data sets. Both DaLage, et al. (2007) and Schawaroch (2002) will be re-analyzed and two new trees will be generated by combining character information for either or all taxa possible even if the character set is not complete or all taxa that overlap.

Sampling/compare four trees

The first tree is a re-analysis of DaLage, et al. (2007) with 74 taxa sampled and a complete character set (other than for two taxa, see “Re-analysis of previous studies” section of Methods).
but only for a single nuclear gene, *Amyrel*. The second tree is a re-analysis of Schawaroch (2002) which contains 49 taxa and a complete sampling of gene regions throughout the genome - two nuclear (*Adh, hb*) and one mitochondrial (*CoII*).

The third tree will combine the DaLage, *et al.* (2007) and Schawaroch (2002) studies creating a large taxon sample (87 species). However, some taxa do not have character information for all 4 gene/gene regions. Each base pair of missing DNA sequence will be coded as a question mark “?” (Platnick, *et al*., 1991). The last tree includes the 36 taxa that overlap between DaLage, *et al.* (2007) and Schawaroch (2002) studies; this tree will have a complete molecular data set for 4 gene regions – three nuclear, partial *Adh*, partial *hb* and complete *Amyrel* plus partial sequence of mitochondrial gene *CoII*.

Beyond the total number of species and DNA characters sampled, there are questions of sampling distribution. For example, is the taxon sample representative of the species groups and subgroups (Table 2), and is the DNA sampling representative for the genome?

**DNA sequence**

Sequences for species of the *melanogaster* and *obscura* groups for the nuclear gene *Amyrel* from the DaLage, *et al.* (2007) study were downloaded from GenBank (NCBI, Bethesda, MD) (Figure 5, on next page). Both the DNA and amino acid sequence files were downloaded. Within the DNA sequence file, exon 1 and exon 2 with an intervening intron were identified. Aligned partial DNA sequence for two nuclear coding regions of *Adh* and *hb* and one mitochondrial coding region of *CoII* from Schawaroch (2000; 2002) was provided by Dr. Valerie Schawaroch (my mentor).
Figure 5. Example of GenBank (NCBI, Bethesda, MD) file. GenBank (NCBI, Bethesda, MD) files include the known taxonomic classification of the organism, the name of the gene, location of the coding region(s) of the gene, the translated amino acid sequence, and the base pair sequence.
Establishing a character matrix

A character matrix contains the hypotheses of homologies used to create a phylogenetic tree. For each character, a hypothesis of homology (primary homology, \textit{sensu} DePinna, 1991), is estimated by the investigator. For molecular (DNA sequence) data, the nucleotide letter (A, T, C, G) is the character. Decisions, on which character in one taxon corresponds to a character in another taxon, are made by performing alignments. An alignment is the first step in estimating primary homology called topological identity (Brower and Schawaroch, 1996) (Figure 6). The second step in estimating primary homology is called the character state identity (Brower and Schawaroch, 1996). The character state identity refers to the specific character (DNA base pair letter) for each taxon at a specific alignment position within the matrix (Figure 6).

![Gene Sequence](image)

**Figure 6. Hypotheses of homology prior to phylogenetic tree construction.** This initial hypothesis of homology referred to as putative or primary homology of DePinna (1991) is subdivided into a two step process by Brower & Schawaroch (1996). In figure A it is shown that determining the positional correspondence of DNA base pairs across species is the first step in hypothesis of homology for DNA sequence prior to phylogenetic analysis referred to as topological identity \textit{(sensu} Brower & Schawaroch, 1996). In figure B it is shown that once the DNA sequence alignment is established, the DNA base pair assigned for each position is placed in the boxes of the matrix. This second step in hypothesis of homology is referred to as the character state identity \textit{(sensu} Brower & Schawaroch, 1996).

The gene regions \textit{Adh} and \textit{CoII} have no insertions/deletions (indels), which means that all taxa sampled for these gene regions have the same length of base pair sequence; therefore it is possible, and sufficient to perform alignments by eye for these gene regions (Schawaroch, 2002). The \textit{Adh} data set contains 290 characters and the \textit{CoII} data set contains 384 characters.
However, not all sequence is the same length for all taxa (i.e., there are indels), as in the case of *Amyrel* and *hb* gene/gene region. Therefore, to bring all sequence to the same length, an alignment must be performed in which gaps are inserted in the shorter length DNA sequence (Figure 7).

<table>
<thead>
<tr>
<th>Taxon 1</th>
<th>AATGCTTCA (9 base pairs)</th>
<th>Taxon 1</th>
<th>AATGCTTCA (9 characters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon 2</td>
<td>ATGCTTCA (8 base pairs)</td>
<td>Taxon 2</td>
<td>A-TGCTTCA (9 characters)</td>
</tr>
<tr>
<td>Taxon 3</td>
<td>ATTGCTTGA (9 base pairs)</td>
<td>Taxon 3</td>
<td>ATTGCTTGA (9 characters)</td>
</tr>
</tbody>
</table>

**Figure 7. Example of sequence before and after hypothetical alignment.** Figure A displays an example of sequence before an alignment. Taxon 2 differs from Taxon 1 and 3 by one base pair. Figure B displays an example of the sequence after a hypothetical alignment. The hypothetical alignment has placed a gap in the second position of the sequence for Taxon 2. Although Taxon 2 still only has 8 base pairs in sequence, it now has 9 characters of sequence, making it match the amount of characters in Taxon 1 and 3.

Alignments were made using the ClustalW method (Thompson, *et al.*, 1994) within the program MegAlign ver. 8.1.4.7 and ver. 9 (DNASTAR, Inc.) (see Appendix B: Protocol 4). Different parameters will cause differences in where the gap is inserted. Such parameters include (1) gap penalties, parameters in which a cost is set to insert a new gap, and (2) gap length penalties, parameters in which a cost is set to insert gaps in areas where gaps have already been inserted. The higher the penalty, the less likely a gap will be inserted. We have chosen to hold gap length penalty (GLP) constant while varying gap penalty (GP) (*sensu* Schawaroch, 2002). Therefore, when GP is lower than GLP, it will be easier for the program to insert new gaps in areas with no gaps than it will be to insert a gap at an area where a gap has already been inserted. When GP is higher than GLP, gaps will be more easily inserted in areas where gaps already exist.

To eliminate experimenter bias, multiple alignments were performed by computer under various parameters. Differences between alignments are shown by differences in where gaps were inserted. Areas with differences between alignments under varied conditions are referred to
as alignment ambiguous areas (Gatsey, et al., 1993) (Figure 8). This study has chosen to eliminate alignment ambiguous regions.

<table>
<thead>
<tr>
<th>Parameter 1</th>
<th>Taxon 1</th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon 2</td>
<td></td>
<td>G</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>T</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>Taxon 3</td>
<td></td>
<td>G</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>T</td>
<td>G</td>
<td>A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter 2</th>
<th>Taxon 1</th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon 2</td>
<td></td>
<td>G</td>
<td>C</td>
<td>-</td>
<td>T</td>
<td>T</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>Taxon 3</td>
<td></td>
<td>G</td>
<td>-</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>-</td>
<td>A</td>
</tr>
</tbody>
</table>

Figure 8. An illustration of alignment ambiguous sites. Alignment ambiguous sites are regions of sequence in which gaps will be inserted in different positions depending on alignment parameters.

Since the DNA sequence was highly variable for both *Amyrel* and *hb*, alignments were initially performed on the amino acid sequence. Once alignment ambiguous sites were removed, then the amino acid sequence was changed back to DNA sequence to increase the number of possible characters (i.e., for every amino acid there are three DNA base pairs of characters).

Once the topological identity and character state identity have been established (which may require the insertion of gaps by alignment), and all sequence is within a single data matrix, the information held within this matrix are the character by character primary homology statements (DePinna, 1991) to be tested by running a tree analysis (Patterson, 1982). If the groupings hypothesized in the primary homology statement form a single node on the cladogram, then this primary homology is now a tested/true homology statement referred to as a synapomorphy.
Tree reconstruction

Phylogenetic tree reconstruction analyses were run using the criterion of parsimony in the computer program PAUP* version 4.0b4a (Swofford, 2000) (see Appendix B: Protocol 5). Parsimony analysis appeals to the concept of Ockham's razor - when selecting one hypothesis from competing hypotheses, the simplest hypothesis is the best selection (Schawaroch, 2000; Schawaroch, 2002). In terms of phylogenetic hypotheses, applying this concept would be to select for a tree with the least number of steps (or evolutionary changes). Because the matrix has more than 10 taxa, it becomes computationally too complex for a computer to efficiently solve it (it is an NP-complete problem); therefore it is necessary to perform an estimated or heuristic search.

Parsimony uninformative characters do not provide grouping information. Therefore, analyses only included parsimony informative characters. Next, the outgroup (species from the obscura group) was defined. All (base pair) characters were equally weighted. Gaps were treated as missing data ("?"), which has been tradition when working with molecular sequence (Schawaroch, 2002). For each analysis, the parsimony settings were kept as default, except that there was a random addition of taxa in which a random seed. A heuristic search was performed in which there were 10 replicates for each tree search. The branch swapping algorithm used was tree-bisection-reconnection (TBR). The character optimization on the tree was ACCTRAN, meaning that any homoplasy is interpreted as a reversal (DePinna, 1991).

Tree statistics of length, ensemble consistency index (CI), and ensemble retention index (RI) will be reported for the equally most parsimonious tree(s). If the analysis results in more than one equally parsimonious tree, each tree will have the same tree statistics. Therefore, it is only necessary to report the statistics for one of the equally most parsimonious trees. The CI measures
how well all characters fit on the tree. It is a measure of homoplasy since it measures how many
times the primary homology statement is incorrect. CI varies from 0 (100% homoplasious) to 1
(100% of all characters fit on tree, no homoplasy, therefore, 100% consistency). Uninformative
characters have been known to inflate the value of CI (Kitching, et al., 1998); hence tree
analyses only included parsimony informative characters. In addition, as the number of taxa
increases, the CI artificially decreases in value (Kitching, et al., 1998). Therefore, the ensemble
retention index (RI), which does not have the same issues as CI, is also reported. RI reports
similarity, which is interpreted as a synapomorphy; therefore, its value represents synapomorphy
levels.

Bootstrap values (Felsenstein, 1985) and decay indices (Bremer, 1994) will be reported for
the nodes of either the single most parsimonious tree or the first tree in the tree file of a series of
equally most parsimonious trees to measure and compare clade stability. Bootstrap analysis is a
statistical analysis that can be performed in PAUP* version 4.0b (Swofford, 2000) (see
Appendix B: Protocol 6). This analysis randomly resamples the entire data set with replacement.
This means that the analysis may resample the same character multiple times and not sample
others at all. The analysis then generates bootstrap values for each node. The larger the bootstrap
value, the more it can be assumed that there are a larger number of characters in the data set that
agree with the node. The bootstrap analysis was set up for 100 replications with a random seed.
However, there was only one replication per heuristic search. The decay index (a.k.a. Bremer
support value (Bremer, 1994)) determines the amount of additional steps required to collapse
each individual node. As the number of steps required to collapse a node increases, we can say
that the node is more stable. Decay index values were calculated using a combination of
computer analysis programs (see Appendix B: Protocol 7): Autodecay 2.9.8. (Eriksson, 1997) to
make a batch file of decay trees, PAUP* to analyze the data on the tree, and *TreeViewPPC* (Page, 2000) to view the resulting decay tree with values.

In some cases, there may be large amounts of equally parsimonious trees; therefore, a consensus tree, such as a strict consensus tree, or a 50% majority rule consensus tree, is created to summarize the information provided by the equally most parsimonious trees. A strict consensus tree is more conservative, being that it only displays nodes that are in agreement among all equally parsimonious trees. Even if a node is displayed on all equally parsimonious trees but one, it will not be shown on a strict consensus tree. A 50% majority rule consensus tree displays nodes that appear on 50% or more of the equally parsimonious trees. A strict consensus tree will inherently lack resolution because nodes collapse for any node with less than 100% agreement. We have chosen to report the strict consensus tree.

