# TESTING THE UTILITY OF DNA BARCODING ON DIPTERA OF ETHEKWINI

by

**SANELISIWE THINASONKE DUZE**

BSc. (*Hons*) Genetics

Submitted in fulfillment of the academic requirements for the degree of Master of Science in

The School of Life Sciences

University of KwaZulu-Natal

Pietermaritzburg campus

November 2016

As the candidate's supervisor I have approved this dissertation for submission.

Signed: Name: Dr S. Willows-Munro Date: 29 September 2016

Signed: CEardley Name: Dr C. Eardley Date: 29 September 2016

Signed: Name: Dr D. Swanevelder Date: 30 September 2016

# ADVISORY COMMITTEE

**Supervisor**

**Dr Sandi Willows-Munro**

School of Life Sciences

Department of Genetics

University of KwaZulu-Natal

**Co-supervisor 1**

**Dr Connal Eardley**

Plant Protection Research Institute

Agricultural Research Council

**Co-supervisor 2**

**Dr Dirk Swanevelder**

Biotechnology Platform

Agricultural Research Council

# PREFACE

The experimental work described in this dissertation was carried out at the University of KwaZulu-Natal, Pietermaritzburg, in the Discipline of Genetics under the School of Life Sciences from March 2013 under the supervision of Dr S. Willows-Munro.

The study represents original work by the author and has not been submitted in any form to another University. Work done by other authors mentioned in this study has been duly acknowledged in the text.

# DECLARATION OF PLAGIARISM

I, Sanelisiwe Thinasonke Duze declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a. Their words have been re-written but the general information attributed to them has been referenced

b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: TSDuze

# ABSTRACT

DNA barcoding is an exciting tool that provides a way of identifying species effectively and rapidly. It involves the use of a short, standardized DNA sequence to assign species name to unknown specimens. For animals, a 658 bp long fragment of the mitochondrial gene, the cytochrome c oxidase subunit I (*COI*) is used as the standard barcode region. The success of DNA barcoding is dependent on the absence of overlap between intraspecific and interspecific variation, i.e. barcoding gap. Although DNA barcoding has been successfully used across a number of insects; its adequacy in identifying Diptera species is still questioned. Also, many of the DNA studies on Diptera don't include any African taxa, thus, it is unknown how successful DNA barcoding will be on our native Southern African taxa. In this study, the efficacy of using the *COI* gene as a barcode for the identification of Diptera species within eThekwini was evaluated by examining the existence of the DNA barcoding gap for the South African Diptera and by testing the identification efficacy of this marker on Dipteran species using three distance-based methods: Near Neighbor (NN), Best Close Match (BCM) and BOLD Identification Criteria (BIC). A total of 844 barcodes from 1060 Diptera specimens collected from 14 localities within the eThekwini and surrounding areas were successfully sequenced. No barcoding gap was observed when intraspecific and interspecific sequence divergences were compared. Furthermore, the identification success of the three distance-based methods was low, ranging between 62% and 68%. The low identification success and the lack of barcoding gap in Diptera suggest that *COI* gene is not a good marker to use for species delimitation in Diptera. The MiSeq sequencer from Illumina was then used in this study to construct a complete mitochondrial genome of one of the Diptera species (*Lucilia cuprina:* Calliphoridae) collected in eThekwini. This complete mitochondrial genome together with 48 complete mitochondrial genomes of other Diptera species obtained from the National Center for Biotechnology Information (NCBI) Genome Database were used to explore other potential mitochondrial genes that can be used as DNA barcodes for the identification of Diptera species. Thirteen mitochondrial protein coding genes from 49 Diptera species were evaluated as potential DNA barcodes that can be used for the identification of Diptera species. The *COI* and the ATPase subunits 6 (*ATP6)* genes are potential barcode markers that can be used in delimitating South African Diptera species.

# ACKNOWLEDGEMENTS

I thank God Almighty for His love and His grace toward me. “For since the creation of the world God’s invisible qualities, His eternal power and divine nature have been clearly seen, being understood from what has been made, so that people are without excuse, Romans 1: 20”.

Writing this dissertation has been one of the most significant academic challenges I’ve ever faced. This study would not have been completed without the support, guidance and encouragement of the following people; it is to them that I owe my deepest gratitude.

I owe my deepest gratitude to my supervisor, Dr. Sandi Willows-Munro. Thank you for not giving up on me and for your wise kind words and encouragement. Thank you for your guidance, motivation and making time for me. I really appreciated the quick feedback, the shouting but mostly encouragement.

Thank you to my co-supervisors Dr Connal Eardely and Dr Dirk Swanevelder

I acknowledge the National Research Foundation (NRF) and the University of KwaZulu-Natal for funding this project, the University of Guelph in Canada for their assistance and the eThekwini KwaZulu-Natal sandstone sourveld for providing support and guidance.

Cornerstone Assemblies of God church for making my stay in Pietermaritzburg fun and most memorable and for helping me grow spiritually and teaching me the ways of the Lord.

The past and present members of the UKZN Conservation Genetics Laboratory: Sohana Singh, Qiniso and Sihle Mthetwa, Andrinajoro Rakotoarivelo, Riel Coetzee and Ashreene Govender for making the lab a fun and a warm environment to be in and for their willingness to help.

Data collection for this project would not have been easy without the help of the SANBI interns. Thank you to Matabara, Patrick, Lorika and Vince for their assistance with field work and sorting of the insects.

Finally, I would like to thank the staff of the Department of Genetics at the University of KwaZulu-Natal – Dr. Pillay, Dr. Meusel, Mr. Dzomba, Mrs. Brunkhorst and Mr. Zondi for their support and assistance.

# TABLE OF CONTENT

# LIST OF ABBREVIATIONS

**AIC** - Akaike Information Criterion

**ATP 6 –** ATPase Subunit 6

**ATP 8 –** ATPase Subunit 8

**BCM** - Best Close Match

**BI** - Bayesian Inference

**BOLD** - Barcode of Life Data System

**BIC** - BOLD Identification Criteria

**BIN** - Barcode Index Number

**iBOL** - International Barcode of Life

**COI** - Cytochrome C Oxidase Subunit I

**COII –** Cytochrome C Oxidase Subunit II

**COIII** - Cytochrome C Oxidase Subunit III

**Cytb** - Cytochrome B Subunit

**CR –** Control Region

**DHU** - Dihydrouridine

**DNA** - Deoxyribonucleic Acid

**D'MOSS** - Durban Metropolitan Open Space System

**ESS** - Effective Sample Size

**ETKD** - eThekwini Diptera

**GTR + I+ G** - Generalized-time Reversible + Invariant Sites + Gamma

**JM –** Jeffries-Matusita

**K2P –** Kimura-Two-Parameter

**matK** - Megakaryocyte-associated Tyrosine Kinase

**MCMC** - Markov Chain Monte Carlo

**Mt –** Mitochondrial

**ML** - Maximum Likelihood

**NAD** - NADH Dehydrogenase Subunits

**NCBI** - National Center for Biotechnology Information

**NGS**- Next Generation Sequencing

**NN** - Nearest Neighbor

**NJ** - Neighbor Joining

**OUTs** - Operational Taxonomic Units

**PCR** - Polymerase Chain Reaction

**PCGs**- Protein Coding Genes

**QC –** Quality Check

**rRNA**- Ribosomal Ribonucleic Acid

**rbcL** - Ribulose-bisphosphate Carboxylase

**SPIDER** - Species Identity and Evolution,

**tRNA** - Transfer Ribonucleic Acid

**UKZN** - University of KwaZulu-Natal

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF APPENDICES

# CHAPTER ONE: LITERATURE REVIEW

# DNA BARCODING: A RAPID ASSESSMENT METHOD OF BIODIVERSITY FOR CONSERVATION

## Abstract

The earth's biodiversity is vanishing at an accelerated rate, this increases the need to protect and catalogue biodiversity. Conservation planning begins with an inclusive evaluation of regional biodiversity. The knowledge of species richness in a particular habitat and understanding the life history, endemism, range of morphological and genetic variability as well as the evolutionary history of the species are all part of biodiversity assessment. Thus, species inventories have become an important factor for conservation planning. The use of key morphological characters has in the past been the main method for species identification. However, this method has a number of limitations and often relies on well-trained taxonomists. DNA barcoding is a method that can be used to facilitate species identification, which can complement the traditional morphology-based approach. DNA barcoding has been successfully used in a number of species and has proven its utility as a tool for a rapid assessment of animal and plant diversity. In animals, a short, standardized mitochondrial cytochrome c oxidase subunit I (*COI*) gene is used as the DNA "barcode" to rapidly and effectively identify species. In plants, barcoding concentrates on the chloroplast's large subunit of the ribulose-bisphosphate carboxylase (*rbcL*) and the megakaryocyte-associated tyrosine kinase (*matK*) genes. This literature review will summarize the protocols involved in DNA barcoding. The review will focus on the *COI* gene as the standard gene for DNA barcoding animals, the methodology of DNA barcoding, the role of next generation sequencing on DNA barcoding as well as the prospects and problems associated with using DNA barcoding in the identification of insects, in particular Diptera.

**Key Words:** Conservation, biodiversity, DNA barcoding, cytochrome c oxidase subunit I

## 1.1 Introduction

### 1.1.1 Biodiversity assessment

An overwhelming proportion of the earth's biodiversity is vanishing (Pimm *et al*. 2014; Ceballos *et al*. 2015) with taxa facing the challenges of habitat destruction, fragmentation and degradation, and also biological invasions (Iii *et al.* 2000; deVere 2008; Butchart *et al.* 2010; Pimm *et al.* 2014; Ceballos *et al.* 2015). All these factors lead to a decrease in biodiversity. This is particularly true for endemic species with small distribution ranges.

Species are the essential building block of natural history (Goerge & Mayden 2001). Therefore, species inventories are an important component of most conservation strategies. Often in conservation related fields, "species" are regarded as the "smallest unit" of biodiversity (Goerge & Mayden 2001; Agapow *et al.* 2004), making it important to correctly identify species. Taxonomic knowledge and identification tools are still limited or absent for many groups, especially in the hyper diverse groups such as arthropoda. Therefore, methods that will accelerate and simplify the process of species identification are required to protect native and threatened species and preserve the natural biodiversity of the ecosystem (Waugh 2007; Renaud *et al.* 2012).

Traditionally, species identification is based on key morphological characters that are taxon specific. These are often presented in the so called taxonomic identification keys, for specific levels. However, this method is usually time consuming and requires well-trained and experienced taxonomists – often with experts focusing only on a specific group of taxa. In addition, there are a number of limitations associated with the use of only morphological characters in taxonomy. First, phenotypic plasticity in some morphological features used for species identification can easily lead to identification errors (Hebert *et al.* 2003a; Waugh 2007). Second, this method often overlooks morphologically cryptic taxa (Jarman & Elliott 2000). Third, the morphological keys used in traditional morphology-based species identification are often effective only for a particular life stage or gender of a species hence many individuals cannot be identified (Hebert *et al.* 2003a; Pili *et al.* 2010). Finally, the taxonomy of many taxa has not been studied sufficiently well to enable accurate species identification. DNA sequencing technology has introduced the possibility of using variations in short sequences of DNA as "labels" or "tags" for species identification, a concept known as DNA barcoding (Hebert *et al.* 2003a).

### 1.1.2 DNA barcoding

DNA barcoding is a molecular method that uses short, standardized genomic fragments to facilitate species identification and discovery. The standard sequence region is called a DNA barcode because, like a "barcode tag" for products in a supermarket, they are unique identifiers for a particular species (Jinbo *et al.* 2011). This method was first proposed by Hebert *et al.* (2003a) where a "universal" primer set was used to amplify a 648 bp region of the mitochondrial cytochrome c oxidase subunit I (*COI)* gene in a group of moths. This small section of the mitochondrial genome has been found to be useful for species level identifications in animals over a broad range of biological specimens (Hebert *et al.* 2003a).

To promote DNA barcoding as a global standard for sequence-based species identification of eukaryotes, the *International Barcode of Life* (iBOL) proposed and initiated the *Barcode of Life* project in 2004. This international barcode of life project seeks to develop a standard protocol for DNA barcoding and to construct a comprehensive DNA barcode reference library of all life on earth (Dasmahapatra & Mallet 2006; Hajibabaei *et al.* 2007; Ratnasingham & Hebert 2007). An online database, the *Barcode of Life Data System* (BOLD; www.barcodinglife.org) is available to acquire, store, analyse and manage DNA barcode records (Ratnasingham & Hebert 2007).

The efficiency of DNA barcoding as well as the need to have a panel of reference species barcodes to which one can compare unidentified specimens has prompted efforts to construct DNA barcode reference libraries for various animal groups (Ekrem *et al.* 2007; Lee *et al.* 2011; Zhou *et al.* 2011; Webb *et al.* 2012). These DNA barcode libraries not only aid the documentation of biodiversity (Janzen *et al.* 2005) including endangered species (Elmeer *et al.* 2012), but can also reveal species endemism (Bossuyt *et al.* 2004; Sourakov & Zakharov 2011).

### 1.1.3 *COI* gene as a standard fragment for DNA barcoding

The ability of DNA barcoding to distinguish between species is largely dependent on the gene region being used as the barcode. This gene region must have a slow enough mutation rate to minimize intraspecific (between individual of the same species) variation but be sufficiently variable to highlight interspecific (between different species) variation (Hebert *et al.* 2003a). Moreover, it must be easy to amplify in a range of taxonomically distinct taxa and should have few insertions and deletions to facilitate sequence alignment (Hebert *et al.* 2003a).

The mitochondrial *COI* gene was identified as a suitable region to be used in DNA barcoding of animals. Present in every eukaryotic organism, it is diverse enough to be able to differentiate between most animal taxa at the species level. Additionally, it can be easily amplified and sequenced using a universal primer set (Hebert *et al.* 2003a). The priming sites of the region are located within highly conserved amino acid sequences, which in turn ensures that primers will be broadly applicable (Moritz & Cicero 2004). The *COI* gene is protein-coding gene and contains few indels which makes it easy to align (Jinbo *et al.* 2011).

### 1.1.4 Methodology of DNA barcoding

The methodology used for DNA barcoding is fairly simple. Genomic DNA is extracted from the organism of interest. The barcode region (*i.e. COI* gene) is then amplified using polymerase chain reaction (PCR) with barcode targeting primers. The amplified barcode is then sequenced, usually using Sanger sequencing. The barcode of the organism of interest is then compared to barcode sequences in a reference library, which were derived from individuals of known species. The unknown organism is identified if its sequence closely matches one in the barcode reference library (Hajibabaei *et al.* 2007; Ratnasingham & Hebert 2007).

The comparison of an unknown barcode with sequences from the reference library is done through the construction of a multiple sequence alignment which is then used to construct a cladogram using the distance-based neighbor-joining (NJ) or similar method (Ratnasingham & Hebert 2007). This method usually uses the Kimura-2-Parameter (K2P) model to calculate genetic distance between taxa. Related individuals are clustered together (Hajibabaei *et al.* 2007; Ratnasingham & Hebert 2007). If a species-level match is not obtained, sequence divergence values can still be used to assign the specimen to a genus or a family (Ratnasingham & Hebert 2007).

The success of DNA barcoding is ultimately dependent on completeness of the DNA reference library and on the absence of overlap between intraspecific and interspecific divergence, a concept known as the “barcoding gap” (Meyer & Paulay 2005; Aliabadian *et al.* 2009). For a DNA barcoding gap to occur, the amount of genetic variation within species should be much smaller than the amount of variation between species and this allows species to be easily distinguished (Meyer & Paulay 2005). The DNA barcoding gap is essential for accurate species discrimination and underlies both specimen identification and species discovery (Chapple & Ritchie 2013). The more overlap there is between intraspecific and interspecific divergences the less effective barcoding becomes (Meyer & Paulay 2005).

An overlap between intraspecific and interspecific divergence could be attributed to recent speciation and interspecific hybridization of taxa (van Velzen *et al.* 2012b). Recently diverged species share similar DNA barcode sequence, which could prevent accurate identification by barcoding using *COI* alone (Nichols 2001; Chapple *et al.* 2012; van Velzen *et al.* 2012a). This situation is not often observed in insects and most studies focused on insect taxa have successfully used DNA barcoding for the molecular identification of a broad variety of insect taxa, including Ephemeroptera (Ball & Hebert 2005; Ebert & Ebb 2005), Trichoptera (Zhou *et al.* 2011), Lepidoptera (Hausmann *et al.* 2011; Strutzenberger *et al.* 2011), Hymenoptera (Smith & Fisher 2009; Zaldívar-riverón *et al.* 2011), Hemiptera (Deister *et al.* 2014; Tembe *et al.* 2014), Coleoptera (Raupach *et al.* 2010; Woodcock *et al.* 2013), Arachnida (Blagoev *et al.* 2013; Blagoev *et al.* 2016) and Diptera (Jordaens *et al.* 2015; Pinto *et al.* 2015). However, very few of these studies have actually checked for the presence of a DNA barcoding gap.

The success of DNA barcoding is also affected by the sequence divergence threshold used to delimit species (Smith *et al.* 2005; Chapple & Ritchie 2013). The use of sequence divergence thresholds is usually effective for species identification because levels of barcode variation within species are highly conserved while deep sequence divergence usually occurs between different species (Hebert *et al.* 2003b). For separating species, sequence divergence thresholds between 1-3% have usually been used (Hebert *et al.* 2003a; Hebert *et al.* 2003b). The BOLD system uses a standard sequence divergence threshold of 1% for the identification of insect species. However, different thresholds can be used to identify species in different taxonomic groups. For example, Hebert *et al.* (2003b) was able to correctly identify approximately 98% of Lepidopteran species identified through conventional morphological taxonomy at a threshold of 3%, while a 2.7% threshold correctly assigned 90% of the 260 recognized bird species from North America (Hebert *et al.* 2004). Furthermore, Smith *et al.* (2005) suggested a 2–3% threshold as suitable for ant species. In some animal groups the use of *COI* sequence divergence threshold has been less successfully applied. This is due to the fact that in some animal groups the mutational rate of the *COI* is either too slow *i.e* Cnidarian and sponges (Shearer *et al.* 2002; Park *et al.* 2007) or too fast *e.g.* aves, gastropods and amphibians (Remigio & Hebert 2003) to accurately delimit species.

### 1.1.5 Next generation sequencing and DNA barcoding

The current protocol used in DNA barcoding is based on the PCR amplification of the *COI* gene followed by Sanger sequencing. This approach has proven robust and effective when applied to a few samples (Galan & Page 2012) and has been successfully used for the construction of sequence libraries such as BOLD (Hajibabaei *et al.* 2011). However, Sanger sequencing becomes inefficient and expensive when scaled up to thousands of samples. Moreover, difficulties such as heteroplasmy (several mitochondrial genomes co-existing within the same cell) or Numts (copies of MtDNA that integrated into nuclear genome) further complicate the task of species identification using the Sanger sequencing techniques (Richly & Leister 2004; Rubinoff *et al.* 2006; Galan & Page 2012).

The introduction of high throughput Next Generation Sequencing (NGS) technology has revolutionized molecular biology in recent years. These technologies allow researchers to generate vast amounts of sequence data in a relatively short space of time (Bybee *et al.* 2011). More importantly, NGS technology has led to the development of a novel barcoding methods, for a fast and accurate identification of large number of species (Hajibabaei *et al.* 2011; Galan & Page 2012). There are currently two major NGS technologies available: SOLiD technology (Life Technologies, USA) and Illumina technology (Illumina, USA) The SOLiD technology was introduced by Applied Biosystems in 2007 as their NGS platform and it utilizes a sequence-by oligo-ligation-technology. Two versions of the SOLiD platform are commercially available: the

5500 system and the 5500xl system with 100 Gb and 250 Gb sequencing capacity respectively (Shokralla *et al.* 2012).

The Illumina sequencing platform was also introduced in 2007. This technology utilizes a sequence-by-synthesis approach, coupled with bridge amplification on the surface of a flow cell. There are currently five versions of the Illumina sequencer commercially available: The MiniSeq, MiSeq, NestSeq, HiSeq and HiSeq X. These Illumina platforms can generate sequence output ranging between 8 Gb to 1800 Gb per run (Shokralla *et al.* 2012). These technologies provide billions of sequence reads in a single experiment, while the traditional Sanger sequencing utilizing a single capillary per sample (Taberlet & Coissac 2012).

Recent increases in fragment sizes amplified using Illumina platforms have made them acceptable for barcode studies. Currently, Illumina (MiSeq) can produce 300 bp paired-end reads with a maximum output of 15 Gb per run. Furthermore, the MiSeq system is cost effective, offers an easy, fast sequencing workflow and turn-around, it has high yields as well as datasets with high quality scores (Illumina 2014).

### 1.1.6 Prospects and problems of DNA barcoding

Several studies have demonstrated the efficiency of DNA barcoding in different animal groups (Hebert *et al.* 2003a; Hebert *et al.* 2003b; Hebert *et al.* 2004; Ward *et al.* 2005; Hajibabaei *et al.* 2006; Cander & Kuntner 2015; Dona *et al.* 2015). In particular, DNA barcoding has been successfully used to identify various insect species and other animal species including mites (Cander & Kuntner 2015), cryptic bee species (Murray *et al.* 2007), mosquito species in Colombia (Rozo-Lopez & Mengual 2015), sand flies in India and Colombia (Kumar *et al.* 2012; Gutierrez *et al.* 2014) and Nearctic black flies (Rivera & Currie 2009). Moreover, DNA barcoding has been used to enhance taxonomic investigations (Droege *et al.* 2010) and investigate the validity of morphological keys (Carolan *et al.* 2012). However, the adequacy of DNA barcoding in the identification of Diptera is still in question (Meier *et al.* 2006; Virgilio *et al.* 2010). In particular, Meier *et al.* (2006) reported a remarkably low identification success in Diptera (<70% identification success). Furthermore, DNA barcoding has been criticized as it relies on a single mitochondrial gene region for identification and can be misleading especially in the face of widespread mitochondrial paraphyly and polyphyly (Blaxter 2004; Will & Rubinoff 2004). Failure of the barcode marker to accurately discriminate species could be attributed to factors such as recent speciation (Nichols 2001; van Velzen *et al.* 2012a), interspecific hybridization (Chapple *et al.* 2012) and incomplete DNA barcode reference libraries (Hebert *et al.* 2004).

The current study will begin the construction of the DNA barcode library of the Diptera of eThekwini. The DNA barcode library will be used to test the efficiency of the *COI* region as a DNA barcode for the identification of the South African Diptera species by examining the existence of a barcoding gap as well as testing the identification efficacy on Diptera species using three distance-based methods: Near Neighbor (NN), Best Close Match (BCM) and Bold Identification Criteria (BIC). The study will also use NGS technology to construct a complete mitochondrial genome of a Diptera species (*Lucilia cuprina*: Calliphoridae). This complete mitochondrial genome together with 48 complete mitochondrial genomes of other Diptera species obtained from the National Center for Biotechnology Information (NCBI) Genome Database will be used to explore other potential mitochondrial genes that can be used as DNA barcodes for the identification of Diptera species.

## 1.2 References

Agapow P.M., Binind-Emonds O.R.P., Crandall K.A., Gittleman J.L., Mace G.M., Marshall J.C. & Purvis A. (2004) The impact of species on biodiversity studies. *The Quarterly Review of Biology* **79**, 161-79.

Aliabadian M., Kaboli M., Nijman V. & Vences M. (2009) Molecular identification of birds: performance of distance based DNA barcoding in three genes to delimit parapatric species. *PLoS ONE* **4**, 1-8.

Ball S.L. & Hebert P.D.N. (2005) Biological identification of mayflies (Ephemeroptera) using DNA barcodes. *Journal of North American Benthological Society* **24**, 508-24.

Blagoev G.A., Dewaard J.R., Ratnasingham S., Stephanie L., Lu L., Robertson J. & Telfer A.C. (2016) Untangling taxonomy: a DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources* **16**, 325-41.

Blagoev G.A., Nikolova N.I., Sobel C.N., Hebert P.D.N. & Adamowicz S.J. (2013) Spiders (Araneae) of Churchill, Manitoba: DNA barcodes and morphology reveal high species diversity and new Canadian records. *BMC Ecology* **13**, 1-17.

Blaxter M.L. (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society, Biological Sciences* **359**, 669-79.

Bossuyt F., Meegaskumbura M., Beenaerts N., Gower D.J., Pethiyagoda R., Roelants K., Mannaert A., Wilkinson M., Bahir M.M., Manamendra-arachchi K., Oommen O.V. & Milinkovitch M.C. (2004) Local endemism within the Western Ghats – Sri Lanka biodiversity hotspot. *Science* **306**, 479-82.

Butchart S.H.M., Walpole W., Collen B., Strien A.V., Scharlemann J.P.W., Almond R.E.A., Baillie J.E., Bomhard B., Brown C., Bruno J., Carpenter K.E., Carr G.M., Chanson J., Chenery A.M., Csirke J., Davidson N.C., Dentener F., Foster M., Galli A., Galloway J.N., Genovesi P., Leverington F., Loh J., McGeoch M.A., McRae L., Minasyan A., Morcillo M.H., Oldfield T.E.E., Pauly D., Quader S., Revenga C., Sauer J.R., Skolnik B., Spear D., Stanwell-Smith D., Stuart S.N., Symes A., Tierney M., Tyrrell T.D., Vie J.C. & Watson R. (2010) Global Biodiversity: indicators of recent declines. *Science* **328**, 1164-8.

Bybee S.M., Bracken-Grissom H.D., Hermansen R.A., Clement M.J., Crandall K.A. & Felder D.L. (2011) Directed next generation sequencing for phylogenetics: an example using Decapoda (Crustacea). *Journal of Comparative Zoology* **250**, 497-506.

Cander K. & Kuntner M. (2015) DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources* **15**, 268-77.

Carolan J.C., Murray T.E., Fitzpatrick Ú., Crossley J., Schmidt H., Cederberg B., McNally L., Paxton R.J., Williams P.H. & Brown M.J.F. (2012) Colour patterns do not diagnose species: quantitative evaluation of a DNA barcoded cryptic bumblebee complex. *PLoS ONE* **7**, 1-11.

Ceballos G., Ehrlch P.R., Barnosky A.D., García A., Pringle R.M. & Palmer T.M. (2015) Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances* 19, e1400253.

Chapple D.G., Birkett A., Miller K.A., Daugherty C.H. & Gleeson D.M. (2012) Phylogeography of the endangered Otago Skink, *Oligosoma otagense*: population structure, hybridisation and genetic diversity in captive populations. *PLoS ONE* **7**, 1-11.

Chapple D.G. & Ritchie P.A. (2013) A retrospective approach to testing the DNA barcoding method. *PLoS ONE* **8**, 1-12.

Dasmahapatra K.K. & Mallet J. (2006) DNA barcodes: recent successes and future prospects. *Heredity* **97**, 254-5.

Deister F., Raupach M.J., Hendrich L., Ku S.M. & Gossner M.M. (2014) Building up of a DNA barcode library for true bugs (Insecta: Hemiptera: Heteroptera) of Germany reveals taxonomic uncertainties and surprises. *PLoS ONE* **9**, 1-13.

deVere N. (2008) Biodiversity. *Modern Taxonomy and Fieldwork*, 1-12.

Dona J., Diaz-Real J., Mironov S., Bazaga P., Serrano D. & Jovani R. (2015) DNA Barcoding and minibarcoding as a powerful took for feather mites studies. *Molecular Ecology Resources*, 1-10.

Droege S.A.M., Rightmyer M.G., Sheffield C.S. & Brady S.G. (2010) New synonymies in the bee genus Nomada from North America (Hymenoptera: Apidae). *Zootaxa* **32**, 1-32.

Ebert P.A.U.L.D.N.H. & Ebb J.E.M.W. (2005) Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of North American Benthological Society* **24**, 508-24.

Ekrem T., Willassen E. & Stur E. (2007) A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution* **43**, 530-42.

Elmeer K., Almalki A. & Mohran K.A. (2012) DNA barcoding of *Oryx leucoryx* using the mitochondrial cytochrome c oxidase I gene. *Genetics and Molecular Research* **11**, 539-47.

Galan M. & Page M. (2012) Next generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. *PLoS ONE* **7**, 1-13.

Goerge A. & Mayden R.L. (2001) Species concepts and the endangered species act: how a valid biological definition of species enhances the legal protection of biodiversity. *Natural Resoures Journal* **32**, 371-405.

Gutierrez M.A.C., Vivero R.J., Velez I.D., Porter C.H. & Uribe S. (2014) DNA barcoding for the identification of snad fly species (Dipera, Psychodidae, Phlebotominae) in Colombia. *PLoS ONE*, 1-9.

Hajibabaei M., Janzen D.H., Burns J.M., Hallwachs W. & Hebert P.D.N. (2006) DNA barcodes distinguish species of tropical Lepidoptera. *PNAS* **103**, 968-71.

Hajibabaei M., Shokralla S., Zhou X., Singer G.A.C. & Baird D.J. (2011) Environmental barcoding: a next generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* **6**, 1-7.

Hajibabaei M., Singer G.A.C., Hebert P.D.N. & Hickey D.A. (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* **23**, 167-72.

Hausmann A., Haszprunar G. & Hebert P.D.N. (2011) DNA Barcoding the Geometrid fauna of Bavaria (Lepidoptera): Successes, Surprises and Questions. *PLoS ONE* **6**, 1-9.

Hebert P.D.N., Cywinska A., Ball S.L. & Waard J.R.D. (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society, Biological Sciences* **270**, 313-21.

Hebert P.D.N., Ratnasingham S. & deWaard J.R. (2003b) Barcoding animal life: cytochrome c oxidase subunit I divergences among closely related species. *Proceedings of the Royal Society, Biological Sciences* **270**, 1-4.

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S. & Francis C.M. (2004) Identification of birds through DNA barcodes. *PLoS Biology* **2**, 1-7.

Iii F.S.C., Zavaleta E.S., Eviner V.T., Naylor R.L., Vitousek P.M., Reynolds H.L., Hooper D.U., Lavorel S., Sala O.E., Hobbie S.E., Mack M.C. & Díaz S. (2000) Consequences of changing biodiversity. *Nature* **405**, 234-42.

Illumina (2014) http://www.illumina.com/systems/miseq/system.ilmn.

Janzen D.H., Hajibabaei M., Burns J.M., Hallwachs W., Remigio E. & Hebert P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society, Biological Sciences* **360**, 1835-45.

Jarman S.N. & Elliott N.G. (2000) DNA evidence for morphological and cryptic Cenozoic speciations in the Anaspididae, ‘living fossils’ from the Triassic. *Journal of Evolutionary Biology* **13**, 624-33.

Jinbo U., Kato T. & Ito M. (2011) Current progress in DNA barcoding and future implications for entomology. *Entomological Science* **14**, 107-24.

Jordaens K., Goergen G., Virgilio M. & Backeljau T. (2015) DNA barcoding to improve the taxonomy of the Afrotropical hoverflies (Insecta: Diptera: Syrphidae). *PLoS ONE*, 1-15.

Kumar N.P., Srinivasan R. & Jambulingam P. (2012) DNA barcoding for identification of sand flies (Diptera: Psychodidae) in India. *Molecular Ecology Resources* **12**, 414-20.

Lee W., Kim H., Lim J., Choi H.R., Kim Y., Kim Y.S., Ji J.Y., Foottit R.G. & Lee S. (2011) Barcoding aphids (Hemiptera: Aphididae) of the Korean Peninsula: updating the global data set. *Molecular Ecology Resources* **11**, 32-7.

Meier R., Shiyang K., Vaidya G. & Ng P.K.L. (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55**, 715-28.

Meyer C.P. & Paulay G. (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* **3**, 422-5.

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

Murray T.E., Fitzpatrick Ú., Brown M.J.F. & Paxton R.J. (2007) Cryptic species diversity in a widespread bumble bee complex revealed using mitochondrial DNA RFLPs. *Conservation Genetics* **9**, 653-66.

Nichols R. (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution* **16**, 358-64.

Park M.H., Sim C.J., Baek J. & Min G.S. (2007) Molecules and identification of genes suitable for DNA barcoding of morphologically indistinguishable Korean Halichondriidae sponges. *Molecules and Cells* **23**, 220-7.

Pili E., Carcangiu L., Oppo M. & Marchi a. (2010) Genetic structure and population dynamics of the biting midges *Culicoides obsoletus* and *Culicoides scoticus*: implications for the transmission and maintenance of bluetongue. *Medical and Veterinary Entomology* **24**, 441-8.

Pimm S.L., Jenkins C.N., Abell R., Brooks T.M., Gittleman J.L., Joppa L.N., Raven P.H., Roberts C.M. & Sexton J. (2014) The biodiversity of species and their rates of extinction, distribution, and protection. *Science*. 201344:1246752

Pinto I.D.S., Dias B., Alencastre A., Rodrigues F., Ferreira A.L., Rezende H.R., Bruno R.V., Falqueto A., Andrade-filho J.D., Aparecida E., Galati B., Helena P., Shimabukuro F., Brazil R.P. & Peixoto A.A. (2015) DNA barcoding of Neotropical sand flies species identification and discovery *PLoS ONE*, 1-18.

Ratnasingham S. & Hebert P.N.D. (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355-64.

Raupach M.J., Astrin J.J., Hannig K., Peters M.K., Stoeckle M.Y. & Wägele J.-w. (2010) Molecular species identification of Central European ground beetles (Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Frontiers in Zoology* **7**, 1-15.

Remigio E. & Hebert P.N.D. (2003) Testing the utility of partial *COI* sequences for phylogenetic estimates of gastropod relationships. *Molecular Phylogenetics and Evolution* **29**, 641-7.

Renaud A.K., Savage J. & Adamowicz S.J. (2012) DNA barcoding of Northern Nearctic Muscidae (Diptera) reveals high correspondence between morphological and molecular species limits. *BMC Ecology* **12**, 1-15.

Richly E. & Leister D. (2004) NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution* **21**, 1081-4.

Rivera J. & Currie D.C. (2009) Identification of Nearctic black flies using DNA barcodes (Diptera: Simuliidae). *Molecular Ecology Resources* **9** 224-36.

Rozo-Lopez P. & Mengual X. (2015) Mosquito species (Diptera, Culicidae) in three ecosystems from Colombian Andes: identification through DNA barcoding and adult morphology. *ZooKeys* **513**, 39-64.

Rubinoff D., Cameron S. & Will K. (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity* **97**, 581-94.

Shearer T.L., Van Oppen M.J.H., Romano S.L. & Wrheide G. (2002) Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Molecular Ecology* **11**, 2475-87.

Shokralla S., Spall J.L., Gibson J.F. & Hajibabaei M. (2012) Next generation sequencing technologies for environmental DNA research. *Molecular Ecology* **21**, 1794-805.

Smith M.A. & Fisher B.L. (2009) Invasions, DNA barcodes and rapid biodiversity assessment using ants of Mauritius. *Frontiers in Zoology* **12**, 1-12.

Smith M.A., Fisher B.L. & Hebert P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society Biological Sciences* **360**, 1825-34.

Sourakov A. & Zakharov E.V. (2011) "Darwin's butterflies"? DNA barcoding and the radiation of the endemic Caribbean butterfly genus Calisto (Lepidoptera, Nymphalidae, Satyrinae). *Comparative Cytogenetics* **5**, 191-210.

Strutzenberger P., Brehm G. & Fiedler K. (2011) DNA barcoding based species delimitation increases species count of Eois (Geometridae) moths in a well studied tropical mountain forest by up to 50%. *Insect Science* **18**, 349-62.

Taberlet P. & Coissac E. (2012) Towards next generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **33**, 2045-50.

Tembe S., Shouche Y. & Ghate H.V. (2014) DNA barcoding of Pentatomomorpha bugs (Hemiptera: Heteroptera ) from Western Ghats of India. *Meta Gene 2* **2**, 737-45.

van Velzen R., Weitschek E., Felici G. & Bakker F.T. (2012a) DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE* **7**, e30490-e.

van Velzen R., Weitschek E., Felici G. & Bakker F.T. (2012b) DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE* **7**, 1-12.

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

Ward R.D., Zemlak T.S., Innes B.H., Last P.R. & Hebert P.D.N. (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society, Biological Sciences* **360**, 1847-57.

Waugh J. (2007) DNA barcoding in animal species: progress, potential and pitfalls. *Bioessays* **29**, 188-97.

Webb J.M., Jacobus L.M., Funk D.H., Zhou X., Kondratieff B., Geraci C.J., DeWalt R.E., Baird D.J., Richard B., Phillips I. & Hebert P.D.N. (2012) A DNA barcode library for North American Ephemeroptera: progress and prospects. *PLoS ONE* **7**.

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

Woodcock T.S., Boyle E.E., Roughley R.E., Kevan P.G., Labbee R.N., Smith A.B.T., Goulet H., Steinke D. & Adamowicz S.J. (2013) The diversity and biogeography of the Coleoptera of Churchill: insights from DNA barcoding. *BMC Ecology* **13**, 1-15.

Wyman S.K., Jansen R.K. & Boore J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252-5.

Zaldívar-riverón A., Martínez J.J., Ceccarelli F.S., Jesús-bonilla V.S.D., Rodríguez-pérez A.C., Smith M.A., Martínez J.J., Ceccarelli F.S., Jesús-bonilla V.S.D., Rodríguez-pérez A.C., Reséndiz-flores A. & Jose J. (2011) DNA barcoding a highly diverse group of parasitoid wasps (Braconidae: Doryctinae) from a Mexican nature reserve *The Journal of DNA Mapping, Sequencing and Analysis* **1736**, 18-23.

Zhou X., Robinson J.L., Geraci C.J., Parker C.R., Flint O.S., Etnier D.A., Ruiter D., DeWalt R.E., Jacobus L.M. & Hebert P.D.N. (2011) Accelerated construction of a regional DNA-barcode reference library: caddisflies (Trichoptera) in the great smoky mountains national park. *Journal of the North American Benthological Society* **30**, 131-62.