To better communicate and compare the relationships among the sp. subgroups on the trees shown, all taxa have been color-coded by sp. group (for obscura)/subgroup (Figure 9). Both *eugracilis* and *flavohirta* are monotypic, meaning there is only one species known within each sp. subgroup. Therefore, no conclusions can be generated about relationships within these subgroups.

<table>
<thead>
<tr>
<th>Color Key:</th>
</tr>
</thead>
<tbody>
<tr>
<td>obscura group</td>
</tr>
<tr>
<td>suzukii sp. subgroup</td>
</tr>
<tr>
<td>takahashii sp. subgroup</td>
</tr>
<tr>
<td>melanogaster sp. subgroup</td>
</tr>
<tr>
<td>elegans sp. subgroup</td>
</tr>
<tr>
<td>eugracilis</td>
</tr>
<tr>
<td>ficusphila sp. subgroup</td>
</tr>
<tr>
<td>flavohirta</td>
</tr>
<tr>
<td>ananassae sp. subgroup</td>
</tr>
<tr>
<td>montium sp. subgroup</td>
</tr>
</tbody>
</table>

Figure 9. Color key for cladograms. All taxa in the following cladograms will be color-coded according to this scheme.
Re-analysis of Previous Studies

A. Re-analysis of DaLage, et al. (2007) data

Methods

Alignments were created for all 74 taxa for exon 1. Because *D. kitumensis* and *D. psuedoananassae* are missing data for exon 2, they were not included in the alignment for exon 2. In addition, DaLage, et al. (2007) did not include *D. pseudoananassae nigrens* in any analyses because they claimed the sequence was identical to *D. pseudoananaassae pseudoananassae*. However, we found that the sequence varied, and that *D. pseudoanaanaza pseudoananassae* was missing sequence for exon 2. There were overall differences in the methodology employed by the DaLage, et al. (2007) investigation and this study (Table 3).

<table>
<thead>
<tr>
<th>Comparison of DaLage, et al. (2007) and this study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DaLage, et al. (2007)</strong></td>
</tr>
<tr>
<td>intron included</td>
</tr>
<tr>
<td>alignment</td>
</tr>
<tr>
<td>alignment ambiguous sites removed</td>
</tr>
<tr>
<td>method of analysis</td>
</tr>
<tr>
<td>type of tree</td>
</tr>
</tbody>
</table>

Table 3. Comparison of DaLage, et al. (2007) and this study. The differences between the original DaLage, et al. (2007) data analysis and analysis used in this study are summarized in this table. DaLage, et al. (2007) is re-analyzed to keep methodology the same for all data compared. For DaLage, et al. (2007), it should be noted that an initial alignment was performed using CLUSTALW, but further aligning was done by eye. The 50% majority rule consensus tree for both the parsimony and Bayesian inference analysis yielded similar topologies at the sp. subgroup level. This study only deals with parsimony analysis, therefore, we will only compare to the DaLage, et al. (2007) 50% majority rule consensus tree of the parsimony analysis. *DaLage, et al. (2007) reported an alignment ambiguous site in the signal peptide (not part of gene, and not included in this study), and made no indication of its removal.*
The *Amyrel* gene studied in DaLage, *et al.* (2007) contains an exon (654 to 657 base pairs), a highly variable intron (52 to 110 base pairs), and another exon (828 to 831 base pairs). The sequence was obtained from the DaLage, *et al.*, (2007) for 74 taxa from GenBank (NCBI, Bethesda, MD). This published sequence not only contained the first exon, intron, and second exon, but also sequences before the first exon, and after the second exon, probably due to the location of the primers utilized by DaLage, *et al.* (2007) for PCR amplification (Figure 10). This re-analysis only included the exons (intron and excess sequence was removed).

![Figure 10. Relative position of primers used by DaLage, *et al.*, (2007).](image-url)

Figure 10. Relative position of primers used by DaLage, *et al.*, (2007). All primer pairs run over the highly variable intron, and some include sequence beyond the exons.

To identify the alignment ambiguous sites, multiple alignments were performed on the amino acid sequence obtained from GenBank for each exon using MegAlign ver. 8.1.4.7 and ver. 9 (DNASTAR, Inc.). The multiple alignment parameters were generated using the CLUSTALW method with the following parameters: the amino acid change cost was according to the Gonnet residue weight table, and the gap length penalty was a constant value of 10, while the gap penalty value varied from 1 to 50 (1, 5, 10, 15, 20, 30, 40, 50). An alignment ambiguous site consisting of 12 amino acids in length was found in exon 1. Although gaps were inserted in the second exon’s sequence it was consistent throughout parameters (not ambiguous).
The DNA sequence corresponding to the amino acids was identified by comparing the DNA and amino acid Genbank files in Sequencher 4.7 (Gene Codes Corporation). The amino acid sequence was converted back to base pair sequence, gaps were inserted, and the alignment ambiguous site was removed (*sensu* Gatesy, *et al.*, 1993). Question marks (?) will be inserted for the two taxa, *D. kitumensis* and *D. p.pseudoananassae*, where there are missing base pair sequences. As a result of this alignment, the Amyrel gene has 1455 characters for each taxon.

**Results and Discussion**

The original DaLage, *et al.* (2007) tree reported was a 50% majority rule consensus tree summarizing the 37 equally most parsimious trees (Figure 11). DaLage, *et al* (2007) also reported a Bayesian analysis tree which maintained the same relationships at a sp. subgroup level. We have decided to employ parsimony analysis, and therefore will only compare our data to the DaLage, *et al.* (2007) parsimony tree. We have displayed the portions of this tree corresponding to the *obscura* and *melanogaster* species group (Figure 11). In this tree, the following well-established sp. subgroups/clades have been recovered (Table 4 on page 33): *obscura* sp. group, *melanogaster* sp. group, *melanogaster* sp. subgroup, *elegans* sp. subgroup, *takahashii* sp. subgroup, *ananassae* sp. subgroup, *montium* sp. subgroup, and the

---

*Figure 11 (on next page). The obscura and melanogaster species groups section of the DaLage, *et al.* (2007) 50% majority rule tree.* The tree statistics reported (length, CI, and RI) are erroneous because the values are taken from the DaLage, *et al.* (2007) analysis that included 164 taxa for the nuclear Amyrel gene (of which the coding sequence varies from 1470-1485 base pairs (DaLage, *et al.* (2007))). In addition, DaLage, *et al.* (2007) did not state if the analyses were run with or without the inclusion of parsimony uninformative characters which would also affect the values.
More Taxa or More Characters

30

50% Majority Rule
sensu DaLage, et al. (2007)
MP tree statistics
Length = 20,607 steps
C.I. = 0.21
R.I. = 0.76
melanogaster + Asian sp. subgroups clade. The following clades were not monophyletic (not recovered): suzukii sp. subgroup, and the ficusphila sp. subgroup. The melanogaster + Asian sp. subgroups clade and montium sp. subgroup were grouped together as sister taxa, and the ananassae sp. subgroup was basal to that clade. Almost all nodes were fully resolved. However, there was lack of resolution within the ananassae sp. subgroup (see “combs” within the the ananassae sp. subgroup (yellow) in Figure 11). The tree statistics reported in DaLage, et al. (2007) can not be compared with the other analyses in this study, because they represent the tree statistics for the full study, which investigated 164 taxa. It should also be noted that DaLage, et al. (2007) stated that the Amyrel sequence for pseudoananassae nigrens and pseudoananassae pseudoananassae was the same. Therefore, they only reported one representative taxa in their tree. We will inlcude both taxa separately. Also, the DaLage, et al. (2007) analysis shows a “nov. sp.” taxa, not included in our data set for re-analysis.

Two data sets were run for the re-analysis of the Amyrel gene. D. kitumensis and D. pseudoananassae pseudoananassae were both missing data for exon 2; therefore, one analysis was performed with and one without these taxa (74 taxa and 72 taxa total respectively) (Figure 12). Both analyses resulted in 32 most equally most prasimonious trees. A strict consensus tree summarizing the information, included D. kitumensis and D. pseudoananassae pseudoananassae with dotted lines (Figure 12). Both of these analyses yielded trees with identical topologies with the following exceptions: (1) D. kitumensis is sister to the D. affinis + D. pseudoobscura clade.
More Taxa or More Characters

[Diagram showing a phylogenetic tree with various species labeled, including:
affinis, pseudoobscura, kitumensis, bifasciata, imaii, subobscura, biarmipes, mimetica, takahashii, pseudotakahashii, lutescens, erecta, orena, santomea, yakuba, teissieri, mauritiana, sechellia, simulans, melanogaster, eugracilis, levii, ficusphila, flavohirta, elegans, subelegans, lucipennis, ananassae, pallidosa, papuensis-like, atripex, ochrogaster, monieri, phaeopleura, bipectinata, parabipectinata, malerkotliana, malerkotliana pal., pseudoan. nigrens, pseudoan. pseudoan., ercepeae, vallismaia, merina, varians, lachaisei, asahinai, rufa, auraria, quadraria, biauraria, triauraria, bakoue, tsacasi, malagassya, vulcana, bocqueti, bocqueti-like, burlai, chauvacae, diplacantha, dossoui, nikananu, davidi, barbarae, bicornuta, cauverii, punjabiensis, nagarholensis, kikkawaii, leontia, lini, serrata, greeni, jambulinaDaLage]

Strict Consensus
Amyrel (74 and 72 taxa)
MP tree statistics
Length = 3469 steps (74 sp.),
3451 steps (72 sp.)
C.I. = 0.329 (74 sp.), 0.331 (72 sp.)
R.I. = 0.745 (74 sp.), 0.742 (72 sp.)
This new clade is sister to the *D. bifasciata* + *D. imanii* clade. (2) *D. pseudoananassae* is sister to *D. pseudoananassae nigrens*. All clade recovery information was the same for these analyses as well (Table 4). The re-analysis of the *Amyrel* data yielded a different tree than the published DaLage, *et al.* (2007) 50% majority-rule consensus tree, most importantly in the relationships between the three major clades (DaLage, *et al.* (2007) places the *melanogaster* + Asian sp. subgroups clade as sister to the *montium* sp. subgroup while the re-analyzed *Amyrel* analyses provide no resolution to these three clades).

A comparison of the clade recovery between the published DaLage, *et al.* (2007) tree and re-analyzed *Amyrel* data is summarized in Table 4.

**Table 4. Clade recovery analysis for the original (1) DaLage, *et al.* (2007) tree and (2) for the re-analyzed *Amyrel* data’s strict consensus tree(s).** This table compares the topologies of the two trees - the re-analysis of *Amyrel* exon’s 1 and 2 for 74 and 72 species, and the *obscura* + *melanogaster* species group region of the DaLage, *et al.* (2007) *Amyrel* 50% majority rule tree derived for a parsimony analysis. The clades listed and used in this comparison were originally established through a taxonomic revision starting with Bock and Wheeler (1972). Different methods of analysis of the *Amyrel* data yielded different results. The strict consensus tree generated by our analysis provided no resolution for the three major clades of the *melanogaster* + Asian species subgroups, *ananassae* species subgroup, and the *montium* species subgroup. The DaLage, *et al.* (2007) 50% majority rule parsimony tree demonstrates resolution, placing the *melanogaster* and *montium* subgroups as sister taxa with the *ananassae* subgroup basal to that clade. However, in both analyses, the well-established species subgroups of *melanogaster*, *elegans*, *takahashii*, *ananassae*, and *montium* and the *melanogaster* + Asian species subgroups clade were all recovered. The *ficusphila* and *suzukii* subgroups were not monophyletic.
B. Re-analysis of Schawaroch (2002) data

Methods

The data set and alignments for Adh and CoII were as described in Schawaroch (2002). However, the treatment of the \( hb \) gene data differed between Schawaroch (2002) and this re-analysis. After removal of alignment ambiguous sites, the \( hb \) gene had gaps coded either as missing or as binary characters, depending on the alignment context (Schawaroch, 2002). This re-analysis codes all gaps as missing which required the re-inclusion of 9 DNA base pair characters into the originally aligned sequence (leaving 441 characters with the 9 base pairs re-inserted). Schawaroch had portions of coding DNA sequence from the following genes Adh (290 characters), mt:CoII (384 characters), and \( hb \) (434 characters). The complete data set for 49 taxa contained 1108 characters, 341 (31%) characters being parsimony informative (Table 5, on page 52).

Results and Discussion

A re-analysis yielded a single most parsimonious tree (Figure 13). This tree very closely resembles the original tree published in Schawaroch (2002) with two differences in the re-analysis. The \textit{obscura} sp. group was monophyletic in the re-analyzed tree while it was unresolved in the Schawaroch (2002) study. Also, the tree length was 1538 steps (two steps shorter than the published Schawaroch (2002) tree), while, the CI (0.348) and the RI (0.666) remained the same as the published Schawaroch (2002) tree (Table 6, on page 52).

Figure 13 (on next page). Schawaroch (2002) data re-analyzed. This re-analyzed data set converted the binary coding for certain gaps back to the traditional “gaps as missing data” (increasing the data set by 6 characters) for the \( hb \) gene region. The topology, the CI and RI were the same as reported in Schawaroch (2002); however, the overall length of the tree was shorter by two steps.
More Taxa or More Characters

Schawaroch (2002)
Length = 1538 steps
C.I. = 0.348
R.I. = 0.666
RESULTS

In the methods section, the validity of using the re-analyses for both the DaLage, et al. (2007) and Schawaroch (2002) data sets was established. Now, each of these re-analyzed trees will be compared to each of the new total evidence trees - 87 taxa with incomplete character sampling and 36 taxa with complete characters sampling.

Re-analysis of DaLage, et al. (2007) data

Parsimony informative characters per analysis

As previously stated in the methods, all tree analyses were run only using parsimony informative characters. The re-analysis of DaLage, et al. (2007) for both the 74 and 72 taxon analyses contained 1455 total base pair characters for Amyrel with 642 characters (44%) being parsimony informative (Table 5, on page 52).

Trees (equally most parsimonious trees and strict consensus)

Both analyses resulted in 32 most equally most parsimonious trees; therefore, a strict consensus was taken. The trees were identical with the exceptions for the two taxa that were either included as question marks for exon 2 in the 74 taxon analysis or omitted in the 72 taxon analysis. Taxa were added to the topology as described earlier under methods of re-analysis of DaLage, et al. (2007) data. Therefore, a single summary tree for both taxon data sets analysis is presented (Figure 12).

Tree statistics (data fit to tree topology – length, CI, and RI)

For each of the equally most parsimonious trees, the tree length was 3469 steps for 74 taxa and 3451 steps for 72 taxa. The CI was 0.329 and 0.331 for the 74 and 72 taxon sampling...
respectively. The RI values were 0.745 and 0.742 for the 74 and 72 taxon sampling respectively (Table 6, on page 52).

Recovery of clades (traditional taxonomic groupings)

In the strict consensus tree (for both 74 and 72 taxa) (Figure 12), the following well-established sp. subgroups/clades have been recovered (Table 7, on page 53): melanogaster sp. group, melanogaster sp. subgroup, elegans sp. subgroup, takahashii sp. subgroup, ananassae sp. subgroup, montium sp. subgroup, and the melanogaster + Asian sp. subgroups clade. The following clades were not monophyletic (not recovered): obscura sp. group, suzukii sp. subgroup, and the ficusphila sp. subgroup. There was no resolution provided for the three major clades of the melanogaster + Asian sp. subgroups clade, montium sp. subgroup and ananassae sp. subgroup.

Tree node statistics (bootstrap and decay indices)

If more than one most equally parsimonious tree results for the analyses, then the bootstrap and decay values were calculated for the first tree in the tree file of equally most parsimonious trees. Only bootstrap values that are 50% or greater are reported.

The bootstrap and decay values were also found for each node (Figures 14 and 15) for the first of the equally most parsimonious trees of both re-analyses of the Amyrel data (74 and 72 sp.).

The topologies and bootstrap values of both analyses (for 74 and 72 taxa) were almost identical. The values all strongly support the existence of the melanogaster sp. group, and the existence of the three major clades (Table 8, on page 54). DaLage, et al. (2007) proposed to elevate the ananassae and montium sp. subgroups to sp. groups. Although the bootstrap and decay values are particularly high (for bootstrap, over 50% is supported, and for the decay
analysis a high decay index would be at least 5) for each clade, they are even higher for the *melanogaster* sp. group. My interpretation is that this means the subgroups within the group all fit well with each other, and therefore, should not be elevated. Further analyses all support this notion (Table 8, on page 54). However, there is no resolution provided in these analyses as to the relationships of the three major clades. The bootstrap analysis displayed no resolution at all. The decay analysis performed on the first equally parsimonious tree displayed the *melanogaster* + Asian sp. subgroups clade as sister to the *montium* sp. subgroup. However, the decay index was 0. It should also be noted that the strict consensus did not display this relationship, meaning, not all equally parsimonious trees even possessed this node at all.

**Figure 14 (on page 39). Bootstrap and decay values for the re-analyzed *Amyrel* data for 74 sp.** The figure displays the first of the equally parsimonious trees (Tree 1) for the re-analyzed *Amyrel* data (for 74 sp.). Bootstrap values are displayed in blue on top of the node and decay values are displayed in red on the bottom of the node. Bootstrap values of less than 50% are not displayed. Not all nodes in this first of the equally parsimonious trees exist in the decay trees (in which case, no decay value is displayed). Although this tree displays resolution of the three major clades, both the bootstrap and decay values were less than 50% and 0 respectively. The bootstrap analysis produced a tree with a different topology from Tree 1. The dotted line indicates where Tree 1 differs from the tree generated by the bootstrap analysis.

**Figure 15 (on page 40). Bootstrap and decay values for the re-analyzed *Amyrel* data for 72 sp.** The figure displays the first of the equally parsimonious trees (Tree 1) for the re-analyzed *Amyrel* data (for 72 sp.). Bootstrap values are displayed in blue on top of the node and decay values are displayed in red on the bottom of the node. Bootstrap values of less than 50% are not displayed. Not all nodes in Tree 1 exist in the tree created by decay analysis (in which case, no decay value is displayed). The values highlighted in yellow indicate a node that was actually more basal in the bootstrap and decay analysis, leaving no resolution among the three major clades. The dotted line indicates where Tree 1 differs from the tree generated by the bootstrap analysis, in which there is one less node.
More Taxa or More Characters

40
Re-analysis of Schawaroch (2002) data

Parsimony informative characters per analysis

Schawaroch had portions of coding DNA sequence from the following nuclear genes, Adh (290 characters), hb (434 characters) and mitochondrial gene, CoII (384 characters). The complete data set for 49 taxa contains 1108 characters, 341 (31%) characters being parsimony informative (Table 5, on page 52).

Trees (equally most parsimonious trees and strict consensus)

A single most parsimonious tree was obtained for this 49 taxon 3 gene region analysis (Figure 13).

Tree statistics (data fit to tree topology – length, CI, and RI)

The single most parsimonious tree length was 1538 steps, with a CI= 0.348 and RI = 0.666, which was obtained for this 49 taxon, 3 gene region analysis (Table 6, on page 52).

Recovery of clades (traditional taxonomic groupings)

In this tree, the following well-established sp. subgroups/clades have been recovered (Table 7, on page 53): obscura sp. group, melanogaster sp. group, melanogaster sp. subgroup, takahashii sp. subgroup, ananassae sp. subgroup, montium sp. subgroup, and the melanogaster + Asian sp. subgroups clade. The suzukii sp. subgroup was not monophyletic. This data set has only one representative from both the elegans and figusphila sp. subgroup; therefore, no conclusions on clade recovery can be made for these subgroups. The ananassae and montium sp. subgroup were grouped together as sister taxa, with the melanogaster + Asian sp. subgroups clade as basal to that clade.
Tree node statistics (bootstrap and decay indices)

The bootstrap and decay values were calculated for the single most parsimonious tree. Only bootstrap values that are 50% or greater are reported.

The bootstrap and decay values were found for each node (Figure 16) for this single most parsimonious tree.

All of the three major clades were strongly supported (Table 8, on page 54). The *melanogaster* sp. group is the most supported when looking at both bootstrap and decay values. This further supports the idea that the *ananassae* and *montium* sp. subgroups should not be elevated to sp. groups. The bootstrap analysis united the *melanogaster* + Asian sp. subgroups clade with the *ananassae* sp. subgroup, which is a different topology than the strict consensus. Although the bootstrap value is relatively low (57%), it is still supported. Contrary to that, the decay analysis displayed the same overall topology as the strict consensus tree for the major clades, but the relation of the *ananassae* and *montium* sp. subgroups as sister taxa was weakly supported with a low decay index of 1.

**Figure 16 (on next page). Bootstrap and decay values for the re-analyzed Schawaroch (2002) data.** The figure displays the first of the equally parsimonious trees (Tree 1) for the re-analyzed Schawaroch (2002) data. Bootstrap values are displayed in blue on top of the node and decay values are displayed in red on the bottom of the node. Bootstrap values of less than 50% are not displayed. Not all nodes in Tree 1 existed in the decay trees (in which case, no decay value is displayed). The dotted line indicates where Tree 1 differs from the tree generated by the bootstrap analysis, in which there is an additional node.
More Taxa or More Characters
43
**Total evidence - 87 taxa with incomplete gene sampling**

**Parsimony informative characters per analysis**

A total evidence analysis (Kluge, 1989) for four gene/gene regions for 87 taxa was performed. This data set combines all taxa in the DaLage, *et al.* (2007) and Schawaroch (2002) studies. In cases where the taxa sampled does not overlap, there is missing character data.

For this four gene/gene region data set there were 2563 total characters, with 983 characters (38%) being parsimony informative (Table 5, on page 52).

**Trees (equally most parsimonious trees and strict consensus)**

A strict consensus was taken of 955 equally parsimonious trees (Figure 17).

**Tree statistics (data fit to tree topology – length, CI, and RI)**

For each most equally most parsimonious tree: length = 5042 steps, the CI = 0.333 and the RI = 0.723 (Table 6, on page 52).

**Recovery of clades (traditional taxonomic groupings)**

The following well-established sp. subgroups/clades have been recovered in the strict consensus tree (Table 7, on page 53): *melanogaster* sp. group, *melanogaster* sp. subgroup, *elegans* sp. subgroup, *ananassae* sp. subgroup, *montium* sp. subgroup, and the *melanogaster* + Asian sp. subgroups clade. The following clades were not monophyletic: *obscura* sp. group, *suzukii* sp. subgroup, *ficusphila* sp. subgroup, and the *takahashii* sp. subgroup. There was no resolution provided for the three major clades (*melanogaster* + Asian sp. subgroups, *montium* and *ananassae* clades).

---

**Figure 17 (on next page). The total evidence analysis for 87 taxa.** This analysis had question marks for many of the 87 taxa because the character data sets (*Adh, hb, CoII, Amyrel*) were not complete. The analyses yielded 955 equally most parsimonious trees; therefore, a strict consensus was taken. The tree statistics are representative of each of the equally most parsimonious trees.
More Taxa or More Characters

Strict Consensus
Total Dataset (87 sp.)
MP Tree Statistics
Length = 5042 Steps
C.I. = 0.333
R.I. = 0.723
Tree node statistics (bootstrap and decay indices)

If more than one most equally parsimonious tree results for the analyses, then the bootstrap and decay values were calculated for the first tree in the tree file of equally most parsimonious trees. Only bootstrap values that are 50% or greater are reported.

The bootstrap values were found for each node (Figure 18) for the first of the equally most parsimonious trees. The decay analysis ran for four consecutive weeks (and was still running) in PAUP* version 4.0b4a (Swofford, 2000), at which point, it was determined that the analysis would not be included in the thesis due to the constraint of time placed on this project.