# CHAPTER TWO

# THE UTILITY OF DNA BARCODING ON DIPTERA OF ETHEKWINI

## Abstract

The mitochondrial cytochrome c oxidase subunit 1 (*COI*) gene has been widely used as a standard DNA barcode sequence for the identification of species that belong to many animal groups, including Diptera. However, some authors have suggested that its performance is of limited use in the separation of Diptera species because of the overlap between intraspecific and interspecific sequence divergence or the lack of a DNA barcoding gap. In this study the performance of the *COI* gene as a barcode for the identification of South African Diptera was evaluated. A total of 844 *COI* fragments were successfully sequenced from 1060 Diptera specimens collected from 14 localities within the eThekwini municipality, South Africa. Two analyses for evaluating the performance of the *COI* region were conducted, the first analysis tested for the presence of the barcoding gap using the Kimura-two-Parameter (K2P) model which is a standard model used by the barcoding community and the General Time Reversible (GTR) + Invariant Sites (I) + Gamma (G) model. The Jeffries-Matusita (JM) was used to test for separability of the intraspecific and interspecific sequence divergences. The second analysis assessed the proportion of correct species identified using three species identification criteria namely: Near Neighbor (NN), Best Close Match (BCM) and the BOLD Identification Criteria (BIC). Our results showed a significant overlap between interspecific and intraspecific sequence divergence using the K2P model (JM = 0.220), however when the GTR + I + G substitution model was applied there was no significant overlap between interspecific and intraspecific sequence divergences (JM = 1.818). This result suggested that the model choice is important in establishing the presence of the DNA barcoding gap and the widely used K2P model may not always be an appropriate model for use in DNA barcoding. Relatively moderate identification success (<70%) was obtained for all three species identification criteria: NN = 577 (68%); BCM = 546 (65%) and BIC= 526 (62%). The overlap between the intraspecific and interspecific sequence divergence (using K2P as utilized on BOLD) together with the low identification success suggests limited performance of the *COI* region for the identification of South African Diptera.

**Key Words:** eThekwini municipality, Diptera, DNA barcoding, barcoding gap

## 2.1 Introduction

The order Diptera contains winged insects commonly known as flies. This order includes all true flies namely, crane flies, robber flies, fruit flies, gnats, midges and mosquitoes. All members have distinctive hind wings that are reduced to small, club-shaped structures called halters (Meyer 2009). This order is extremely diverse and is one of the three largest and most taxon-rich animal groups in the world, with 188 currently recognized families and about 120,000 described species (Skevington & Dang 2002; Rambaut 2009; Wiegmann *et al.* 2011). Diptera is divided into two main sub-orders: Nematocera and Brachycera (Szymura *et al.* 1996). The Nematocera are small delicate flies with segmented antennae, long legs and thin wings, while members of Brachycera are much larger and more robust (Serna *et al.* 2004).

Dipterans play a significant role in the healthy functioning of ecosystems. They are involved in the decomposition of plants and animals. Diptera also facilitate the breakdown and release of nutrients back to the soil (Meyer 2009). They are also among the most important groups of pollinating organisms second only to the Hymenoptera (Ssymank *et al.* 2008) and were among the first pollinators that are linked to the early angiosperm radiation (Endress 2001). Diptera species are often linked to the plants they visit (Menges 1991). A number of publications have reported large-scale parallel declines of pollinators and plants, reinforcing the concern that pollination, as an important ecosystem services, is at risk (Menges 1991; Biesmeijer *et al.* 2006; Potts *et al.* 2010).

The first step in conservation planning is an inclusive evaluation of regional biodiversity, starting with species identification. However, morphology-based identifications are often difficult on taxon-rich animal groups such as Diptera. Indeed, morphology-based species identification in Diptera is mainly based on differences in male genitalia (Pili *et al.* 2010). The larvae and females of closely related species are usually difficult to separate by morphology alone (Ekrem *et al.*

2007). Therefore, well-trained taxonomists with a high level of expertise are often required to correctly identify species belonging to this order.

DNA barcoding provides an alternative way of identifying species effectively and rapidly (Hebert *et al.* 2003b). In animals, the short, standardized mitochondrial cytochrome c oxidase subunit I (*COI)* gene is used as the DNA "barcode". This DNA barcode has been successful used in a broad range of invertebrates including Lepidoptera (Hebert *et al.* 2004a), Formicidae (Smith *et al.* 2005), Hemiptera (Park *et al.* 2011), Hymenoptera (Magnacca & Brown 2012) and Arachnida (Blagoev *et al.* 2013). The *COI* barcode has also been successfully used in the identification of species belonging to the order Diptera (Rivera & Currie 2009; Kumar *et al.* 2012; Versteirt *et al.* 2014), however, it's wide-spread reliability has been questioned (Meier *et al.* 2006; Virgilio *et al.* 2010). In particular, Meier *et al.* (2006) reported a remarkably low identification success when using barcoding in Diptera (<70% identification success) when compared to other insect orders.

This low identification success of DNA barcoding in Diptera using *COI* barcode was mainly attributed to an overlap between intraspecific and interspecific sequence divergence (Meier *et al.* 2006). Indeed, species identification using DNA barcoding is only reliable and effective if there is a significant difference between the average intraspecific and interspecific sequence divergence (Meyer & Paulay 2005; Chapple & Ritchie 2013). Furthermore, successful DNA barcoding is affected by the sequence divergence threshold used to differentiate species. For separating species, sequence divergence thresholds between 1-3% has been used (Hebert *et al.* 2003a; Hebert *et al.* 2003b). However, the literature suggests that no single threshold is optimal for all species (Ferguson 2002; Hebert *et al.* 2003a; Hebert *et al.* 2004a; Smith *et al.* 2005; Chapple & Ritchie 2013). Therefore, it is important to choose an appropriate threshold for each taxonomic group studied.

Spatial scale is another factor that affects species identifications using DNA barcoding. A species sampled throughout its geographical range will have a larger genetic variation than when only a few individuals are sampled from a single geographic locality (Bergsten *et al.* 2012). This is because in DNA barcoding, as spatial sampling increases intraspecific divergence also increases.

This may also increase if the species shows phylogeographic structure and could result in the overlap of intra- and interspecific sequence divergence values.

Most DNA barcoding studies on Diptera were done in Europe, with no studies done in Africa and in particular South Africa. Therefore, it was unclear how successful the *COI* region would be at delimiting South African species collected from a relatively small geographic region. In this study, the performance of the standard *COI* region as a barcode for the identification of South African taxa was evaluated by testing for the existence of barcoding gaps using two substitution models, the standard K2P model, which is used widely in the barcoding community and a more parameter rich model (GTR + I + G) which best fits our data.

This study also offers a unique case study to test the effect of small regional sampling on the utility of a DNA reference library. To test the effect of spatial scale, barcode gap analyses were performed on two different data sets. First, the DNA barcode gap was evaluated on all the available South African Diptera barcode data is available from the global Barcode of Life Database (BOLD). Second, analyses were repeated for a group of taxa collected at a much smaller spatial scale namely the eThekwini municipality area. To determine if the presence of the DNA barcode gap was taxonomically linked, analyses were repeated separately on ten Diptera families. The performance of the *COI* region was also evaluated by estimating the proportion of correct species identifications obtained using three species identification criteria: Near Neighbor (NN), Best Close Match (BCM) following Meier *et al.* (2006), and the BOLD Identification Criteria (BIC) which is used in BOLD (www.boldsystems.org*)*.

## 2.2 Materials and methods

### 2.2.1 Study area: The eThekwini municipality

The eThekwini municipality is in the province of KwaZulu-Natal, South Africa. It covers a land area of 2, 297 km$^2$ and includes the city of Durban and surrounding areas (Outer West Durban) making it one of the largest municipalities in South Africa. The eThekwini region is a sub-tropical coastal region, characterized by high temperatures, humidity and summer rainfall

(Fairbanks, Reyers, & van Jaarsveld, 2001). It contains three of South Africa's eight terrestrial biomes (savanna, forest and grassland), as well as eight broad vegetation types: the Eastern Valley Bushveld, KwaZulu-Natal Coastal Belt, KwaZulu-Natal Hinterland Thornveld, KwaZulu-Natal Sandstone Sourveld, Ngongoni Veld, Northern Coastal Forest, Scarp Forest and Mangrove Forest (eThekwini state of biodiversity report, 2014/2015).

The eThekwini municipality falls within the Maputaland-Pondoland-Albany Region of endemism, which is a biodiversity hotspot of global importance (Scott-Shaw 2011). However, there is increasing pressure on the natural habitats within the region due to urbanization. One significant impact of urbanization is that it leads to a dramatic decrease in native species (McKinney, 2002). This makes it necessary to conserve the remaining urban green spaces.

Urban green spaces are defined as public or private open spaces in urban areas, primarily covered by vegetation which are directly or indirectly available to the public (Tuzin *et al.* 2002). These can protect and enhance biodiversity in urban areas as they function as protection centers for the conservation of both plants and animals (Haq 2011). The significance of urban green spaces within cities is important from both socio-economic and a biodiversity point of view (Haq 2011). Within the eThekwini municipality, there are a number of open green spaces. These open green spaces are managed through the Durban Metropolitan Open Space System (D'MOSS), which is used as a key planning tool to help achieve provincial and national biodiversity targets. The D'MOSS system incorporates areas of high biodiversity and also prioritizes areas which provide essential ecosystem goods and services (eThekwini state of biodiversity report, 2014/2015).

To investigate the biodiversity present within these urban green spaces, Diptera were collected from twelve sites within the eThekwini municipality and two sites in Pietermaritzburg (Figure 2.1; Table 2.1).

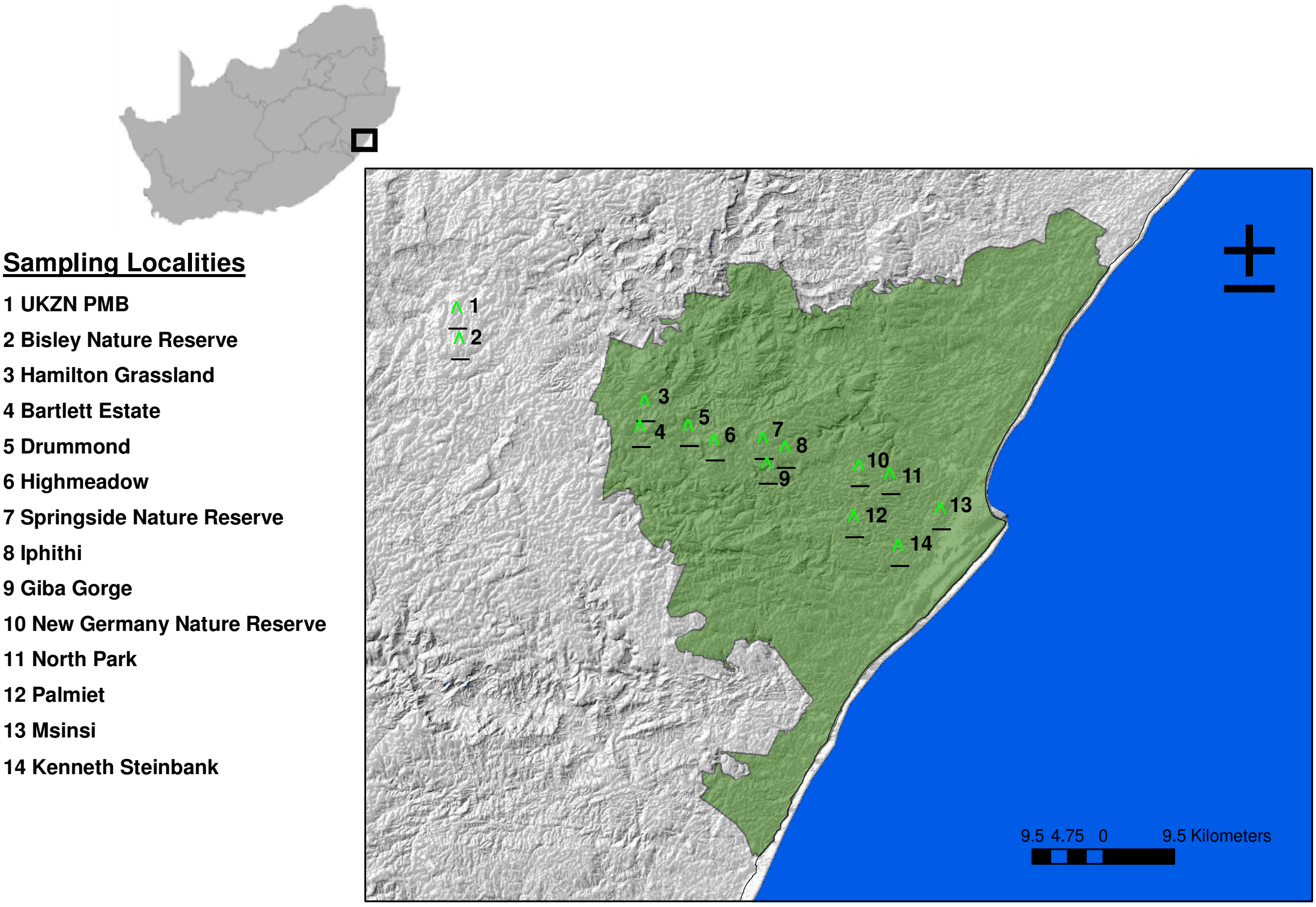


Figure 2.1 Geographic distribution of sampling localities used in the current study. The eThekwini municipal area is highlighted in green. The GPS coordinates of sites are listed in Table 2.1.

Diptera specimens were sampled using sweep nets in summer and spring 2012 and 2013. Specimens were collected into glass jars with 96% ethanol and stored at low temperature (4$^{o}$C) to preserve the material for molecular analysis. Geographic coordinates for all sampled sites, as well as the vegetation characteristic of each site were recorded (Table 2.1). In most cases a site was visited multiple times, although five sites were visited only once.

Table 2.1 List of sampling localities included in the study. Details of prominent vegetation type and number of separate sampling events are provided. NR – Nature Reserve.

| **Site** | **Vegetation type** | **Latitude** | **Longitude** | **Sampling events** |
|---|---|---|---|---|
| Msinsi | Grassland and forest | 29°51'48"S | 30°59'13"E | 3 |
| Palmiet | Grassland and forest | 29°49.35"S | 30°55'58"E | 3 |
| New Germany NR* | Grassland and forest | 29°48'45"S | 30°53'19"E | 2 |
| Iphithi NR* | Grassland and forest | 29°47'26"S | 30°47'59"E | 3 |
| North Park | Coastal bush | 29°52'23"S | 30°52'57"E | 1 |
| Kenneth Steinbank | Grassland and forest | 29°54'25"S | 30°56'12"E | 1 |
| Springside NR | Grassland and forest | 29°46'49"S | 30°46'23"E | 4 |
| Giba Gorge NR* | Grassland and forest | 29°48'36"S | 30°46'40"E | 2 |
| Erf Drummond | Grassland | 29°45'55"S | 30°40'58"E | 4 |
| Hamilton grassland | Grassland | 29°44'09"S | 30°37'50"E | 2 |
| High Meadows | Grassland | 29°46'56"S | 30°42'52"E | 1 |
| UKZN -PMB | Botanical garden | 29°37'32"S | 30°24'13"E | 2 |
| Bisley NR* | Savanna | 29°39'44"S | 30°23'25"E | 1 |
| Bartlett Estate | Grassland | 29°76'63"S | 30°62'46"E | 1 |

### 2.2.2 DNA extraction and *COI* amplification

Specimens were sorted into morphospecies using available taxonomic literature (Picker *et al.* 2004). For molecular analyses, five individuals from each morphospecies per locality were selected. Each specimen was given a unique collection code for identification. The code was made up of the location where the specimen was sampled, vegetation type and a unique number was assigned to the specimen. For example, a specimen collected from Msinsi forest was assigned an ID code of MsiFoDipt01. All specimens were photographed at the Centre for Electron Microscopy at the University of KwaZulu-Natal (UKZN), using a Leica DFC450C digital camera and Leica stereo dissecting light microscope (Leica microsystems).

DNA extraction, PCR and sequencing of the *COI* gene were performed using standard barcoding protocols (Hajibabaei *et al.* 2005). The primers C_LepFolF 5'-ATTCAACCAATCATAAAGATATTGG-3' and C_LepFolR 5'-TAAACTTCTGGATGTCCAAAAAATCA-3' were used to amplify 658 bp fragment of the *COI* gene. Both forward and reverse sequences were generated and were combined to form a good quality consensus sequence. This data, together with photographs, corresponding GPS coordinates, taxonomic information and collection information were uploaded onto BOLD under the Diptera of eThekwini subproject (ETKD). Voucher specimens were retained and are stored in a designated storage facility in the Department of Genetics at the University of KwaZulu-Natal, Pietermaritzburg.

### 2.2.3 Data analysis

***Assigning barcode index numbers (BIN)***

All available Diptera sequences from eThekwini were selected in BOLD and aligned using the amino acid based BOLD Aligner (Eddy 1998). The sequences were then analysed using the Barcode Index Number (BIN) system implemented in BOLD. The BIN system clusters DNA sequences to calculate operational taxonomic units (OTUs) that closely correspond to species (Ratnasingham & Hebert 2013). BIN clusters are indexed in such a way that all genetically similar taxa reside under a shared identifier (Ratnasingham & Hebert 2007; Ratnasingham & Hebert 2013). The assignment of taxa to BINs, however, relies on the existence of the "DNA barcoding gap". The DNA barcoding gap is the difference between interspecific and intraspecific genetic distances within a group of organisms (Meyer & Paulay 2005). Thus, where the DNA barcoding gaps do not exist, BINs cannot reliably represent species. Therefore, a neighbor joining (NJ) cluster analysis was performed in the BOLD workbench using the K2P model (Kimura 1980) to construct a graphical representation of the nucleotide divergences and the respective BIN clusters (Appendix 1).

### *The contribution of the eThekwini project to the global barcoding initiative*

To determine the contribution that the eThekwini project had on the global barcode initiative, one sequence from each BIN was selected and blasted against the BOLD database using the BOLD identification search engine. The BOLD identification search engine works by matching the query sequences to the already identified species in the BOLD database with the corresponding sequence similarity. Sequence similarity values were recorded, and species/BINs with sequence similarity values above 95% were considered to already be present in BOLD. Using sequence similarity values eThekwini specimens were also given species names or provisional genus-level names.

### *The effect of geographic scale of sampling on DNA barcoding*

To test for the effect of geographical scale of sampling on the utility of the DNA reference library, barcode analyses were performed using all the South African Diptera barcode data available on BOLD. Unfortunately, there has not been much barcoding of Diptera in other regions of South Africa and BOLD database consisted of only 1112 Diptera specimens, of which 52 of these specimens were collected in the Gauteng province and the remaining 1060 specimens were contributed by this eThekwini municipality study. To determine whether the DNA barcode gap was taxonomically linked, the interspecific and intraspecific sequence divergences were also calculated separately for ten Diptera families namely: Agromyzidae, Anthomyiidae, Asilidae, Calliphoridae, Chloropidae, Culicidae, Drosophilidae, Muscidae, Syrphidae and Tephritidae. These families were selected as they contained more than 20 *COI* sequences each.

### *Testing for the presence of the DNA barcoding gap*

In total 844 *COI* sequences from specimens collected from eThekwini were retrieved from BOLD in FASTA format for further analysis. Sequence alignments of these *COI* sequences were performed using the ClustalW (Thompson *et al.* 1994) module in BioEdit (Hall 1999). The alignments were also optimized manually to ensure homology. The aligned sequences contained no insertions, deletions or stop codons. The software program MEGA version 5 (Peterson *et al.*

2011) was used to describe nucleotide composition and the DNASP software version 5.10.01 (Librado & Rozas 2009) was used to describe haplotype and nucleotide diversity.

Tests for the presence of the DNA barcoding gap analyses were conducted at two different spatial and taxonomic levels. First, the presence of the barcoding gap in Diptera data collected from across South Africa was compared to the data collected within a much smaller spatial area of eThekwini. To check what effect taxonomy may have on the presence of barcoding gaps, analyses were also conducted independently on ten Diptera families. The intraspecific and interspecific sequence divergences for all datasets were calculated using two different substitution models. The first model was the K2P model, which is the standard substitution model used for DNA barcoding. The genetic distances generated using this model were calculated using the software SPIDER version 1.1-1 in R (SPecies IDentity and Evolution, (Brown *et al.* 2012).

Although the use of the K2P model may be appropriate when nucleotide sequence divergences are very low (Nei & Kumar 2000), in cases where *COI* sequences are more variable, this model is not the most appropriate and could lead to an underestimation of the genetic divergences amongst taxa. Despite this problem there are few studies which have examined how sensitive DNA barcoding is to model choice (Collins & Cruickshank 2012).

In the field of phylogenetics, selecting the substitution model that describes the mutational process in a set of data is an important step during the inference process. Applying an incorrect model can bias genetic distance calculations, which in turn can affect nodal support values and tree topology (Buckley & Cunningham 2002; Lemmon & Moriarty 2004). It is therefore, important to use the model that best fits the data (Posada, 1998). For each data set (SA data, eThekwini data and each family) the best fit substitution model was selected using the Akaike information criterion (AIC) in jModelTest (Posada, 2008). For all datasets, the GTR + I + G model was selected, this is the most parameter rich model available in jModelTest. This model was then used to estimate the intraspecific and interspecific genetic distances using raxmLGUI version 1.3 (Silvestro & Michalak 2011).

To statistically test for the existence of the gap between GTR + I + G intraspecific and interspecific divergences, the Jeffries-Matusita, JM (Dabboor *et al.* 2014) distance was calculated in R. It takes into account the distance between the two means, as well as the distribution of values from the means. This criterion can be used to pairwise measure the separability between classes, in this case the interspecific and intraspecific divergences, allowing for the statistical assessment of the barcoding gap. The JM distance calculations are always between 0 and 2, a JM value above 1.447 suggests that the two classes are separable (*i.e* a barcoding gap is present) and a JM value below 1.447 suggests that there was overlap between the two classes (*i.e* a barcoding gap is absent). This novel statistical approach has not been applied to barcode data before.

### 2.2.4 Species identification

The accuracy of the *COI* gene to delimit species was tested using two approaches: a distance-based approach and a tree-based approach.

***Distance-based approach***

All the *COI* sequences were first labeled according to species name or genus name (where possible) based on similarity scores from BOLD (see above). When testing the accuracy of DNA barcoding, each sequence is considered as an unknown (query) and the remaining sequences in the dataset are considered as the DNA barcode reference database which is used for identification. If the identification of the query is the same as the prior identification (the sequence labels) then that identification is scored as “correct”. In this study, three criteria where used to test accuracy of DNA barcoding using the *COI* region: the NN, the BCM which was used by Meier *et al.* (2006), and the BIC used by BOLD. All three methods were implemented using the software SPIDER version 1.1-1 in R (Brown *et al.* 2012).

The NN and BCM analyses measure the identification efficacy by searching for the closest individuals (neighbors); the near neighbor focuses on a single nearest neighbor match, whereas the best close match considers all matches under a specific threshold. The BIC method performs species delimitation based on a standard distance threshold of 1%. The near neighbor has only two possible outcomes: true or false while the best close match and the BOLD species identification methods have four possible outcomes: “correct”, “no id”, “ambiguous” and “incorrect”. Moreover, the identification threshold for both methods is user defined and so by changing this variable the optimal threshold can be estimated. Previous studies suggested the 2-3% sequence divergence as the threshold above which a query sequence is considered as distinct from a reference sequence (Hebert *et al.* 2003a; Hebert *et al.* 2003b). In order to determine the suitable threshold for discriminating amongst the Diptera species for this dataset, the Threshold Optimization method (Meyer & Paulay 2005) was used in conjunction with the BCM and the BIC in SPIDER to compare the identification success against a series of genetic distance thresholds ranging from 1% to 5%.

***Tree-based approach: Phylogenetic analysis***

The tree-based approach assigns unidentified (query) barcodes to species based on the clustering of taxa on a phylogenetic tree. The BOLD workbench uses the neighbor joining (NJ) algorithm, which uses the K2P model, to construct phylogenetic trees and assign taxa to BINs. Given that the jModelTest did not recover K2P as the best-fit substitution model the phylogenetic trees were constructed using two model-based approaches not implemented in BOLD: the Bayesian inference (BI) and the maximum likelihood (ML) methods. In both cases, the GTR + I + G model was selected.

The program Garli 0.96 win32 (Zwickl 2006) was used to perform the ML analysis. To assess branch support 100 bootstrap replicates were performed. The BI was performed using MrBayes 3.1.2 (Huelsenbeck & Ronquist 2001). Two independent runs each consisting of four parallel Markov Chain Monte Carlo (MCMC) chains were launched from random starting trees and run for 100 million generations, with the cold Markov chain sampled every 300 generations. Priors were set to nst = 6, invariant sites and gamma. The convergence of the MCMC chains in the BI analysis was assessed using the program Tracer v1.5 (Drummond & Rambaut 2007). This program was used to calculate the Effective Sample Size (ESS) values. Values above 200 indicate that the MCMC chains had converged (Sahlin 2011). The first 25% of trees from each run were discarded as burn-in. All the tree files (BI and ML tree files) were first converted into Phylip format using Mesquite 2.75 (Jühling *et al.* 2012) and consensus trees were constructed using the consensus program which is part of the Phylip v3.69 package (Felsenstein 2005). Phylogenetic trees were viewed in Figtree v1.3.1 (Rambaut 2009).

## 2.3 Results

### 2.3.1 DNA Barcode Library for eThekwini Diptera

A portion of the *COI* gene was successfully sequenced for 850 specimens collected during this study. Of these, 844 *COI* sequences are between 630 – 658 bp in length and meet the criteria for being DNA barcode compliant. The remaining six sequences are ± 200 bp and were removed from all subsequent analyses. Multiple sequence alignment of the 844 *COI* sequences was easily achieved with no insertions, deletions or stop codons. The 844 sequences from specimens collected within the eThekwini and surrounding areas clustered into 400 BINs on the NJ tree generated through BOLD (Appendix 1). A total of 66 (17% of total) BINs exhibit 95% and higher sequence similarly with other records stored on BOLD. These species belong to 47 genera and 14 families (Table 2.2). The Syrphidae family is the best represented family in our eThekwini dataset with 75 species/BINs, followed by the Tephritidae with 23 species/BINs, and the Drosophilidae and Calliphoridae with 13 species/BINs.

Table 2.2 Taxonomic assignments for BINs that have a sequence similarity match greater than 95% based on the BOLD identification search engine.

| BIN number | Taxonomic assignment | Family | Total no. of individuals in BIN | Sequence similarity (%) |
|---|---|---|---|---|
| AAF6797 | *Liriomyza sativae* | Agromyzidae | 1 | 100 |
| ACF4607 | *Lucilia cuprina* | Calliphoridae | 2 | 100 |
| AAA1831 | *Drosophila simulans* | Drosophilidae | 2 | 100 |
| AAG7056 | *Lonchaeid* | Lonchaeidae | 1 | 100 |
| AAA6020 | *Musca domestica* | Muscidae | 2 | 100 |
| AAZ7054 | *Allograpta nasuta* | Syrphidae | 1 | 100 |
| ACF4574 | *Dioxyna cf.sororcula* | Tephritidae | 12 | 99.85 |
| ACC3953 | *Hemipyrellia fernandica* | Calliphoridae | 2 | 99.84 |
| ACB1793 | *Chrysomya inclinata* | Calliphoridae | 1 | 99.75 |
| AAG4663 | *Syritta flaviventris* | Syrphidae | 5 | 99.69 |

| | | | | |
|---|---|---|---|---|
| AAY9761 | *Microdon brevicornis* | Syrphidae | 1 | 99.69 |
| AAE2099 | *Stegomyia simpsoni* | Culicidae | 1 | 99.54 |
| AAW7902 | *Trirhithrum quadrimaculatum* | Tephritidae | 2 | 99.54 |
| AAA4210 | *Stegomyia aegypti* | Culicidae | 1 | 99.46 |
| AAV6733 | *Drosophila vulcana* | Drosophilidae | 2 | 99.39 |
| AAZ4941 | *Sphaeniscus sexmaculatus* | Tephritidae | 1 | 99.36 |
| AAZ8125 | *Paragus borbonicus* | Syrphidae | 22 | 99.24 |
| AAZ8126 | *Paragus borbonicus* | Syrphidae | 22 | 99.24 |
| AAZ8127 | *Paragus borbonicus* | Syrphidae | 22 | 99.24 |
| AAZ8128 | *Paragus borbonicus* | Syrphidae | 22 | 99.24 |
| ACD4493 | *Bengalia depressa* | Calliphoridae | 2 | 99.23 |
| AAX3121 | *Musca asiatica* | Muscidae | 2 | 99.08 |
| ACA6833 | *Hermya* | Tachinidae | 1 | 99.08 |
| ABX9741 | *Paraspheniscoides binaries* | Tephritidae | 1 | 99.08 |
| AAK6361 | *Microcephalops* | Pipunculidae | 2 | 98.92 |
| AAY9765 | *Syritta longiseta* | Syrphidae | 1 | 98.78 |
| AAZ1345 | *Mesembrius* | Syrphidae | 4 | 98.78 |
| AAW1904 | *Allobaccha* | Syrphidae | 1 | 98.76 |
| AAV6732 | *Drosophila vulcana* | Drosophilidae | 6 | 98.62 |
| AAZ3622 | *Culex nebulosusnebulosus* | Culicidae | 1 | 98.61 |
| AAD7633 | *Caenosia attenuate* | Muscidae | 2 | 98.6 |
| AAG6786 | *Rhingia coerulescens* | Syrphidae | 1 | 98.47 |
| ACE7845 | *Episyrphus balteatus* | Syrphidae | 7 | 98.47 |
| AAW3995 | *Chironomus transvaalensis* | Chironomidae | 1 | 98.32 |
| ACE7845 | *Episyrphus balteatus* | Syrphidae | 7 | 98.22 |
| ACH1578 | *Melanostoma annulipes* | Syrphidae | 1 | 98.22 |
| ABX8273 | *Tabanocella denticornis* | Tabanidae | 1 | 98.21 |
| ACH1712 | *Melanostoma annulipes* | Syrphidae | 1 | 97.96 |
| ACH1793 | *Asarkina fulva* | Syrphidae | 1 | 97.69 |
| ACB1846 | *Eretmapodites intermedius* | Culicidae | 1 | 97.47 |
| ACH0903 | *Chrysomya marginalis* | Calliphoridae | 1 | 97.34 |

| | | | | |
|---|---|---|---|---|
| ACB1789 | *Catageiomyia phyllolabis* | Culicidae | 1 | 97.25 |
| ACK5926 | *Centrioncus* | Diopsidae | 1 | 97.25 |
| ACA6779 | *Coenosia acuticornis* | Muscidae | 1 | 97.09 |
| ACB1873 | *Coenosia attenuate* | Muscidae | 1 | 97.07 |
| ACH1075 | *Microdon brevicornis* | Syrphidae | 1 | 96.94 |
| ABX4376 | *Cephalops* | Pipunculidae | 1 | 96.79 |
| ACH0908 | *Syritta bulbus* | Syrphidae | 4 | 96.78 |
| ACA6470 | *Melanostoma univittatum* | Syrphidae | 8 | 96.69 |
| ACB1874 | *Hemigymnochaeta unicolor* | Calliphoridae | 1 | 96.64 |
| AAG6788 | *Rhinia* | Calliphoridae | 4 | 96.48 |
| ACH1506 | *Drosophila jambulina* | Drosophilidae | 1 | 96.48 |
| ACB1857 | *Stomoxys calcitrans* | Muscidae | 1 | 96.41 |
| ACH1579 | *Melanostoma mellinum* | Syrphidae | 2 | 96.22 |
| ABV1242 | *Coenosia attenuate* | Muscidae | 3 | 96.15 |
| ABW4190 | *Oscinella* | Chloropidae | 6 | 95.97 |
| ACH1696 | *Phorinia aurifrons* | Tachinidae | 1 | 95.85 |
| ACC3937 | *Stomorhina lunata* | Calliphoridae | 2 | 95.26 |
| ACC3967 | *Atherigona* | Muscidae | 2 | 95.26 |
| ABW2517 | *Paracantha* | Tephritidae | 7 | 95.26 |
| ACA6862 | *Melanagromyza metallica* | Agromyzidae | 3 | 95.11 |
| ACH0911 | *Conops chinensis* | Conopidae | 1 | 95.11 |
| ACC1790 | *Drosophila jambulina* | Drosophilidae | 1 | 95.11 |
| ACH1574 | *Ceracia* | Tachinidae | 1 | 95.11 |
| ACC3898 | *Leucophenga today* | Drosophilidae | 1 | 95.06 |
| ACB1872 | *Coenosia attenuate* | Muscidae | 3 | 95.05 |

From the remaining 334 BINs, 193 BINs shared between 90 – 95% sequence similarities with BOLD taxa while 141 BINs shared a sequence similarity less than 90%. These BINs include representatives of 30 families and 167 genera. The large number of the BINs with sequence similarity < 95% indicates that the South African taxa are underrepresented in the global database.

### 2.3.2 Sequence analysis

The *COI* sequence alignment of 844 sequences had a high average AT content of 67.98%, which is characteristic of insect mitochondrial DNA (Crozier & Crozier 1993). These values are also well within the average values (66.67 - 70.7%) reported for other species of the order Diptera, including suborders Nematocera and Brachycera (Szymura *et al.* 1996). There was a total of 608 haplotypes from the 844 sequences. The observed haplotype diversity for the Diptera *COI* sequences (Hd > 0.98) and nucleotide diversity (π = 0.151) was high (Table 2.3).

Table 2.3. Summary sequence statistics and diversity indices for alignment of 844 Diptera of *COI* sequences as well as the sequence alignment comprising one representative of each BIN of the 400 BINs.

| | **Complete data (N=844)** | **BINs (N=400)** |
|---|---|---|
| Number of taxa | 844 | 400 |
| Conserved characters | 274 | 275 |
| Variable characters | 384 | 383 |
| Parsimony informative characters | 374 | 364 |
| Number of haplotypes | 608 | 395 |
| Haplotype diversity | 0.998 | 0.999 |
| Standard deviation of haplotype diversity | 0.001 | 0.004 |
| Nucleotide diversity | 0.151 | 0.158 |
| Standard deviation of nucleotide diversity | 0.001 | 0.004 |
| Average number of nucleotide differences | 88.60 | 80.31 |

### 2.3.3 Barcoding gap

The mean interspecific distance of the K2P model (eThekwini = 0.094; SA = 0.092) was smaller than that recovered by the GTR + I + G model (eThekwini = 0.332; SA = 0.332). In contrast the K2P corrected mean intraspecific distances (eThekwini = 0.058; SA = 0.056) were larger than that recovered by the GTR + I + G model (eThekwini = 0.0127; SA = 0.0125). The models also differ in their ability to separate inter- and intraspecific genetic distances. The K2P model had JM values less than 1.447 for both the eThekwini data (JM = 0.220) and South African data (JM = 0.216) suggesting that there is an overlap between the interspecific and intraspecific sequence divergences (Figure 2.2). In contrast the JM values for GTR + I + G corrected distances for both eThekwini (JM = 1.818) and South African data (JM = 1.820) were above 1.447. This means that the interspecific and intraspecific sequence divergences are separable and a barcoding gap does exist (Figure 2.2). This result highlights the importance of model choice in future barcoding studies on South African Diptera.

Plots of the density distributions of intraspecific and interspecific divergences for ten Diptera families were used to test whether the lack of barcoding gap is taxonomically linked. Two Diptera families namely Anthomyiidae (JM = 1.998) and Asilidae (JM = 1.944) have JM values above 1.447 suggesting that there is a gap between the interspecific and intraspecific sequence divergences (Figure 2.3). The remaining eight families namely, Agromyzidae (JM = 0.782), Calliphoridae (JM = 0.474), Chlorophidae (JM = 0.259), Culicidae (JM = 0.770), Drosophilidae (JM = 0.243), Muscidae (JM = 0.206), Syrphidae (JM = 0.547) and Tephritidae (JM = 0.249) have JM values below 1.447 suggesting an overlap between the interspecific and intraspecific sequence divergences (Figure 2.3). This interesting trend will need to be tested using further sampling.