All of the three major clades were strongly supported (Table 8, on page 54). The melanogaster sp. group is the most supported when looking at the bootstrap values. This further supports the idea that the ananassae and montium sp. subgroups should not be elevated to sp. groups. Conflicting with the strict consensus tree and the first equally parsimonious tree, the bootstrap analysis united the melanogaster + Asian sp. subgroups clade with the monium sp. subgroup. However, this node was weakly supported (57%). Also, clades within the montium sp. subgroup change their topology in the bootstrap analysis. This may point to some rare character that is not being resampled during the analysis.

Figure 18 (on next page). Bootstrap Tree for the Total Evidence Analysis of 87 sp. The figure displays the first of the equally parsimonious trees (Tree 1) for the Total Evidence Analysis (for 87 sp.). Bootstrap values are displayed in blue on top of the node. A decay analysis was performed on this analysis, but did not provide any results in a reasonable time (the analysis ran for over a month). Bootstrap values of less than 50% are not displayed. It should be noted that the bootstrap analysis displayed a topology change (from Tree 1) for the following groups of taxa: (1) auraria, quadraria, triauraria, and (2) barbarae, mayri, birchii, bicornuta, cauverii, jambulinaVS, punjabiensis, nagarholensis, serrata.
Total evidence - 36 taxa with complete characters sampling

Parsimony informative characters per analysis

The total evidence analysis (Kluge, 1989) for the 36 overlapping taxa and complete character (gene/gene region) sampling was performed. This data set did not include any taxa with missing data.

For this four gene/gene region data set, there were 2563 total characters with 983 characters (38 %) were parsimony informative (Table 5, on page 52).

Trees (equally most parsimonious trees and strict consensus)

The analysis resulted in a single most parsimonious tree (Figure 19).

Tree statistics (data fit to tree topology – length, CI, and RI)

The tree length was 3647 steps. The CI was 0.413 and the RI was 0.652 (Table 6, on page 52).

Recovery of clades (traditional taxonomic groupings)

In this tree, the following well-established sp. subgroups/clades have been recovered (Table 7, on page 53): melanogaster sp. group, suzukii sp. subgroup, melanogaster sp. subgroup, takahashii sp. subgroup, ananassae sp. subgroup, montium sp. subgroup, and the melanogaster + Asian sp. subgroups clade. The obscura sp. group was not monophyletic-. This data set has only one representative from both the elegans and ficusphila sp. subgroups; therefore, no conclusions on clade recovery can be made for these subgroups. The ananassae and montium sp. subgroup were grouped together as sister taxa, with the melanogaster + Asian sp. subgroups clade as basal to that clade.

Figure 19 (on next page). The total evidence analysis for 36 taxa. This data set has complete characters sampling for all three gene regions of Schawaroch (2002) plus the complete Amyrel gene sequence. This analysis resulted in a fully resolved single most parsimonious tree.
Single Most Parsimonious Tree
Total Dataset (36 sp.)
Length = 3467 Steps
C.I. = 0.413
R.I. = 0.652
Tree node statistics (bootstrap and decay indices)

The bootstrap and decay values were calculated for the single most parsimonious tree. Only bootstrap values that are 50% or greater are reported.

The bootstrap and decay values were found for each node (Figure 20) for this single most parsimonious tree.

All of the three major clades were strongly supported (Table 8, on page 54). The *melanogaster* sp. group is the most supported when looking at the bootstrap values. This further supports the idea that the *ananassae* and *montium* sp. subgroups should not be elevated to sp. groups. Although the *ananassae* and *montium* sp. subgroups have been united in both the bootstrap and decay analysis, the values are relatively low (bootstrap 65%, decay index of 1). It should also be noted that the topology within the basal relationships of the *melanogaster* + Asian sp. subgroups clade of the bootstrap analysis changes from the single most parsimonious tree. However, one of the bootstrap values (within the clade) was relatively low (50%). That same node also had a well supported decay index of 15. Also, the overall clade is strongly supported (bootstrap 96%, and decay index of 15).

---

**Figure 20 (on next page).** Bootstrap and decay values for the Total Evidence Analysis (36 sp.). The figure displays the first of the equally parsimonious trees (Tree 1) for the Total Evidence Analysis of the 36 overlapping sp. Bootstrap values are displayed in blue on top of the node and decay values are displayed in red on the bottom of the node. Bootstrap values of less than 50% are not displayed. Not all nodes in Tree 1 existed in the decay tree (in which case, no decay value is displayed). The dotted line indicates where Tree 1 differs from the tree generated by the bootstrap analysis, in which there is an additional node.
The results are summarized in the following tables.

<table>
<thead>
<tr>
<th></th>
<th>Amyrel Exons</th>
<th>VS data re-analyzed</th>
<th>Total Dataset (all 87 sp.)</th>
<th>Total Dataset, only overlapping taxa (36 sp.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Characters</td>
<td>1455</td>
<td>1108</td>
<td>2563</td>
<td>2563</td>
</tr>
<tr>
<td>Parsimony Informative Characters</td>
<td>642 to 640</td>
<td>341</td>
<td>983</td>
<td>983</td>
</tr>
<tr>
<td>Percent Parsimony Informative Characters</td>
<td>44%</td>
<td>31%</td>
<td>38%</td>
<td>38%</td>
</tr>
</tbody>
</table>

**Table 5. Parsimony informative characters.** Although the data set may contain a large amount of characters, upon parsimony analysis, less than half are parsimony informative. None of the analyses had more than 44% parsimony informative characters.

<table>
<thead>
<tr>
<th></th>
<th>Amyrel re-analyzed (74 taxa)</th>
<th>Amyrel re-analyzed (72 taxa)</th>
<th>Schawaroch (2002) re-analyzed</th>
<th>Total evidence with &quot;?&quot; (87 taxa)</th>
<th>Total evidence without &quot;?&quot; only overlap sp. (36 taxa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number most equally parsimonious trees</td>
<td>32</td>
<td>32</td>
<td>1</td>
<td>955</td>
<td>1</td>
</tr>
<tr>
<td>Tree Length</td>
<td>3469</td>
<td>3451</td>
<td>1538</td>
<td>5042</td>
<td>3647</td>
</tr>
<tr>
<td>Tree stat CI</td>
<td>0.329</td>
<td>0.331</td>
<td>0.348</td>
<td>0.333</td>
<td>0.413</td>
</tr>
<tr>
<td>Tree stat RI</td>
<td>0.745</td>
<td>0.742</td>
<td>0.666</td>
<td>0.723</td>
<td>0.652</td>
</tr>
<tr>
<td>Number of parsimony informative characters</td>
<td>642</td>
<td>640</td>
<td>341</td>
<td>983</td>
<td>983</td>
</tr>
</tbody>
</table>

**Table 6. Summary of tree statistics for strict consensus or single most parsimonious tree.** Reported are the number of equally most parsimonious trees, tree length, CI, RI, number of included characters, number of parsimony informative characters, nodes within the melanogaster species group, number of unresolved nodes within the melanogaster species group, and the percentage of resolved nodes within the melanogaster species group.
<table>
<thead>
<tr>
<th>CLADE</th>
<th>DaLage et al. 2007</th>
<th>Amyrel Exons 1&amp;2 (74/72 sp)</th>
<th>VS data re-analyzed</th>
<th>Total Dataset (all 87 sp.)</th>
<th>Total Dataset, only overlapping taxa (36 sp.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>obscura sp. group</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>melanogaster sp. group</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>suzuki sp. subgr.</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>melanogaster sp. subgr.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>elegans sp. subgr.</td>
<td>yes</td>
<td>yes</td>
<td>N/A</td>
<td>yes</td>
<td>N/A</td>
</tr>
<tr>
<td>ficusphila sp. subgr.</td>
<td>no</td>
<td>no</td>
<td>N/A</td>
<td>no</td>
<td>N/A</td>
</tr>
<tr>
<td>takahashii sp. subgr.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>ananassae sp. subgr.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>montium sp. subgr.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>mel. + Asian sp. subgroups</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>mel. + Asian sp. subgroups + ananassae sp. subgr</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>mel. + Asian sp. subgroups + montium sp. subgr</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ananassae + montium sp. subgr</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 7. Clade Recovery. This table compares the topologies of the five trees shown above: (1) 5) The obscura + melanogaster species group region of the DaLage, et al. (2007) Amyrel exon 1 and 2 50% majority rule tree derived for a parsimony analysis, (2) the re-analysis of Amyrel exon 1 and 2 for 74 and 72 species, (3) the re-analysis of Schawaroch (2002), (4) the total evidence tree for 87 species where there is missing character data, and (5) the total evidence tree for 36 species and complete character data. The clades listed and used in this comparison were originally established through a taxonomic revision starting with Bock and Wheeler (1972).
More Taxa or More Characters

<table>
<thead>
<tr>
<th>CLADE</th>
<th>Amyrel Exons 1&amp;2 (74/72 sp)</th>
<th>VS data re-analyzed</th>
<th>Total Dataset (all 87 sp.)</th>
<th>Total Dataset, only overlapping taxa (36 sp.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bootstrap</td>
<td>decay</td>
<td>bootstrap</td>
<td>decay</td>
</tr>
<tr>
<td>melanogaster sp. group</td>
<td>100%</td>
<td>90</td>
<td>100%</td>
<td>35</td>
</tr>
<tr>
<td>ananassae sp. subgroup</td>
<td>98-100%</td>
<td>12</td>
<td>98%</td>
<td>9</td>
</tr>
<tr>
<td>montium sp. subgroup</td>
<td>98-100%</td>
<td>22-23</td>
<td>100%</td>
<td>14</td>
</tr>
<tr>
<td>mel. + Asian (Asian)sp. subgroup</td>
<td>95-96%</td>
<td>16</td>
<td>89%</td>
<td>3</td>
</tr>
<tr>
<td>Asian + ananassae sp. subgr</td>
<td>X</td>
<td>X</td>
<td>57%</td>
<td>x</td>
</tr>
<tr>
<td>Asian + montium sp. subgr</td>
<td>X</td>
<td>0</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>ananassae + montium sp. subgr</td>
<td>X</td>
<td>X</td>
<td>x</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 8. Summary of bootstrap and decay values for all analyses. The table displays the bootstrap and decay values for the major three clades and their relationships. The bootstrap and Bremer tree were not always in agreement on these relationships, in which cases there will be a bootstrap value for one relationship, and a decay index for another. Wherever an “X” is displays indicates that this relationship was not present, or that there was no resolution for the relationship described. The re-analyzed Amyrel analysis values are displayed as a range, representing both analyses (for 74 and 72 sp.). It is sufficient to display this range because these are statistical analyses, and different values could have been generated each time the analyses were performed. For the Total Evidence of the 36 overlapping sp. analysis, the bootstrap analysis displayed a node uniting the melanogaster + Asian sp. subgroups clade while the decay analysis did not display this node.
DISCUSSION

Among all analyses, it was consistently found that the following clades were recovered as monophyletic (Table 7): the *melanogaster* sp. group, the *melanogaster* sp. subgroup, the *ananassae* sp. subgroup, the *montium* sp. subgroup, and the *melanogaster* + Asian sp. subgroups clade.

It should be noted that for all analyses, the monophyly of the *flavohirta* and *eugracilis* sp. subgroups could not be tested since they are monotypic subgroups. Also, in both the Schawaroch (2002) data re-analyzed and total evidence for 36 taxa analyses, the monophyly of the *elegans* and *ficusphila* species subgroups was not possible because there was only one representative taxon sampled in each case. For the re-analysis of DaLage, et al. (2007) and the total evidence 87 taxon trees, the two representative taxa of the *elegans* species subgroup were recovered as monophyletic (although there are more sp. within this sp. subgroup). However, the *ficusphila* species subgroup, with two representative taxa, was not monophyletic in these analyses. The question as to the monophyly for the *ficusphila* sp. subgroup needs to be further investigated. Perhaps, more taxa could be tested since there are six known species within the subgroup and this study only analyzed two representatives (Table 2).

All four phylogenetic analyses in this study did not recover the *suzukii* sp. subgroup as monophyletic (Table 7). The monophyly of the *suzukii* sp. subgroup has long been questioned, starting with Bock & Wheeler (1972). The members of the subgroup were originally placed based on generalized morphological male genitalia characteristics. However, there are many disparities in other morphological characteristics commonly used to classify *Drosophila*, such as the sex comb and other phallic structures (Schawaroch, 2002). This lack of monophyly is also
seen in molecular studies, including DaLage, *et al.* (2007) and Schawaroch (2002). This study’s findings agree that the suzukii sp. subgroup needs a redefinition/revision.

The total evidence analyses for 36 taxa provided the most resolution and most closely resembles the topology generated from the Schawaroch (2002) data re-analyzed. Both analyses place the ananassae + montium species subgroups as sisters and the melanogaster + Asian subgroups clade as basal; whereas, the DaLage, *et al.* (2007) 50% majority rule parsimony tree places the melanogaster + Asian sp. subgroups as sister to the montium species subgroup with the ananassae species subgroup in a basal position. Both the Amyrel data re-analyzed and the total evidence analysis for the 87 taxa provided no resolution for these three major clades. However, both of these data sets included missing data.

The bootstrap and decay values all strongly support the existence of the three major clades (Table 8). Even more supported is the entire melanogaster sp. group, demonstrating that the three clades should all be grouped together, not as separate sp. groups as proposed by DaLage, *et al.* (2007). It seems that for any analyses that displayed resolution as to the relationships of the three major clades, these relationships were weakly supported. Therefore, it is my belief that there are some rare characters that are sometimes uniting these groups together. This calls for more characters to provide resolution at the sp. subgroup level. This demonstrates that not all genes are suited to provide clarity at the level desired. Some characters may be better at resolving disputes of taxa within a certain sp. subgroup, while others may vary enough to settle disputes on the relationships of the overall sp. subgroups. Choosing a gene/gene region to study should take the taxonomic level resolved into consideration. With this in mind, the concept of total evidence (Kluge, 1989) can be applied in an attempt to resolve more than one taxonomic level.
The total evidence analyses for 36 taxa resulted in a fully resolved single most parsimonious tree. In contrast, the total evidence analysis for the 87 taxa yielded 955 equally most parsimonious trees as a strict consensus of the data which lacked resolution (Table 6). This lack of resolution in a strict consensus is understandable because a strict consensus only shows nodes that are unambiguously supported among all the most equally parsimonious trees. However, it appears that the lack of characters (missing data) results in conflicting (ambiguously supported) nodes and contributes to the overall homoplasy of the data set. In conclusion, data sets with full character sampling for all taxa are preferred over data sets with more taxa but less complete character data sets.

To address the original question posed, “Is it better to increase taxa or characters?” it is necessary to introduce the current state of Systematics. Initially, analyses were limited in the number of taxa that could be sampled because of computational limitations. Computer analysis (e.g., PAUP*) and more heuristic methods of sampling large data sets (such as parsimony rachet) have eliminated the taxon sampling limits on data sets. Characters are now the limiting factor in phylogenetic analyses. With this said, there is one main conclusion that can be drawn from these experiments: Full character sampling allows for more supportive statements on the relationships of taxa better than the addition of taxa with incomplete data. This is most evident when comparing the total evidence analysis for 87 taxa (which has missing data) and the total evidence analysis for the 36 overlapping taxa (which has a complete data set). The total evidence analysis for 87 taxa produced 955 equally most parsimonious trees, while the total evidence analysis for the 36 taxa produced only a single most parsimonious tree (Table 6). This clearly demonstrates that the missing data causes disagreement in the topologies provided by the characters.
Therefore, it is better to have more characters with complete data than to add more taxa to the data set.

**Future Work**

Future work can be focused on providing resolution in the relationships amongst taxa within the *melanogaster* sp. group. In doing so, missing data must be minimized, as this study has concluded that missing data increases the amount of conflicting topologies possible for a data set. More sequence data must be generated for the taxa sampled in this study area. Therefore, all near future efforts should be focused on molecular work in obtaining new sequence data for the non-overlapping taxa utilized in this study. This will have implications for current investigations where whole genomes are being sequenced for several model organisms.
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentor Dr. Valerie Schawaroch of Baruch College and the American Museum of Natural History. She is truly the most patient person I know, and this trait was much needed such projects. Above all, Dr. Schawaroch has granted me with a quality experience. I am extremely thankful for the countless hours she has taken from her life in my quest to gain knowledge. Every question I ever had related to the sciences was more than fully answered and supported with background information. I have witnessed first-hand the dedication she puts into every single student, and I am happy for the time I was lucky enough to spend with her.

Above all, Dr. Schawaroch has a genuine grasp on what a student research project entails. She provided the foundation of the principles that supported the project. She was there to guide me through new procedures. As the project moved forward, she also allowed me to branch into the work on my own (with her close by if needed). She graciously allowed me to make my own mistakes and learn from them, and was there to help me collect the pieces and re-evaluate my methods.

What I would like to thank Dr. Schawaroch for the most is the knowledge gained that will be applicable in all my future work. Not only did she spend time ensuring I understood scientific concepts, but she also dedicated huge amounts of effort guiding me on how to manage a scientific investigation. I only hope that she understands the value of what I have learned from working with her.
I appreciate the scientific endeavors of the investigators in the frequently cited DaLage, *et al.* (2007) paper - J.-L. DaLage, G.J., Kergoat, F. Maczkowial, J.-F. Silvain, Marie-Louise Cariou, and Daniel Lachaise - whose data was an integral part of my project. I am indebted to the NIH for the GenBank (NCBI, Bethesda, MD) archive, which made this investigation and other data mining projects possible.

The Metropolitan Association of College and University Biologists (MACUB) hosts an annual conference providing undergraduate investigators such as myself an opportunity to attend a professional meeting and to present my research findings in the form of a poster.

My wet laboratory experiences were supported by a Benjamin Cummings/MACUB 2010 Student Research Award and matching grant from the Weissman School of Arts and Sciences of Baruch College. This project was also made possible by a Eugene M. Lang 2008-2009 Junior Faculty Research Fellowship to Dr. Valerie Schawaroch.

I would like to acknowledge the dedication and efforts of Dr. Helene Eisenman, Dr. Edyta Greer, Dr. Keith Ramig, and Dr. Seymour Schulman who took their time to read through my work and make useful comments.

Dr. Susan Locke of Baruch College was essential, guiding and supporting me through this honors thesis and honors degree. I would also like to give a very special thank you to Keri Bertino, Director, and Heather Samples of the Writing Center, and Stephen Francoeur of the William and Anita Newman Library of Baruch College whose experience in writing and literature research provided me with a foundation for writing my thesis.

I am grateful to Dr. Angela Klaus of Seton Hall University, Dr. Mary Egan of the American Museum of Natural History, and Dr. Joan Japha (professor emeritus) of Baruch College for
providing background knowledge and technical assistance throughout my time working with Dr. Valerie Schawaroch.

I would like to thank the very special people within the Department of Natural Sciences at Baruch College. First, the laboratory staff, Dalchand (Neil) Rampaul, Merton Lewis, Doris Law, and Beremis Perez were all extremely helpful any time I needed assistance or materials from the department. Sonia Donaldson was essential for her constant help in coordinating with individuals in the department. I would like to sincerely thank the student volunteers, Ekaterina Migunova and Eugene Kharonov, in the Schawaroch Laboratory. A fellow undergraduate investigator in the Schawaroch Laboratory, Chuen Yan (Jamie) Lau was extremely helpful and supportive in the final stages of my project –Thank you.

Most importantly is the dedication of the professors of the Baruch College Natural Science Department, who caused me to change my career goal from accounting to the sciences. First, I am extremely grateful for Dr. Emil Gernert, the man I hold responsible for sparking my interest in the sciences with my first Biology course. If it was not for this class and how it was taught, my life may have taken a very different course. I would also like to thank Dr. John Wahlert for his interest in all of the students, and for the interest he has shown in my work and progress. Last, I would like to thank the professors of my courses who provided me with an excellent education in the sciences - Dr. Joel Brind, Dr. Robert Butler, Dr. Helene Eisenman, Dr. Emil Gernert, Dr. Edyta Greer, Dr. Ramzi Khuri, Dr. Chandrika Kulatilleke, Dr. Jason Munshi-South, Dr. Keith Ramig, Dr. Valerie Schawaroch, Dr. David Szalda, and Dr. John Wahlert.
REFERENCES


More Taxa or More Characters

MegAlign ver. 8.1.4.7 and ver. 9. 2010-1011. Part of the Lasergene Core Suite 9. DNASTAR, Inc., Madison, WI, USA


APPENDIX A: WET LAB PROCEDURES

The original course of this project was to obtain sequence data for the taxa that did not overlap between the Schawaroch (2002) and DaLage, et al. (2007) studies, generating a complete data set for all four gene/gene regions. The project was actually too large for an honor’s thesis. Therefore, it was reconceived as a data mining project (working with previously published DNA sequence). However, I have learned how to culture (rear) flies, isolate DNA, and amplify DNA by the polymerase chain reaction (PCR) (Figure 21).

Figure 21. Lab techniques for this project. Lab techniques utilized in this project include (A) DNA extraction, (B) PCR amplification, (C) Gel Electrophoresis, and (D) DNA sequencing. The center image displays the lab workspace.

The following is a description of the techniques I also learned during my thesis project which were not used in the thesis: (1) extract DNA from taxa either directly raised in culture, or frozen samples, (2) select a primer and amplify fragments of the DNA by PCR, (3) isolate these fragments by gel electrophoresis, and (4) use these isolated fragments for a sequencing process (various methods of sequencing).
Culturing fruit flies

Cultures of fruit flies were obtained from the National *Drosophila* Species Stock Center, San Diego, California. Requirements for raising *Drosophila* species in culture include preparation of food medium and control of the physical environment. Special food ingredients beyond the standard cornmeal recipe were bananas and noni fruit. Environmental requirements are temperature (some required 18°C, while most other taxa can be raised at room temperature), lighting cycles (12-hour day/night cycle), and humidity.

Additional species were previously raised in culture and frozen at -80°C. Issues with frozen samples arise because of the degradation of DNA that can occur (Figure 22). This limited us to sampling, via PCR, relatively small DNA fragments (approx. 400 bp in length).

**Figure 22.** Experiment showing DNA degradation. A. A model of the mt: *ColI* gene primer positions. The primer sequences are published in Brower (1994). Numbers correspond to the position from the 3’ end of *Drosophila yakuba* mtDNA sequence (Clary & Wolstenholme, 1985). B. This inverted gel image compares results from mt:*ColI* PCR amplification for (i) an external primer pair (2792-3772) in contrast to (ii) an internal primer pair (3291-3661). The following samples were loaded in the top and bottom wells in the same order: (1) Ladder, (2) DNA Extraction Negative Control, (3) Species A, (4) Species B, (5) Species C, (6) Species D, (7) Species E, (8) PCR Negative Control (dH₂O), (9) Positive Control (recently collected representative *Drosophila sp.*), (10) no sample. Species A-E were frozen at -80°C for approximately 10 years. For Species B, no PCR amplification occurred when using the external primer pair (total fragment length = 980 bp), while it did occur when using the internal primer pair (total fragment length = 370 bp). The older DNA appears to be degraded and less reliable when using primer pairs with a large fragment length. For Species D and E, subsequent PCR assays, showed a greater amount of sample DNA was needed to achieve amplification.
DNA isolation

DNA samples from both frozen and freshly reared individuals were extracted using the QIAamp DNA Micro Kit (QIAGEN, Valencia, CA) (Protocol 1). To avoid contamination, UV equipment and aseptic technique were used. If contamination occurs, it may also be useful to check the stocks of materials (chemicals in the kit) for contamination as well. When working with stocks, it is best to work with aliquots of the chemicals to minimize contact with the stock material.