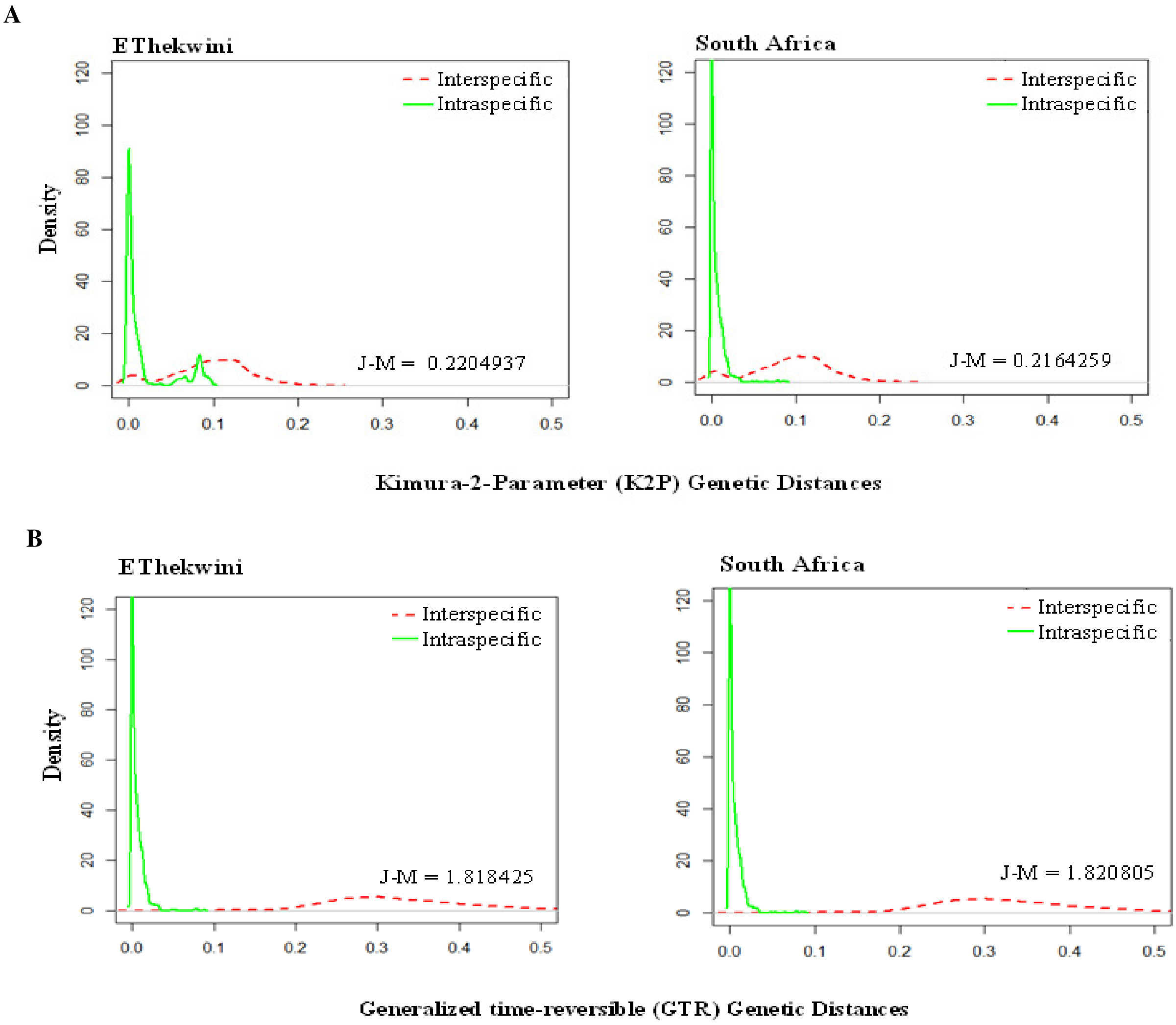


Figure 2.2 Results for the presence of the barcoding gap on the eThekwini data (left panel) and South African data (right panel). Density distributions of intraspecific and interspecific genetic divergences were calculated using the K2P model [A] and the GTR + I + G model [B] for both datasets.

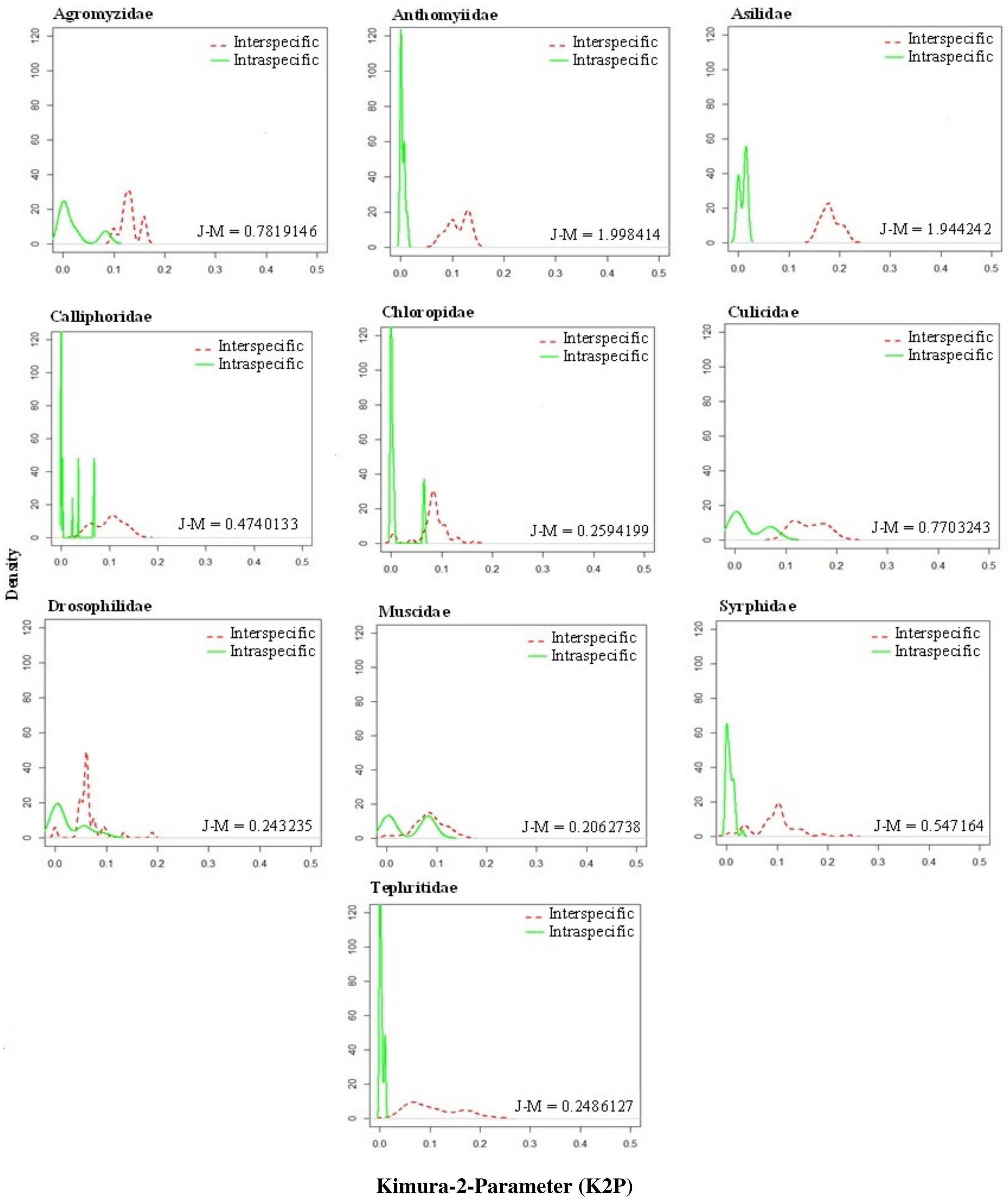


Figure 2.3 Density distributions of intraspecific and interspecific divergences of ten Diptera families. These families were chosen as they have at least 20 *COI* sequences. In each case, the K2P model was used to calculate genetic distances.

### 2.3.4 Species identification

***Distance-based approach***

The threshold optimization identification criterion analysis (Figure 2.4) was performed in order to determine the threshold *COI* sequence divergence, which would be considered as the optimal threshold for the DNA barcode library for the Diptera of eThekwini.

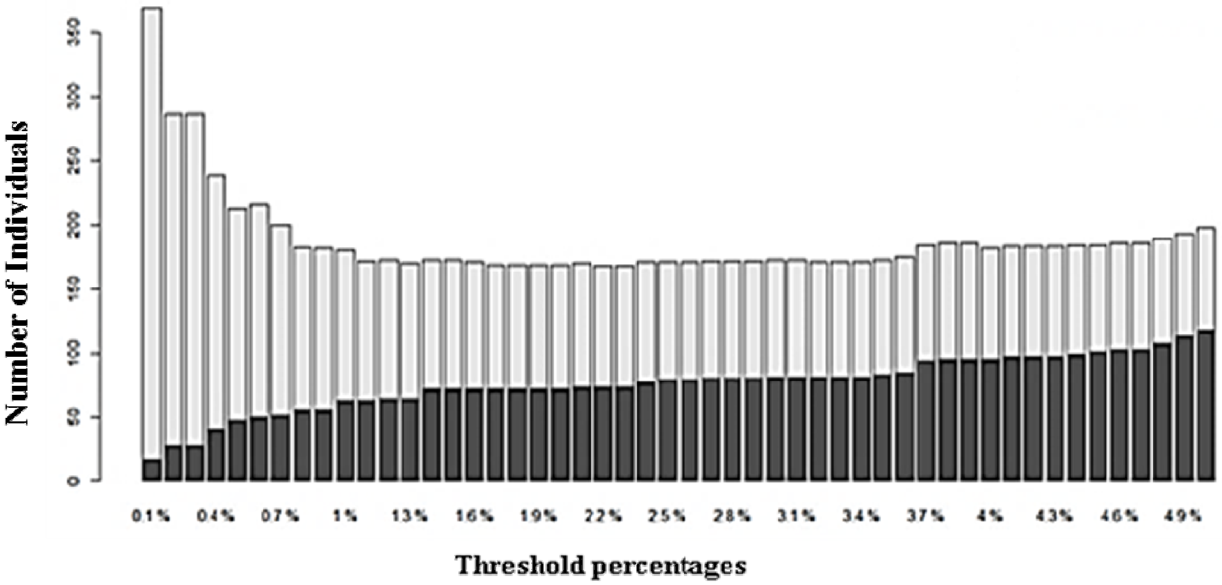


Figure 2.4 Bar-plot showing the false positive (light grey) and false negative (dark grey) rate of identification success containing Diptera species collected in eThekwini as the threshold (0.1 - 4.9%) is changed.

The 3% sequence divergence threshold had the lowest accumulation error (169) while the 5% threshold had the highest accumulation error (198) (Figure 2.4 and Table 2.4). Therefore, the optimal threshold used for delimiting Diptera species from eThekwini was 3%.

Table 2.4 Threshold optimization analysis of the Diptera *COI* sequences at a range of thresholds from 1 –to 5%.

| **Threshold (%)** | **True negative** | **True positive** | **False negative** | **False positive** | **Cumulative error** |
|---|---|---|---|---|---|
| 1.0 | 152 | 512 | 63 | 117 | 180 |
| 2.0 | 149 | 526 | 71 | 98 | 169 |
| 3.0 | 142 | 529 | 81 | 92 | 173 |
| 4.0 | 137 | 526 | 95 | 86 | 181 |
| 5.0 | 131 | 515 | 118 | 80 | 198 |

To test the performance of the *COI* region in the identification of Diptera species, three distance-based species identification criteria were used. Identification success using the Near Neighbor method was 68% with 577 "TRUE" and 267 "FALSE" identifications. At a 3% threshold, the BIC method had the lowest success rate (62%) with a total of 526 individuals correctly identified while the BCM method had a success of 65% with 546 individuals correctly identified (Table 2.5).

Table 2.5 Comparison between the three distance-based methods for measuring the identification success of the *COI* in discriminating against Diptera species.

| **Method** | **Near Neighbor** | | **Best Close Match (3%)** | | | | **BOLD Identification criteria (3%)** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Output | TRUE | FALSE | Correct | Incorrect | Ambiguous | No ID | Correct | Incorrect | Ambiguous | No ID |
| Score n | 577 | 267 | 546 | 55 | 9 | 234 | 526 | 49 | 32 | 234 |
| (%) | (68) | (32) | (65) | (6) | (1) | (28) | (62) | (6) | (4) | (28) |

As the sequence divergence threshold increases there was a decrease in the identification success using the BOLD identification criteria and an increase in the number of "Incorrect", "Ambiguous" and "No ID" results recovered. The opposite was observed with the BCM, the identification success was highest at 5% threshold with 66% of the *COI* sequences correctly assigned to their species names, but, there was also an increase in the number of “Incorrect” assignments with a decrease in the number of “No ID” (Figure 2.5).

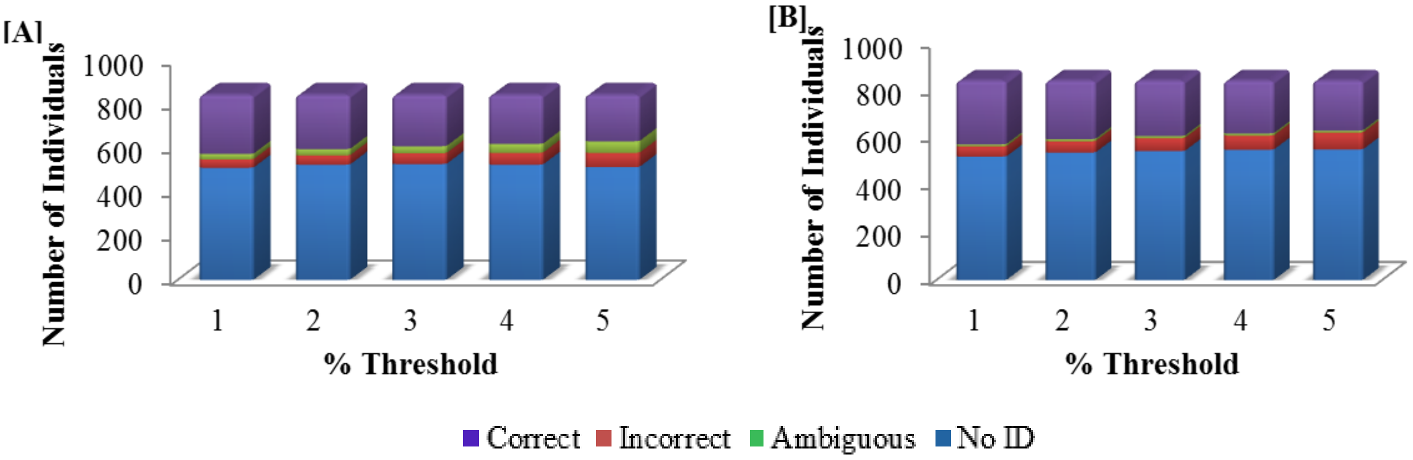


Figure 2.5 Bar-plots of the identification success accuracy measured using the BOLD Identification Criteria [A] and the Best Close Match [B] at a range of thresholds (1 – 5%).

***Tree-based approach: Phylogenetic analysis***

There was consistency between the maximum likelihood and the Bayesian topologies, recovered from analyses of the dataset, including only one representative of each of the 400 BINS. A maximum likelihood tree without species names and additional information (bootstrap values and posterior probabilities), is shown in the figure below (Figure 2.6). The tree shows the relationships between representatives of each of the 400 BINs (annotated tree in Appendix 2). Eighteen distinct Diptera family clusters were recovered; these family clusters are well represented with three or more individuals found within each family (Figure 2.6).

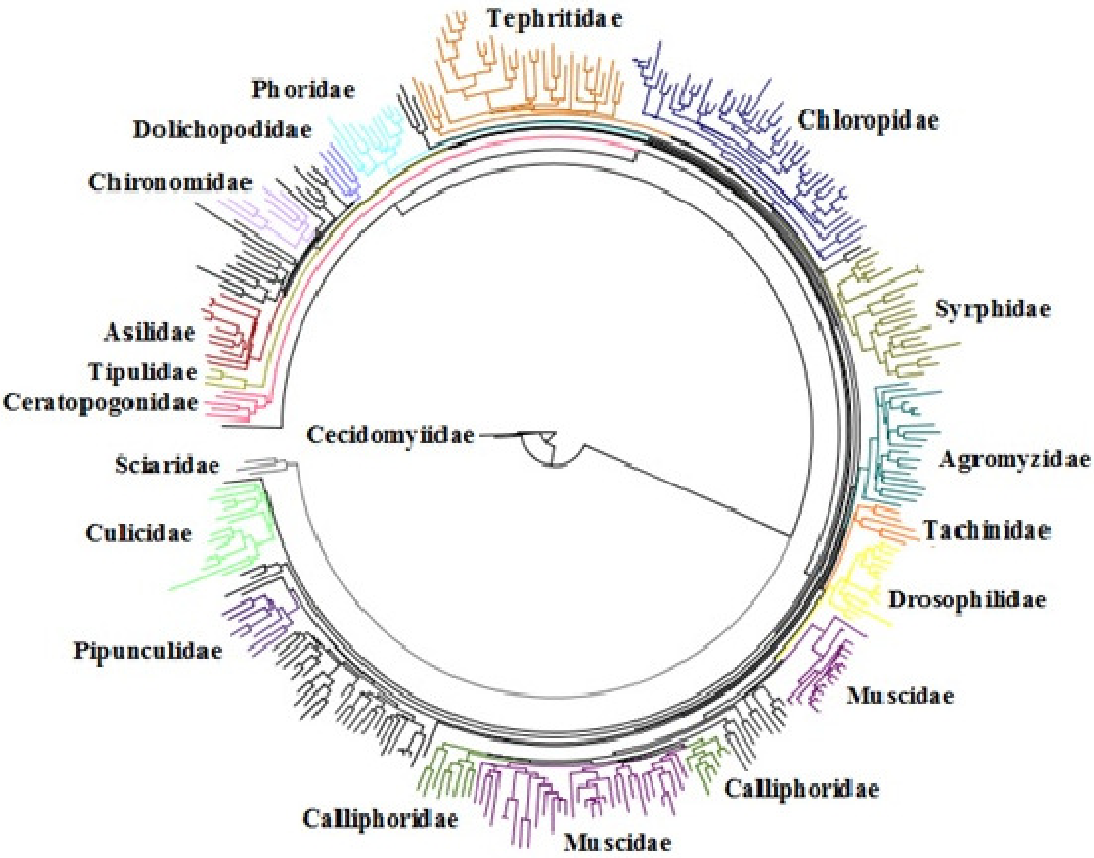


Figure 2.6 Maximum likelihood tree of one representative individual per BIN, only the major families are colour coded on the tree. The branches in black represent Diptera families that had less than three representatives and families that were not monophyletic. The tree was midpoint rooted. For clarification no taxon names and branch support values are provided on this tree. For the full annotated tree see Appendix 2.

Almost all the species-level groups close to or at terminal nodes were generally well supported (Bayes posterior probability: >0.7; ML bootstrap: 70). Within the Muscidae family a monophyletic relationship was observed for the genus *Atherigon*a (Bayes posterior probability: >0.85; ML bootstrap: 68) and *Coenosia* (Bayes posterior probability: >0.75; ML bootstrap: 75). The same was observed for other families such as Chloropidae, Tephritidae, Drosophilidae, Calliphoridae and Syrphidae.

## 2.4 Discussion

### 2.4.1 DNA barcode library for eThekwini Diptera

This study added significantly to the DNA barcode reference library of South African Diptera. Although the sampling for this study was conducted on a small regional scale, the barcode data had a large impact on the data availability on BOLD, with 95% of the DNA barcodes available for SA generated in the present study. Of a total of 400 distinct barcode clusters that were recovered from a NJ tree computed from the 844 *COI* sequences. Using the global BOLD database, only 66 (17%) of the 400 barcode clusters matched records already available on BOLD. These species belonged to 47 genera and 14 families. This means that 83% of the barcode clusters added in the present study are new to BOLD. Currently, the DNA barcode library for the Diptera of eThekwini consists of 844 sequences, which belong to 44 families, 214 genera and 400 BINs.

### 2.4.2 DNA barcoding gap

The success of species identifications based on genetic distances ultimately depended on the absence of overlap between interspecific and intraspecific divergence (Meyer & Paulay 2005; Aliabadian *et al.* 2009). Our results showed a significant overlap between the interspecific and intraspecific sequence divergences when using the K2P model (JM = 0.220; Figure 2.2) for barcode data collected from species within eThekwini. The lack of barcoding gap observed supported the results of a study by Meier *et al.* (2006), which found an extensive overlap between the interspecific and intraspecific sequence divergences in Diptera (15.5%). They reported that 99% of the pairwise distances for congeneric sequences fall into the area of overlap. The absence of the barcoding gap could, first, be due to recent speciation in Diptera - recently diverged species would have accumulated fewer genetic differences, and therefore would have fewer diagnostic characters to separate them (van Velzen *et al.* 2012). Also, recently diverged taxa are likely to have a most recent common ancestor pre-dating the speciation event resulting in overlapping intraspecific and interspecific sequence divergences (Nichols 2001). Second, the overlap between intraspecific and interspecific sequence divergences can also be attributed to uncertainty in species identification using morphology, or the quality of the reference dataset

(Hebert *et al.* 2004b), interspecific hybridization (Chapple *et al.* 2012) and the presence of unrecognized species complexes (Chapple & Ritchie 2013). Repeating the analyses using the best-fit GTR + I + G substitution model, however, resulted in JM values above 1.447 (JM = 1.818) (Figure 2.2), suggesting the presence of the barcoding gap. These results suggested that the K2P model should not be used as a standard model for DNA barcoding as it may produce inaccurate results. Therefore, model selection should be included as an important step in biodiversity studies using barcodes to identify species, and the best-fit substitution model should always be used when estimating genetic distances.

To determine the effect of spatial sampling on the utility of a DNA barcode reference library, a similar barcode analysis was performed using South African Diptera barcode data. Again the results from K2P and GTR + I + G models were compared. The results were similar to those obtained when only the eThekwini data is used. The JM values were above 1.447 for the GTR+ I + G model (JM = 1.821), suggesting the presence of a barcoding gap and below 1.447 for the K2P model (JM = 0.216), suggesting overlap (Figure 2.2). The South African data only had 52 more *COI* sequences compared to the eThekwini data and this could be the reason for the similarities in the results. Therefore, more geographic sampling is required in order to accurately assess the effect of spatial sampling on DNA barcoding.

The analyses also suggest that the presence of a DNA barcoding gap could be taxonomically linked. The Anthomyiidae and Asilidae families showed the presence of a barcoding gap, while the other eight families show an overlap between inter- and intraspecific sequence divergences (Figure 2.3).

### 2.4.3 Distance-based species identification approach

The concept of a barcoding gap is directly linked to the search for an optimal threshold at which to delimit species within DNA barcoding studies (Smith *et al.* 2005; Chapple & Ritchie 2013). The major criticism of DNA barcoding is that the integrity of identifications are compromised by false positives which could overestimate the true number of species, and false negatives which could underestimate the number of true species (Packer *et al.* 2009). However, this can be minimized by using a threshold optimization to find a threshold that has the lowest cumulative error (Meyer & Paulay 2005).

In order to determine the most suitable threshold for discriminating amongst the Diptera species in this study, the threshold optimization method in conjunction with the Best Close Match and the BIC was used to compare the identification success against a series of genetic distance thresholds (1-5%). Our results found the lowest cumulative error at 3% threshold and the highest cumulative error at 5% threshold (Figure 2.4 and Table 2.4). This suggested that 3% is the optimal threshold that can be used for the identification of Diptera species in the eThekwini region. This threshold is comparable to the *COI* sequence divergence considered to be the standard threshold for delimiting insect species (Hebert *et al.* 2003b) and the 3% threshold recommended by Meier *et al.* (2006) as a suitable threshold for identifying species belonging to Diptera.

The performance of the *COI* region as a barcode for the identification of Diptera was also evaluated by assessing the proportion of correct species identifications made using three species identification methods: the near neighbor method and the best close match method, both following Meier *et al.* (2006) and the BIC. The three species identification methods yielded different proportions of correctly matched sequences: NN = 577; BCM = 546 at 3% threshold and BIC = 526 at 3% threshold (Table 2.5 and Figure 2.5). The overall species identification success using all three criterion was (<70%). A low identification success in Diptera was also reported by Meier *et al.* (2006) and their identification success in Diptera also never exceeded 70%. The low identification success in Diptera could be attributed to recent speciation. Recently diverged species share very similar DNA barcode sequences and this prevents accurate identification of these species (van Velzen *et al.* 2012; Deister *et al.* 2014). The low identification success could also be attributed to the quality of the reference database's identifications which would depend on the accuracy of the taxonomy used (Chapple & Ritchie 2013).

### 2.4.4 Tree-based species identification approach: Phylogenetic analysis

The performance of the *COI* region in delimiting species was also tested using the tree-based approach. Accurate species identification and discovery using the tree-based approach require monophyletic species. In our study, maximum likelihood and bayesian inference were used for the tree-based species identification. Both ML and BI recovered consistent tree topologies. We were able to confidently recover 18 families from our phylogenetic tree (Figure 2.6). Some monophyletic relationships were observed within species and genus belonging to the following families: Muscidae, Chloropidae, Tephritidae, Drosophilidae, Calliphoridae and Syrphidae. However, in many families, paraphyletic and polyphyletic relationships were observed (Figure 2.6). The paraphyletic and polyphyletic relationships observed in the phylogenetic tree could be attributed to the recent speciation in Diptera. Recently diverged species have few genetic differences; therefore, there are few characters to discriminate them and this could result in the lack of monophyletic relationships (Nichols 2001; van Velzen *et al.* 2012).

## 2.5 Conclusion

Rapid access to biodiversity data is essential for conservation of biodiversity within the eThekwini municipality area. However, appropriate measures can only be commenced once the biodiversity data is inclusive, accurate and up to date. This study begun the construction of an extensive DNA barcode reference library of the South African Diptera, with 95% DNA barcodes currently available on BOLD generated in this study.

The study provides mixed support for the utility of the *COI* gene as a barcode for identifying South African Diptera species. The analysis using the standard K2P model suggests that the *COI* region has limited performance due to the absence of barcoding gap while the GTR + I + G model was able to detect a barcoding gap. Therefore, the use of an appropriate substitution model is an important consideration in DNA barcoding.

The overlap between the intraspecific and interspecific sequence divergence when using K2P as well as the low identification success observed with all three species identification criterions limits the performance of the *COI* region as the barcode for the identification of our South African Diptera as this could lead to misidentifications and unacceptable errors. This doesn't mean that DNA barcoding is not an effective tool for species identification. However, in order to effectively use DNA barcoding for a rapid identification of South African Diptera, additional DNA regions that can be used solely or in combination with the standard *COI* barcode should be investigated.

## 2.6 References

Aliabadian M., Kaboli M., Nijman V. & Vences M. (2009) Molecular identification of birds: performance of distance based DNA barcoding in three genes to delimit parapatric species. *PLoS ONE* **4**, 1-8.

Bergsten J., Bilton D.T., Fujisawa T., Elliott M., Monaghan M.T., Balke M., Hendrich L., Geijer J., Herrmann J., Foster G.N., Ribera I., Nilsson A.N., Barraclough T.G. & Vogler A.P. (2012) The Effect of geographical scale of sampling on DNA barcoding. *Society of Systematic Biology*, 1-52.

Biesmeijer J.C., Roberts S.P.M., Reemer M., Ohlemuller R., Edwards M., Peeters T., Schaffers A.P., Potts S.G., Kleukers R., Thomas C.D., Settele J. & Kunin W.E. (2006) Parallel declines in pollinators and insect pollinated plants in Britain and the Netherlands. *Science* **313**, 351-4.

Blagoev G.A., Nikolova N.I., Sobel C.N., Hebert P.D.N. & Adamowicz S.J. (2013) Spiders (Araneae) of Churchill, Manitoba: DNA barcodes and morphology reveal high species diversity and new Canadian records. *BMC Ecology* **13**, 1-17.

Brown S.D.J., Collins R.A., Boyer S., Lefort M.C., Malumbres-Olarte J., Vink C.J. & Cruickshank R.H. (2012) SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources* **12**, 562-5.

Buckley T.R. & Cunningham C.W. (2002) The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution* **19**, 394-405.

Chapple D.G., Birkett A., Miller K.A., Daugherty C.H. & Gleeson D.M. (2012) Phylogeography of the endangered Otago Skink, *Oligosoma otagense*: population structure, hybridisation and genetic diversity in captive populations. *PLoS ONE* **7**, 1-11.

Chapple D.G. & Ritchie P.A. (2013) A retrospective approach to testing the DNA barcoding method. *PLoS ONE* **8**, 1-12.

Collins R.A. & Cruickshank R.H. (2012) The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, 1-7.

Crozier R.H. & Crozier Y.C. (1993) The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics* **133**, 97-117.

Dabboor M., Howell S., Shokr M. & Yackel J. (2014) The Jeffries Matusita distance for the case of complex Wishart distribution as a separability criterion for fully polarimetric SAR data. *International Journal of Remote Sensing*, 37-41.

Deister F., Raupach M.J., Hendrich L., Ku S.M. & Gossner M.M. (2014) Building up of a DNA barcode library for true bugs (Insecta: Hemiptera: Heteroptera) of Germany reveals taxonomic uncertainties and surprises. *PLoS ONE* **9**, 1-13.

Drummond A.J. & Rambaut A. (2007) Tracer v. 1.5. Computer program and documentation distributed by the authors at http://beast.bio.ed.ac.uk/Tracer.

Eddy S.R. (1998) Profile hidden markov models. *Bioinformatics Review* **14**, 755-63.

Ekrem T., Willassen E. & Stur E. (2007) A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution* **43**, 530-42.

Endress P.K. (2001) The flowers in extant basal angiosperms and inferences on ancestral flowers. *International Journal of Plant Sciences* **162**, 1111-40.

Felsenstein J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *Department of Genome Sciences,* University of Washington, Seattle;.

Ferguson J.W.H. (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society* **75**, 509-16.

Hajibabaei M., deWaard J.R., Ivanova N.V., Ratnasingham S., Dooh R.T., Kirk S.L., Mackie P.M. & Hebert P.D.N. (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society, Biological sciences* **360**, 1959-67.

Hall T.A. (1999) BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows. *Nucleic Acids Symposium Series* **41**, 95-8.

Haq S.M.A. (2011) Urban green space and integrative approach to sustainable environment. *Journal of Environmental Protection* **2**, 601-8.

Hebert P.D.N., Cywinska A., Ball S.L. & Waard J.R.D. (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society, Biological Sciences* **270**, 313-21.

Hebert P.D.N., Penton E.H., Burns J.M., Janzen D.H. & Hallwachs W. (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *PNAS* **101**, 1-6.

Hebert P.D.N., Ratnasingham S. & deWaard J.R. (2003b) Barcoding animal life: cytochrome c oxidase subunit I divergences among closely related species. *Proceedings of the Royal Society, Biological Sciences* **270**, 1-4.

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S. & Francis C.M. (2004b) Identification of birds through DNA barcodes. *PLoS Biology* **2**, 1-7.

Huelsenbeck J.P. & Ronquist F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5.

Jühling F., Pütz J., Bernt M., Donath A., Middendorf M., Florentz C. & Stadler P.F. (2012) Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Research* **40**, 2833-45.

Kimura M. (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111-20.

Kumar N.P., Srinivasan R. & Jambulingam P. (2012) DNA barcoding for identification of sand flies (Diptera: Psychodidae) in India. *Molecular Ecology Resources* **12**, 414-20.

Lemmon A.R. & Moriarty E.C. (2004) The importance of proper model assuption in bayesian phylogenetics. *Systematic Biology* **53**, 265-77.

Librado P. & Rozas J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-2.

Magnacca K.N. & Brown M.J.F. (2012) DNA barcoding a regional fauna: Irish solitary bees. *Molecular Ecology Resources* **12**, 990-8.

Meier R., Shiyang K., Vaidya G. & Ng P.K.L. (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55**, 715-28.

Menges E.S. (1991) Seed germination percentage increases with population size in a fragmented prairie species. *Conservation Biology* **5**, 158-64.

Meyer C.P. & Paulay G. (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* **3**, 422-5.

Meyer J.R. (2009) Diptera. URL (http://www.cals.ncsu.edu/course/ent425/library/compendium/diptera.html).

Nei M. & Kumar S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nichols R. (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution* **16**, 358-64.

Packer L., Gibbs J., Sheffield C. & Hanner R. (2009) DNA barcoding and the mediocrity of morphology. *Molecular Ecology Resources* **9**, 42-50.

Park D.S., Foottit R., Maw E. & Hebert P.D.N. (2011) Barcoding bugs: DNA-based identification of the true bugs (Insecta: Hemiptera: Heteroptera). *PLoS ONE* **6**, e18749-e.

Peterson N., Stecher G., Nei M. & Kumar S. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731-9.

Picker M., Griffiths C. & Weaving A. (2004) Field guide to insects of South Africa. Struik, Cape Town.

Pili E., Carcangiu L., Oppo M. & Marchi A. (2010) Genetic structure and population dynamics of the biting midges *Culicoides obsoletus* and *Culicoides scoticus*: implications for the transmission and maintenance of bluetongue. *Medical and Veterinary Entomology* **24**, 441-8.

Potts S.G., Biesmiejer J.C., Kemen C., Neumann P., Schweiger O. & Kunin W.E. (2010) Global pollinator declines: trends, impacts and drives. *Trends in Ecology and Evolution* **6**, 345-51.

Rambaut A. (2009) FigTree v. 1.3.1. Computer program and documentation distributed by the author at http://tree.bio.ed.ac.uk/software.

Ratnasingham S. & Hebert P.D.N. (2013) A DNA based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* **8**, 1-16.

Ratnasingham S. & Hebert P.N.D. (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355-64.

Rivera J. & Currie D.C. (2009) Identification of Nearctic black flies using DNA barcodes (Diptera: Simuliidae). *Molecular Ecology Resources* **9** 224-36.

Sahlin K. (2011) Estimating convergence of markov chain monte carlo simulations. In: *Mathematical Statistics*, pp. 16-20. Stockholms Universitet.

Scott-Shaw C.R. (2011) Descriptions of new vegetation types for KwaZulu-Natal., p. 66. Ezemvelo KZN Wildlife Biodiversity Research Division Internal Report., Pietermaritzburg.

Serna E., Gorab E., Ruiz M.F., Goday C., Eirín-López J.M. & Sánchez L. (2004) The gene sex lethal of the Sciaridae family (order Diptera, suborder Nematocera) and its phylogeny in dipteran insects. *Genetics* **168**, 907-21.

Silvestro D. & Michalak I. (2011) raxmlGUI: a graphical front end for RAxML. *Organisms Diversity and Evolution* **12**, 335-7.

Skevington J.H. & Dang P.T. (2002) Exploring the diversity of flies (Diptera). *Biodiversity* **3**, 3-27.

Smith M.A., Fisher B.L. & Hebert P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society Biological Sciences* **360**, 1825-34.

Ssymank A., Kearns C.A., Pape T. & Thompson F. (2008) Pollinating flies (Diptera): a major contribution to plant diversity and agricultural production. *Conservation Biology* **9**, 1-2.

Szymura J.M., Hewitt G.M., Lunt D.H. & Zhang D.X. (1996) The insect cytochrome c oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. *Insect Molecular Biology* **5**, 153-65.

Thompson J.D., Higgins D.G. & Gibson T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-80.

Tuzin B.L., van Leeuwen E., Rodenburg C. & Nijkamp P. (2002) Development and management of green spaces in European cities. In: *The pulsar effect planning with peaks* (ed. by Economics DoS), Amsterdam.

van Velzen R., Weitschek E., Felici G. & Bakker F.T. (2012) DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE* **7**, e30490-e.

Versteirt V., Nagy Z.T., Roelants P., Denis L., Breman F.C., Damiens D., Dekoninck W., Backeljau T., Coosemans M. & Van Bortel W. (2014) Identification of Belgian mosquito species (Diptera: Culicidae) by DNA barcoding. *Molecular Ecology Resources*.

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

Wiegmann B.M., Trautwein M.D., Winkler I.S., Barr N.B., Kim J.W., Blagoderov V., Caravas J., Narayanan S., Schmidt-ott U., Kampmeier G.E., Meier R. & Yeates D.K. (2011) Episodic radiations in the fly tree of life. *PNAS* **108**, 5690-5.

Zwickl D.J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. *Ph.D. Dissertation,* The University of Texas at Austin.

# CHAPTER THREE

# DIPTERA MITOCHONDRIAL GENOME RECONSTRUCTION USING NEXT GENERATION SEQUENCING

## Abstract

Next generation sequencing (NGS) allows for the sequencing of genomic scale data. This analysis is quicker, cheaper and does not require prior knowledge of regions compared to the traditional Sanger sequencing. This is particularly important in non-model organisms where sequence data is often limited or absent. In this study, the complete mitochondrial genome (MtDNA) of a blowfly species (*Lucilia cuprina:* Calliphoridae) was reconstructed directly from NGS data. The circular genome is 15 079 bp long, with an AT content of 75.8%. The gene order and orientation are identical to that of *Drosophila yakuba*. The genome encodes for 13 protein-coding genes (PCGs), two ribosomal RNA genes (rRNA), 22 transfer RNA genes (tRNA) and contains a control region. All protein-coding genes use standard mitochondrial initiation codons (methionine and isoleucine), and the usual TAA and TAG termination codon. All 22 tRNA genes show a typical clover-leaf structure, except for *tRNASer (AGN*) which forms a simple loop on the dihydrouridine (DHU) arm.

**Key Words:** Next generation sequencing, mitochondrial genome, *Lucilia cuprina*, Calliphoridae

## 3.1 Introduction

Mitochondrial (MtDNA) regions are usually the markers of choice in studies that include species level analysis such as population genetic, phylogeographic and phylogenetic studies (Lyra *et al.* 2005; Beckenbach 2011). The strict maternal transmission, high mutation rate, and the simple genetic structure of the MtDNA are some of the features that make it a valuable source of genotypic characters (Avise *et al.* 1987; Moore 1995; Singh 2015). Furthermore, the MtDNA has high copy numbers, lacks introns, lacks recombination and contains conserved regions - all features allow for the development of universal primers (Avise *et al.* 1987; Singh 2015).

Animal MtDNA are small, circular molecules with genomic sizes ranging between 14 – 19 Kb (Boore 1999). They usually contain a set of 37 genes, which are divided into 13 protein-coding genes (PCGs), 22 transfer RNA (tRNA) genes and two ribosomal RNA (rRNA) genes. Non-coding control elements are also found in the MtDNA, which play a role in regulation of transcription and translation (Boore 1999; Taanman 1999). The major non-coding element in insect MtDNA is the control region also known as AT region in insects. This region is AT-rich and is located between the conserved small rRNA, and the tRNA that carries anticodon for isoleucine (*tRNA-Ile*). The other non-coding elements are small intergenic spacers located between the genes.