Polymerase chain reaction (PCR)

The polymerase chain reaction (PCR) makes many copies of a portion of the DNA. Besides performing PCR amplification, I also learned how primers are designed to amplify areas where the sequence is unknown. A primer consists of about twenty base pairs of sequence. To develop a primer, sequence from species more distantly related than the species in question are aligned and compared to find the most conserved areas of the gene. Once you have the sequence of these conserved areas, the primer can be made. It should be noted that you will need to amplify both strands of the DNA double helix. Also, in other studies, such as DaLage, et al. (2007), the investigators sometimes choose to have internal and external primers.

The PCR has many variables that can affect the quality and efficiency of amplification. For example, if a primer sequence does not exactly match (complement) the actual gene sequence, then it is important to optimize the annealing temperature- low enough so the primers can copy (amplify) a region of DNA, but high enough so that only one PCR band is created (indicating primer specificity). Once primers are developed and PCR protocols are found to be optimal, extracted DNA will be amplified by PCR using Taq 2X Master Mix (New England BioLabs, Ipswich, MA). Directions for PCR was created to streamline procedures (Protocol 2).
Agarose gel electrophoresis

To check that PCR product was produced, or to isolate the PCR product, it is necessary to run a gel electrophoresis. Both 1% TBE (Tris-borate-EDTA) and 1% TAE (Tris-acetate-EDTA) gels were run. The PCR product is put into a well of a gel matrix, which sits in a buffer.

Gels were stained with EtBr in order to visualize the PCR’s amplified DNA. Images are taken of the gel as a record of the PCR amplification experiment. The brightness of the bands is some indicator as to the amounts of PCR product produced (the brighter, the more concentrated the PCR product). Also, it is here where one can be alerted to contamination (Protocol 3).

DNA sequencing

PCR products were to be purified using the QIAquick Gel Extraction Kit (QIAGEN, Valencia, CA). The purified PCR product was to be used for the cycle sequencing reaction according to the Beckman Coulter CEQ 8000 GenomeLab™ DNA Sequencing Kit (Brea, CA) protocol. The purified PCR product could also have been sent out to other labs for sequencing. Once sequencing was completed for both directions of the DNA fragment, a consensus sequence would be made, and at this point the data would match the data gathered from GenBank.

Addendum: Advice to the student researcher

Besides the biological knowledge that was gained from my mentor and working on this project, there were many aspects of what I learned that could be applied to any thesis project. The first and most important advice is for the investigator to maintain organized records. Scientific research will call for large amounts of similar files, all being re-worked countless times. You should begin organizing files from Day 1 in a fashion that is clear and logical. Also, if multiple steps were necessary in a file’s creation, make another file outlining the steps. It is suggested to keep a text document in every folder explaining the contents of the folder and the
source of the files, for example, GenBank file, previously analyzed file, and soon to be analyzed file. Another useful tool when organizing folders and files is to name each with a date at the end of the file. This will preserve information sometimes lost in file transfers, and will help you maintain the progress of your work.

Take detailed, clear notes as if you were in class. Each evening, go back to these notes (which can be written as brief bullet points) and organize them under the correct headings and topics. It is humanly impossible to retain and remember the details of every procedure or issue you encountered throughout the day. Going back to this journal will also help solidify what you learned for the day and where you are at in your project. Check your revised notes with your mentor to ensure you understood everything correctly, and that you understood how the information should be organized. This running journal should include what you did for the day, why, and any issues that arose. Clear step-by-step directions, checklists, and reviews of what you will be doing are very important before performing any procedures. The importance of organization is arguably the most valuable piece of knowledge I gained from this project.

Last, I want to comment on what really helped me to understand the project - asking questions. If told to perform certain methods, to truly learn, one must ask, “Why have we chosen this method?” The difference between what is convention and what is the opinion of the mentor is not always stated directly. It was when I was attentive to this that I truly began to understand the decisions I was making. I began to ask questions and learned the alternatives to the methods we had chosen, and why we chose to go about the project in the fashion we did. It was only then that I really felt that I had a part in making the decisions that shaped this project. Always listen, and ask when something does not make sense. A true, quality mentor will take the time to explain their decisions, and in doing so, there are enormous amounts of knowledge to be gained.
APPENDIX B: PROTOCOLS

PROTOCOL 1: DNA isolation template and directions

Date: ___________________

Remember to Wear Gloves and have clean kimwipe on bench area!

Early Preparation:

Task Check
Set water bath to 70 degrees Celsius
Ice Bucket for dead flies and Proteinase K (liquid at 4°C)
Set heating block on rocker to 56 degrees Celsius.

Samples Used Freshly Dead/Frozen Stock
1. Negative Control 10 λ sdwater run as last sample
2. ___________________ ___________________ ___________________
3. ___________________ ___________________ ___________________
... 
10. ___________________ ___________________ ___________________

What you need:

- three micropipettors: p45, p200, and p1000
- two types of aerosol barrier tips: (1) for p45 and p200 (1-200 microliter (λ) tips) and (2) for p1000 (100-1000 λ tips)
- plus complete the chart:

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Number of Samples</th>
<th>Total Amt Needed= (Quantity) X (# samples)</th>
<th>Needs UV'ing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grinder*</td>
<td>1</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>1.5 ml Centrifuge Tubes**</td>
<td>2</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Column (attached to collecting tube [CT#1])</td>
<td>1</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>2 ml collecting tube [CT2, CT3, CT4]</td>
<td>3</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

All items, except the aerosol barrier tips, need to be decontaminated using UV radiation. UV Stratalinker (note: UV light source above) 10 minutes (600 sec) or three optimal crosslinks should be sufficient. Remember to first label the tubes before crosslinking. The sterile filter with collection tube should not be separated.

*Directions to make a homemade fly grinder (i.e., a pestle to grind up fly in a microcentrifuge tube): (a) Wear gloves. Hold a 0.6 ml microcentrifuge tube with new small kimwipe. (b) Use a new razor blade to cut off the lid and rim of a 0.6 ml microcentrifuge tube. (c) Insert a p1000 blue tip into the 0.6 ml microcentrifuge tube. (d) place this grinder with the microcentrifuge up and the blue tip opening in the tube rack. This position will be optimal for UV cleaning prior to use.

** (1 holds dead fly + 1 for final DNA collection. if dead fly already in vial then change quantity to 1).

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Volume per Sample</th>
<th>number of samples</th>
<th>Actual Use = (val) X (#samples)</th>
<th>What to aliquot *** (add 10 extra for pipetting error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buffer ATL</td>
<td>50 λ X 2 = 100 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteinase K</td>
<td>10 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buffer AL</td>
<td>100 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethanol (100%)</td>
<td>50 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AW1</td>
<td>500 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AW2</td>
<td>500 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water (to suspend DNA instead of AE buffer)</td>
<td>20 λ X 2 = 40 λ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*** You will need to UV tubes to hold aliquots or if prefer check that you have enough volume of the chemicals/solutions in your personal stock (NOT lab stock) for that days extraction.

NOTES during DNA isolation:
More Taxa or More Characters


NOTE: All black font is taken directly from the handbook. All blue font are directions modified for how the Schawaroch lab does single fly preps.

Important point before starting
- Perform all centrifugation steps at room temperature (15–25°C).

Things to do before starting
- Equilibrate samples to room temperature (15–25°C).
- Equilibrate Buffer AE or distilled water for elution to room temperature.
- Set a thermomixer or heated orbital incubator to 56°C for use in step 5.
- Ensure that Buffers AW1 and AW2 have been prepared according to the instructions on page 14.
- If Buffer AL or Buffer ATL contains precipitates, dissolve by heating to 70°C with gentle agitation.
- Optional: Handbook offers to use a carrier RAN for small samples – I do NOT use carrier RNA so all such steps have been omitted.

Procedure
1. Take a frozen single fly in 1.5 ml microcentrifuge tube off the ice and place it on a tube rack. For negative control use 10 µl of dd water.

NOTE: all remaining procedures at room temperature unless otherwise noted.
2. Add 50 µl Buffer ATL to fly in tube. Use homemade pestal to grind fly along the side. The eye pigment go into solution.
3. Add 50 µl more Buffer ATL to ground up fly in tube. Therefore total volume ATL is 100 µl. Make sure all fly parts are in tube and not on pestal
4. Add 10 µl proteinase K.
5. Add 100 µl Buffer AL, close the lid, and mix by pulse-vortexing for 15 s. To ensure efficient lysis, it is essential that the sample, Buffer ATL, proteinase K, and Buffer AL are thoroughly mixed to yield a homogeneous solution.
6. Incubate at 56°C for 10 min. Use the heating block on the rocker. Note: If samples are shaken during the incubation, DNA yields can be increased.
7. Briefly centrifuge the 1.5 ml tube to remove drops from inside the lid.
8. Add 50 µl ethanol (96–100%), close the lid, and mix thoroughly by pulse-vortexing for 15 s. Incubate for 3 min at room temperature. Note: If room temperature exceeds 25°C, cool the ethanol on ice before adding to the tube.
9. Briefly centrifuge the 1.5 ml tube to remove drops from inside the lid.
10. Carefully transfer the entire lysate (Leave fly body parts [do not transfer] in minimal solution in the 1.5 ml centrifuge tube) from step 9 to the QIAamp MinElute column (in a 2 ml collection tube [CT#1]) without wetting the rim. For the Negative Control transfer 200 µl of the solution. Close the lid, and centrifuge at 6000 x g (8000 rpm) for 1 min. Place the QIAamp MinElute column in a clean 2 ml collection tube [CT#2], and discard the collection tube containing the flow-through.
   If the lysate has not completely passed through the membrane after centrifugation, centrifuge again at a higher speed until the QIAamp MinElute column is empty.
11. Carefully open the QIAamp MinElute column and add 500 µl Buffer AW1 without wetting the rim. Close the lid, and centrifuge at 6000 x g (8000 rpm) for 1 min. Place the QIAamp MinElute column in a clean 2 ml collection tube [CT#3], and discard the collection tube containing the flow-through.
12. Carefully open the QIAamp MinElute column and add 500 µl Buffer AW2 without wetting the rim. Close the lid, and centrifuge at 6000 x g (8000 rpm) for 1 min. Place the QIAamp MinElute column in a clean 2 ml collection tube [CT#4], and discard the collection tube containing the flow-through. Contact between the QIAamp MinElute column and the flow-through should be avoided. Some centrifuge rotors may vibrate upon deceleration, resulting in the flow-through, which contains ethanol, coming into contact with the QIAamp MinElute column. Take care when removing the QIAamp MinElute column and collection tube from the rotor, so that flow-through does not come into contact with the QIAamp MinElute column.

13. Centrifuge at full speed (20,000 x g; 14,000 rpm. Note: our centrifuge max speed is 13,200 rpm – works fine) for 3 min to dry the membrane completely. This step is necessary, since ethanol carryover into the eluate may interfere with some downstream applications.

14. Place the QIAamp MinElute column in a clean 1.5 ml microcentrifuge tube (not provided) and discard the collection tube containing the flow-through. Carefully open the lid of the QIAamp MinElute column and apply 20 µl distilled water to the center of the membrane. Important: Ensure that distilled water is equilibrated to room temperature (15–25°C). Remember to add distilled water onto the center of the membrane to ensure complete elution of bound DNA.

QIAamp MinElute columns provide flexibility in the choice of elution volume. Choose a volume according to the requirements of the downstream application. Remember that the volume of eluate will be up to 5 µl less than the volume of elution solution applied to the column.

15. Close the lid and incubate at room temperature (15–25°C) for 5 min. Centrifuge at full speed (20,000 x g; 14,000 rpm. Note: our centrifuge max speed is 13,200 rpm – works fine) for 1 min. Incubating the QIAamp MinElute column loaded with water for 5 min at room temperature before centrifugation generally increases DNA yield.

16. Carefully open the lid of the QIAamp MinElute column and apply 20 µl distilled water to the center of the membrane.

17. Close the lid and incubate at room temperature (15–25°C) for 5 min. Centrifuge at full speed (20,000 x g; 14,000 rpm. Note: our centrifuge max speed is 13,200 rpm – works fine) for 1 min. Note: 16 and 17 are almost repeats of steps 14 and 15 this is to increase DNA yield (better to repeat steps with less water than to add more water and do step only once.