The mitochondrial cytochrome c oxidase subunit I (*COI*) gene is a protein-coding gene, extensively used as the standard DNA barcode marker in animals. The *COI* gene has also been successfully used in DNA barcoding studies of a number of insect taxa (Virgilio *et al.* 2010; Park *et al.* 2011; Zhou *et al.* 2011; Tembe *et al.* 2014; Versteirt *et al.* 2014). Despite this success, the *COI* gene has had a relatively low discrimination rate in identifying Diptera species, owing to an overlap between intraspecific and interspecific sequence divergences (Meier *et al.* 2006; Virgilio *et al.* 2010) and as described in Chapter Two of this work. This low species differentiation ability of the *COI* gene limits its use in Diptera and supports the need to search for alternative DNA barcodes – not only to avoid an exclusive reliance on the *COI* gene, but also to allow multiple DNA barcodes to be used for better identification. Unfortunately there are limited genomic resources available for non-model organisms to evaluate possible regions for new DNA barcodes.

Until recently, sequencing mitochondrial genomes has been challenging. However, the introduction of high-throughput Next Generation Sequencing (NGS) technologies have revolutionized molecular biology, allowing researchers to rapidly generate vast amount of sequence data economically (Bybee *et al.* 2011). The ability of NGS to process thousands to millions - or even billions - of sequence reads at low cost, sets it apart from the conventional single fragment capillary-based Sanger sequencing systems (Mardis 2008). Currently, there are two major NGS technologies commercially available, these are SOLiD and Illumina technologies. The Illumina platform is the most successful and widely adopted NGS platform worldwide. It was first introduced in 2007 and prior to Illumina, this technology was known as

Solex. Illumina enables economical, scalable, high throughput NGS application and supports both single and paired-end reads (Illumina 2014). Amongst the four Illumina sequencer versions commercially available, the preferred version for more focused applications such as targeted gene sequencing, metagenomics, small genome sequencing, targeted gene expression, amplicon sequencing is the MiSeq platform. This platform produces 2x 300 paired-end reads in a single run (Illumina 2014).

The aim of this study is to sequence and reconstruct the complete mitochondrial genome of a blowfly, *Lucilia cuprina,* using NGS technology from the Illumina platform.

## 3.2 Materials and methods

### 3.2.1 Specimen selection and DNA extraction

Three specimens (PalGr3Dipt0701, IphiGrDipt06 and NewGGr2Dipt0401) of *Lucilia cuprina* (family Calliphoridae) were included in the present study. These specimens were selected randomly from the BOLD database based on large body size and frequency of sampling.. Total genomic DNA was extracted from all three fly specimens using the Nucleospin®Tissue Kit (Macherey-Nagel), following the animal tissue protocol. To maximize the amount of DNA extracted from individuals, the abdomen and thorax of the fly specimens were used. The DNA concentration for the specimens was measured using a Nanodrop 2000 spectrophotometer (Thermo Scientific), and the specimen (PalGr3Dipt0701) which had the highest DNA concentration was selected for next generation sequencing (Table 3.1).

Table 3.1 DNA concentrations measured using both the Nanodrop 2000 spectrophotometer (Thermo Scientific), and the Qubit fluorometer (Thermo Fisher). The fly specimen highlighted in green was the specimen used for NGS since it had the highest DNA concentration. The asterisk (*) is the DNA concentration of PalGr3Dipt0701, measured using the Qubit.

| | **DNA Concentrations (ng/µl)** | |
|---|---|---|
| **Fly Specimen** | **Nonodrop** | **Qubit** |
| **PalGr3Dipt0701** | **72.5** | **21.6**[*] |
| **IphiGr1Dipt06** | 28.3 | - |
| **NewGr2Dipt0401** | 71.1 | - |

### 3.2.2 Nextera DNA sample preparation

The correct quantification of DNA concentration is essential for Nextera (Illumina) NGS sample preparation. The DNA concentration of the selected specimen was therefore also measured using a Qubit fluorometer (Thermo Fisher) to ensure correct quantification of the input DNA. Genomic DNA of the fly specimen PalGr3Dipt0701 was prepared according to the Nextera Sample Preparation protocol (Illumina). The Nextera DNA Sample Preparation Kit, enables the generation of sequencing-ready libraries in less than 90 minutes, with less than 15 minutes of hands-on time (Illumina 2012).

Briefly, 25 ng of the fly genomic DNA was fragmented and tagged (tagmented) using transposomes. The Nextera XT transposome simultaneously fragments the input DNA and adds two adapter sequences – a unique one to each end. This was followed by a PCR amplification step in which the tagmented-added adapter sequences were used as priming sites for indexing primers through a limited-cycle PCR program. This step adds two indexes, index 1 (i7) and index 2 (i5), onto each end of the fragments, as well as the adapters required for clustering in the flow cell. The PCR products were cleaned using AMPure XP beads, thus removing excess adapters, primers, enzymes, etc. from the library DNA. Additionally, during this cleaning process, very short library fragments are removed from library, leaving fragments longer than

300 bp. The library was sequenced using the Illumina MiSeq sequencer at the Agricultural Research Council (ARC) Biotechnology Platform using Illumina reagents and in accordance with their recommendations.

### 3.2.3 Analysis of the next generation sequencing data

The CLCBio Genomics Workbench (Qiagen), was used to analyze and visualize the NGS data. First, sequence reads with low quality and ambiguous nucleotides were trimmed from the data set using default parameters. Second, overlapping paired-end sequence reads were merged to form one single read where reads overlapped. The cleaned, merged and un-merged sequence reads were then used for all downstream analyses.

### 3.2.4 Mapping of sequence reads to a reference sequence

Mapping is the assemblage of the sequence reads against an existing backbone sequence, or reference sequence. This involves building a consensus sequence that is similar but not necessarily identical to that of the reference sequence. The genomic data available for *Lucilia cuprina* was used as the reference sequence in the present study. The complete mitochondrial genome of *Lucilia cuprina* was downloaded from GenBank (Accession number: NC_019573). The sequence reads generated in the present study were then mapped to the complete mitochondrial reference genome using the CLCBio Genomics Workbench. A BLASTn search was then performed on the resulting consensus sequence to confirm that it was indeed a *Lucilia cuprina* MtDNA.

### 3.2.5 *De novo* assembly of the sequence reads

No laboratory mitochondrial enrichment procedure was followed due to limited *Lucilia cuprina* sample DNA. Therefore, an *in silico* separation approach in which total DNA is sequenced and used within the assembly process was used. Multiple contigs are expected after a *de novo* assembly that would provide coverage for both nuclear and MtDNA. However, higher copy numbers of MtDNA to nuclear genomes exists in cells, and we've therefore argued that NGS of the whole DNA extraction would provide more coverage of the MtDNA than the larger insect genome. This approach was used in this study. Sequence reads were *de novo* assembled, thereby producing a full-length sequence of the MtDNA. The "*de novo* assembly" module in the CLCBio Genomics Workbench was used for this. The parameters were set to an automatic bubble size of 250 with a word size of 64, a minimum contig length of 1 000 with, scaffolding automatically performed and the "autos detect paired distances" setting selected.

### 3.2.6 Annotation of the MtDNA

The 15 079 bp long sequence (assembled as "*Contig 13*", see results) was produced during the *de novo* assembly step and identified as MtDNA based on its size and the BLASTn results. This contig was annotated using a free web server, the MITOS webserver (Bernt *et al.* 2013). The annotated MtDNA in a GFF format was imported and visualized on the CLCBio Genomics Workbench. The gene order and orientation of the mitochondrial genome, was examined using a complete MtDNA of *L. cuprina* strain DI213.5 (Genbank accession number: JX913753) and a fruit fly *Drosophila yakuba* (Genbank accession number: NC_001322*)* family Drosophilidae. *Drosophila yakuba* was the first invertebrate MtDNA to be completely sequenced and is therefore, used as the primary model organism for mitogenomic research in Diptera (Clary & Wolstenholme 1985). Protein-coding genes were identified using MITOS, and by aligning the annotated MtDNA structure, with that of the *L. cuprina* strains DI213.5 on the CLCBio Genomics Workbench. The location of the start and termination codons of protein-coding genes was determined by examining the nucleotide sequences of each gene in the annotated MtDNA using DOGMA.

### 3.2.7 Nucleotide composition, strand asymmetry and codon usage

Nucleotide composition for the whole *L. cuprina* MtDNA, including protein-coding regions, *tRNA* and *rRNA* genes and the control region were calculated using MEGA5 (Tumura *et al.* 2011). Strand asymmetry was calculated using the formulas: AT skew = [A-T]/ [A+T] and GC skew= [G-C]/ [G+C]. The codon usage for protein-coding genes, was also calculated in MEGA5 and in the sequence analysis module of the online Sequence Manipulation Suite (Stothard 2000).

### 3.2.8 Secondary structure prediction for tRNA genes

The tRNA genes were identified by uploading the entire mitochondrial sequence onto the tRNAscan-SE search server v.1.21 (Schattner *et al.* 2005). The settings used were as follows: search mode = tRNAscan only, source = mito/chloroplast, and genetic code for tRNA isotype prediction = invertebrate mito. The *tRNA* genes (*tRNA-Arg* and *tRNA-Ser*(AGN) that could not be identified in tRNAscan-SE were identified using ARWEM v1.2 (Laslett & Canbäck 2008) with the default settings. The same was done for the mitochondrial sequences of *D. yakuba* and *L. cuprina* strain DI213.5, in order to compare the tRNA secondary structures.

### 3.2.9 Control region

The control region is responsible for the regulation of transcription and control of DNA replication. In insects, the control region is generally rich in adenine and thymine nucleotides, with more than 85% of the region being AT-rich (Zhang & Hewitt 1997; Bruhn 2011). The size of the control region in insects varies in different taxa and even within the same species. The location of the control region within the mitochondrial genome also varies as a result of tRNA transposition. In Diptera, the control region is generally located between the conserved small rRNA (*12S*), and *tRNA-Ile* (Zhang & Hewitt 1997). The control region was not annotated on the sequenced mitochondrial genome (*Contig 13*) when using various bioinformatics softwares already mentioned, therefore, annotation was done manually using the information mentioned above and by aligning the nucleotides in the region between the small rRNA (*12S*) and the *tRNA-Ile* with the control regions of four strains of *L. cuprina strain* DI 213.2 – DI213.5 (Genbank accession numbers: JX913750; JX913751; JX913752 and JX913753 (Nelson, et al., 2012) using BioEdit v5.0.9 (Hall 1999). Additionally, the nucleotide composition of this region was calculated using MEGA 5.

## 3.3 Results

### 3.3.1 Sequencing, quality check and merging of overlapping sequence reads

The fly specimen PalGr3Dipt0701 was successfully sequenced with the MiSeq NGS platform and a total of 3 636 306 paired-end sequence reads were obtained. Of these reads, 864 581 were trimmed according to our quality and adapter parameter settings and 105 sequence reads were trimmed based on the ambiguous nucleotides setting (Table 3.2). Only a small percentage (1.55%) of the 56 544 reads overlapped and were merged. These longer reads, together with the remaining unmerged paired-end reads were used in the assembly and mapping processes.

Table 3.2 Number of paired-end sequence reads trimmed based on the quality parameters (QC30 and adapter sequences). Removal of low quality sequences (limit =0.05) and removal of ambiguous nucleotides (maximum of 2 nucleotides in sequence allowed). About 24% of the sequence reads were trimmed based on quality and adapter sequences present, while 0.003% of the sequence reads were trimmed based on ambiguous nucleotides present.

| Trim | Input reads | Not trimmed | Trimmed |
|---|---|---|---|
| **Trim on quality** | 3 633 445 | 2 768 864 | 864 581 |
| **Ambiguity trim** | 3 637 992 | 3 637 887 | 105 |

### 3.3.2 Mapping Sequence Reads to a Reference Sequence

Only 12 548 of the sequence reads were successfully mapped to the MtDNA of *L. cuprina*. The low number of reads mapped to the reference sequences was not unexpected, since a total DNA extraction was used for the sample preparation with most of the sequence data generated representing the nuclear genome and since we've only used the mitochondrial genome as a reference, only a small percentage of the reads sequences, mapped. The top BLASTn hit of the consensus mitochondrial sequence matched a complete mitochondrial genome of *L. cuprina* species strain DI213.4 (89% sequence identity and E-value of 0.0).

### 3.3.3 *De novo* assembly of the sequence reads

A total of 872 contigs and scaffolds were obtained from the *de novo* assembly. The longest contig retrieved was 15 079 bp long (*Contig 13*) and was consistent with the expected MtDNA sizes of insects. A BLASTn search of the contig matched the complete mitochondrial genome of *L. cuprina* strain DI213.5 (Genbank accession number: JX913756.1; E-value: 0.0; 19654 bits) and confirmed the contig as being mitochondrial in nature. Based on its size, *Contig 13* is a complete mitochondrial genome of the sequenced fly specimen (PalGr3Dipt0701).

The high number of contigs and scaffolds and their small average sizes (Table 3.3) indicates that the nuclear genome was only sequenced at a very low coverage, thereby preventing longer assemblies being formed.

Table 3.3 Contig measurements generated in CLCBio Genomics Workbench, after the *de novo* assembly of all the trimmed, merged and unmerged paired-end reads. The largest contig *i.e.* the mitochondrial genome is indicated in italics.

| | **Including scaffolding** | **Excluding scaffolding** |
|---|---|---|
| Number of Contigs | 872 | 885 |
| N50 | 1 459 bp | 1 452 bp |
| Minimum | 996 bp | 37 bp |
| *Maximum* | *15 079 bp* | *15 079 bp* |
| Average | 1 649 bp | 1 625 bp |
| Total | 1 438 235 bp | 1 438 152 bp |

### 3.3.4 Features of assembled MtDNA (*Contig 13*)

The *de novo* assembled *Contig 13* forms a circular molecule with a length of 15 079 bp, this observed length is well within the observed range (14 – 19 Kb) of an animal MtDNA. The contig also contains the typical 37 genes (13 protein-coding genes, 22 tRNA genes and two rRNA genes) found in all MtDNAs of animal species (Figure 3.1). However, it is smaller compared to the published MtDNAs of *L. cuprina* strains (Table 3.4). These genomes range from 15 226 bp (*L. cuprina* strain *DI213.5*), to 15952 bp (*L. cuprina* strain *DI190.1).* Gene overlaps at 20 gene junctions were identified; these overlaps involve a total of 531 bp. The longest overlap is 144 bp, which is between *COI* and *COII* genes.

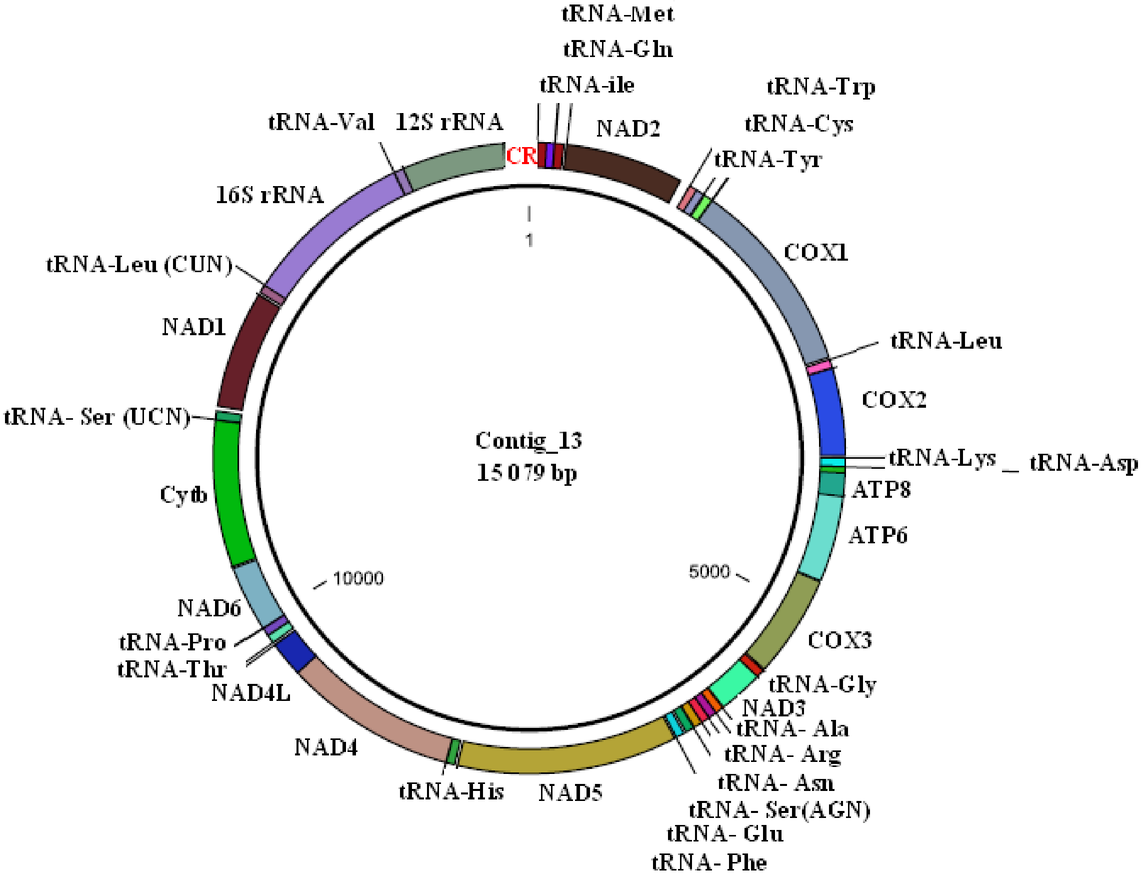


Figure 3.1 Map of the mitochondrial genome of *L. cuprina* obtained from the NGS assembly's *Contig 13*. All 13 protein-coding, 2 ribosomal and 22 transfer RNA genes are present. The control region is located between *12S* rRNA and *tRNA-ile*.

### 3.3.5 Gene order, nucleotide composition and asymmetry of MtDNA (*Contig 13*)

Mitochondrial gene order is generally conserved in most closely related taxa. The gene arrangement of the protein-coding, tRNA and rRNA genes in *Contig 13* is similar to that of *L. cuprina* strain DI213.5 (Figure 3.2) and is also similar to that seen in *D. yakuba*. As expected, the nucleotide composition of the *Contig 13* sequence is biased towards adenine and thymine , with adenine being the most favoured nucleotide (39.1%) and guanine the least favoured (9.8%), this is in accordance with the MtDNAs of the four strains of *L. cuprina* and that of *D. yakuba* (Clary & Wolstenholme 1985). The AT content is 75.8% (39.1% A, 36.7% T), while the GC content is 24.2% (9.8% G and 14.4% C).

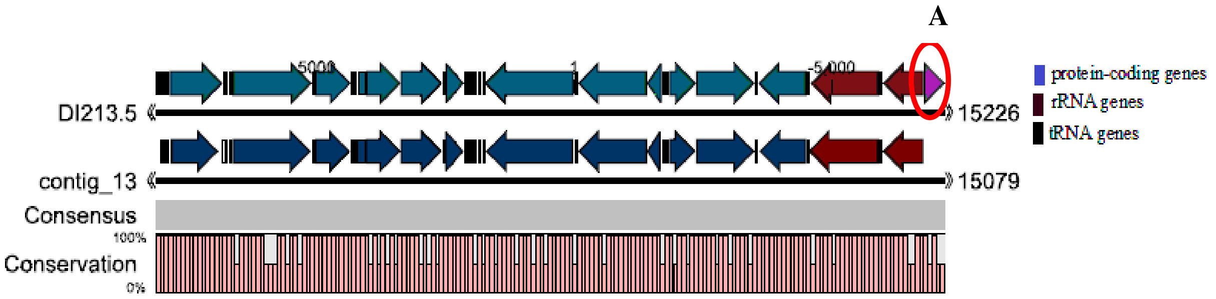


Figure 3.2 Alignment of the MtDNA of *Contig 13* and *L. cuprina* strain D213.5 to compare the order and orientation of the genes. The circled segment [A], is the control region which was not annotated in the MtDNA, but present. The arrows on the genes indicate the direction of transcription. As seen in the figure above, the gene order and orientation of the MtDNA is similar to that of *L. cuprina* strain D213.5.

The strand asymmetry is reflected by the skewness which is calculated as, (A-T)/(A+T) for AT skew and (G-C)/(G+C) for GC skew. AT-skews and GC-skews were calculated for the new MtDNA and the four strains of *L. cuprina* species (Table 3.4). The GC-skew is suggested to be the best indicator of strand asymmetry. As seen in Table 3.4 the four strains of the *L. cuprina* species show obvious strand asymmetry (GC-skew between -0.165 and -0.170). The GC-skew of the *Contig 13* MtDNA is -0.190. In all the four strains of *L. cuprina* and sequenced MtDNA, the GC-skew is negative due to the significantly low G content observed in the MtDNAs.

Table 3.4 Comparison of *Contig 13* MtDNA and four published *L. cuprina* strains (DI213.2-5).

| **Species** | **Whole genome** | | | | **Protein-coding genes** | |
|---|---|---|---|---|---|---|
| | **Size (bp)** | **AT%** | **AT-skew** | **GC-skew** | **No. of codons** | **AT%** |
| Contig 13 | 15079 | 75.8 | 0.031 | -0.190 | 3636 | 74.4 |
| *Lucilia cuprina DI213.5* | 15226 | 77.0 | 0.013 | -0.165 | 3726 | 75.8 |
| *Lucilia cuprina DI213.4* | 15268 | 77.1 | 0.014 | -0.170 | 3726 | 75.9 |
| *Lucilia cuprina DI213.3* | 15289 | 77.1 | 0.014 | -0.170 | 3726 | 75.9 |
| *Lucilia cuprina DI213.2* | 15310 | 77.1 | 0.014 | -0.170 | 3726 | 75.9 |

### 3.3.6 Protein-coding genes

The MtDNA of the sequenced fly contains all 13 protein-coding genes that are normally found in metazoan mitochondrial genomes. The location of the start and termination codons of these protein-coding genes is shown in Figure 3.3.

Nine of the 13 genes are found on the positive strand, while the other four are found on the negative strand (Table 3.5). The 13 PCGs have either the methionine (ATG or ATT), or isoleucine (ATT or ATC) codon as start signal. Nine of these genes *(NAD2*, *COI*, *ATP8*, *NAD3*, *NAD4*, *NAD4L*, *NAD5*, *NAD6* and *NAD1*) have ATT as their start codon, while the other four (*COII*, *COIII*, *ATP6* and *Cytb*), start with ATG (Table 3.5). Twelve protein-coding genes have complete termination codons (nine have TAA and three have TAG as their stop codon), while one of the genes (*NAD4*) has an incomplete termination codon T (Table 3.5).

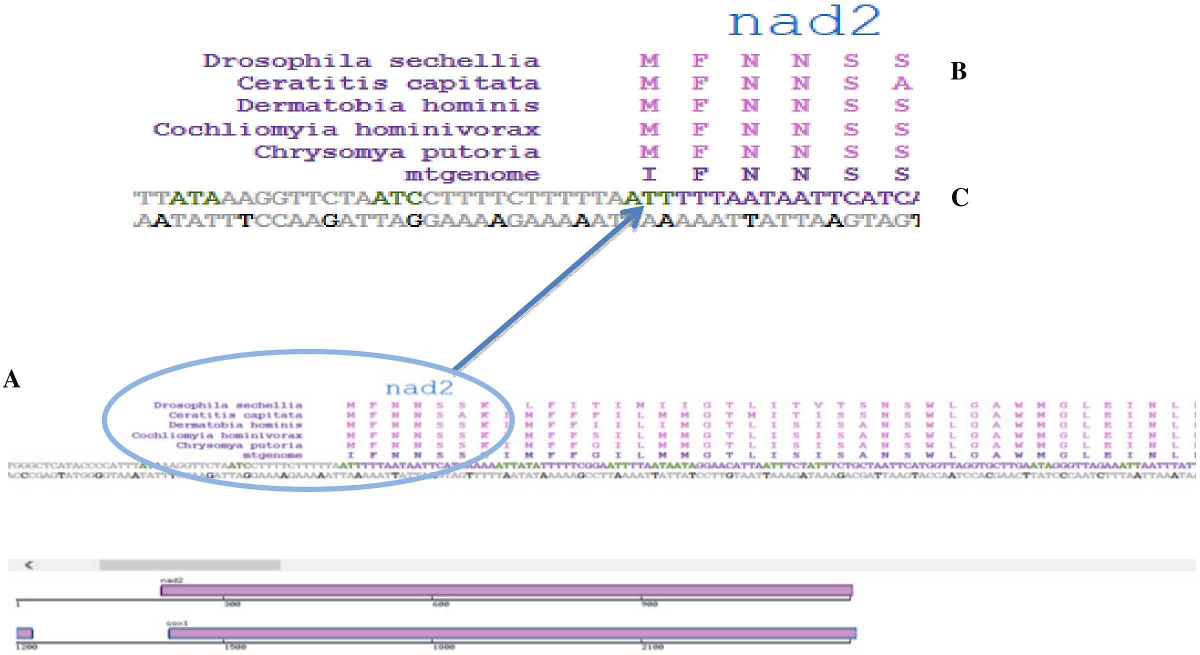

nad2
Drosophila sechellia M F N N S S
Ceratitis capitata M F N N S A
Dermatobia hominis M F N N S S
Cochliomyia hominivorax M F N N S S
Chrysomya putoria M F N N S S
mtgenome I F N N S S
B
TTATAAAGGTTCTAATCCTTTTCTTTTTAATTTTTAATAATTCATCA
AATATTTCCAAGATTAGGAAAAGAAAAATTAAAAATTATTAAGTAGT
C
A
nad2
cox1

Figure 3.3 Illustration of how the start and stop codons of all the protein-coding genes were identified in DOGMA. [A] is the actual illustration of what is seen in DOGMA, [B] is the enlarged circled segment of [A], to show clearly the start codon of *NAD2* and [C] is the *NAD2* sequence of MtDNA. As seen in the diagram, the start codon of *NAD2* is ATT (green colour). The species name in [B] shows the *NAD2* protein sequences that matched the *NAD2* sequence of the MtDNA.

Table 3.5 Summary of all the genes found on the newly sequenced *Contig 13* MtDNA and how they are arranged. * Incomplete stop codon of *NAD4*. The "+" is the positive strand and "-" is the negative strand of the MtDNA.

| Gene | Strand | Span (bp) | Size (bp) | Anticodon | Start | Stop |
|---|---|---|---|---|---|---|
| tRNA-ile | + | 63-129 | 66 | GAT (31-33) | | |
| tRNA-Gln | - | 126-195 | 69 | TTG(165-163) | | |
| tRNA-Met | + | 194-263 | 69 | CAT(225-227) | | |
| NAD2 | + | 208-1221 | 912 | | ATT | TAA |

| | | | | | | |
|---|---|---|---|---|---|---|
| tRNA-Trp | + | 1278-1346 | 68 | TCA (1309-1311) | | |
| tRNA-Cys | - | 1338-1401 | 63 | GCA (1372-1370) | | |
| tRNA-Tyr | - | 1410-1476 | 66 | GTA (1445-1443) | | |
| COX1 | + | 1415-2950 | 1509 | | ATT | TAA |
| tRNA-Leu(UUR) | + | 3008-3074 | 66 | TAA(3038-3040) | | |
| COX2 | + | 3021-3707 | 672 | | ATG | TAA |
| tRNA-Lys | + | 3766-3837 | 71 | CTT(3797-3799) | | |
| tRNA-Asp | + | 3836-3903 | 67 | GTC(3868-3870) | | |
| ATP8 | + | 3846-4007 | 180 | | ATT | TAA |
| ATP6 | + | 4004-4678 | 675 | | ATG | TAA |
| COX3 | + | 4681-5466 | 792 | | ATG | TAA |
| tRNA-Gly | + | 5541-5606 | 65 | TCC(5572-5574) | | |
| NAD3 | + | 5539-5892 | 351 | | ATT | TAA |
| tRNA-Ala | + | 5961-6026 | 65 | TGC(5991-5993) | | |
| tRNA-Arg | + | 6026-6088 | 63 | TCG | | |
| tRNA-Asn | + | 6089-6154 | 65 | GTT(6120-6122) | | |
| tRNA-Ser(AGN) | + | 6155-6222 | 68 | GCT | | |
| tRNA-Glu | + | 6228-6295 | 67 | TTC(6259-6261) | | |
| tRNA-Phe | - | 6313-6379 | 66 | GAA(6347-6345) | | |
| NAD5 | - | 6320-8035 | 1695 | | ATT | TAG |
| tRNA-His | - | 8114-8179 | 65 | GTG(8149-8147) | | |
| NAD4 | - | 8117-9454 | 1335 | | ATT | T* |
| NAD4L | - | 9451-9744 | 273 | | ATT | TAG |
| tRNA-Thr | + | 9810-9875 | 65 | TGT(9841-9843) | | |
| tRNA-Pro | - | 9875-9941 | 66 | TGG(9911-9909) | | |
| NAD6 | + | 9880-10401 | 510 | | ATT | TAA |
| Cytb | + | 10404-11537 | 1131 | | ATG | TAG |
| tRNA-Ser(UCN) | + | 11602-11670 | 68 | TGA(11633-11635) | | |
| NAD1 | - | 11626-12561 | 906 | | ATT | TAG |
| tRNA-Leu(CUN) | - | 12635-12700 | 65 | TAG(12671-12669) | | |
| 16S rRNA | - | 12639-13962 | 1361 | | | |
| tRNA13-Val | + | 14023-14095 | 72 | TAC(14062-14060) | | |
| 12S rRNA | - | 14035-14819 | 787 | | | |
| Control region | | 14820-15079 | 259 | | | |

### 3.3.7 Codon usage

The total amino acids (aa) used in the predicted proteins of *Contig 13* is 3636 aa (Table 3.6). The pattern of codon usage of *Contig 13* was studied (Table 3.6). The most frequently used amino acids are: Leu (12.54%), Trp (11.63%), Ile (11.47%), Lys (11.22%) and Asn (11.14%) and the least used amino acids are Arg (1.46%) and Cys (2.03%). *Contig 13* employs TTA (leucine) 367 times for protein synthesis (Table 3.6), making this codon the most frequently used.

Table 3.6 Codon usage of each amino acid in *Contig 13* MtDNA protein synthesis. The percentages for each amino acid are the percentages of the amino acids found among all the predicted amino acids (3636 aa). This analysis also includes stop codons.

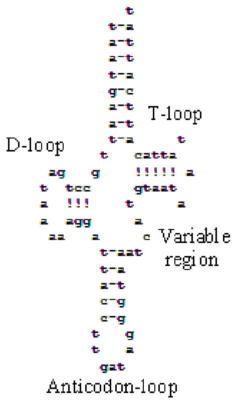
| **Amino Acid/Percentage (%)** | **Codon** | **Occurrence** | **Amino Acid/Percentage (%)** | **Codon** | **Occurrence** |
|---|---|---|---|---|---|
| Ala/2.83 | GCG | 1 | Pro/4.89 | CCG | 4 |
| | GCA | 40 | | CCA | 65 |
| | GCT | 49 | | CCT | 76 |
| | GCC | 13 | | CCC | 33 |
| Cys/2.03 | TGT | 49 | Gln/3.60 | CAG | 25 |
| | TGC | 25 | | CAA | 106 |
| Asp/2.42 | GAT | 63 | Arg/1.46 | CGG | 6 |
| | GAC | 25 | | CGA | 31 |
| Glu/3.16 | GAG | 14 | | CGT | 13 |
| | GAA | 101 | | CGC | 3 |
| Phe/10.67 | TTT | 282 | Ser/5.47 | AGG | 47 |
| | TTC | 106 | | AGA | 50 |
| Gly/2.58 | GGG | 7 | | AGT | 49 |
| | GGA | 53 | | AGC | 53 |
| | GGT | 29 | Ser/6.52 | TCG | 20 |
| | GGC | 5 | | TCA | 100 |
| Hist/3.41 | CAT | 95 | | TCT | 86 |
| | CAC | 29 | | TCC | 31 |
| Ile/11.47 | ATT | 345 | Thr/6.16 | ACG | 11 |
| | ATC | 72 | | ACA | 86 |
| Lys/11.22 | AAG | 67 | | ACT | 90 |
| | AAA | 341 | | ACC | 37 |

| | | | | | |
|---|---|---|---|---|---|
| Leu/12.54 | TTG | 89 | Val/4.12 | GTG | 13 |
| | TTA | 367 | | GTA | 71 |
| Leu/7.01 | CTG | 28 | | GTT | 51 |
| | CTA | 86 | | GTC | 15 |
| | CTT | 111 | Trp/11.63 | TGG | 34 |
| | CTC | 30 | | TGA | 75 |
| Met/10.56 | ATG | 55 | | TAT | 224 |
| | ATA | 329 | | TAC | 90 |
| Asn/11.17 | AAT | 329 | End/8.80 | TAG | 64 |
| | AAC | 77 | | TAA | 256 |

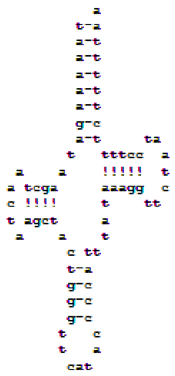
### 3.3.8 Transfer RNAs and ribosomal RNAs

The MtDNA (*Contig 13*) has a complete set of 22 tRNA genes (Figure 3.1). The 20 tRNA genes were identified by the tRNAscan-SE software, and the other two were identified by the ARWEN software. The tRNA genes vary in length from 63 – 72 bp, which is within the observed range for other insects (Laiho & Ståhls 2013). All the tRNA genes form a typical clover-leaf structure except for the tRNA-Ser (AGN), in which the dihydrouracil arm forms a simple loop (Figure 3.5).

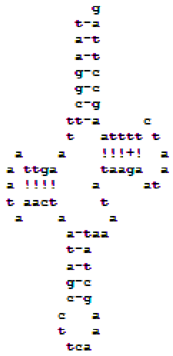
All the tRNA secondary structures and their anticodons are identical to that of *D. yakuba* and *L. cuprina* strain DI213.5. The predicted secondary structures of the 22 tRNA genes are shown in Figure 3.5. The A + T content of all the tRNA is 76.7% which is slightly less than that of the four published *L. cuprina* strains (Table 3.4). Sixteen of the tRNA genes are found on the positive strand, and the other seven are on the negative strand. The order and the orientation of the tRNA genes of this MtDNA are identical to that of *D. yakuba* and *L. cuprina* strain DI213.5. Moreover, the secondary structures of the tRNA genes have a similar leaf-like structure suggesting similar function.



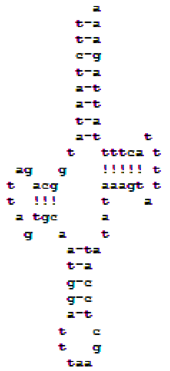
Isoleucine (Ile)



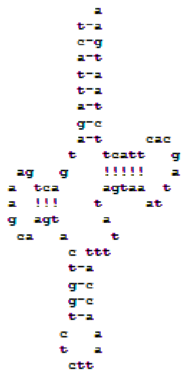
Methionine (met)



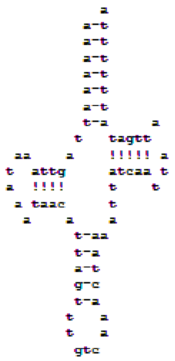
Tryptophan (Trp)



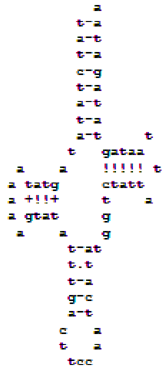
Leucine (Leu- UUR)



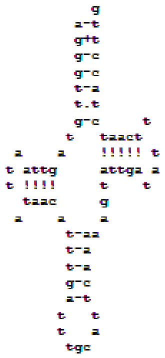
Lysine (Lys)



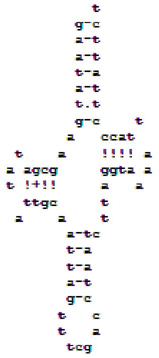
Aspartate (Asp)



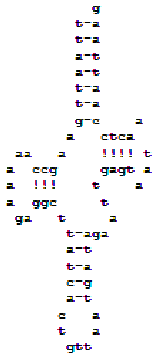
Glycine (Gly)



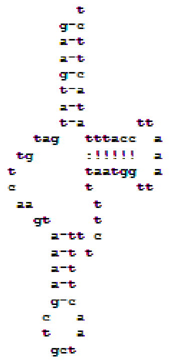
Alanine (Ala)



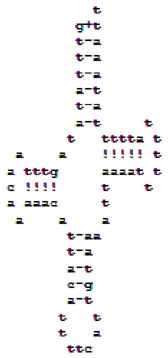
Arginine (Arg)



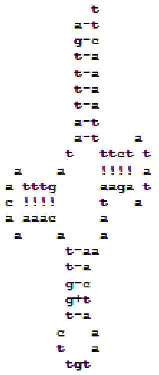
Asparagine (Asn)



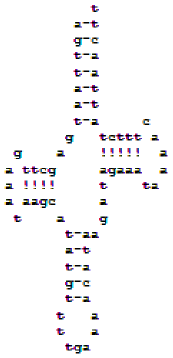
Serine (Ser -AGN)



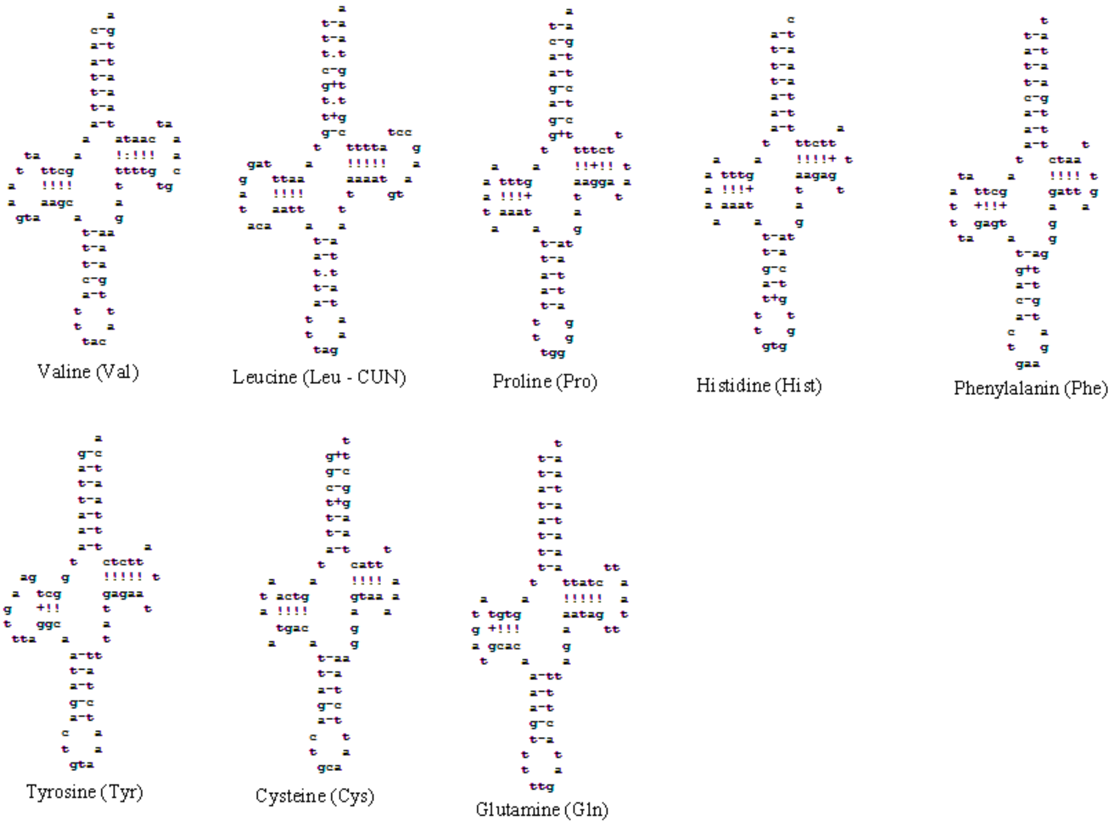
Glutamate (Glu)



Threonine (Thr)



Serine (Ser - UCN)



Valine (Val)

Leucine (Leu - CUN)

Proline (Pro)

Histidine (Hist)

Phenylalanin (Phe)

Tyrosine (Tyr)

Cysteine (Cys)

Glutamine (Gln)

Figure 3.4 Putative secondary structure folds for the tRNAs of MtDNA (*Contig 13*) using ARWEN software. Watson-Crick base pairs designated by “-” or “!” and G–T base pairs by “+”.

Like in all other sequenced MtDNAs, two genes of rRNA (*16S* and *12S*) are also present in this sequenced MtDNA. The large rRNA is 1361 bp long, and is located between tRNA –Leu (CUN) and tRNA-Val while the 787 bp long small rRNA is located between tRNA-Val and the control region. The A + T content for *16S* and *12S* rRNA is 80.9% and 76.8% respectively (Table 3.7)

Table 3.7 Mitochondrial genome comparison of tRNA and rRNA genes between the newly sequenced MtDNA (*Contig 13*) and four published *L. cuprina* strains.

| **Species** | **tRNAs** | | **lrRNA** | | **srRNA** | |
|---|---|---|---|---|---|---|
| | **Size** | **AT%** | **Size(bp)** | **AT%** | **Size(bp)** | **AT%** |
| *Contig 13* | 1471 | 76.7 | 1361 | 80.9 | 787 | 76.8 |
| *Lucilia cuprina DI213.5* | 1471 | 76.8 | 1327 | 81.7 | 785 | 77.3 |
| *Lucilia cuprina DI213.4* | 1471 | 76.8 | 1328 | 81.7 | 786 | 77.3 |
| *Lucilia cuprina DI213.3* | 1471 | 76.8 | 1327 | 81.7 | 785 | 77.3 |
| *Lucilia cuprina DI213.2* | 1471 | 76.8 | 1327 | 81.7 | 785 | 77.3 |

### 3.3.9 Non-coding elements

The non-coding elements of the metazoan MtDNA consist of the control region and small intergenic spacers. In the newly sequenced MtDNA, a 259 bp long control region and 13 intergenic spacers were identified. Most of the intergenic spacers range in length from 1 - 16 bp. However, there is a much longer intergenic spacer located between *NAD3* and *NAD5* which is 57 bp long.

Table 3.8 Intergenic spacers found between the genes of the newly sequenced MtDNA (*Contig 13*) and their respective sizes. The genes surrounding the intergenic spacers are indicated with the spacer region represented by a "-".

| **Intergenic spacer located within the MtDNA** | **Size (bp)** | **Intergenic spacer located within the MtDNA** | **Size (bp)** |
|---|---|---|---|
| tRNA (Met) - NAD2 | 3 | NAD3 - NAD5 | 57 |
| NAD2 - COI | 4 | NAD5 - NAD4 | 16 |
| COX1 - COX2 | 4 | NAD4L - NAD6 | 4 |
| COII - ATP8 | 1 | NAD 6 - Cytb | 2 |
| ATP6 - COX3 | 2 | NAD1 - 16S rRNA | 12 |
| COX3 - NAD3 | 7 | 16S rRNA - 12S rRNA | 1 |

The control region of the sequenced MtDNA is located between the conserved small rRNA (12S) and *tRNA-Ile*. This control region is much smaller (259 bp) than the control regions of the four *L. cuprina* strains (Table 3.9). This region aligns perfectly with the control region sequences of the four *L. cuprina* strains, except for a couple of point mutations (Figure 3.5).

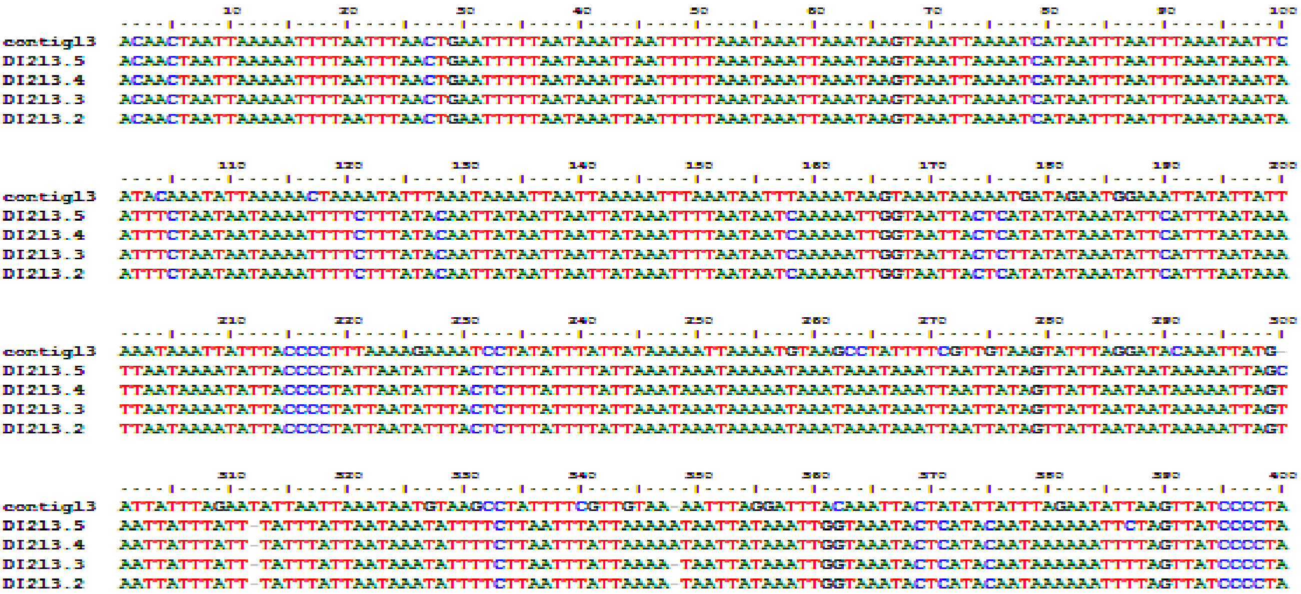


Figure 3.5 Alignment of the nucleotide sequences of the control region of sequenced MtDNA (*Contig 13*) and four of the published *Lucilia cuprina* strains (DI213.2-5).

The nucleotide composition of this control region has an AT-content of 88.4% (A: 49.4% and T: 39%) and a GC-content of 11. 6% (G: 7.5% and C: 4.1 %). The AT-content of this control region is much smaller compared to the other four strains (Table 3.9). This could be attributed to the size of the control region as it half the size of that in the four *L. cuprina* strains. The reduced GC-content is one of the most diagnostic features of the insect control region (Boore 1999).

Table 3.9 Nucleotide composition of the control region of the newly sequenced MtDNA compared to the four published *L.cuprina* strains.

| **Taxon** | **Size (bp)** | **T** | **C** | **A** | **G** | **AT (%)** |
|---|---|---|---|---|---|---|
| New MtDNA | 259 | 39 | 4.1 | 49.4 | 7.5 | 88.4 |
| *Lucilia cuprina DI213.5* | 407 | 42.8 | 6.9 | 48.2 | 2.2 | 91 |
| *Lucilia cuprina DI213.4* | 447 | 43.2 | 6.4 | 48.2 | 2.2 | 91.4 |
| *Lucilia cuprina DI213.3* | 470 | 44 | 6.4 | 47.7 | 1.9 | 91.7 |
| *Lucilia cuprina DI213.2* | 491 | 44.1 | 5.9 | 48 | 2 | 92.1 |

## 3.4 Discussion

### 3.4.1 Consensus sequence generation

The MiSeq platform generated 3 636 306 paired-end (250 bp) reads. The sequence quality score of the NGS reads ranged between Q29 – Q40, this is a base call accuracy above 99%. The quality check summary (QC) reported 24% reads showing low quality, and only 0.003% reads showing some ambiguity (Table 3.3). The low quality sequence reads were clipped and/or removed from the sequence data and were not included in any of the downstream analyses. A *de novo* assembly of the sequence reads generated 872 contigs/scaffolds from which a 15 kb consensus sequence was identified to be the complete MtDNA of the fly specimen

PalGr3Dipt0701. This MtDNA was sequenced at a 183.4-fold average coverage, producing a consensus sequence with strong statistical support**.**

### 3.4.2 Species identification using the consensus sequence

The consensus sequences generated from both the *de novo* assembly and the mapping of sequencing reads to the reference sequence downloaded from Genbank, validated the species identification and the consensus sequences matched a previously published MtDNA of *Lucilia cuprina* species available in Genbank.

### 3.4.3 Organization and characteristics of the MtDNA of PalGr3Dipt0701

The size (15 079 bp) of the complete MtDNA of PalGr3Dipt0701 (*Contig 13*), is well within the observed range of insect MtDNAs which is between 14 – 19 kb long(Boore 1999). The architecture of this genome including the genome content, gene order and orientation is consistent with that of *Drosophila yakuba* which is regarded as the primary model organism for insect mitogenomic research (Clary & Wolstenholme 1985; Boore 1999), as well as the published *Lucilia cuprina* strains (DI213.1-5, DI190.1-5). The gene order of this mitochondrial genome shows that the gene order in Diptera species is highly conserved.

Moreover, other known Diptera species all have the same gene order as *D. yakuba* except the family Cecidomyiidae where there is rearrangement in tRNA -Ala and tRNA-Arg (Zhang *et al.* 2015). The genome produced in this study contains all 37 genes: 13 protein-coding genes, 22 tRNA and 2 rRNA genes and a control region. The genes overlap at 20 gene junctions. These overlaps involve a total of 531 bp, with the longest overlap (144 bp) found between *COI* and *COII*. The overlapping of adjacent genes is common in many animal MtDNAs, although the extent of overlaps may vary amongst taxa (Cai *et al.* 2012; Liu *et al.* 2012).

The MtDNA of PalGr3Dipt0701 is much smaller compared to that of other previously published *Lucilia cuprina* strains. There was variation in MtDNA size amongst the published *Lucilia cuprina* strains; this size difference among the strains could be due to the variation of intergenetic regions, and the control region. For example, *Lucilia cuprina* strain DI213.2 has the largest genome size (15 310 bp) and this strain also has the largest control region (491 bp), whilst *Lucilia cuprina* strain DI213.5 has a smaller genome size (15 226 bp) and a smaller control region (Table 3.4 and Table 3.9). This could be the case with PalGr3Dipt0701, which has the smallest genome size and also has the smallest control region compared to the other *Lucilia cuprina* strains.

The small genome of PalGr3Dipt0701 could also be prone with next generation sequencing challenges associated with sequencing and mapping repetitive DNA regions. This is the biggest technical challenge associated with next generation sequencing (Alkan *et al.* 2011). During computational analysis, these repeats create ambiguities in alignment and in genome assembly, which in turn can produce errors when interpreting results (Alkan *et al.*, 2011; Treangen and Salzberg 2011). For *de novo* assembly, repeats that are longer than the read length create gaps in the assembly thus, producing fragmented assemblies (Treangen and Salzberg 2011). Also, repeats can be erroneously collapsed on top of one another and can cause complex, misassembled rearrangements and near-identical repeats are often collapsed into fewer copies resulting in reduced or lost genomic complexity (Alkan *et al.* 2011; Alkan *et al.* 2011).

The 13 protein-coding genes have the standard mitochondrial initiation codons (methionine and isoleucine). Nine of these protein-coding genes (*NAD2*, *COI*, *ATP8*, *NAD1*, *NAD3*, *NAD4*, *NAD4L*, *NAD5*, and *NAD6*) have isoleucine as a start codon (ATT) and the rest of the proteins, namely *COII*, *COIII*, *Cytb* and *ATP6* have methionine as a start codon (ATG). All the protein-coding genes have complete termination codons (TAA or TAG), except for *NAD4* which has an incomplete termination codon (T). Incomplete termination codons (T or TA), is common in many animal MtDNAs including that of insects (Bae *et al.* 2004). The presence of incomplete stop codons is not an unusual phenomenon in protein-coding genes, but is found in a number of invertebrate MtDNAs (Ojala *et al.* 1981). Ojala *et al.* (1981) explained this phenomenon as being created by polyadenylation of mRNA.

All 22 tRNA genes are present in the mitochondrial genome of PalGr3Dipt0701. The tRNA genes vary in length from 63 – 72 bp and these values are within the observed range for other insects (Boore 1999; Bae *et al.* 2004; Song *et al.* 2010). All 22 tRNA genes secondary structures and their anticodons are identical to that of *D. yakuba* and *L. cuprina* strain DI213.5. They form a typical clover-leaf structure except for the *tRNA-Ser* (AGN), in which the dihydrouracil arm forms a simple loop (Figure 3.5). This is not unusual as it is seen in several animal Mt DNAs, including insects (Bae *et al.* 2004; Song *et al.* 2010). Moreover, the secondary structures of the tRNA genes are similar to each other, and this suggests that these tRNA genes have similar functions.

Like most animal MtDNA, the MtDNA sequenced here has non-coding elements, including the 259 bp control region and the 13 small intergenic spacers (Table 3.9). The majority of the small intergenic spacers ranged between 1 – 16 bp, with one longer intergenic spacer (57 bp) which is located between *NAD3* and *NAD5*. This is not unusual in insects as small intergenic spacers can be longer than 50 bp in size (Song & Liang 2009).

## 3.5 Conclusion

The MtDNA of PalGr3Dipt0701 (15 079 bp), a *L. cuprina* species, was constructed from sequence reads generated by the MiSeq platform (Illumina). The circular genome contains all 37 mitochondrial genes (13 protein-coding genes, 2 ribosomal RNA, 22 transfer RNA genes) and a control region. All protein-coding genes use standard mitochondrial initiation codons (methionine and isoleucine), and the usual TAA and TAG termination codon. All 22 transfer RNA show a typical clover-leaf structure, except for *tRNASer (AGN*) which forms a simple loop on the dihydrouridine arm. The gene order and orientation are identical to that of *Lucilia cuprina* and *Drosophila yakuba*. The newly sequenced MtDNA PalGr3Dipt0701, together with other Diptera species with complete MtDNAs will be used to search for genes (other than *COI* gene) that can be used as potential barcodes to facilitate Diptera species identification.

## 3.6 References

Alkan C., Coe B.P. & Eichler E.E (2011) Genome structure variation discovery and genotyping. *Nature Reviews Genetics* **12** (5), 363-376.

Alkan C., Sajjadian S. & Echler E.E. (2011) Limitations of next generation genome sequence assembly. *Nature Methods* **8**, 62-65.

Avise J.C., Arnold J., Ball J.E., Bermingham E. & Saunders N.C. (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology, Evolution and Systematics* **18**, 489-522.

Bae J.S., Kim I., Sohn H.D. & Jin B.R. (2004) The mitochondrial genome of the firefly, *Pyrocoelia rufa*: complete DNA sequence, genome organization, and phylogenetic analysis with other insects. *Molecular Phylogenetics and Evolution* **32**, 978-85.

Beckenbach A.T. (2011) Mitochonrial genome sequences of Nematocera (Lower Diptera): Evidence of rearrangement following a complete genome duplication in a winter crane fly. *Genome Biology and Evolution* **4**, 89-101.

Bernt M., Donath A., Juhling F., Externbrink F., Florentz C., Fritzsch G., Putz J., Middendorf M. & Stadler P.F. (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* **69**, 313-9.

Boore J.L. (1999) Animal mitochondrial genomes. *Nucleic acids research* **27**, 1767-80.

Bruhn T. (2011) Sequence and analysis of the mitochondrial DNA control region of nine Australian species of the genus Chrysomya (Diptera: Calliphoidae). In: *School of Biological Sciences*. University of Wollongong

Bybee S.M., Bracken-Grissom H.D., Hermansen R.A., Clement M.J., Crandall K.A. & Felder D.L. (2011) Directed next generation sequencing for phylogenetics: an example using Decapoda (Crustacea). *Journal of Comparative Zoology* **250**, 497-506.

Cai X.Q., Liu G.H., Song H.Q., Wu C.Y., Zou F.C., Yan H.K., Yuan Z.G., Lin R.Q. & Zhu X.Q. (2012) Sequences and gene organization of the mitochondrial genomes of the liver flukes *Opisthorchis viverrini* and *Clonorchis sinensis* (Trematoda). *Parasitology Research* **110**, 235-43.

Clary D.O. & Wolstenholme D.R. (1985) The Mitochondrial DNA molecule of *Drosophila yakuba:* Nucleotide sequence, gene organisation and genetic code. **22**, 252-71.

Hall T.A. (1999) BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows. *Nucleic Acids Symposium Series* **41**, 95-8.

Illumina (2012) Nextera XT DNA Sample Preparation Guide (15031942).

Illumina (2014) http://www.illumina.com/systems/miseq/system.ilmn.

Laiho J. & Ståhls G. (2013) DNA barcodes identify Central Asian Colias butterflies. *ZooKeys* **365**, 175-96.

Laslett D. & Canbäck B. (2008) ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* **24**, 172-5.

Liu G.H., Wang S.Y., Huang W.Y., Zhao G.H., Wei S.J., Song H.Q., Xu M.J., Lin R.Q., Zhou D.H. & Zhu X.Q. (2012) The complete mitochondrial genome of *Galba pervia* (Gastropoda: Mollusca), an intermediate host snail of *Fasciola spp*. *PLoS ONE* **7**, 1-9.

Lyra M.L., Fresia P., Gama S., Cristina J., Klaczko L.B. & Azeredo-Espin A.M.L. (2005) Analysis of mitochondrial DNA variablity and genetic structure in population of new world screwworm flies (Diptera:Calliphoridae) from Uruguay. *Journal of Medical Evolution* **42**, 589-95.

Mardis E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133-41.

Meier R., Shiyang K., Vaidya G. & Ng P.K.L. (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55**, 715-28.

Moore W.S. (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees vs nuclear-gene trees. *Evolution* **49**, 718-26.

Ojala D., Montoya J. & Attardi G. (1981) tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**, 470-4.

Park D.S., Foottit R., Maw E. & Hebert P.D.N. (2011) Barcoding bugs: DNA-based identification of the true bugs (Insecta: Hemiptera: Heteroptera). *PLoS ONE* **6**, e18749-e.

Schattner P., Brooks A.N. & Lowe T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**, 686-9.

Singh D.K.N. (2015) Lucrative potentials of mitochondrial DNA: a review accentuaing particularly blow flies beyond forensic importance. *Journal of Entomology and Zoology Studies* **3**, 1-8.

Song N., Liang A.P. & Ma C. (2010) The complete mitochondrial genome sequence of the planthopper, *Sivaloka damnosus*. *Journal of Insect Science* **10**, 76-.

Song N. & Liang A. (2009) The complete mitochondrial genome sequence of *Geisha distinctissima* (Hemiptera: Flatidae) and comparison with other hemipteran insects. *Acta Biochemica Biophysica Sinica* **41**, 206-16.

Stothard P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102-4.

Taanman J.W. (1999) The mitochondrial genome: structure, transcription, translation and replication. *Bioenergetics* **1410**, 103-23.

Tembe S., Shouche Y. & Ghate H.V. (2014) DNA barcoding of Pentatomomorpha bugs (Hemiptera: Heteroptera ) from Western Ghats of India. *Meta Gene 2* **2**, 737-45.

Tumura K., Peterson D., Peterson N., Steder G., Nei M. & Kumar S. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony method **28**, 2731-9.

Treangen T.J. & Salzberg S.L. (2011) Repetitive DNA and next generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13** (1), 36-46.

Versteirt V., Nagy Z.T., Roelants P., Denis L., Breman F.C., Damiens D., Dekoninck W., Backeljau T., Coosemans M. & Van Bortel W. (2014) Identification of Belgian mosquito species (Diptera: Culicidae) by DNA barcoding. *Molecular Ecology Resources*.

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

Wyman S.K., Jansen R.K. & Boore J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252-5.

Zhang D.X. & Hewitt G.M. (1997) Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochemical Systematics and Ecology* **25**, 99-120.

Zhang N.X., Yu G., Li T.J., He Q.Y., Zhou Y., Si F.L, Ren S. & Chen B. (2015) The complete mitochondrial genome of *Delia antiqua* and its implications in Dipteran phylogenetics. *PLoS ONE* **10**, 1-16.

Zhou X., Robinson J.L., Geraci C.J., Parker C.R., Flint O.S., Etnier D.A., Ruiter D., DeWalt R.E., Jacobus L.M. & Hebert P.D.N. (2011) Accelerated construction of a regional DNA-barcode reference library: caddisflies (Trichoptera) in the great smoky mountains national park. *Journal of the North American Benthological Society* **30**, 131-62.

# CHAPTER FOUR

# EVALUATION OF MITOCHONDRIAL GENES AS POTENTIAL DNA BARCODES FOR DIPTERA

## Abstract

The mitochondrial genome is well suited for the study of closely related taxa, and has been used as the source of DNA barcode information in animals owing to its small effective population size and faster mutation rate compared to the nuclear genome. A 658 bp fragment of the mitochondrial cytochrome *c* oxidase subunit I gene (*COI*), is widely used as a universal barcode for the identification of animal taxa with more than 95% resolution in most animal groups. However, the *COI* gene appears to be problematic when identifying Diptera species with a relatively low success rate (< 70%). Therefore, alternative DNA barcodes should be considered for use in Diptera. Our study evaluated the potential of 13 protein-coding genes (PCGs) from the mitochondrial genome of 49 Diptera species by comparing the phylogenetic trees produced by each PCG as well as testing for the presence of the barcoding gap in each gene. In our results, the *COI* gene performed better than expected, and the *ATP6* gene showed promise as an alternative barcode marker.

**Key Words:** Mitochondrial genome, protein-coding genes, DNA barcodes

## 4.1 Introduction

The mitochondrial genome (MtDNA), is suitable for studying closely related taxa and has been used as the source of DNA barcode information in many different animal taxa (Hebert *et al.* 2003; Hebert *et al.* 2004b; Ward *et al.* 2005; Park *et al.* 2011). The utility of the MtDNA in resolving relationships amongst closely related taxa is based on its simple circular genetic structure, small effective population size and high copy number (Brown *et al.* 1979; Moore 1995; Nichols 2001). Furthermore, the MtDNA has limited exposure to genetic recombination, no introns, and has a higher rate of evolution compared to the nuclear genome (Xu & Singh 2005), making it well suited for phylogenetic and population genetic studies of closely related taxa (Mandal *et al.* 2014).

Typical animal MtDNAs are small and circular with size ranging between 14 – 19 Kb (Boore 1999). They contain a set of 37 genes that are divided into 22 transfer RNA (tRNA) genes, 13 protein-coding genes (PCGs), and 2 ribosomal RNA (rRNA) genes. The two rRNA are: small subunit of ribosomal RNA (*12S* rRNA), and large subunit of ribosomal RNA (*16S* rRNA).

The *12S* rRNA gene is highly conserved in insects and is used to study genetic diversity in higher categorical levels such as phyla and subphyla (Qui-Hong *et al.* 2004). While the *16S* rRNA gene is often used for studies at lower, intermediate levels such as resolving relationships amongst families or genera (Hickson *et al.* 1996; Gerber *et al.* 2001). Compared to the rRNA, the mitochondrial PCGs evolve at a much faster rate making them useful markers for studying lower categorical levels such as relationships amongst families, genera and species (Qui-Hong *et al.* 2004; Mandal *et al.* 2014).

The PCGs encode proteins involved in the oxidative phosphorylation machinery: cytochrome c oxidase subunits 1, 2, and 3 (*COI*, *COII* and *COIII*); cytochrome b subunit (*Cytb*), NADH dehydrogenase subunits 1, 2, 3, 4, 4L, 5, and 6 *(NAD1* to *NAD6*, *NAD4L*), and ATPase subunits 6 and 8 (*ATP6* and *ATP8*). Amongst all the PCGs, the *COI* has been extensively used as a molecular marker in DNA barcoding studies.

In DNA barcoding, a 658 bp region of the *COI* gene is used as a universal barcode for the identification of most animal taxa (Hebert *et al.* 2003; Hebert *et al.* 2004a; Smith *et al.* 2006; Lee *et al.* 2011). This gene has > 95% species identification success in most animal groups (Hebert *et al.* 2003; Hebert *et al.* 2004a; Hebert *et al.* 2004b). However, a relatively low success rate (< 70%) was achieved when identifying our South African Diptera (See Chapter Two). The low success rate in the identification of Diptera species has also been reported in other studies (Meier *et al.* 2006; Virgilio *et al.* 2010). The low success rate of the *COI* gene in identifying Diptera species makes it necessary to identify alternative genes that can be used as DNA barcodes for accurate identification of Diptera, to avoid an exclusive reliance on *COI* gene.

For a gene to be selected as a barcode for DNA barcoding it must meet certain criteria. First, its mutation rate must be slow enough to minimize intraspecific variation, but sufficiently rapid to highlight interspecific variation. Second, it must be easy to amplify in a range of taxonomically distinct taxa. Third, the gene should have few insertions and deletions to facilitate sequence alignment (Hebert *et al.* 2003). Among the 37 mitochondrial genes, the protein-coding genes are potentially good targets for DNA barcoding owing to lower levels of insertion and deletion events, which can complicate the process of sequence alignment. Furthermore, PCCs have a much faster evolutionary rate compared to ribosomal RNA genes which have also been proposed as species-level markers (Vences *et al.* 2005).

This study will evaluate the potential of 13 PCGs from the MtDNA of 49 Diptera species, as DNA barcodes for the identification of Diptera species. This will be done by comparing the phylogenies retrieved by each gene as well as by testing for the presence of the barcoding gap which is essential for the success of DNA barcoding.

## 4.2 Materials and methods

### 4.2.1 Recovery of MtDNA sequences and protein-coding genes

A complete mitochondrial genome of *Lucilia cuprina* was obtained using next generation sequencing in Chapter Three. Additionally, complete and annotated MtDNAs of 48 other Diptera species (Table 4.1) were obtained from the National Center for Biotechnology Information (NCBI) Genomes databases (http://www.ncbi.nlm.nih.gov/). Taxa were chosen according to the available complete mitochondrial genomes in NCBI database, and taxa with at least two complete mitochondrial genomes per species were selected to allow for between species comparisons. The 49 Diptera species are distributed among 11 Diptera families and 19 genera. Thirteen PCG sequences were extracted from each of 49 complete mitochondrial genomes (these PCGs are listed in Table 4.1.).

Table 4.1 List of the Diptera species with complete and annotated mitochondrial genomes retrieved from the NCBI database.

| **Family** | **Genus** | **Species name** | **Access number** | **References** |
|---|---|---|---|---|
| Culicidae | Anopheles | *Anopheles quadrimaculatus* | NC_000875.1 | Cockburn *et al.*, 1990 |
| Culicidae | Anopheles | *Anopheles farauti* | NC_020770.1 | Logue *et a*l., 2013 |
| Culicidae | Culex | *Culex quinquefasciatus* | GU188856 | Behura *et al.*, 2011 |
| Culicidae | Culex | *Culex quinquefasciatus* | NC_014574 | Behura *et al.*, 2011 |
| Culicidae | Anopheles | *Anopheles gambiae* | NC_022084.1 | Beard *et al.*, 1993 |
| Culicidae | Anopheles | *Anopheles gambiae* | L20934 | Beard *et al.*, 1993 |
| Calliphoridae | Lucilia | *Lucilia sericata* | AJ422212.1 | Stevens *et al.*, 2001 |
| Calliphoridae | Calliphora | *Calliphora vicina* | JX913760.1 | Nelson *et al.*, 2012 |
| Calliphoridae | Calliphora | *Calliphora vicina* | NC_0.19639.1 | Nelson *et al.*, 2012 |
| Calliphoridae | Lucilia | *Lucilia cuprina* | NC019573.1 | Nelson *et al.*, 2012 |
| Calliphoridae | Lucilia | *Lucilia cuprina* | Present Study | Chapter Three |
| Drosophilidae | Drosophila | *Drosophila melanogaster* | KJ947872.2 | Wan and Celniker, 2014 |
| Drosophilidae | Drosophila | *Drosophila melanogaster* | NC_001709 | Lewis *et al.*, 1995 |
| Drosophilidae | Drosophila | *Drosophila yakuba* | NC_001322 | Clary and Wolstenholme, 1985 |
| Drosophilidae | Drosophila | *Drosophila incompta* | KM275233.1 | De re *et al.*, 2014 |
| Drosophilidae | Drosophila | *Drosophila incompta* | NC_025936.1 | De re *et al.*, 2014 |
| Tephritidae | Bactrocera | *Bactrocera correcta* | NC_018787 | Wu *et al.*, 2012 |
| Tephritidae | Bactrocera | *Bactrocera correcta* | JX456552.1 | Wu *et al.*, 2012 |
| Tephritidae | Bactrocera | *Bactrocera minax* | HM776033.1 | Zhang *et al.*, 2014 |
| Tephritidae | Bactrocera | *Bactrocera minax* | NC_014402.1 | Zhang *et a*l., 2010 |
| Tephritidae | Ceratitis | *Ceratitis capitata* | NC_000857.1 | Spana *et al.*, 2000 |
| Tachinidae | Elodia | *Elodia flavipalpis* | JQ348961.1 | Zhao *et al.*, 2012 |
| Tachinidae | Elodia | *Elodia flavipalpis* | NC_018118.1 | Zhao *et al.*, 2012 |

| | | | | |
|---|---|---|---|---|
| Tachinidae | Exorista | *Exorista sorbillans* | NC_014704.1 | Shao *et al*., 2010 |
| Tachinidae | Exorista | *Exorista sorbillans* | HQ322500 | Shao *et al*., 2010 |
| Tachinidae | Rutilia | *Rutilia goerlingiana* | NC_019640.1 | Nelson *et al*., 2012 |
| Tachinidae | Rutilia | *Rutilia goerlingiana* | JX913762.1 | Nelson *et al*., 2012 |
| Tabanidae | Cydistomyia | *Cydistomyia duplonotata* | NC_008756.1 | Cameron *et al*., 2007 |
| Tabanidae | Cydistomyia | *Cydistomyia duplonotata* | DQ866052.1 | Cameron *et al*., 2007 |
| Chironomidae | Chironomus | *Chironomus tepperi* | NC_016167.1 | Beckenbach, 2012 |
| Chironomidae | Chironomus | *Chironomus tepperi* | JN861749.1 | Beckenbach, 2013 |
| Muscidae | Scathophaga | *Scathophaga stercararia* | KM200724.1 | Li *et al*., 2014 |
| Muscidae | Scathophaga | *Scathophaga stercararia* | NC_024856.1 | Li *et al*., 2014 |
| Muscidae | Muscina | *Muscina stabulans* | NC_026292.1 | Zha *et al*., 2015 |
| Muscidae | Muscina | *Muscina stabulans* | KM676394.1 | Zha *et al*., 2014 |
| Muscidae | Musca | *Musca domestica* | NC_024855.1 | Li *et al*., 2014 |
| Muscidae | Musca | *Musca domestica* | KM200723.1 | Li *et al*., 2014 |
| Agromyzidae | Liriomyza | *Liriomyza sativae* | JQ862475.1 | Wang *et al*., 2012 |
| Agromyzidae | Liriomyza | *Liriomyza sativae* | NC_015926.1 | Yang *et al*., 2011 |
| Agromyzidae | Liriomyza | *Liriomyza sativae* | HQ333260.1 | Du *et al*., 2011 |
| Agromyzidae | Liriomyza | *Liriomyza trifolii* | NC_014283.1 | Wang *et al*., 2011 |
| Agromyzidae | Liriomyza | *Liriomyza trifolii* | GU327644.1 | Wang *et al*., 2011 |
| Agromyzidae | Liriomyza | *Liriomyza trifolii* | JN570506.1 | Yang *et al*., 2011 |
| Cecidomyiidae | Mayetiola | *Mayetiola destructor* | GQ387648.1 | Beckenbach and Joy, 2009 |
| Cecidomyiidae | Mayetiola | *Mayetiola destructor* | NC_013066 | Beckenbach and Joy, 2009 |
| Cecidomyiidae | Rhopalomyia | *Rhopalomyia pomum* | GQ3887649.1 | Beckenbach and Joy, 2009 |
| Cecidomyiidae | Rhopalomyia | *Rhopalomyia pomum* | NC_013063 | Beckenbach and Joy, 2009 |
| Syrphidae | Simosyrphus | *Simosyrphus grandicornis* | NC_008754.1 | Cameron *et al*., 2007 |
| Syrphidae | Simosyrphus | *Simosyrphus grandicornis* | DQ866050 | Cameron *et al*., 2007 |

### 4.2.2 Sequence alignment and phylogenetics

Alignments of each of the 13 PCGs across 49 complete mitochondrial genomes were done using ClustalW (Thompson *et al.* 1994) module, in BioEdit (Hall 1999). The alignments were optimized manually to ensure homology and the aligned sequences contained no insertions, deletions or stop codons. The program MEGA version 5 (Tumura *et al.* 2011), was used to describe the sequence length, number of conserved characters, variable characters and parsimony informative characters for each of the 13 PCGs.

An appropriate model of evolution was then determined for each of the 13 PCG's using jModelTest (Posada 2008), implementing the Akaike Information Criterion (Cavanaugh 2007). In all cases the GTR + I + G model was selected as the optimal substitution model. Phylogenies were then constructed for each PCG using two methods: maximum likelihood (ML) and bayesian inference (BI). The program Garli 0.96 win32 (Zwickl 2006) was used to perform the ML analysis, with branch support accessed by performing 100 bootstrap replicates for each gene.

The BI was performed using MrBayes 3.1.2 (Huelsenbeck & Ronquist 2001). Two independent runs each consisting of four parallel Markov Chain Monte Carlo (MCMC) chains, were launched from random starting trees and run for 10 million generations each, with the cold Markov chain sampled every 300 generations. Priors were set to nst = 6, invariant sites and gamma. The convergence of the MCMC chains from the BI analyses was accessed in Tracer v1.5 (Drummond & Rambaut 2007). The first 25% of trees from each run were discarded as burn-in after analyses in Tracer.

All tree files (BI and ML tree files), were first converted into Phylip format using Mesquite (Jühling *et al.* 2012), and consensus trees were constructed using the consensus program which is part of the Phylip v3.69 package (Felsenstein 2005). Phylogenetic trees were viewed in Figtree v1.3.1 (Rambaut 2009).

### 4.2.3 DNA barcoding gap for each protein-coding gene

Since the success of species identification using DNA barcoding is largely dependent on the absence of overlap between intraspecific and interspecific divergence, the presence of the DNA barcoding gap was tested in each of the 13 PCG data sets. Intraspecific and interspecific genetic distances were calculated using the K2P model in R, using the software SPIDER version 1.1-1 SPecies IDentity and evolution, (Brown *et al.* 2012).

Although this model is not best-fit for the data (see above) this model is routinely used by the barcoding community. Using a simple model such as K2P rather than a more parameter rich model such as GTR will lead to the underestimation of genetic distances. Using the K2P model in this study to test for the presence of the DNA barcoding gap, is thus the more conservative approach. The Jeffries-Matusita (JM) distances were then calculated for each of the 13 PCGs to check for separability between inter- and intraspecific sequence divergence values. A J-M value above 1.447 suggests that the two sequence divergences are separable and a value below 1.447 suggests an overlap between the two.

## 4.3 Results

### 4.3.1 Sequence analysis

One of the criteria for a gene to be selected as a barcode marker for DNA barcoding, is that its mutation rate must be slow enough to minimize intraspecific variation but sufficiently rapid to highlight interspecific variation. The *COI* gene has a relatively slow mutation rate. This marker has been used in many DNA barcoding studies. This gene has a high number of conserved characters (51%), compared to the rest of the PCGs from the 49 Diptera species. Other PCGs with high number of conserved characters include *COIII* (42%), *Cytb* (40%), *ATP 6* (39%) and *COII* (39%). While *NAD4* (15%) and *NAD4L* (14%) have the lowest number of conserved characters (Table 4.2).