Keep ALL tubes covered in parafilm until you are sure the DNA isolation worked. Dispose of tubes with solution according to lab safety.

Check if procedure worked via PCR.

Safety Information from QIAamp DNA Micro Handbook second edition May 2010 pages 6

When working with chemicals, always wear a suitable lab coat, disposable gloves, and protective goggles. For more information, please consult the appropriate material safety data sheets (MSDSs). These are available online in convenient and compact PDF format at www.qiagen.com/ts/msds.asp where you can find, view, and print the MSDS for each QIAGEN kit and kit component. CAUTION: DO NOT add bleach or acidic solutions directly to the sample-preparation waste. Buffer PB contains guanidine hydrochloride, which can form highly reactive compounds when combined with bleach. In case liquid containing this buffer is split, clean with suitable laboratory detergent and water. If the spill liquid contains potentially infectious agents, clean the affected area first with laboratory detergent and water, and then with 1% (v/v) sodium hypochlorite. The following risk and safety phrases apply to the components of the QIAquick system. Buffer PB: Contains guanidine hydrochloride and isopropanol: harmful, irritant, flammable. Risk and safety phrases*: R10-22-36/38. S23-26-37/39-46. Buffer PN: Contains sodium perchlorate and isopropanol: harmful, highly flammable. Risk and safety phrases*: R11-22. S13-16-23-26-36-46. Buffer GG: Contains guanidine thiocyanate: harmful. Risk and safety phrases*: R20/21/22-32. S13-26-36-46, 24-hour emergency information Emergency medical information in English, French, and German can be obtained 24 hours a day from: Poison Information Center Mainz, Germany, Tel: +49-6131-19240*. R10: Flammable. R11: Highly Flammable. R22: Harmful if swallowed. R20/21/22: Harmful by inhalation, in contact with skin and if swallowed. R32: Contact with acids liberates very toxic gas. R36/38: Irritating to eyes and skin. S13: Keep away from food, drink and animal feeding stuffs. S16: Explosive when mixed with oxidizing substances. S23: Do not breathe vapour/spray. S26: In case of contact with eyes, rinse immediately with plenty of water and seek medical advice. S36: Wear suitable protective clothing. S37/39: Wear suitable protective clothing, gloves and eye/face protection. S46: If swallowed, seek medical advice immediately and show the container or label.
**PROTOCOL 2: PCR directions and recipe**

We are using Taq 2X Master Mix from New England BioLabs. My PCR products are small (900 bp or less) so that this is the optimal Taq. Primers arrive lyophilized (freeze dried) and need to be hydrated to make a 100 microMolar mM stock. Keep half of this stock in lab Freezer and half in minus 80 storage. The working stock of the primer is 10mM. Therefore dilute primer stock by 1/10 to make the working stock tube kept in your experiment box in freezer. The 25 λ Total Volume recipe is when you have everything optimized and once checked by agarose gel electrophoresis this product will be used for sequencing.

### 25 microLiter (λ) Total Volume PCR reaction:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (10 mM)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Primer B (10 mM)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Taq 2X Master Mix</td>
<td>12.5 λ</td>
</tr>
</tbody>
</table>
| DNA                | 1 λ    *
| sdWater            | 9.5 λ  *
| **Total Volume**   | 25 λ   |

*Note:* DNA volume can be increased or decreased and then you will need to compensate by adjusting the sdWater (actually nanopure water or purified water from Fisher etc.)

Therefore here are our two most popular recipes:

### For 1 λ sample DNA

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Primer B (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Taq 2X Master Mix</td>
<td>12.5 λ</td>
</tr>
<tr>
<td>DNA</td>
<td>1 λ</td>
</tr>
<tr>
<td>sdWater</td>
<td>9.5 λ</td>
</tr>
<tr>
<td><strong>Total Volume</strong></td>
<td>25 λ</td>
</tr>
</tbody>
</table>

### For 2 λ sample DNA

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Primer B (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Taq 2X Master Mix</td>
<td>12.5 λ</td>
</tr>
<tr>
<td>DNA</td>
<td>2 λ</td>
</tr>
<tr>
<td>sdWater</td>
<td>8.5 λ</td>
</tr>
<tr>
<td><strong>Total Volume</strong></td>
<td>25 λ</td>
</tr>
</tbody>
</table>

Cycling conditions as per New England Biolabs protocol as optimized for the Taq 2X

- 95 degrees for 2 minutes
- Then 35 repeats of:
  - 95 degrees for 30 seconds
  - **50 degrees annealing for 30 seconds**
  - 68 degrees for 1 minute
- After cycling then 68 degrees for 5 minutes
- To 4 degrees (ice bucket) forever

**Note:** annealing temperature varies with primer pair and species

The 10λ Total Volume PCR reaction recipe is for initial screening (1) to see if the DNA isolation worked and (2) to optimize the PCR reaction (usually in this case annealing temperature since Taq 2X has its own cycling protocol).

For the 10 total volume reaction you are actually decreasing the above PCR reaction by 2.5 fold or in other words 40% of the original reaction.

First, make a diluted working stock of the primer to specifically use for the 10 Total Volume reactions

### To dilute each primer:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (10 mM)</td>
<td>12 λ</td>
</tr>
<tr>
<td>sdWater</td>
<td>18 λ</td>
</tr>
<tr>
<td><strong>Total Volume</strong></td>
<td>30 λ</td>
</tr>
</tbody>
</table>

### 10 microLiter (λ) Total Volume PCR reaction:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Primer B (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Taq 2X Master Mix</td>
<td>5 λ</td>
</tr>
</tbody>
</table>
| DNA                | 1 λ    *
| sdWater            | 2 λ    *
| **Total Volume**   | 10 λ   |
Note: DNA volume and can be increased or decreased and then you will need to compensate by adjusting the sdWater (actually nanopure water or purified water from Fisher etc.)

Therefore here are our two most popular recipes:

For 1 λ sample DNA

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (λ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Primer B (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Taq 2X Master Mix</td>
<td>5 λ</td>
</tr>
<tr>
<td>DNA</td>
<td>1 λ</td>
</tr>
<tr>
<td>sdWater</td>
<td>2 λ</td>
</tr>
<tr>
<td>Total Volume</td>
<td>10 λ</td>
</tr>
</tbody>
</table>

For 2 λ sample DNA

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (λ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Primer B (diluted)</td>
<td>1 λ</td>
</tr>
<tr>
<td>Taq 2X Master Mix</td>
<td>5 λ</td>
</tr>
<tr>
<td>DNA</td>
<td>2 λ</td>
</tr>
<tr>
<td>sdWater</td>
<td>1 λ</td>
</tr>
<tr>
<td>Total Volume</td>
<td>10 λ</td>
</tr>
</tbody>
</table>

Cycling conditions as per New England Biolabs protocol as optimized for the Taq 2X

- 95 degrees for 2 minutes
- Then 35 repeats of:
  - 95 degrees for 30 seconds
  - **50 degrees annealing for 30 seconds**
  - 68 degrees for 1 minute
- After cycling then 68 degrees for 5 minutes
- To 4 degrees (ice bucket) forever

** (Note: annealing temperature varies with primer pair and species)
# PROTOCOL 3: Gel template and directions

**Date:** _______________

## A) Top Row

<table>
<thead>
<tr>
<th>Well, Samples Used</th>
<th>Primer Pair</th>
<th>Band</th>
<th>Date of PCR</th>
<th>Date of Extra.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. _________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
<tr>
<td>2. _________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
<tr>
<td>3. _________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. ________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
</tbody>
</table>

## B) Bottom Row

<table>
<thead>
<tr>
<th>Well, Samples Used</th>
<th>Primer Pair</th>
<th>Band</th>
<th>Date of PCR</th>
<th>Date of Extra.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. _________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
<tr>
<td>2. _________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
<tr>
<td>3. _________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. ________________</td>
<td>_____________</td>
<td>_______</td>
<td>____________</td>
<td>____________</td>
</tr>
</tbody>
</table>

### Details

- **Gel Type:** __%____
- **Gel TV:** ______ ml
- **Amt of Gel Buffer:** ______ ml
- **Amt of Agarose:** ______ g
- **Voltage Setting:** ______ V
- **Total Running Time:** ______ min

### After Gel is Run

- **Stain Used:** ____________
- **Stained for:** ______ min
- **De-stained for:** ______ min

### Picture File Name(s)

- ___________________

### Notes

__________________________________________________________________________
__________________________________________________________________________

---

**I. Preparation:**

- UV p-10
- UV yellow tips
- Get Ice
- Prepare Gel box for making a gel (special setup if small total volume)

**Making Gel**

1) Depending on what total volume you would like the gel to be, prepare your materials. We usually use a 1% TAE Gel.

   - Eg: A 1% TAE Gel with a desired total volume of 35 ml will require 35 ml of TAE buffer, and 0.35 g of Agarose (use from aliquot given to you by Dr. Schawaroeh).

2) In an Erlenmeyer Flask (125 ml Erlenmeyer for TV 35 ml, 50 ml Erlenmeyer for anything less than TV 35 ml), first pour the measured agarose.
3) Next pour the measured TAE Buffer, making sure all of the agarose on the sides of the Erlenmeyer are flushed to the bottom of the Erlenmeyer.
4) Slightly swirl the solution. Measure the Erlenmeyer with its contents on the analytical balance. Tare the measurement.
5) Put a paper towel stopper into the top of the Erlenmeyer.
6) After putting on safety glasses and oven glove, put the Erlenmeyer and contents into the microwave.
7) Set the microwave to:
   - 1 min for 35 ml TV gels
   - 15 s for any TV below 35 ml
   NOTE: You do not want to put in the solution for too long (for smaller total volumes, because the solution will boil out of the Erlenmeyer)
8) Let it stand in closed microwave for at least 20 sec.
9) Take out the Erlenmeyer and contents (while wearing an oven mit), and swirl.
10) Keep putting in microwave for 15-30 second intervals (less time for smaller total volumes) (do not forget to let it stand for at least 20 seconds after microwaving) until the solution is absolutely clear.
11) Take off the paper towel stopper, and measure on the analytical balance.
12) Add back the ddH2O until the balance reads 1 g.
13) Put the paper towel stopper on the Erlenmeyer again.
14) Put in the microwave and set it to 15 s (or 10 s with a smaller total volume).
15) Let it stand, then take it out and swirl. Put on the countertop surface and let it cool until the Erlenmeyer is of bearable temperature on your skin.
16) Once the Erlenmeyer and contents have cooled to a bearable temperature, pour into the prepared gel box.
   NOTE: You should pour the gel into what will be the bottom right corner of the gel to decrease errors if bubbles pour out.
17) Immediately put the comb into the gel.
18) Allow it to cool for at least 15 minutes (gels with higher Total Volumes will need more time to cool).
19) Once gel is cool and solid, carefully remove the comb by lifting straight up. Also, carefully remove any other molding apparatus.
20) Lift the gel case out of the gel box and orient it so that the wells are on the side of the black electrode.

Running gel
1) Make sure the gel is in the gel case in the gel box, and that the wells are on the side of the black electrode.
   NOTE: Remember, DNA always Runs to Red. This is because DNA is negatively charged and it is attracted to the red cathode, and repelled by the black electrode.
2) Pour TAE running buffer until the whole gel is fully submersed in buffer (there should be approximately 1 cm of buffer above the gel).
3) After putting the Loading Dye, ladder, and PCR Products on ice, cut a piece of parafilm.
4) Using the UV’d p-10 and yellow tips, pipet 3 λ of Loading Dye onto the parafilm for each well that will be in use (# of PCR products + Ladder).
5) Pipet 1 λ of Ladder and mix with the Loading Dye, then load the mix into the first well.
6) Pipet 5 λ of the PCR product and mix with the loading dye, then load the mix into the next well. Repeat this process for all PCR products.
7) After all samples have been loaded in the wells, place the Gel Box cover on the box, matching the colors of electrodes.
8) Plug the wires from the cover into the voltage source.
9) Turn on the voltage source. Check the gel box for bubbles to make sure the source is on. Also check the amount of voltage, and make sure it is your desired amount.
10) Let the gel run for a specified amount of time (25 minutes for 35 ml TV gels, less time for smaller TV).
11) Check the progress of the gel during its run time.
12) Stop the gel electrophoresis process by turning off the voltage source.
   NOTE: If you will not be able to work with the gel after stopping it from running, set the voltage to the lowest setting. It is always better to have electricity running through gel.
**Staining/De-Staining Gel**

CAUTION: Gloves should be worn throughout this process and contact with objects should be minimal. The EtBr is potentially hazardous. Its contact with surfaces and equipment should be minimized. Also, when using the UV light, face shields should be worn.