Table 4.2 Sequence lengths, number of conserved characters, variable characters and parsimony informative characters for each of the 13 PCGs extracted from the 49 Diptera species. The conserved characters, variable characters and parsimony informative characters are expressed as percentages in order to compare the differences between the genes.

| **Gene name** | **Sequence length (bp)** | **Percentage conserved characters** | **Percentage variable characters** | **Percentage parsimony informative characters** |
|---|---|---|---|---|
| *COI* | 1534 | 51 | 49 | 40 |
| *COII* | 688 | 39 | 61 | 49 |
| *COIII* | 789 | 42 | 58 | 46 |
| *Cyt b* | 1135 | 40 | 60 | 51 |
| *ATP6* | 678 | 39 | 61 | 50 |
| *ATP8* | 163 | 23 | 77 | 66 |
| *NAD1* | 939 | 39 | 61 | 46 |
| *NAD2* | 1017 | 23 | 77 | 65 |
| *NAD3* | 354 | 35 | 65 | 55 |
| *NAD4* | 1339 | 15 | 85 | 56 |
| *NAD4L* | 297 | 14 | 86 | 56 |
| *NAD5* | 1727 | 17 | 83 | 53 |
| *NAD6* | 521 | 22 | 78 | 68 |

### 4.3.2 Phylogenetic analysis

The ML and BI analyses generated similar phylogenies for the respective PCGs except for *NAD4L* and *NAD5*. Also, the relationships amongst the different species differed among the PCGs (Figure 4.1 - 4.4).

In the *COI* gene phylogeny, the species *Lucilia cuprina*, *Musca domestica* and *Drosophila yakuba* were recovered as monophyletic with strong supporting posterior probabilities (1.0) and bootstrap values (100%). Species from the same family and genus formed monophyletic lineages with well supported branches. For example, the families Drosophilidae (1.0/87), Culicidae (1.0/95), Agromyzidae (1.0/99) and Cecidomyiidae (1.0/100) were recovered as monophyletic (Figure 4.1).

The *COII* gene retrieved a phylogeny different from that of the *COI* gene. Diptera families including Drosophilidae (1.0/95), Cecidomyiidae (1.0/100), Culicidae (1.0/95) and Agromyziidae (1.0/88) were still recovered as monophyletic. However, in this phylogeny the species *Lucilia sericata* and *Lucilia cuprina* were placed together as sister taxa (1.0/92). Furthermore, a monophyletic relationship was expected for *Scathophaga stercoraria, Musca domestica* and *Muscina stabulan* since they belong to the same family, instead *Musca domestica* formed a monophyletic relationship with Drosophilidae family but this association no supporting bootstrap values and probability posteriors (Figure 4.2).

In the *ATP6* gene, a monophyletic relationship between *Lucilia sericata* and *Scathophaga stercoraria* was observed. However, this relationship is not well supported (<0.5/50 posterior probabilities and bootstrap values respectively). The Muscidae species (*Scathophaga stercoraria*, *Musca domestica* and *Muscina stabulan)* were still placed in different clades (Figure 4.3). There were some monophyletic relationships recovered for species from the same family and genus with well supported branches. For example, the families Culicidae (0.65/64), Agromyzidae (1.0/100) and Tephritidae (1.0/95) were recovered as monophyletic.

In the *NAD2* phylogeny, all the species belonging to the Culicidae family were recovered as monophyletic with a strong supporting (1.0/99 posterior probability and bootstrap value respectively). Also, species belonging to the Agromyzidae (1.0/100), Tephritidae (1.0/100) and Drosophilidae (1.0/100) families were recovered as monophyletic with strong supporting bootstrap values and posterior probabilities. The species *Lucilia sericata* however, was recovered as paraphyletic with species belonging to this family (*Lucilia cuprina* and *Calliphona vicina*) with low supporting bootstrap value and posterior probability. Furthermore, species *Muscina stabulans* was recovered as monophyletic with no supporting bootstrap value and posterior probability (Figure 4.4).

*NAD4L* performed poorly compared to the other PCGs .The ML and BI analyses generated very different phylogenies with no monophyletic families recovered in both trees (Figure 4.5). Not all 49 Diptera species were recovered in both trees. There is an incorrect clustering of species *Lucilia cuprina* and *Lucilia sericata* in the BI tree and species *Rutilia goerlingana* and *Elodia flavipalpis* in the ML tree. However, both these lineages are not supported. The rest of the PCGs (*COIII*, *Cytb*, *ATP8* and six NADs (*NAD 1,3-6*,) had different phylogenies from the *COI* phylogeny with unexpected clusters (Appendix 4.1- 4.8). These PCGs also had few Diptera families that showed monophyletic relationships compared to the *COI*, *ATP6* and *NAD2* genes (Table 4.3). Furthermore, these PCGs lack a DNA barcoding gap (Figure 4.6-4.7).

Table 4.3 Total number of monophyletic relationship recovered for each of the 13 protein coding genes. Currently the 49 MtDNA sequences used in the comparisons represent 11 families, 10 genera and 49 species.

| **Gene** | **Monophyletic families** | **Monophyletic genera** | **Monophyletic species** |
|---|---|---|---|
| *COI* | 9 | 19 | 46 |
| *COII* | 7 | 17 | 30 |
| *COIII* | 6 | 7 | 24 |
| *ATP6* | 7 | 18 | 32 |
| *ATP8* | 5 | 14 | 24 |
| *Cytb* | 8 | 15 | 44 |
| *NAD1* | 6 | 6 | 18 |
| *NAD2* | 8 | 17 | 32 |
| *NAD3* | 5 | 7 | 18 |
| *NAD4* | 6 | 9 | 24 |
| *NAD4L* | 0 | 0 | 4 |
| *NAD5* | 7 | 10 | 28 |
| *NAD6* | 4 | 8 | 22 |

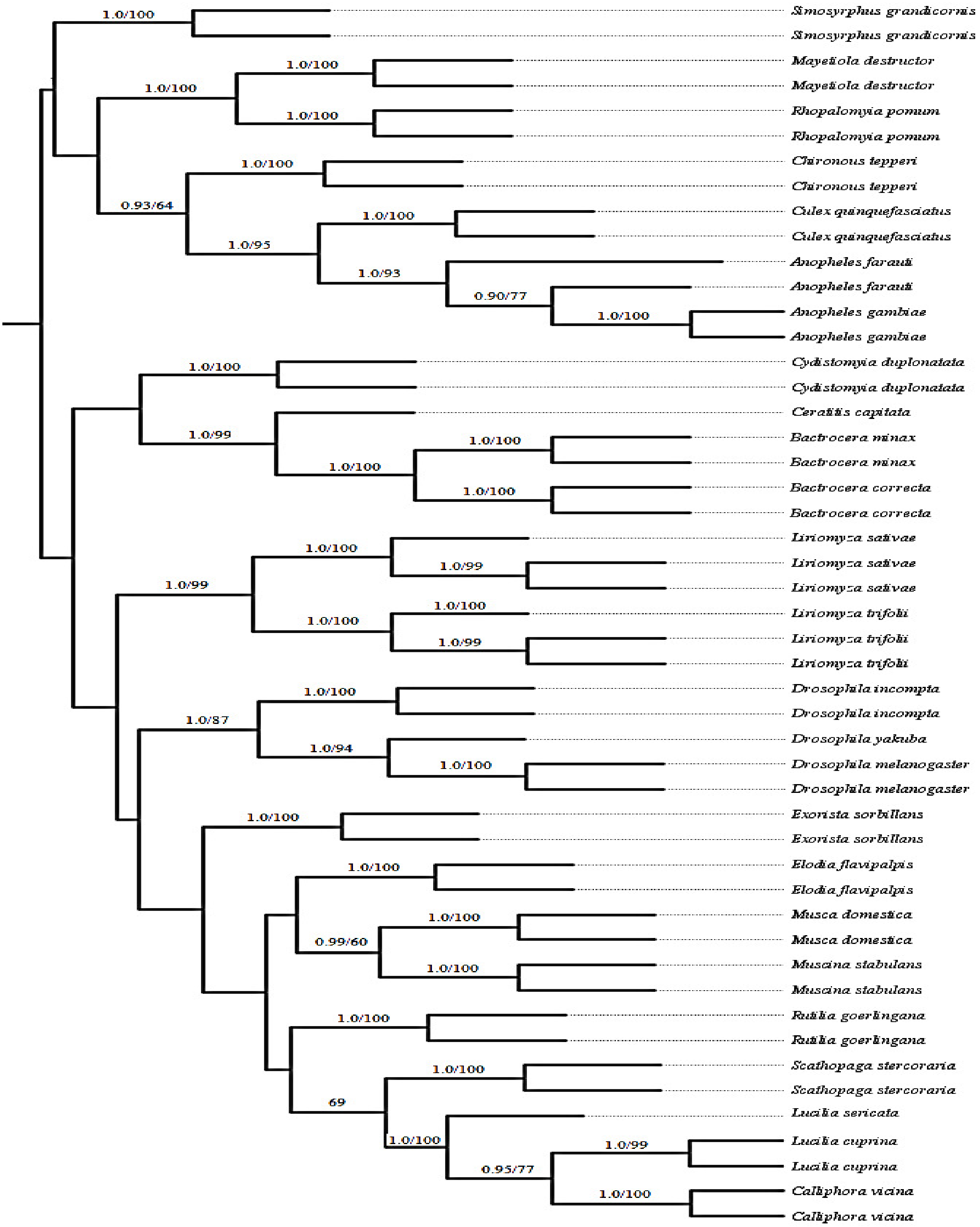


Figure 4.1 Maximum likelihood tree of the *COI* gene. Only bootstrap values above 50 and Bayesian posterior probabilities above 0.5 are shown.

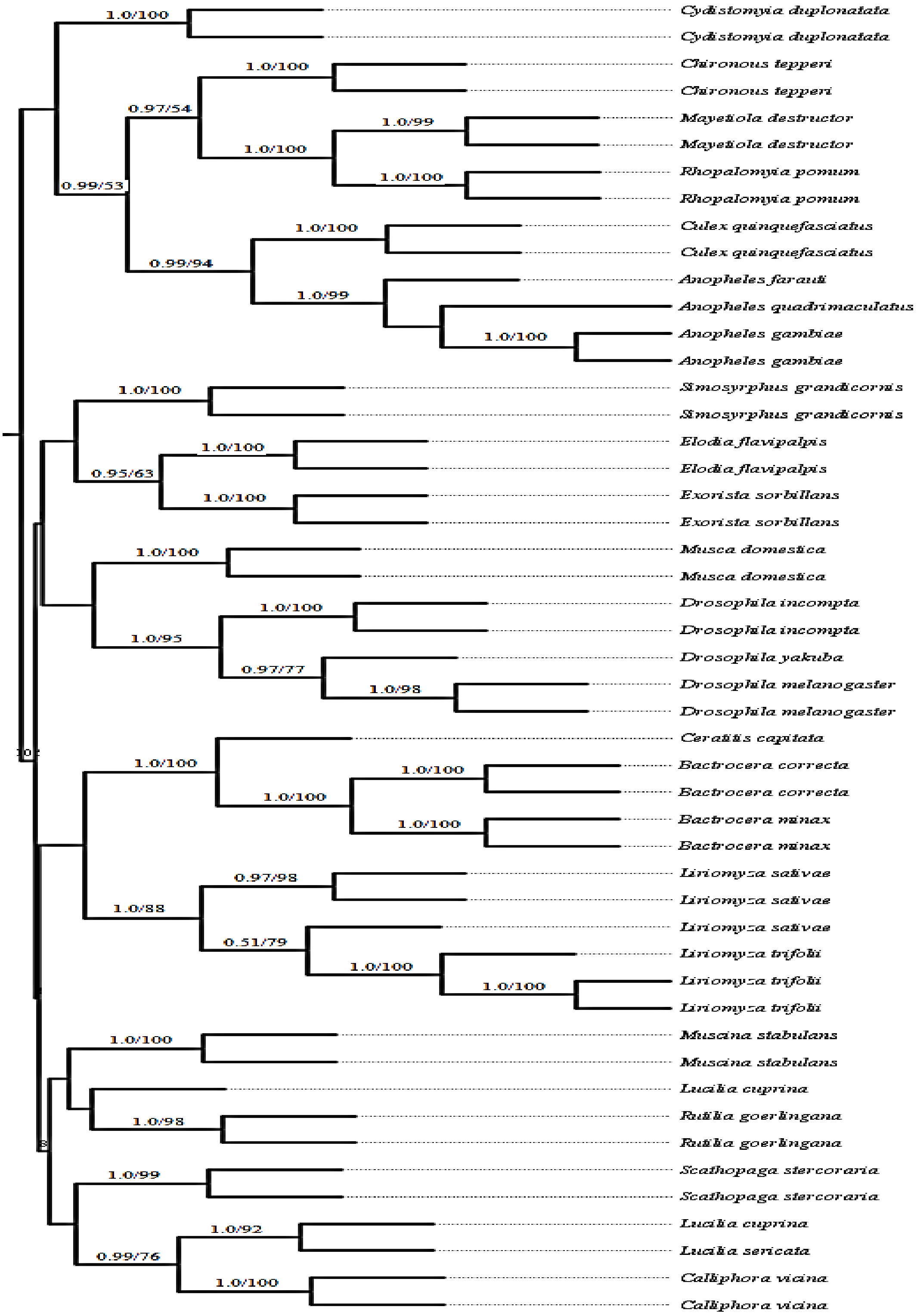


Figure 4.2 Maximum likelihood tree of the *COII* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.

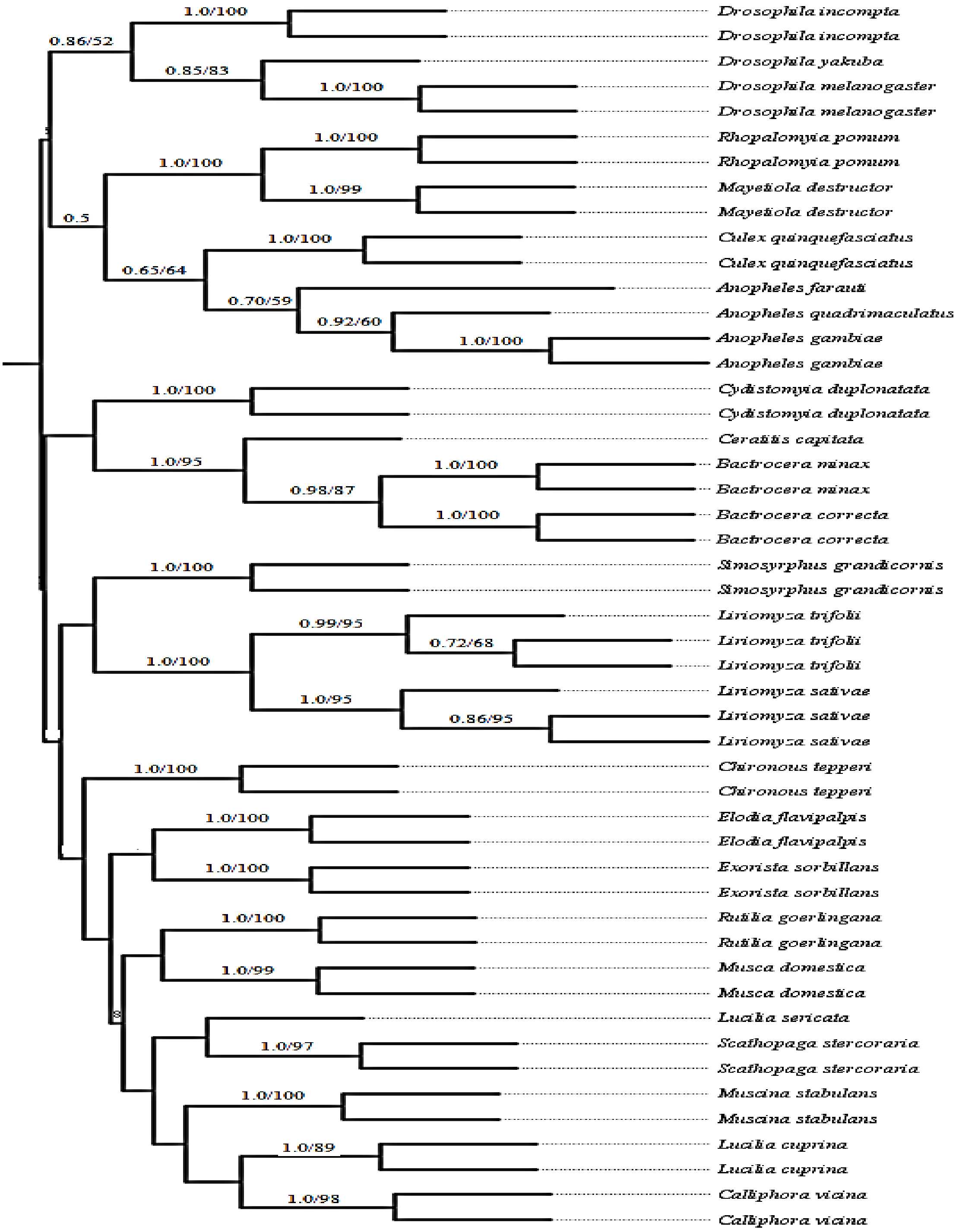


Figure 4.3 Maximum likelihood tree of the *ATP6* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.

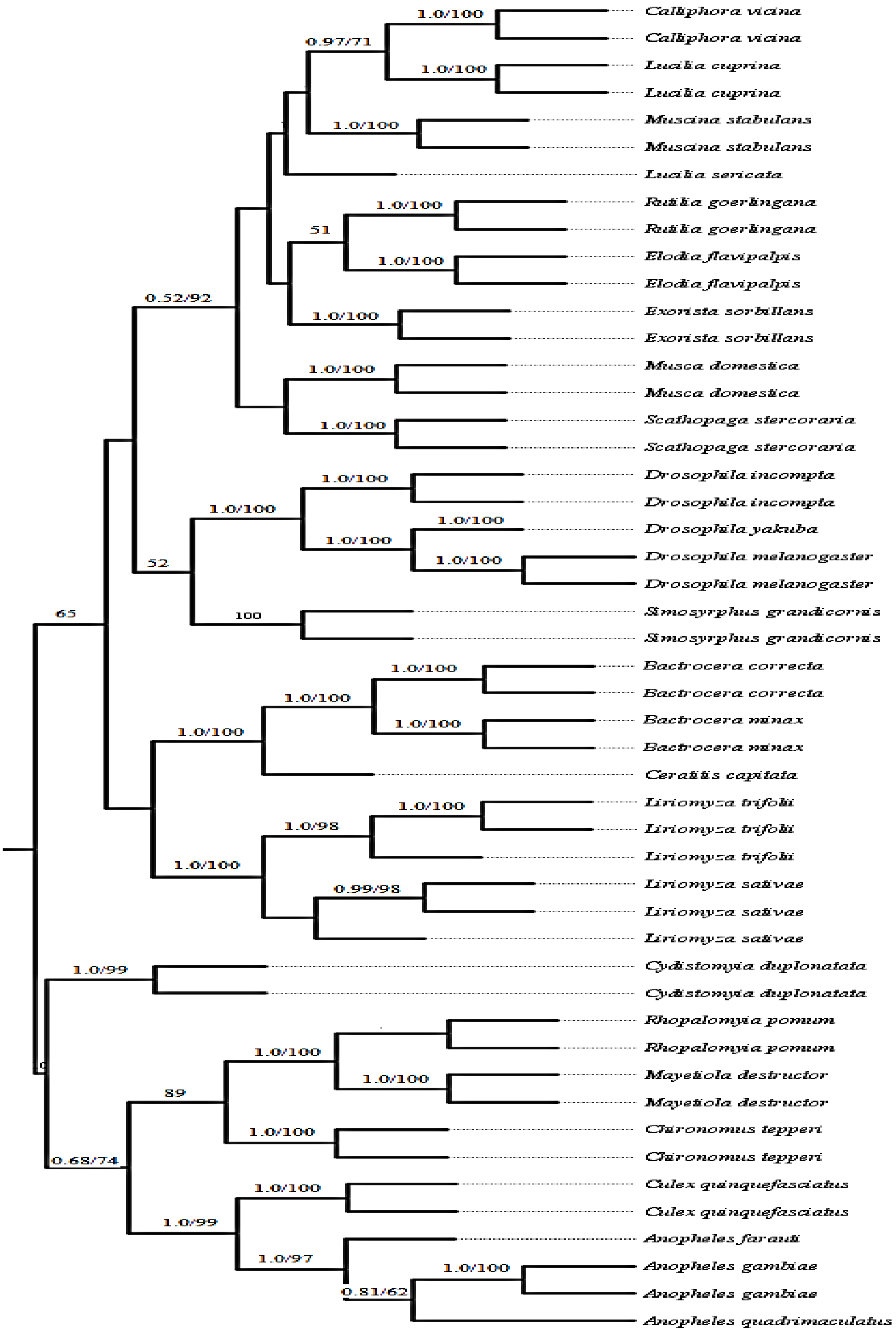


Figure 4.4 Maximum likelihood tree of the *NAD2* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.

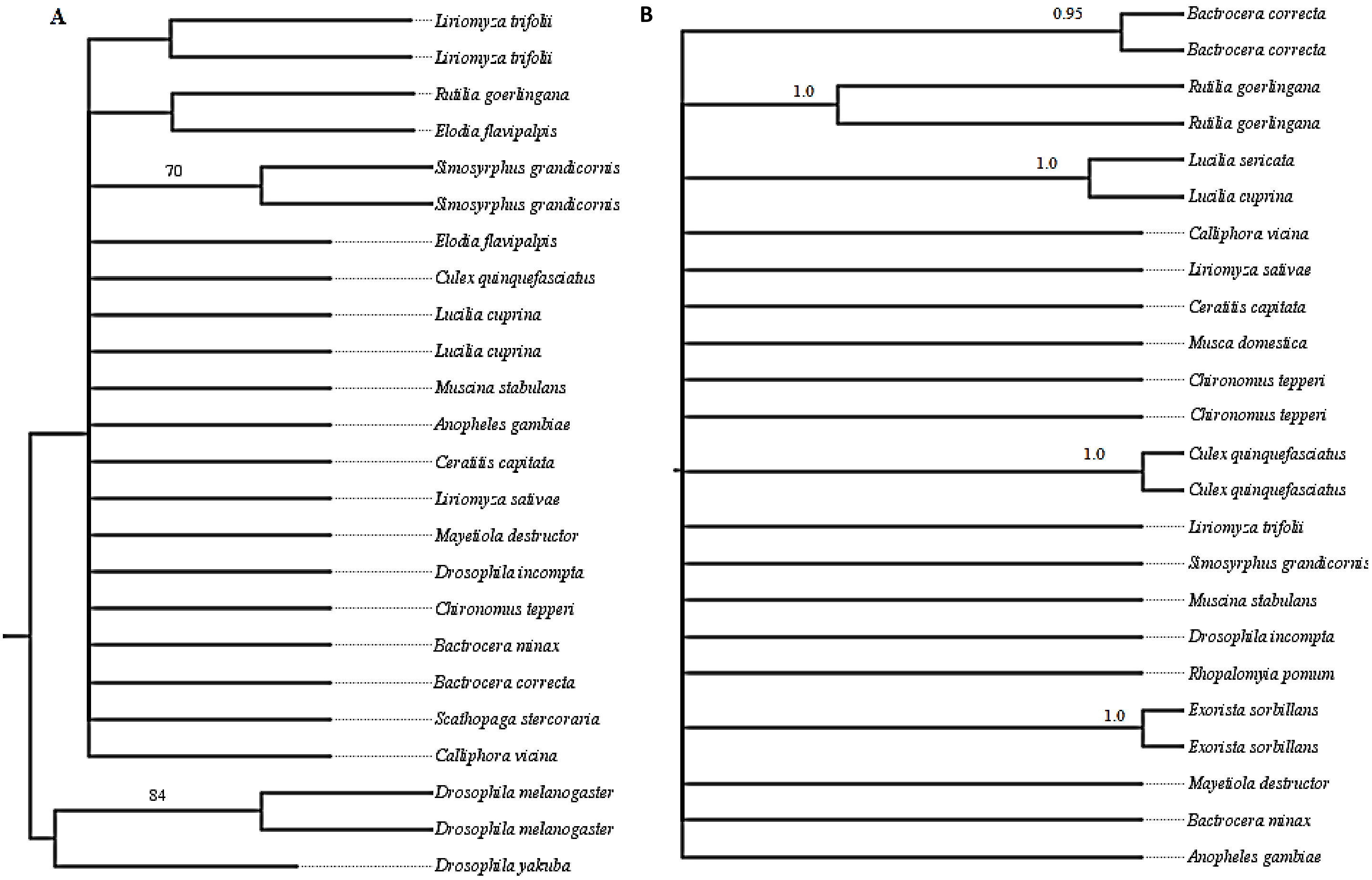


Figure 4.5 Maximum likelihood tree [A] and Bayesian tree [B] of the *NAD4L* gene. Only bootstrap values above 50 and Bayesian posterior probabilities above 0.5 are shown.

### 4.3.3 DNA barcoding gap

Plots of the intraspecific and interspecific divergences for the 13 PCGs is shown in Figures 4.6 to 4.8. The barcoding gap is only observed in two (*COI* and *ATP6*) out of the 13 PCGs examined with JM values greater than 1.447 (Figure 4.5).

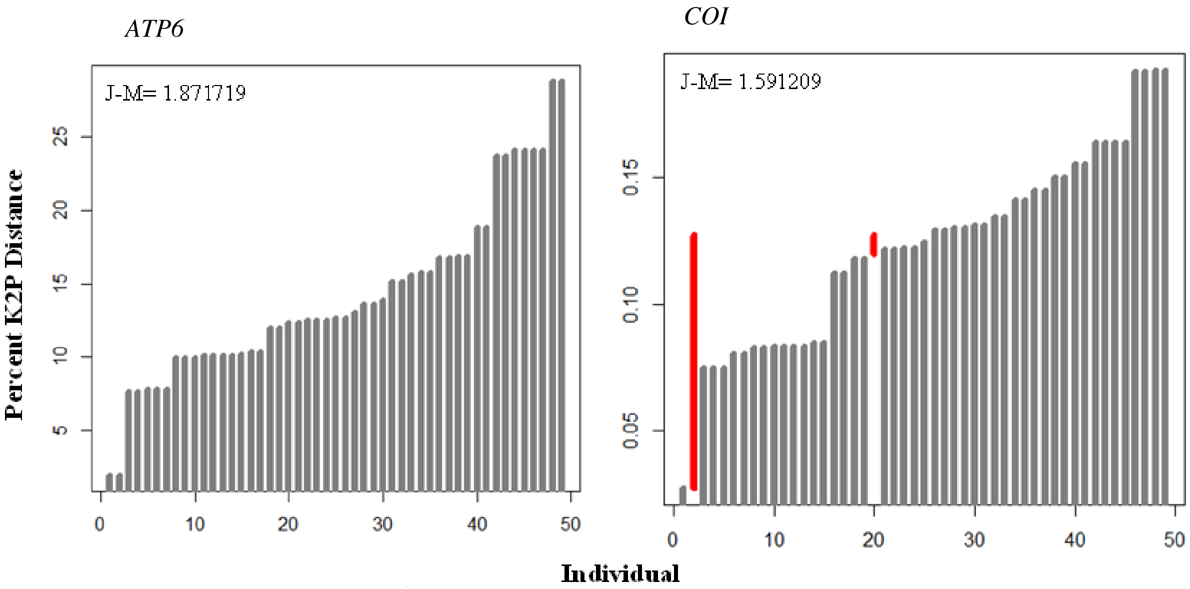


Figure 4.6 Line plot of the barcode gap for the *ATP6* and the *COI* genes. The grey lines represent the furthest intraspecific distance (bottom of line value), and the closest interspecific distance (top of the line value). The red lines show where this relationship is reversed.

The *ATP6* gene shows no overlap between the intraspecific and interspecific sequence divergences. This gap is supported by a JM value of 1.8717. The *COI* gene also had a JM value above 1.447 (Figure 4.6). However, on the graph there is a slight overlap is seen between intraspecific and interspecific sequence divergences (Figure 4.6).

The other two cytochrome oxidase subunits (*COII* and *COIII*), together with ATPase subunit 8 (*ATP8*) and cytochrome b subunit (*Cytb*), showed an overlap between the intraspecific and interspecific sequence divergences with JM values less than 1.447 (Figure 4.7). Therefore these four genes cannot be used with confidence as barcodes for identifying Diptera species.

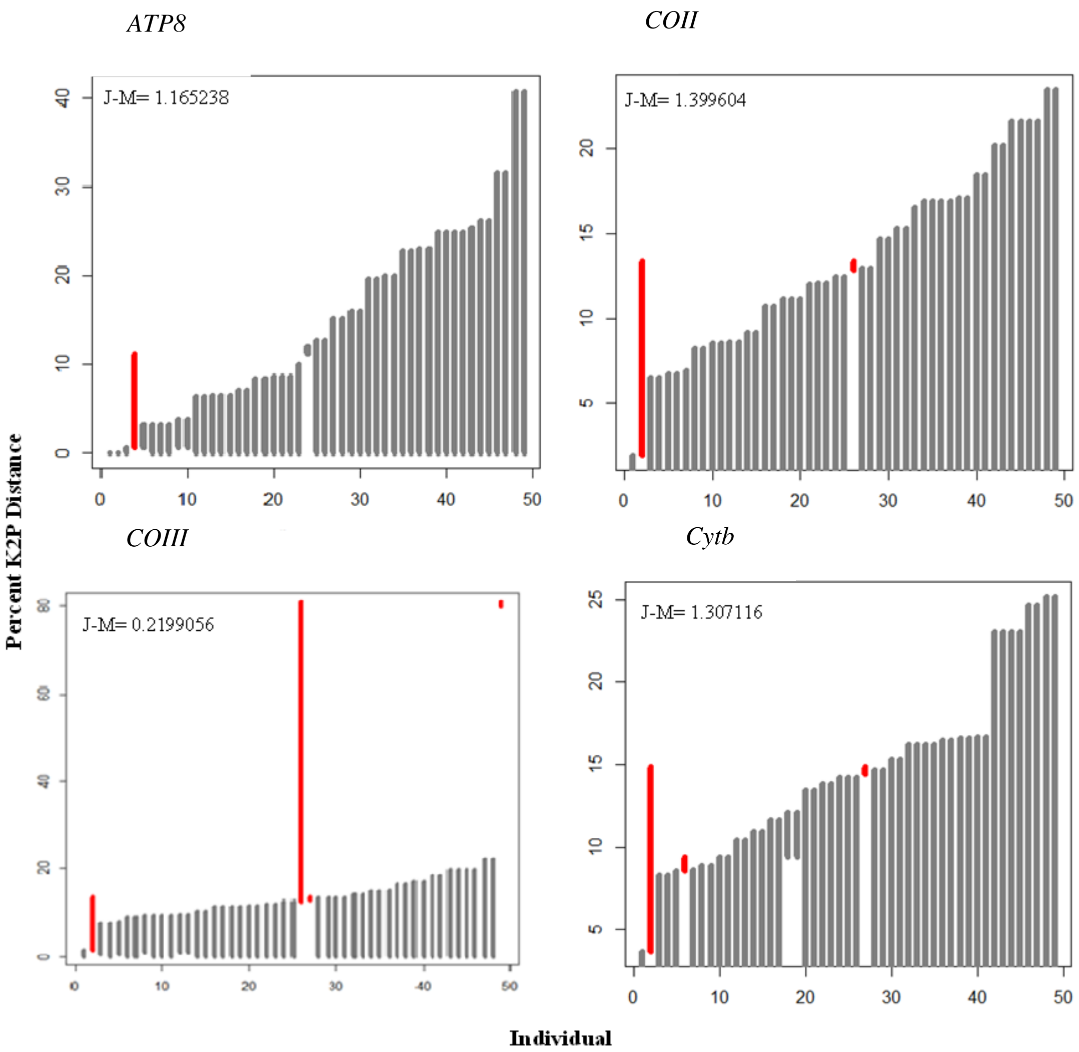


Figure 4.7 Line plot of the barcode gap for the *ATP8*, *COII*, *COIII* and *Cytb* genes. The grey lines represent the furthest intraspecific distance (bottom of line value), and the closest interspecific distance (top of the line value). The red lines show where this relationship is reversed.

The remaining seven PCGs, the NADH dehydrogenase subunits 1, 2, 3, 4, 4L, 5, and 6 (*NAD1* to *NAD6* and *NAD4L*), also showed an overlap between the intraspecific and interspecific sequence divergences with JM values less than 1.447 (Figure 4.8).

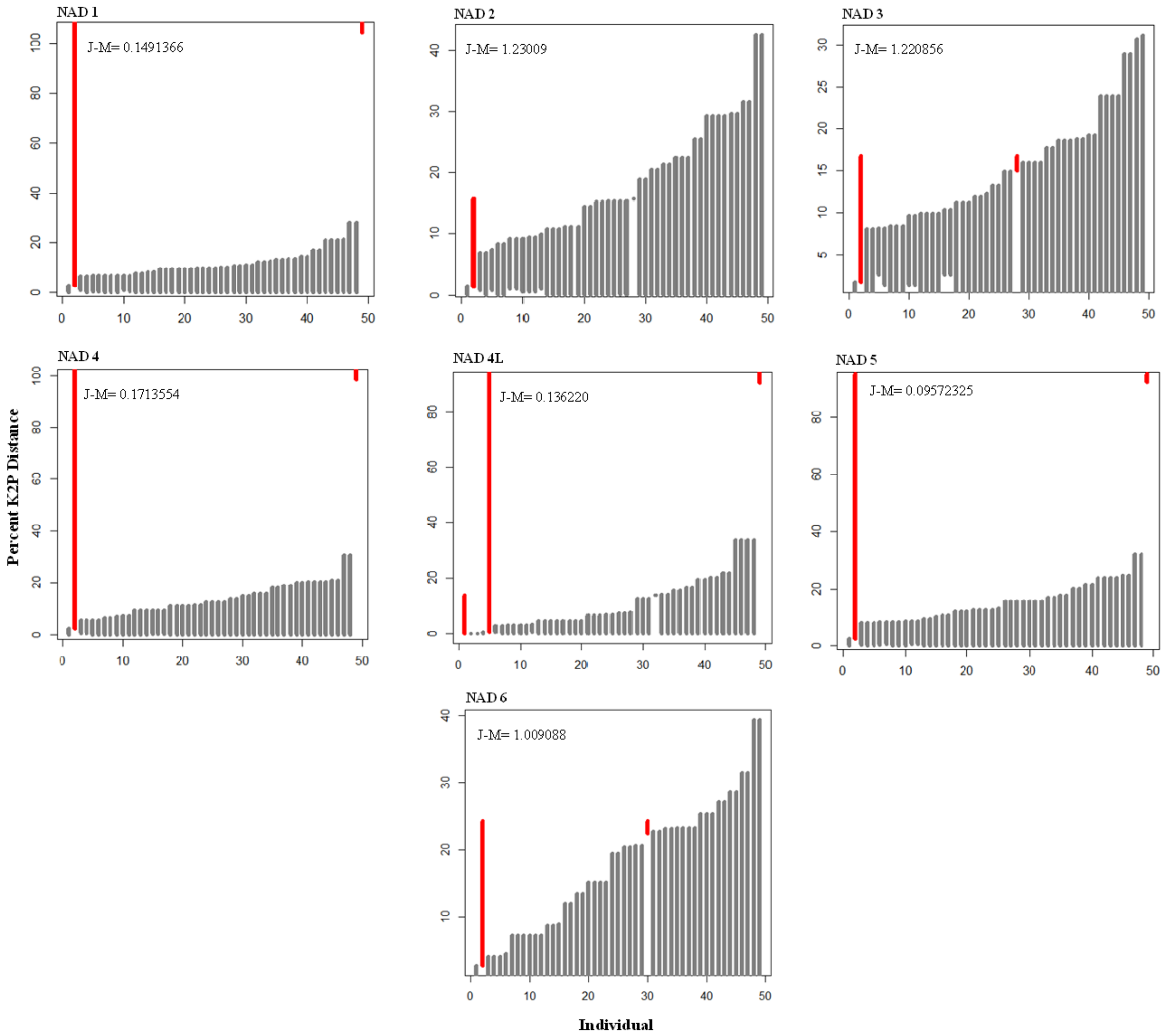


Figure 4.8 Line plot of the barcode gap for the six NADH dehydrogenase subunits (*NAD1* to *NAD6*, *NAD4L*). The grey lines represent the furthest intraspecific distance (bottom of line value), and the closest interspecific distance (top of the line value). The red lines show where this relationship is reversed.

## 4.4 Discussion

The mitochondrial protein-coding genes can be used to study the relationships between family, genera and species. These genes are classified into four groups namely: cytochrome oxidase subunits 1, 2, and 3, cytochrome b subunit, NADH dehydrogenase subunits 1, 2, 3, 4, 4L, 5, and 6 and ATPase subunits 6 and 8 (*ATP6* and *ATP8*). Amongst all the PCGs, a 658 bp *COI* gene region has been found to be the most important gene for DNA barcoding. However, the success of this gene region in identifying Diptera species is limited due to its lack of DNA barcoding gap and low identification success (See Chapter Two). The aim of this study was to identify other informative genes within the PCGs that can be used to facilitate species identification in Diptera. This was done by comparing the phylogenies retrieved by each gene as well as by testing for the presence of the barcoding gap which is essential for the success of DNA barcoding.

### 4.4.1 Phylogenetic analysis

Phylogenetic analysis of the 13 PGCs was performed using the ML and BI analyses. The ML and BI provided different levels of resolution, this was expected, since different MtDNA genes can generate different trees (Zardoya & Meyer 1996). The differences among gene trees arise mainly from short or highly conserved sequences, that may lack the phylogenetic information to detect short, deep branches that distinguish the correct phylogeny from an incorrect one (Zardoya & Meyer 1996). Accurate species identification and discovery using the tree-based methods require species to be monophyletic. Therefore, the aim of the analysis was to find the genes that produce reliable phylogenies, based on the clustering of species and the high supporting bootstrap and posterior probability values.

*COI* and *NAD2* gene recovered the expected phylogenies with strong bootstrap and posterior probabilities (Figure 4.1 and 4.4). The Diptera species clustered as expected, with species belonging to the same family forming monophyletic relationships with well-supported branches. However, there were differences between the two phylogenies produced by the different markers. The *NAD2* incorrectly placed *Lucilia sericata* and *Muscina stabulans* (Figure 4.4). Both *NAD1* and *COI* were classified by Zardoya and Meyer as good phylogenetic performers in recovering expected trees amongst phylogenetically distant relatives (Zardoya & Meyer 1996). This was also true in our analysis as these two genes performed better compared the other 11 protein-coding genes.

The *ATP6* and *COII* phylogenies were very different from the *COI* and *NAD2* phylogenies. They were able to recover some monophyletic relationships between some species including species belonging to the Drosophilidae, Cecidomyiidae, Culicidae and Agromyziidae with well supported branches. They were unable to correctly place the species *Lucilia sericata*, *Scathophaga stercoraria, Musca domestica* and *Muscina stabulan* in the correct family (Figure 4.2 and Figure 4.3).

*NAD4L* performed the worst compared to the other PCGs .The ML and BI analyses generated very different phylogenies with no monophyletic families recovered as monophyletic in both trees (Figure 4.5). Only 24 Diptera species out of the 49 Diptera species were recovered in both trees. There is an incorrect clustering of species *Lucilia cuprina* and *Lucilia sericata* in the BI tree and species *Rutilia goerlingana* and *Elodia flavipalpis* in the ML tree. However, both these clustering are not supported. Even though some correct clustering were observed in both tress such as the Drosopila genus in the ML tree (bootstrap value: 84%) and Batrocera genus in the BI tree (posterior probability: 0.95), the clustering did not include all the species of the respective genera (Figure 4.5).

The other PCGs, including the five *NAD* (1, 3, 4, 5 and 6), *ATP8*, *Cytb*, and *COIII* retrieved different phylogenies compared to the *COI* with unexpected clustering of species on the trees. Furthermore, the number of non-monophyletic species is greater and the supporting bootstrap values and posterior probabilities are smaller (Appendix 4.1-4.8).

### 4.4.2 DNA barcoding gap

DNA barcoding relies on intraspecific genetic variation being less than interspecific genetic variation. When this condition is met, a barcoding gap exists and this allows for unknown species to be identified (Meyer & Paulay 2005). The presence of a barcoding gap for each of the 13 PCGs was assessed by plotting the intraspecific and intraspecific genetic variations, as well as calculating the JM distances between the two genetic distance classes.

The *ATP6* and *COI* gene have a JM value above 1.447, which means that there is a gap between the intraspecific and interspecific distances (Figure 4.6). Therefore, these two genes can be used as potential barcodes for the identification of Diptera species. The rest of the Mt PCGs (*ATP8*, *COII, COIII*, *Cytb* and *NAD 1-6*) had a JM value less than1.447, suggesting an overlap between the intraspecific and interspecific genetic distances (Figure 4.7 and 4.8). In previous studies, the overlap between intraspecific and interspecific variation has shown to be problematic for all the mitochondrial genes in Diptera species, limiting the use of these genes for DNA barcoding (Luo *et al.* 2011).

A 658 bp long *COI* gene region is already used as a universal barcode for the identification of animal taxa, but needs to be used with caution in the identification of Diptera species. In chapter two (Figure 2.6), we saw that some Diptera families including Drosophilidae, Muscidae, Calliphoridae and Tephritidae, lack the DNA barcoding gap and only a few families such as; Asilidae and Anthomyiidae have a barcoding gap.

The *ATP6* gene on the other hand has a DNA barcoding gap which is critical for DNA barcoding. Other studies on DNA barcoding in Fungi, have also suggested the *ATP6* as a potential barcode for species identification (Vialle *et al.* 2009). Despite the presence of a barcoding gap, *ATP6* should be used with caution especially in tree-based method due to some discrepancies in the placement of taxa.

## 4.5 Conclusion

DNA barcoding using mitochondrial genes as barcodes remains problematic for the order Diptera including our South African Diptera due to lack of reliable genes. Currently a 658 bp *COI* gene region is the barcode marker of choice in DNA barcoding of animal species. However, this gene region has proven not to be reliable in identifying Diptera species due to the lack of DNA barcoding gap, and a low identification success. Furthermore, the *COI* gene region is biased by the presence of Diptera families that lack the DNA barcoding gap. The overlap between intraspecific and interspecific variation has shown to be problematic for most mitochondrial genes on the MtDNA, and this limits the use of these genes for DNA barcoding. The only gene that shows some potential characteristics for DNA barcoding is the *ATP6* gene. This gene has the barcoding gap, which is essential in DNA barcoding. However, it should be used with caution especially in tree-based methods as it can produce some discrepancies which can lead to misidentifications. More studies are still required in DNA barcoding Diptera species. These can include looking at the use of combination of genes and/or even the used of whole genome sequencing to aid species identification in this order.

## 4.6 References

Boore J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Research* **27**, 1767-80.

Brown S.D.J., Collins R.A., Boyer S., Lefort M.C., Malumbres-Olarte J., Vink C.J. & Cruickshank R.H. (2012) SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources* **12**, 562-5.

Brown W.M., Goerge M. & Wilson A.C. (1979) Rapid evolution of animal mitochondrial DNA. *PNAS* **76**, 1967-71.

Cavanaugh J. (2007) Akaike information criterion. In: (ed. by N. Salkind N), pp. 16-8. SAGE Publications, Thousand Oaks, CA.

Drummond A.J. & Rambaut A. (2007) Tracer v. 1.5. Computer program and documentation distributed by the authors at http://beast.bio.ed.ac.uk/Tracer.

Felsenstein J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *Department of Genome Sciences,* University of Washington, Seattle;.

Gerber A.S., Loggins R., Kumar S. & Dowling T.E. (2001) Does non-neutral evolution shape observed patterns of DNA variation in animal mitochondrial genome? *Annual Review of Genomics* **35**, 539-66.

Hall T.A. (1999) BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows. *Nucleic Acids Symposium Series* **41**, 95-8.

Hebert P.D.N., Cywinska A., Ball S.L. & Waard J.R.D. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society, Biological Sciences* **270**, 313-21.

Hebert P.D.N., Penton E.H., Burns J.M., Janzen D.H. & Hallwachs W. (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *PNAS* **101**, 1-6.

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S. & Francis C.M. (2004b) Identification of birds through DNA barcodes. *PLoS Biology* **2**, 1-7.

Hickson R.E., Simon C., Cooper A. & Spicer G.S. (1996) Conserved sequence motifs, alignment and secondary structure for the third domain in animal 12S rRNA. *Molecular Biology and Evolution* **13**, 150-69.

Huelsenbeck J.P. & Ronquist F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5.

Jühling F., Pütz J., Bernt M., Donath A., Middendorf M., Florentz C. & Stadler P.F. (2012) Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Research* **40**, 2833-45.

Lee W., Kim H., Lim J., Choi H.R., Kim Y., Kim Y.S., Ji J.Y., Foottit R.G. & Lee S. (2011) Barcoding aphids (Hemiptera: Aphididae) of the Korean Peninsula: updating the global data set. *Molecular Ecology Resources* **11**, 32-7.

Luo A., Zhang A., Ho S.Y.W., Xu W., Zhang Y., Shi W. & Cameron S.L. (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. *BMC Genomics* **12**, 84-.

Mandal S., Chhakchhuak L., Gurusubramanian G. & Kumar N.S. (2014) Mitochondrial markers for identification and phylogenetic studies in insects – a review.  **2**, 1-9.

Meier R., Shiyang K., Vaidya G. & Ng P.K.L. (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55**, 715-28.

Meyer C.P. & Paulay G. (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* **3**, 422-.

Moore W.S. (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees vs nuclear-gene trees. *Evolution* **49**, 718-26.

Nichols R. (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution* **16**, 358-64.

Park D.S., Foottit R., Maw E. & Hebert P.D.N. (2011) Barcoding bugs: DNA-based identification of the true bugs (Insecta: Hemiptera: Heteroptera). *PLoS ONE* **6**, e18749-e.

Posada D. (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253-6.

Qui-Hong W., Wu H., Fujihara T. & Fang S. (2004) Which genetic marker for which conservation genetic issue? *Electrophoresis* **25**, 2165-76.

Rambaut A. (2009) FigTree v. 1.3.1. Computer program and documentation distributed by the author at http://tree.bio.ed.ac.uk/software.

Smith M.A., Woodley N.E., Janzen D.H., Hallwachs W. & Hebert P.D.N. (2006) DNA barcodes reveal cryptic host specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *PNAS* **103**, 3657-62.

Thompson J.D., Higgins D.G. & Gibson T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-80.

Tumura K., Peterson D., Peterson N., Steder G., Nei M. & Kumar S. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony method. **28**, 2731-9.

Vences M., Thomas M., Van der Meijden A., Chiari Y. & Vieites O.R. (2005) Comparative perfomance of the 16S rRNA gene in DNA barcoding of amphibians. *Front Zoology* **2**, 5-.

Vialle A., Feau N., Allaire M., Didukh M., Martin F., Moncalvos J. & Hamelin R.C. (2009) Evaluation of mitochondrial genes as DNA barcode for Basidiomycota. *Molecular Ecology Resources* **9**, 99-113.

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

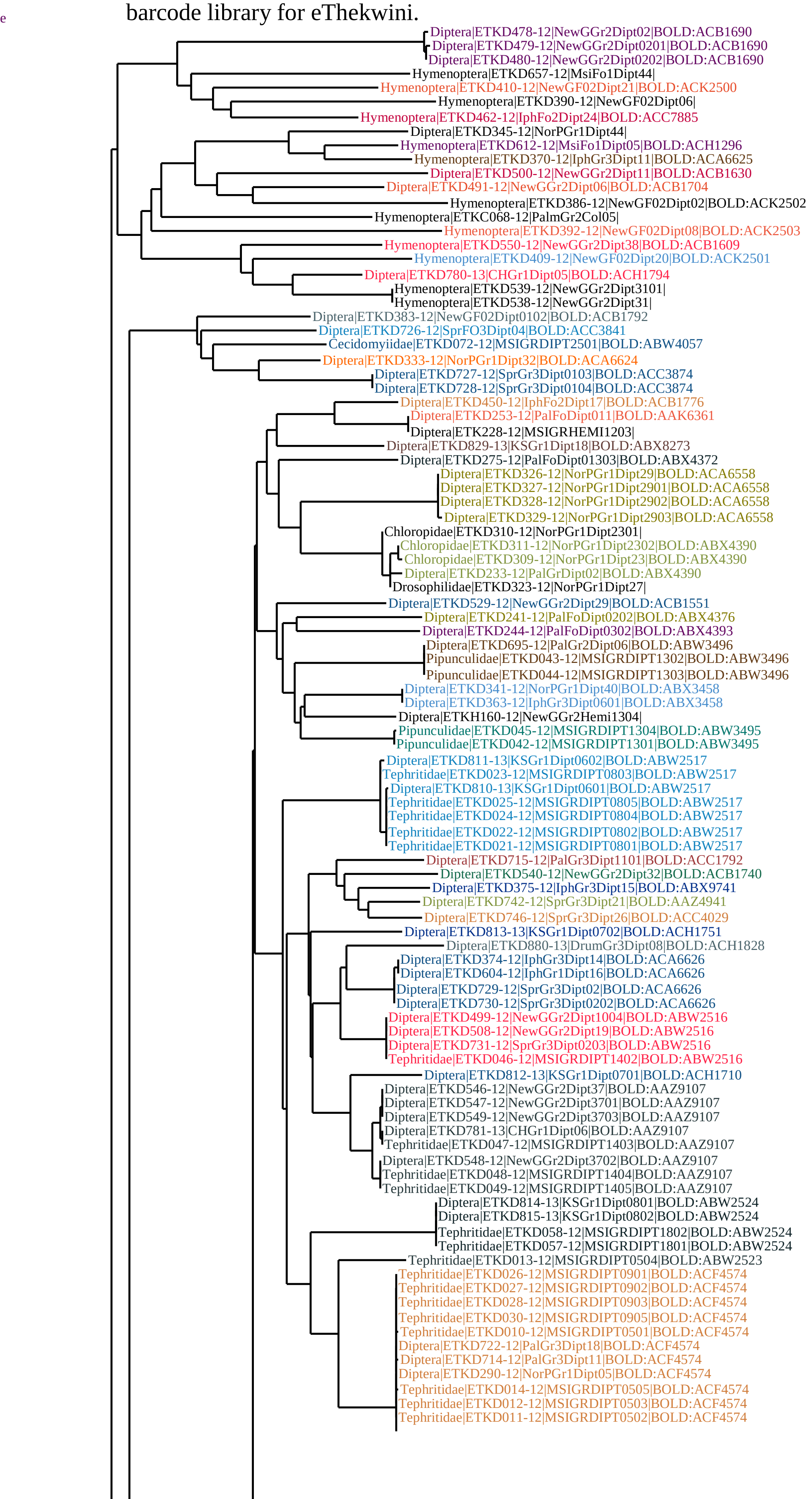
Ward R.D., Zemlak T.S., Innes B.H., Last P.R. & Hebert P.D.N. (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society, Biological sciences* **360**, 1847-57.

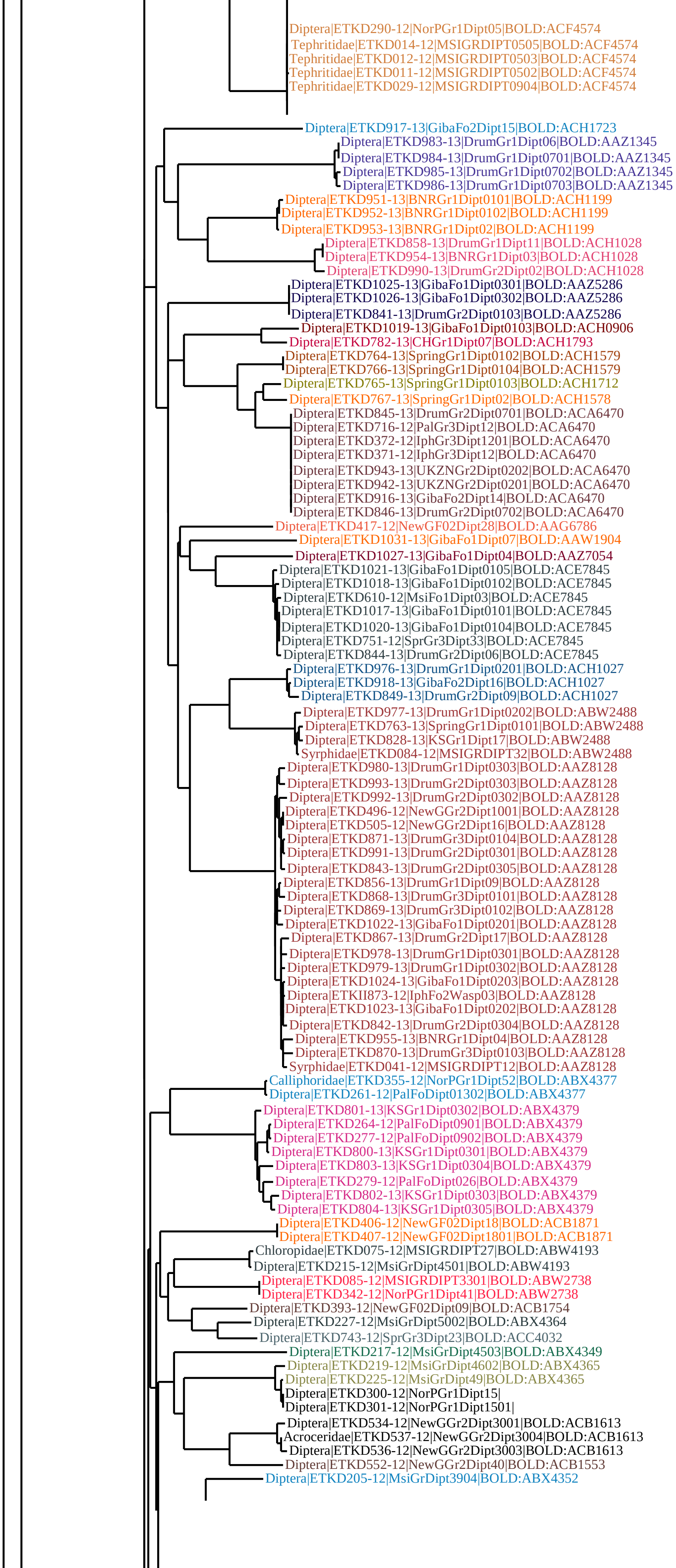
Xu J. & Singh R.S. (2005) The inheritance of organelle genes and genome: patters and mechanisms. *Genome* **48**, 951-8.

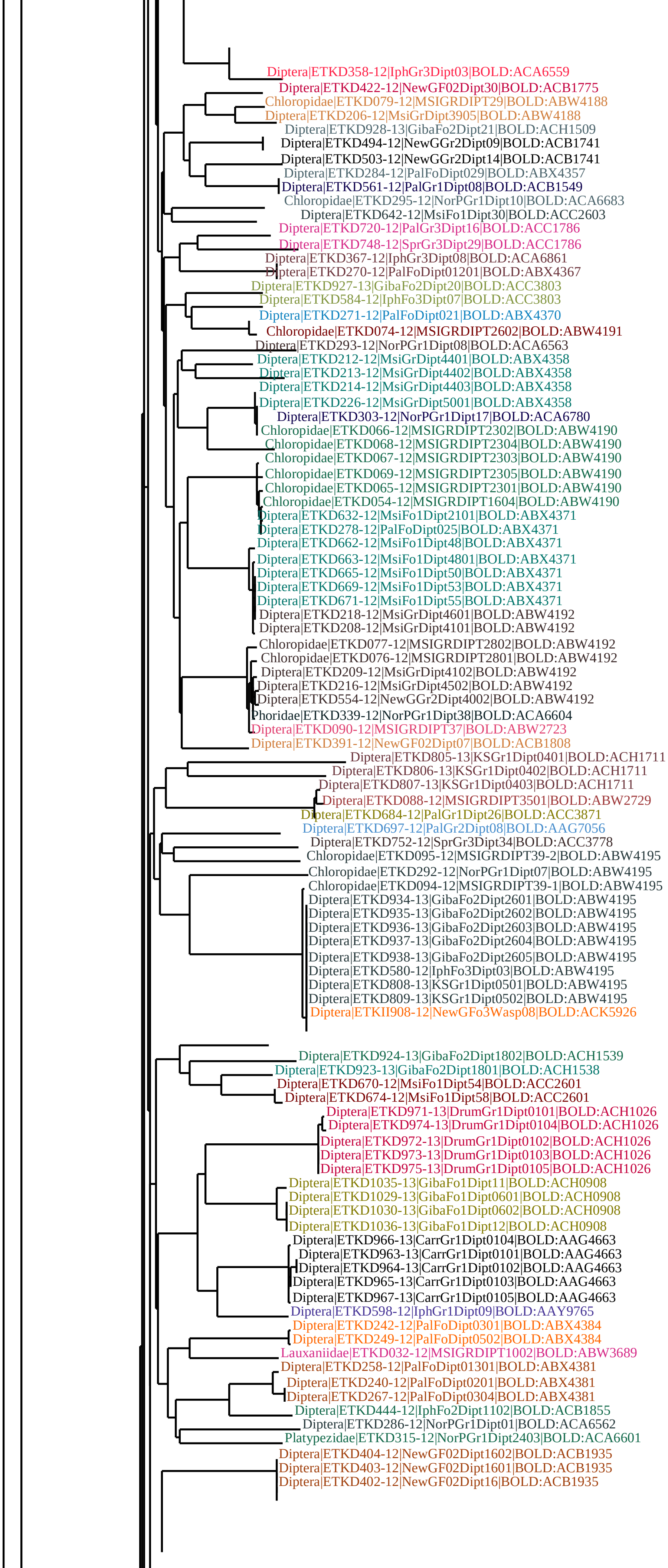
Zardoya R. & Meyer A. (1996) Phylogenetic performance of mitochondrial protein coding genes in resolving relationships amongst vertebrates. *Molecular Biology and Evolution* **13**, 933-42.

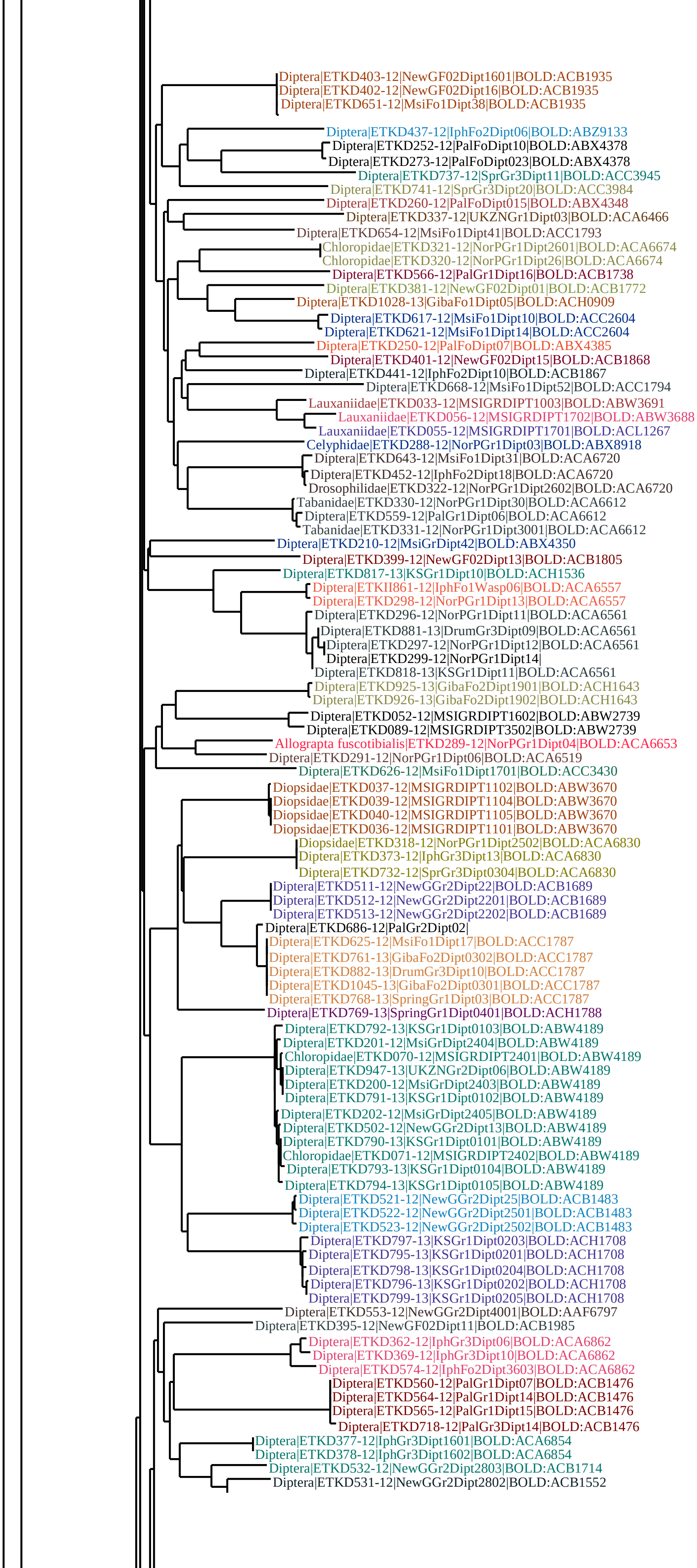
Zwickl D.J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. *Ph.D. Dissertation,* The University of Texas at Austin.

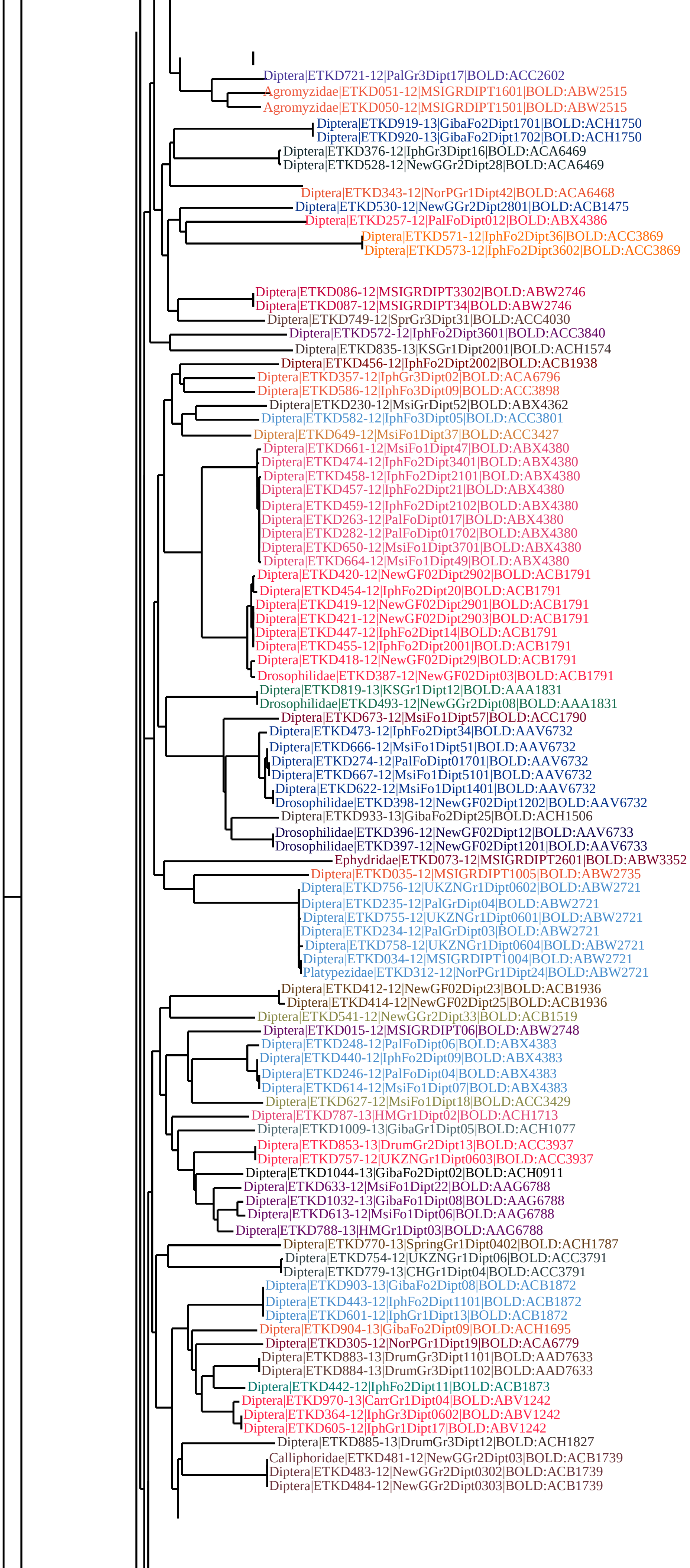
**Appendix 2.1** Neighbour joining tree (K2P) of all 844 specimens of Diptera represented in the DNA barcode library for eThekwini.

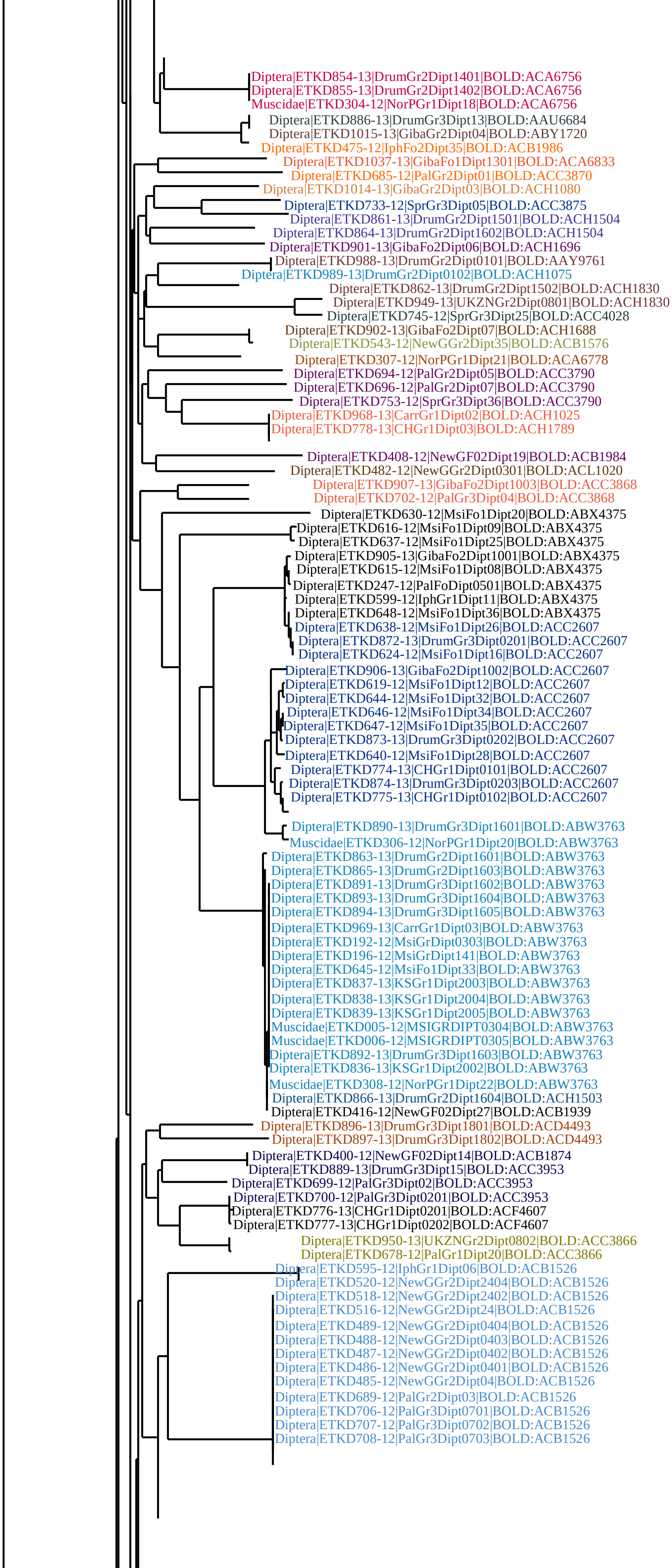

e
Diptera|ETKD478-12|NewGGr2Dipt02|BOLD:ACB1690
Diptera|ETKD479-12|NewGGr2Dipt0201|BOLD:ACB1690
Diptera|ETKD480-12|NewGGr2Dipt0202|BOLD:ACB1690
Hymenoptera|ETKD657-12|MsiFo1Dipt44|
Hymenoptera|ETKD410-12|NewGF02Dipt21|BOLD:ACK2500
Hymenoptera|ETKD390-12|NewGF02Dipt06|
Hymenoptera|ETKD462-12|IphFo2Dipt24|BOLD:ACC7885
Diptera|ETKD345-12|NorPGr1Dipt44|
Hymenoptera|ETKD612-12|MsiFo1Dipt05|BOLD:ACH1296
Hymenoptera|ETKD370-12|IphGr3Dipt11|BOLD:ACA6625
Diptera|ETKD500-12|NewGGr2Dipt11|BOLD:ACB1630
Diptera|ETKD491-12|NewGGr2Dipt06|BOLD:ACB1704
Hymenoptera|ETKD386-12|NewGF02Dipt02|BOLD:ACK2502
Hymenoptera|ETKC068-12|PalmGr2Col05|
Hymenoptera|ETKD392-12|NewGF02Dipt08|BOLD:ACK2503
Hymenoptera|ETKD550-12|NewGGr2Dipt38|BOLD:ACB1609
Hymenoptera|ETKD409-12|NewGF02Dipt20|BOLD:ACK2501
Diptera|ETKD780-13|CHGr1Dipt05|BOLD:ACH1794
Hymenoptera|ETKD539-12|NewGGr2Dipt3101|
Hymenoptera|ETKD538-12|NewGGr2Dipt31|
Diptera|ETKD383-12|NewGF02Dipt0102|BOLD:ACB1792
Diptera|ETKD726-12|SprFO3Dipt04|BOLD:ACC3841
Cecidomyiidae|ETKD072-12|MSIGRDIPT2501|BOLD:ABW4057
Diptera|ETKD333-12|NorPGr1Dipt32|BOLD:ACA6624
Diptera|ETKD727-12|SprGr3Dipt0103|BOLD:ACC3874
Diptera|ETKD728-12|SprGr3Dipt0104|BOLD:ACC3874
Diptera|ETKD450-12|IphFo2Dipt17|BOLD:ACB1776
Diptera|ETKD253-12|PalFoDipt011|BOLD:AAK6361
Diptera|ETK228-12|MSIGRHEMI1203|
Diptera|ETKD829-13|KSGr1Dipt18|BOLD:ABX8273
Diptera|ETKD275-12|PalFoDipt01303|BOLD:ABX4372
Diptera|ETKD326-12|NorPGr1Dipt29|BOLD:ACA6558
Diptera|ETKD327-12|NorPGr1Dipt2901|BOLD:ACA6558
Diptera|ETKD328-12|NorPGr1Dipt2902|BOLD:ACA6558
Diptera|ETKD329-12|NorPGr1Dipt2903|BOLD:ACA6558
Chloropidae|ETKD310-12|NorPGr1Dipt2301|
Chloropidae|ETKD311-12|NorPGr1Dipt2302|BOLD:ABX4390
Chloropidae|ETKD309-12|NorPGr1Dipt23|BOLD:ABX4390
Diptera|ETKD233-12|PalGrDipt02|BOLD:ABX4390
Drosophilidae|ETKD323-12|NorPGr1Dipt27|
Diptera|ETKD529-12|NewGGr2Dipt29|BOLD:ACB1551
Diptera|ETKD241-12|PalFoDipt0202|BOLD:ABX4376
Diptera|ETKD244-12|PalFoDipt0302|BOLD:ABX4393
Diptera|ETKD695-12|PalGr2Dipt06|BOLD:ABW3496
Pipunculidae|ETKD043-12|MSIGRDIPT1302|BOLD:ABW3496
Pipunculidae|ETKD044-12|MSIGRDIPT1303|BOLD:ABW3496
Diptera|ETKD341-12|NorPGr1Dipt40|BOLD:ABX3458
Diptera|ETKD363-12|IphGr3Dipt0601|BOLD:ABX3458
Diptera|ETKH160-12|NewGGr2Hemi1304|
Pipunculidae|ETKD045-12|MSIGRDIPT1304|BOLD:ABW3495
Pipunculidae|ETKD042-12|MSIGRDIPT1301|BOLD:ABW3495
Diptera|ETKD811-13|KSGr1Dipt0602|BOLD:ABW2517
Tephritidae|ETKD023-12|MSIGRDIPT0803|BOLD:ABW2517
Diptera|ETKD810-13|KSGr1Dipt0601|BOLD:ABW2517
Tephritidae|ETKD025-12|MSIGRDIPT0805|BOLD:ABW2517
Tephritidae|ETKD024-12|MSIGRDIPT0804|BOLD:ABW2517
Tephritidae|ETKD022-12|MSIGRDIPT0802|BOLD:ABW2517
Tephritidae|ETKD021-12|MSIGRDIPT0801|BOLD:ABW2517
Diptera|ETKD715-12|PalGr3Dipt1101|BOLD:ACC1792
Diptera|ETKD540-12|NewGGr2Dipt32|BOLD:ACB1740
Diptera|ETKD375-12|IphGr3Dipt15|BOLD:ABX9741
Diptera|ETKD742-12|SprGr3Dipt21|BOLD:AAZ4941
Diptera|ETKD746-12|SprGr3Dipt26|BOLD:ACC4029
Diptera|ETKD813-13|KSGr1Dipt0702|BOLD:ACH1751
Diptera|ETKD880-13|DrumGr3Dipt08|BOLD:ACH1828
Diptera|ETKD374-12|IphGr3Dipt14|BOLD:ACA6626
Diptera|ETKD604-12|IphGr1Dipt16|BOLD:ACA6626
Diptera|ETKD729-12|SprGr3Dipt02|BOLD:ACA6626
Diptera|ETKD730-12|SprGr3Dipt0202|BOLD:ACA6626
Diptera|ETKD499-12|NewGGr2Dipt1004|BOLD:ABW2516
Diptera|ETKD508-12|NewGGr2Dipt19|BOLD:ABW2516
Diptera|ETKD731-12|SprGr3Dipt0203|BOLD:ABW2516
Tephritidae|ETKD046-12|MSIGRDIPT1402|BOLD:ABW2516
Diptera|ETKD812-13|KSGr1Dipt0701|BOLD:ACH1710
Diptera|ETKD546-12|NewGGr2Dipt37|BOLD:AAZ9107
Diptera|ETKD547-12|NewGGr2Dipt3701|BOLD:AAZ9107
Diptera|ETKD549-12|NewGGr2Dipt3703|BOLD:AAZ9107
Diptera|ETKD781-13|CHGr1Dipt06|BOLD:AAZ9107
Tephritidae|ETKD047-12|MSIGRDIPT1403|BOLD:AAZ9107
Diptera|ETKD548-12|NewGGr2Dipt3702|BOLD:AAZ9107
Tephritidae|ETKD048-12|MSIGRDIPT1404|BOLD:AAZ9107
Tephritidae|ETKD049-12|MSIGRDIPT1405|BOLD:AAZ9107
Diptera|ETKD814-13|KSGr1Dipt0801|BOLD:ABW2524
Diptera|ETKD815-13|KSGr1Dipt0802|BOLD:ABW2524
Tephritidae|ETKD058-12|MSIGRDIPT1802|BOLD:ABW2524
Tephritidae|ETKD057-12|MSIGRDIPT1801|BOLD:ABW2524
Tephritidae|ETKD013-12|MSIGRDIPT0504|BOLD:ABW2523
Tephritidae|ETKD026-12|MSIGRDIPT0901|BOLD:ACF4574
Tephritidae|ETKD027-12|MSIGRDIPT0902|BOLD:ACF4574
Tephritidae|ETKD028-12|MSIGRDIPT0903|BOLD:ACF4574
Tephritidae|ETKD030-12|MSIGRDIPT0905|BOLD:ACF4574
Tephritidae|ETKD010-12|MSIGRDIPT0501|BOLD:ACF4574
Diptera|ETKD722-12|PalGr3Dipt18|BOLD:ACF4574
Diptera|ETKD714-12|PalGr3Dipt11|BOLD:ACF4574
Diptera|ETKD290-12|NorPGr1Dipt05|BOLD:ACF4574
Tephritidae|ETKD014-12|MSIGRDIPT0505|BOLD:ACF4574
Tephritidae|ETKD012-12|MSIGRDIPT0503|BOLD:ACF4574
Tephritidae|ETKD011-12|MSIGRDIPT0502|BOLD:ACF4574