1) Put the Gel in the desired stain’s Staining Tray (we usually use EtBr).
2) Carefully pour out the stain into the tray (enough to cover entire gel).
3) Put the tray on a rocker. Let it stain and rock for approximately 2 min.
4) Check the stain under the UV light. If the gel needs more staining time, stain it longer.
5) Once the gel is properly stained, put the gel into the De-Staining Tray.
6) Pour ddH₂O over the gel (enough to cover gel).
7) Put on the rocker and de-stain for approximately 30 min (time is flexible).
8) While de-staining, pour back the stain from the Staining Tray to its proper container.
9) After the proper amount of De-Staining, take the gel and check under the UV light.
10) Properly dispose of the remaining liquid in the de-staining tray into a waste container.

**Taking Pictures of the Gel**

Use the Kodak Logic 100 Imaging System to image the gel under UV light.
More Taxa or More Characters

PROTOCOL 4: Performing alignments with MegAlign
(by DNA Star and part of Lasergene Core Suite)

I. Loading the sequences into MegAlign
All sequences loaded into MegAlign
A) must either already be FastA files (OPTION 1)
   - To do this, the sequences must first be converted using Edit Seq, and then opened with MegAlign
B) OR you can send the Edit Seq File directly to MegAlign (OPTION 2)

A) OPTION 1: Converting text files to FastA files
1) Open EditSeq
2) Go to File → Import
3) On my external hard drive go to:
   Stefan main → RESEARCH → New Project 2010 → HONORS THESIS AUG2011 →
   Amyrel Project_ALL_made15Apr2011 → AmyrelAA →
   AmyrelAAExon# (# = whichever exon you are working with) →
   "the text file you want to work with"
   NOTE: You may want to know exactly how these edited (intron removed) amino acid sequences were generated. This will be in a separate set of directions, but is briefly explained in the 23Oct2010 MACUB poster.
4) Before hitting “Open”, be sure to select Protein.
5) Hit “Open”
6) Once open, go to File → Export
7) Choose a folder to export the FastA file to. For this exercise (exercise of writing the directions), I saved the FastA files on my external hard drive to:
   Stefan main → RESEARCH → New Project 2010 → Testing How to Use MegAlign Again 10Aug11
8) Make sure the file type is a FastA file (this should be the default).
9) Hit Save.
Next: Loading sequences into MegAlign
1) Open MegAlign
2) Go to File → Enter Sequences (CTL + E)
3) Once the dialog box opens, go to the folder with the FastA files you created. For this exercise, it is on my external hard drive on:
   Stefan main → RESEARCH → New Project 2010 → Testing How to Use MegAlign Again 10Aug11
4) Click each one you want to open, and hit open once clicked. It should appear in the selected sequences box. Continue to do this for each sequence you would like to include in the alignment.
5) Double Check the Selected Sequences box and add or remove any sequences you wish.
6) When the appropriate list is complete, hit “Done”.
7) MegAlign will display all of the sequences on what it calls a worktable.
   NOTE: We are not completely sure if the input order affects the alignment, but since the program uses a CLUSTAL framework for the alignments, and in CLUSTALX the input order affected the alignment, then it can be inferred that input order affects the alignment in MegAlign. So, if you need to move sequences around, you need at least 3 sequences loaded, and then you can simply drag the sequences to the desired location.
   HOWEVER, it is easier to input the sequences in the order you intend to align them.
B) OPTION 2:  
*Sending the Edit Seq File directly to MegAlign*

1) Follow Steps 1-5 of Option 1.
2) Go to File \(\rightarrow\) Send Sequence to \(\rightarrow\) MegAlign
3) MegAlign will open and the sequence will appear in the MegAlign workspace
4) Load all sequences by following Steps 1-3 of Option 2 and you may change the order in MegAlign later.

**NOTE:** We are not completely sure if the input order affects the alignment, but since the program uses a CLUSTAL framework for the alignments, and in CLUSTALX the input order affected the alignment, then it can be inferred that input order affects the alignment in MegAlign. So, if you need to move sequences around, you need at least 3 sequences loaded, and then you can simply drag the sequences to the desired location.

**HOWEVER,** it is easier to input the sequences in the order you intend to align them.

II. Performing alignments with MegAlign

1) Either load a MegAlign project or input sequences into MegAlign.
2) Go to Align \(\rightarrow\) Method Parameters
3) We are working with Clustal W, so make sure to be in the CLustal W tab once the dialogue box opens. Also, we are working with Multiple Alignment Parameters, so we will change parameters on this side of the dialogue box.
4) The Gap Length Penalty should be held constant at 10 (Schawaroch, 2002). The Gap Penalty was varied (1, 5, 10, 15, 20, 30, 40, 50).
5) Save each alignment in all possible formats.
PROTOCOL 5: Directions for performing tree analysis with PAUP*

Once on Mac, make sure to make folders for each analysis you plan to run for the day. Be organized, but concise with the naming. The folder name will be carried throughout all file naming related to each analysis. For example, when analyzing the data for the Amyrel gene for Exons 1 and 2, for 74 species, I named the folder “AmEx1_2_74”.

1) Go to the Apple ➔ Recent Applications ➔ PAUP and open PAUP
2) Once PAUP is open, go to File ➔ Setup log
3) It will bring you to a window where you choose the folder to save to and the name of the log. Make sure you are saving the log file in the folder for the correct analysis. The file should be name “LogNAMEOFFOLDER”.
4) The command to save a log should be shown on the actual log.
5) Open up the Nexus file you plan to execute, either through PAUP, or simply by clicking on it from the specified folder.
6) Once PAUP opens this Nexus file, go to File ➔ Execute
7) Under Data, go to Include/Exclude characters
   - got to charset, exclude whatever you plan to exclude (ie. Exon 2)
   - get rid of parsimony uninformative characters (otherwise it will inflate some numbers), so exclude it
8) Next, define outgroup, (ie. affinis, bifasciata, imaii, psuedoobscura, subobscura, kitumnesis)
   select outgroup, then OK
9) Now data set is set up, go to Analysis, and go to Parsimony settings
   - keep all defaults EXCEPT
     - in the STEPWISE ADDITION options to RANDOM, 10 reps, type random seed
10) Hit “Search”, it does search and have to give it time, it may want more max trees (automatically increase by 100, hit OK)
11) Once done with search, it will give some info about the search
12) Under trees, save tree to file
   - name it the same-the log
13) Since this is more than one tree, we do not want to look at all trees, we want to look at tree scores in parsimony, don’t do it for all trees (just the 1st, since they are all same)
   - pick all top 4 options, then hit OK
14) You want to compute consensus, choose strict, you have to look at all trees, hit OK
15) Go to trees, Print consensus trees, choose rectangular, Helvetica 9, and then look at it as preview
16) Save it as a pic file from the print preview
   - name it STRICT with same name
17) Convert this pic file to a tiff using Photoshop
More Taxa or More Characters

80

PROTOCOL 6: Directions for bootstrap analysis

1) First, make a folder within the NAMEOFANALYSIS folder called BootNAME
   - make a copy of original data set and put it in here
2) Open PAUP from Recent Applications
3) Set up a log in the BootNAME folder and call it LogBootNAME
4) Execute original data file (with all data in it)
   - setup same conditions (include/exclude char, insert/delete taxa, outgroup)
5) Since we keep all parsimony settings on default, you do not need to go to parsimony settings (but if we were
   going to change something like how we treat gaps from missing data to a 5th character, we would have to change
   this). If we are not changing any parsimony settings:
   - go to Analysis -> Bootstrap/Jackknife
6) The Bootstrap analysis window comes up and we keep all of the defaults (100 reps) except for the seed number,
   which we input randomly
7) We hit continue and it will bring us to the heuristic search options. We keep all defaults except in Stepwise
   Additions options, where we change it to Random, and do 1 rep. We also create a random seed here.
8) After going through all defaults, hit search. So we are now doing 100 reps for each bootstrap but 1 rep per
   heuristic search.
9) When search is complete, go to Trees -> Save Trees to file and name the tree file as BootTreeNAME
10) Next, go to Trees -> Print Trees
11) In the printing options, set printing to a Rectangular Cladogram, and the font to 9. Hit preview
12) In the preview menu, click save trees to file, and save the pict of the tree as BootTreeNAMEpict
13) You can close the program, and now you must save the BootTreeNAMEpict as a tiff in Photoshop
   - open photoshop first from recent applications
   - open the pict with photoshop
   - go to save as
   - save it as a tiff
PROTOCOL 7: Directions for Bremer analysis

These directions are applicable to performing an Bremer analysis on the PAUP analysis of various data sets. With these directions, it is assumed that you have already performed an analysis and tree construction on some data set (NAME), and that you have each analysis in a NAMEOFANALYSIS folder, which should contain (but not limited to these) an original data file (Nexus with all data) and a Tree file (which should be named as TreeNAME).

Preliminary
1) You should first make a folder within the NAMEOFANALYSIS folder called BremNAME
2) Next put in a copy of the tree file from the NAME folder into the BremNAME folder.
3) You must open the Tree file, resave it as:
   Tree1_(#of equallyparsimoneoustrees)NAME
   - then delete all but the first tree from the file (make sure you leave semicolon, end, semicolon as the end command)
   - save it again
4) Create an AutoDecayRun folder within the BremNAME folder. Put in a copy of a) the original data file NAME folder and b) of the tree file (from Step 3) from the BremNAME folder into the AutoDecayRun folder.
You now have what you need to do Bremer analysis.

Actual Directions
1) In PAUP
   - make a Log file for the run you will do and make sure it is in the AutoDecayRun folder
   - call it LogAutoDecayNAME
2) Keep PAUP open and open Autodecay from in recent applications
   - it should not matter at this point, but you may want to set up length of optimal tree here (the default originally had nothing, but now it displays the number entered from the last run)
   - once open, go to the Autodecay menu → setup command file
   - now get Nexus tree file you want to use (the one you created with only one tree)
   - open tree file
   - a new window comes up to name the decay file (it names it for you). Save it within AutoDecayRun folder
   - it will make constraint trees (trees with single nodes)
   - you can leave the program running
3) Now go run it in PAUP
   - execute original data file (with all data in it)
   - setup same conditions (include/exclude char, insert/delete taxa, outgroup)
   - open the adc file that Autodecay created. Open it by executing it
   - when it starts to execute, it will go through a bunch of searches
   - it is going through different nodes and making replicates (100 for each search), does it 100 per node.
   - when it is done searching it will show some table in PAUP - save the trees from PAUP to the Autodecay folder as PAUPautodecayNAME
4) Go back to AutoDecay to extract the decay values
   - set the most optimal tree length
   - make sure it says output tree and list
   - got to AutoDecay→extract decay values
   - open the adc file
   - it will perform a number of analyses equivalent to how many nodes were there
5) Now go to Tree View from recent applications
   - once open go to file open, and open the tree file that AutoDecay made (it will have .tre file extension)
6) In treeview
   - Tree→ show internal edge layers, set font of decay values (I did 6 size)
   - go to trees→ print trees
   - make sure sure to check show internal labels
   - save as a pict, call it DecayTreeNAME
   - save it
7) Make that pict a tiff in Photoshop