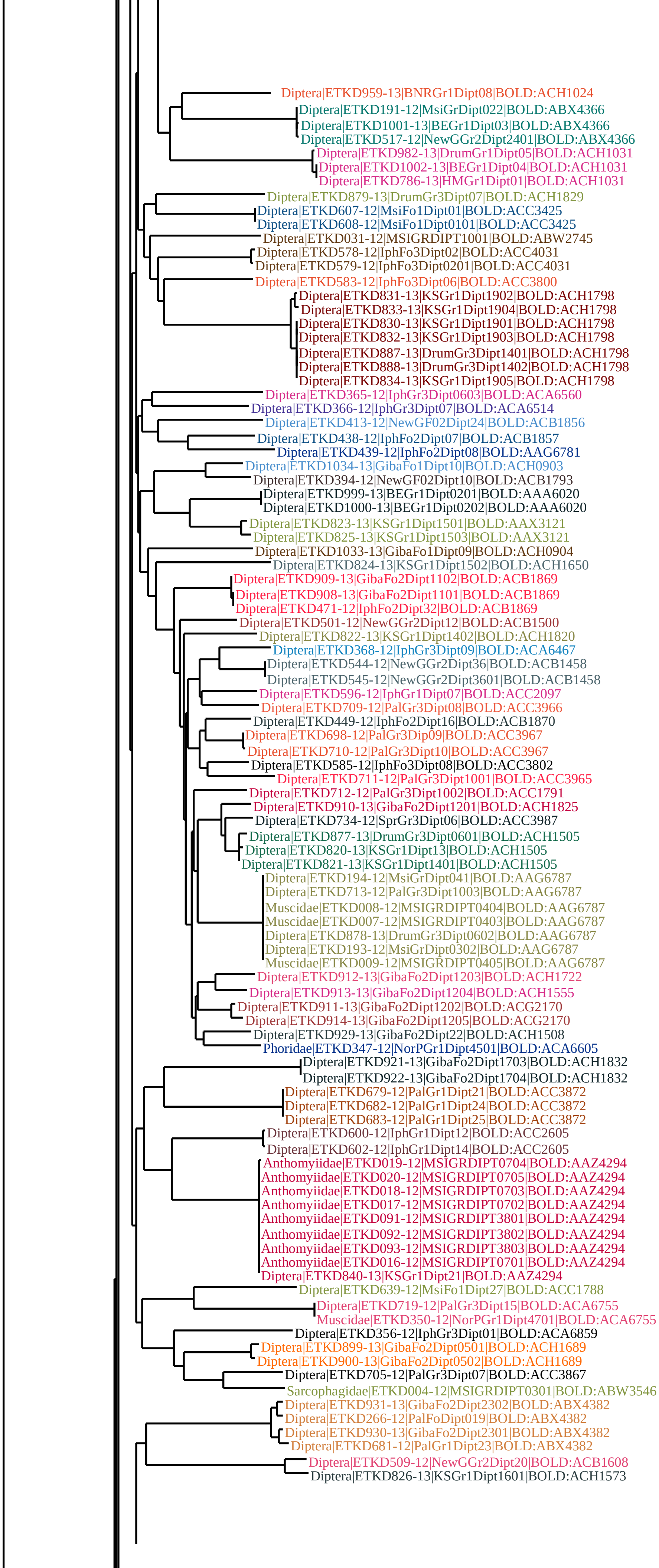

Diptera|ETKD290-12|NorPGr1Dipt05|BOLD:ACF4574
Tephritidae|ETKD014-12|MSIGRDIPT0505|BOLD:ACF4574
Tephritidae|ETKD012-12|MSIGRDIPT0503|BOLD:ACF4574
Tephritidae|ETKD011-12|MSIGRDIPT0502|BOLD:ACF4574
Tephritidae|ETKD029-12|MSIGRDIPT0904|BOLD:ACF4574
Diptera|ETKD917-13|GibaFo2Dipt15|BOLD:ACH1723
Diptera|ETKD983-13|DrumGr1Dipt06|BOLD:AAZ1345
Diptera|ETKD984-13|DrumGr1Dipt0701|BOLD:AAZ1345
Diptera|ETKD985-13|DrumGr1Dipt0702|BOLD:AAZ1345
Diptera|ETKD986-13|DrumGr1Dipt0703|BOLD:AAZ1345
Diptera|ETKD951-13|BNRGr1Dipt0101|BOLD:ACH1199
Diptera|ETKD952-13|BNRGr1Dipt0102|BOLD:ACH1199
Diptera|ETKD953-13|BNRGr1Dipt02|BOLD:ACH1199
Diptera|ETKD858-13|DrumGr1Dipt11|BOLD:ACH1028
Diptera|ETKD954-13|BNRGr1Dipt03|BOLD:ACH1028
Diptera|ETKD990-13|DrumGr2Dipt02|BOLD:ACH1028
Diptera|ETKD1025-13|GibaFo1Dipt0301|BOLD:AAZ5286
Diptera|ETKD1026-13|GibaFo1Dipt0302|BOLD:AAZ5286
Diptera|ETKD841-13|DrumGr2Dipt0103|BOLD:AAZ5286
Diptera|ETKD1019-13|GibaFo1Dipt0103|BOLD:ACH0906
Diptera|ETKD782-13|CHGr1Dipt07|BOLD:ACH1793
Diptera|ETKD764-13|SpringGr1Dipt0102|BOLD:ACH1579
Diptera|ETKD766-13|SpringGr1Dipt0104|BOLD:ACH1579
Diptera|ETKD765-13|SpringGr1Dipt0103|BOLD:ACH1712
Diptera|ETKD767-13|SpringGr1Dipt02|BOLD:ACH1578
Diptera|ETKD845-13|DrumGr2Dipt0701|BOLD:ACA6470
Diptera|ETKD716-12|PalGr3Dipt12|BOLD:ACA6470
Diptera|ETKD372-12|IphGr3Dipt1201|BOLD:ACA6470
Diptera|ETKD371-12|IphGr3Dipt12|BOLD:ACA6470
Diptera|ETKD943-13|UKZNGr2Dipt0202|BOLD:ACA6470
Diptera|ETKD942-13|UKZNGr2Dipt0201|BOLD:ACA6470
Diptera|ETKD916-13|GibaFo2Dipt14|BOLD:ACA6470
Diptera|ETKD846-13|DrumGr2Dipt0702|BOLD:ACA6470
Diptera|ETKD417-12|NewGF02Dipt28|BOLD:AAG6786
Diptera|ETKD1031-13|GibaFo1Dipt07|BOLD:AAW1904
Diptera|ETKD1027-13|GibaFo1Dipt04|BOLD:AAZ7054
Diptera|ETKD1021-13|GibaFo1Dipt0105|BOLD:ACE7845
Diptera|ETKD1018-13|GibaFo1Dipt0102|BOLD:ACE7845
Diptera|ETKD610-12|MsiFo1Dipt03|BOLD:ACE7845
Diptera|ETKD1017-13|GibaFo1Dipt0101|BOLD:ACE7845
Diptera|ETKD1020-13|GibaFo1Dipt0104|BOLD:ACE7845
Diptera|ETKD751-12|SprGr3Dipt33|BOLD:ACE7845
Diptera|ETKD844-13|DrumGr2Dipt06|BOLD:ACE7845
Diptera|ETKD976-13|DrumGr1Dipt0201|BOLD:ACH1027
Diptera|ETKD918-13|GibaFo2Dipt16|BOLD:ACH1027
Diptera|ETKD849-13|DrumGr2Dipt09|BOLD:ACH1027
Diptera|ETKD977-13|DrumGr1Dipt0202|BOLD:ABW2488
Diptera|ETKD763-13|SpringGr1Dipt0101|BOLD:ABW2488
Diptera|ETKD828-13|KSGr1Dipt17|BOLD:ABW2488
Syrphidae|ETKD084-12|MSIGRDIPT32|BOLD:ABW2488
Diptera|ETKD980-13|DrumGr1Dipt0303|BOLD:AAZ8128
Diptera|ETKD993-13|DrumGr2Dipt0303|BOLD:AAZ8128
Diptera|ETKD992-13|DrumGr2Dipt0302|BOLD:AAZ8128
Diptera|ETKD496-12|NewGGr2Dipt1001|BOLD:AAZ8128
Diptera|ETKD505-12|NewGGr2Dipt16|BOLD:AAZ8128
Diptera|ETKD871-13|DrumGr3Dipt0104|BOLD:AAZ8128
Diptera|ETKD991-13|DrumGr2Dipt0301|BOLD:AAZ8128
Diptera|ETKD843-13|DrumGr2Dipt0305|BOLD:AAZ8128
Diptera|ETKD856-13|DrumGr1Dipt09|BOLD:AAZ8128
Diptera|ETKD868-13|DrumGr3Dipt0101|BOLD:AAZ8128
Diptera|ETKD869-13|DrumGr3Dipt0102|BOLD:AAZ8128
Diptera|ETKD1022-13|GibaFo1Dipt0201|BOLD:AAZ8128
Diptera|ETKD867-13|DrumGr2Dipt17|BOLD:AAZ8128
Diptera|ETKD978-13|DrumGr1Dipt0301|BOLD:AAZ8128
Diptera|ETKD979-13|DrumGr1Dipt0302|BOLD:AAZ8128
Diptera|ETKD1024-13|GibaFo1Dipt0203|BOLD:AAZ8128
Diptera|ETKII873-12|IphFo2Wasp03|BOLD:AAZ8128
Diptera|ETKD1023-13|GibaFo1Dipt0202|BOLD:AAZ8128
Diptera|ETKD842-13|DrumGr2Dipt0304|BOLD:AAZ8128
Diptera|ETKD955-13|BNRGr1Dipt04|BOLD:AAZ8128
Diptera|ETKD870-13|DrumGr3Dipt0103|BOLD:AAZ8128
Syrphidae|ETKD041-12|MSIGRDIPT12|BOLD:AAZ8128
Calliphoridae|ETKD355-12|NorPGr1Dipt52|BOLD:ABX4377
Diptera|ETKD261-12|PalFoDipt01302|BOLD:ABX4377
Diptera|ETKD801-13|KSGr1Dipt0302|BOLD:ABX4379
Diptera|ETKD264-12|PalFoDipt0901|BOLD:ABX4379
Diptera|ETKD277-12|PalFoDipt0902|BOLD:ABX4379
Diptera|ETKD800-13|KSGr1Dipt0301|BOLD:ABX4379
Diptera|ETKD803-13|KSGr1Dipt0304|BOLD:ABX4379
Diptera|ETKD279-12|PalFoDipt026|BOLD:ABX4379
Diptera|ETKD802-13|KSGr1Dipt0303|BOLD:ABX4379
Diptera|ETKD804-13|KSGr1Dipt0305|BOLD:ABX4379
Diptera|ETKD406-12|NewGF02Dipt18|BOLD:ACB1871
Diptera|ETKD407-12|NewGF02Dipt1801|BOLD:ACB1871
Chloropidae|ETKD075-12|MSIGRDIPT27|BOLD:ABW4193
Diptera|ETKD215-12|MsiGrDipt4501|BOLD:ABW4193
Diptera|ETKD085-12|MSIGRDIPT3301|BOLD:ABW2738
Diptera|ETKD342-12|NorPGr1Dipt41|BOLD:ABW2738
Diptera|ETKD393-12|NewGF02Dipt09|BOLD:ACB1754
Diptera|ETKD227-12|MsiGrDipt5002|BOLD:ABX4364
Diptera|ETKD743-12|SprGr3Dipt23|BOLD:ACC4032
Diptera|ETKD217-12|MsiGrDipt4503|BOLD:ABX4349
Diptera|ETKD219-12|MsiGrDipt4602|BOLD:ABX4365
Diptera|ETKD225-12|MsiGrDipt49|BOLD:ABX4365
Diptera|ETKD300-12|NorPGr1Dipt15|
Diptera|ETKD301-12|NorPGr1Dipt1501|
Diptera|ETKD534-12|NewGGr2Dipt3001|BOLD:ACB1613
Acroceridae|ETKD537-12|NewGGr2Dipt3004|BOLD:ACB1613
Diptera|ETKD536-12|NewGGr2Dipt3003|BOLD:ACB1613
Diptera|ETKD552-12|NewGGr2Dipt40|BOLD:ACB1553
Diptera|ETKD205-12|MsiGrDipt3904|BOLD:ABX4352

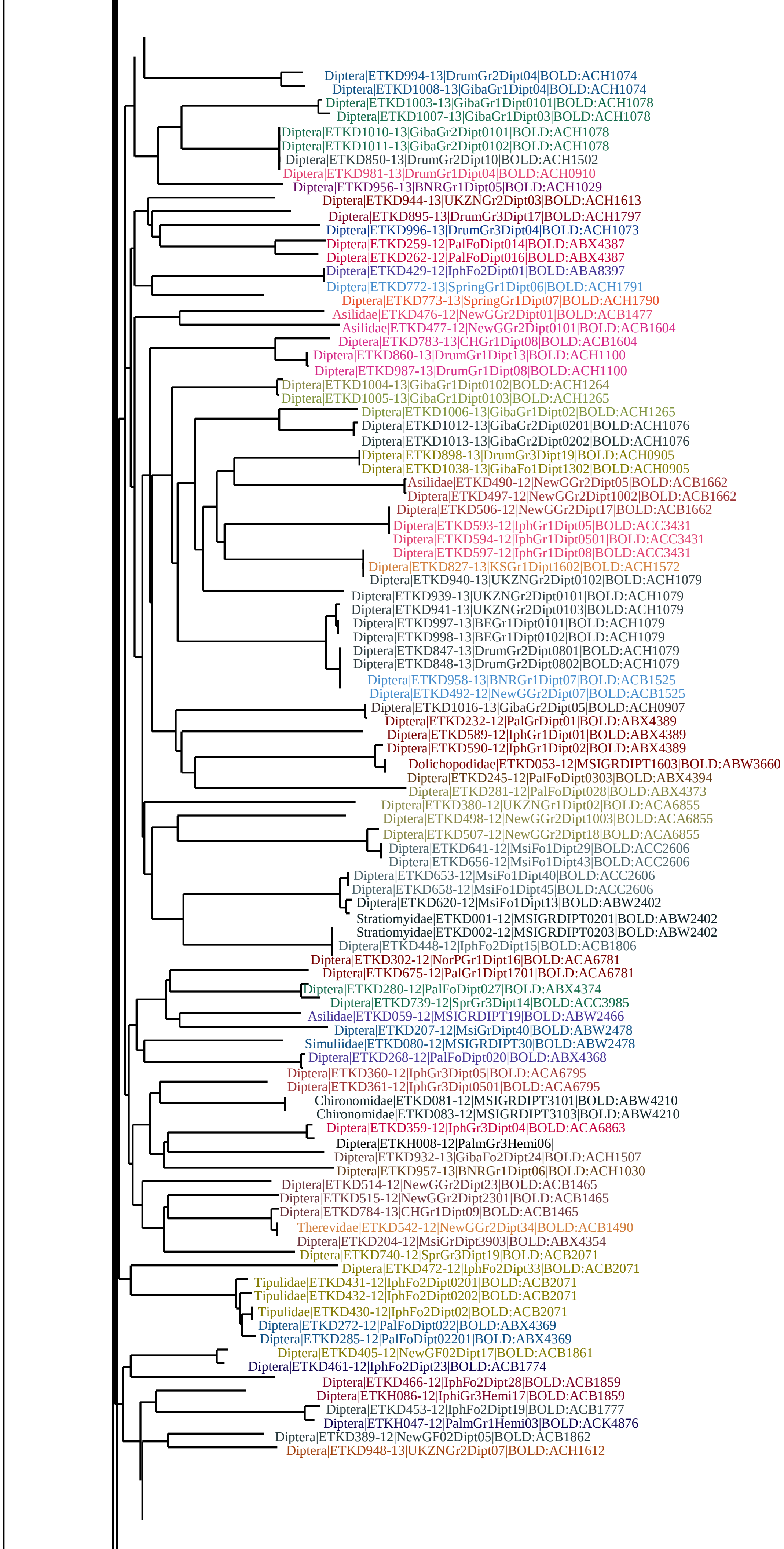

Diptera|ETKD358-12|IphGr3Dipt03|BOLD:ACA6559
Diptera|ETKD422-12|NewGF02Dipt30|BOLD:ACB1775
Chloropidae|ETKD079-12|MSIGRDIPT29|BOLD:ABW4188
Diptera|ETKD206-12|MsiGrDipt3905|BOLD:ABW4188
Diptera|ETKD928-13|GibaFo2Dipt21|BOLD:ACH1509
Diptera|ETKD494-12|NewGGr2Dipt09|BOLD:ACB1741
Diptera|ETKD503-12|NewGGr2Dipt14|BOLD:ACB1741
Diptera|ETKD284-12|PalFoDipt029|BOLD:ABX4357
Diptera|ETKD561-12|PalGr1Dipt08|BOLD:ACB1549
Chloropidae|ETKD295-12|NorPGr1Dipt10|BOLD:ACA6683
Diptera|ETKD642-12|MsiFo1Dipt30|BOLD:ACC2603
Diptera|ETKD720-12|PalGr3Dipt16|BOLD:ACC1786
Diptera|ETKD748-12|SprGr3Dipt29|BOLD:ACC1786
Diptera|ETKD367-12|IphGr3Dipt08|BOLD:ACA6861
Diptera|ETKD270-12|PalFoDipt01201|BOLD:ABX4367
Diptera|ETKD927-13|GibaFo2Dipt20|BOLD:ACC3803
Diptera|ETKD584-12|IphFo3Dipt07|BOLD:ACC3803
Diptera|ETKD271-12|PalFoDipt021|BOLD:ABX4370
Chloropidae|ETKD074-12|MSIGRDIPT2602|BOLD:ABW4191
Diptera|ETKD293-12|NorPGr1Dipt08|BOLD:ACA6563
Diptera|ETKD212-12|MsiGrDipt4401|BOLD:ABX4358
Diptera|ETKD213-12|MsiGrDipt4402|BOLD:ABX4358
Diptera|ETKD214-12|MsiGrDipt4403|BOLD:ABX4358
Diptera|ETKD226-12|MsiGrDipt5001|BOLD:ABX4358
Diptera|ETKD303-12|NorPGr1Dipt17|BOLD:ACA6780
Chloropidae|ETKD066-12|MSIGRDIPT2302|BOLD:ABW4190
Chloropidae|ETKD068-12|MSIGRDIPT2304|BOLD:ABW4190
Chloropidae|ETKD067-12|MSIGRDIPT2303|BOLD:ABW4190
Chloropidae|ETKD069-12|MSIGRDIPT2305|BOLD:ABW4190
Chloropidae|ETKD065-12|MSIGRDIPT2301|BOLD:ABW4190
Chloropidae|ETKD054-12|MSIGRDIPT1604|BOLD:ABW4190
Diptera|ETKD632-12|MsiFo1Dipt2101|BOLD:ABX4371
Diptera|ETKD278-12|PalFoDipt025|BOLD:ABX4371
Diptera|ETKD662-12|MsiFo1Dipt48|BOLD:ABX4371
Diptera|ETKD663-12|MsiFo1Dipt4801|BOLD:ABX4371
Diptera|ETKD665-12|MsiFo1Dipt50|BOLD:ABX4371
Diptera|ETKD669-12|MsiFo1Dipt53|BOLD:ABX4371
Diptera|ETKD671-12|MsiFo1Dipt55|BOLD:ABX4371
Diptera|ETKD218-12|MsiGrDipt4601|BOLD:ABW4192
Diptera|ETKD208-12|MsiGrDipt4101|BOLD:ABW4192
Chloropidae|ETKD077-12|MSIGRDIPT2802|BOLD:ABW4192
Chloropidae|ETKD076-12|MSIGRDIPT2801|BOLD:ABW4192
Diptera|ETKD209-12|MsiGrDipt4102|BOLD:ABW4192
Diptera|ETKD216-12|MsiGrDipt4502|BOLD:ABW4192
Diptera|ETKD554-12|NewGGr2Dipt4002|BOLD:ABW4192
Phoridae|ETKD339-12|NorPGr1Dipt38|BOLD:ACA6604
Diptera|ETKD090-12|MSIGRDIPT37|BOLD:ABW2723
Diptera|ETKD391-12|NewGF02Dipt07|BOLD:ACB1808
Diptera|ETKD805-13|KSGr1Dipt0401|BOLD:ACH1711
Diptera|ETKD806-13|KSGr1Dipt0402|BOLD:ACH1711
Diptera|ETKD807-13|KSGr1Dipt0403|BOLD:ACH1711
Diptera|ETKD088-12|MSIGRDIPT3501|BOLD:ABW2729
Diptera|ETKD684-12|PalGr1Dipt26|BOLD:ACC3871
Diptera|ETKD697-12|PalGr2Dipt08|BOLD:AAG7056
Diptera|ETKD752-12|SprGr3Dipt34|BOLD:ACC3778
Chloropidae|ETKD095-12|MSIGRDIPT39-2|BOLD:ABW4195
Chloropidae|ETKD292-12|NorPGr1Dipt07|BOLD:ABW4195
Chloropidae|ETKD094-12|MSIGRDIPT39-1|BOLD:ABW4195
Diptera|ETKD934-13|GibaFo2Dipt2601|BOLD:ABW4195
Diptera|ETKD935-13|GibaFo2Dipt2602|BOLD:ABW4195
Diptera|ETKD936-13|GibaFo2Dipt2603|BOLD:ABW4195
Diptera|ETKD937-13|GibaFo2Dipt2604|BOLD:ABW4195
Diptera|ETKD938-13|GibaFo2Dipt2605|BOLD:ABW4195
Diptera|ETKD580-12|IphFo3Dipt03|BOLD:ABW4195
Diptera|ETKD808-13|KSGr1Dipt0501|BOLD:ABW4195
Diptera|ETKD809-13|KSGr1Dipt0502|BOLD:ABW4195
Diptera|ETKII908-12|NewGFo3Wasp08|BOLD:ACK5926
Diptera|ETKD924-13|GibaFo2Dipt1802|BOLD:ACH1539
Diptera|ETKD923-13|GibaFo2Dipt1801|BOLD:ACH1538
Diptera|ETKD670-12|MsiFo1Dipt54|BOLD:ACC2601
Diptera|ETKD674-12|MsiFo1Dipt58|BOLD:ACC2601
Diptera|ETKD971-13|DrumGr1Dipt0101|BOLD:ACH1026
Diptera|ETKD974-13|DrumGr1Dipt0104|BOLD:ACH1026
Diptera|ETKD972-13|DrumGr1Dipt0102|BOLD:ACH1026
Diptera|ETKD973-13|DrumGr1Dipt0103|BOLD:ACH1026
Diptera|ETKD975-13|DrumGr1Dipt0105|BOLD:ACH1026
Diptera|ETKD1035-13|GibaFo1Dipt11|BOLD:ACH0908
Diptera|ETKD1029-13|GibaFo1Dipt0601|BOLD:ACH0908
Diptera|ETKD1030-13|GibaFo1Dipt0602|BOLD:ACH0908
Diptera|ETKD1036-13|GibaFo1Dipt12|BOLD:ACH0908
Diptera|ETKD966-13|CarrGr1Dipt0104|BOLD:AAG4663
Diptera|ETKD963-13|CarrGr1Dipt0101|BOLD:AAG4663
Diptera|ETKD964-13|CarrGr1Dipt0102|BOLD:AAG4663
Diptera|ETKD965-13|CarrGr1Dipt0103|BOLD:AAG4663
Diptera|ETKD967-13|CarrGr1Dipt0105|BOLD:AAG4663
Diptera|ETKD598-12|IphGr1Dipt09|BOLD:AAY9765
Diptera|ETKD242-12|PalFoDipt0301|BOLD:ABX4384
Diptera|ETKD249-12|PalFoDipt0502|BOLD:ABX4384
Lauxaniidae|ETKD032-12|MSIGRDIPT1002|BOLD:ABW3689
Diptera|ETKD258-12|PalFoDipt01301|BOLD:ABX4381
Diptera|ETKD240-12|PalFoDipt0201|BOLD:ABX4381
Diptera|ETKD267-12|PalFoDipt0304|BOLD:ABX4381
Diptera|ETKD444-12|IphFo2Dipt1102|BOLD:ACB1855
Diptera|ETKD286-12|NorPGr1Dipt01|BOLD:ACA6562
Platypezidae|ETKD315-12|NorPGr1Dipt2403|BOLD:ACA6601
Diptera|ETKD404-12|NewGF02Dipt1602|BOLD:ACB1935
Diptera|ETKD403-12|NewGF02Dipt1601|BOLD:ACB1935
Diptera|ETKD402-12|NewGF02Dipt16|BOLD:ACB1935

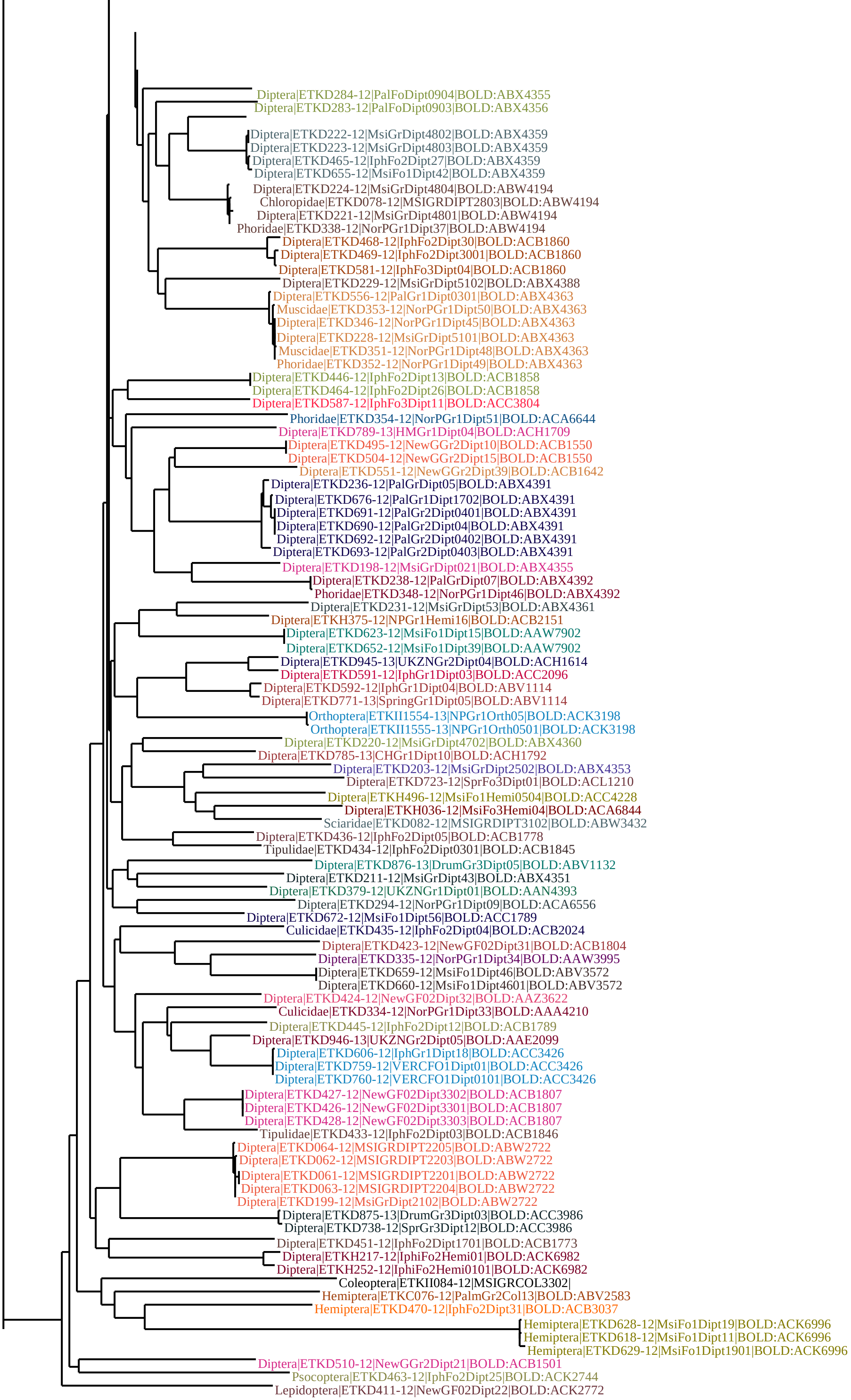

Diptera|ETKD403-12|NewGF02Dipt1601|BOLD:ACB1935
Diptera|ETKD402-12|NewGF02Dipt16|BOLD:ACB1935
Diptera|ETKD651-12|MsiFo1Dipt38|BOLD:ACB1935
Diptera|ETKD437-12|IphFo2Dipt06|BOLD:ABZ9133
Diptera|ETKD252-12|PalFoDipt10|BOLD:ABX4378
Diptera|ETKD273-12|PalFoDipt023|BOLD:ABX4378
Diptera|ETKD737-12|SprGr3Dipt11|BOLD:ACC3945
Diptera|ETKD741-12|SprGr3Dipt20|BOLD:ACC3984
Diptera|ETKD260-12|PalFoDipt015|BOLD:ABX4348
Diptera|ETKD337-12|UKZNGr1Dipt03|BOLD:ACA6466
Diptera|ETKD654-12|MsiFo1Dipt41|BOLD:ACC1793
Chloropidae|ETKD321-12|NorPGr1Dipt2601|BOLD:ACA6674
Chloropidae|ETKD320-12|NorPGr1Dipt26|BOLD:ACA6674
Diptera|ETKD566-12|PalGr1Dipt16|BOLD:ACB1738
Diptera|ETKD381-12|NewGF02Dipt01|BOLD:ACB1772
Diptera|ETKD1028-13|GibaFo1Dipt05|BOLD:ACH0909
Diptera|ETKD617-12|MsiFo1Dipt10|BOLD:ACC2604
Diptera|ETKD621-12|MsiFo1Dipt14|BOLD:ACC2604
Diptera|ETKD250-12|PalFoDipt07|BOLD:ABX4385
Diptera|ETKD401-12|NewGF02Dipt15|BOLD:ACB1868
Diptera|ETKD441-12|IphFo2Dipt10|BOLD:ACB1867
Diptera|ETKD668-12|MsiFo1Dipt52|BOLD:ACC1794
Lauxaniidae|ETKD033-12|MSIGRDIPT1003|BOLD:ABW3691
Lauxaniidae|ETKD056-12|MSIGRDIPT1702|BOLD:ABW3688
Lauxaniidae|ETKD055-12|MSIGRDIPT1701|BOLD:ACL1267
Celyphidae|ETKD288-12|NorPGr1Dipt03|BOLD:ABX8918
Diptera|ETKD643-12|MsiFo1Dipt31|BOLD:ACA6720
Diptera|ETKD452-12|IphFo2Dipt18|BOLD:ACA6720
Drosophilidae|ETKD322-12|NorPGr1Dipt2602|BOLD:ACA6720
Tabanidae|ETKD330-12|NorPGr1Dipt30|BOLD:ACA6612
Diptera|ETKD559-12|PalGr1Dipt06|BOLD:ACA6612
Tabanidae|ETKD331-12|NorPGr1Dipt3001|BOLD:ACA6612
Diptera|ETKD210-12|MsiGrDipt42|BOLD:ABX4350
Diptera|ETKD399-12|NewGF02Dipt13|BOLD:ACB1805
Diptera|ETKD817-13|KSGr1Dipt10|BOLD:ACH1536
Diptera|ETKII861-12|IphFo1Wasp06|BOLD:ACA6557
Diptera|ETKD298-12|NorPGr1Dipt13|BOLD:ACA6557
Diptera|ETKD296-12|NorPGr1Dipt11|BOLD:ACA6561
Diptera|ETKD881-13|DrumGr3Dipt09|BOLD:ACA6561
Diptera|ETKD297-12|NorPGr1Dipt12|BOLD:ACA6561
Diptera|ETKD299-12|NorPGr1Dipt14|
Diptera|ETKD818-13|KSGr1Dipt11|BOLD:ACA6561
Diptera|ETKD925-13|GibaFo2Dipt1901|BOLD:ACH1643
Diptera|ETKD926-13|GibaFo2Dipt1902|BOLD:ACH1643
Diptera|ETKD052-12|MSIGRDIPT1602|BOLD:ABW2739
Diptera|ETKD089-12|MSIGRDIPT3502|BOLD:ABW2739
Allograpta fuscotibialis|ETKD289-12|NorPGr1Dipt04|BOLD:ACA6653
Diptera|ETKD291-12|NorPGr1Dipt06|BOLD:ACA6519
Diptera|ETKD626-12|MsiFo1Dipt1701|BOLD:ACC3430
Diopsidae|ETKD037-12|MSIGRDIPT1102|BOLD:ABW3670
Diopsidae|ETKD039-12|MSIGRDIPT1104|BOLD:ABW3670
Diopsidae|ETKD040-12|MSIGRDIPT1105|BOLD:ABW3670
Diopsidae|ETKD036-12|MSIGRDIPT1101|BOLD:ABW3670
Diopsidae|ETKD318-12|NorPGr1Dipt2502|BOLD:ACA6830
Diptera|ETKD373-12|IphGr3Dipt13|BOLD:ACA6830
Diptera|ETKD732-12|SprGr3Dipt0304|BOLD:ACA6830
Diptera|ETKD511-12|NewGGr2Dipt22|BOLD:ACB1689
Diptera|ETKD512-12|NewGGr2Dipt2201|BOLD:ACB1689
Diptera|ETKD513-12|NewGGr2Dipt2202|BOLD:ACB1689
Diptera|ETKD686-12|PalGr2Dipt02|
Diptera|ETKD625-12|MsiFo1Dipt17|BOLD:ACC1787
Diptera|ETKD761-13|GibaFo2Dipt0302|BOLD:ACC1787
Diptera|ETKD882-13|DrumGr3Dipt10|BOLD:ACC1787
Diptera|ETKD1045-13|GibaFo2Dipt0301|BOLD:ACC1787
Diptera|ETKD768-13|SpringGr1Dipt03|BOLD:ACC1787
Diptera|ETKD769-13|SpringGr1Dipt0401|BOLD:ACH1788
Diptera|ETKD792-13|KSGr1Dipt0103|BOLD:ABW4189
Diptera|ETKD201-12|MsiGrDipt2404|BOLD:ABW4189
Chloropidae|ETKD070-12|MSIGRDIPT2401|BOLD:ABW4189
Diptera|ETKD947-13|UKZNGr2Dipt06|BOLD:ABW4189
Diptera|ETKD200-12|MsiGrDipt2403|BOLD:ABW4189
Diptera|ETKD791-13|KSGr1Dipt0102|BOLD:ABW4189
Diptera|ETKD202-12|MsiGrDipt2405|BOLD:ABW4189
Diptera|ETKD502-12|NewGGr2Dipt13|BOLD:ABW4189
Diptera|ETKD790-13|KSGr1Dipt0101|BOLD:ABW4189
Chloropidae|ETKD071-12|MSIGRDIPT2402|BOLD:ABW4189
Diptera|ETKD793-13|KSGr1Dipt0104|BOLD:ABW4189
Diptera|ETKD794-13|KSGr1Dipt0105|BOLD:ABW4189
Diptera|ETKD521-12|NewGGr2Dipt25|BOLD:ACB1483
Diptera|ETKD522-12|NewGGr2Dipt2501|BOLD:ACB1483
Diptera|ETKD523-12|NewGGr2Dipt2502|BOLD:ACB1483
Diptera|ETKD797-13|KSGr1Dipt0203|BOLD:ACH1708
Diptera|ETKD795-13|KSGr1Dipt0201|BOLD:ACH1708
Diptera|ETKD798-13|KSGr1Dipt0204|BOLD:ACH1708
Diptera|ETKD796-13|KSGr1Dipt0202|BOLD:ACH1708
Diptera|ETKD799-13|KSGr1Dipt0205|BOLD:ACH1708
Diptera|ETKD553-12|NewGGr2Dipt4001|BOLD:AAF6797
Diptera|ETKD395-12|NewGF02Dipt11|BOLD:ACB1985
Diptera|ETKD362-12|IphGr3Dipt06|BOLD:ACA6862
Diptera|ETKD369-12|IphGr3Dipt10|BOLD:ACA6862
Diptera|ETKD574-12|IphFo2Dipt3603|BOLD:ACA6862
Diptera|ETKD560-12|PalGr1Dipt07|BOLD:ACB1476
Diptera|ETKD564-12|PalGr1Dipt14|BOLD:ACB1476
Diptera|ETKD565-12|PalGr1Dipt15|BOLD:ACB1476
Diptera|ETKD718-12|PalGr3Dipt14|BOLD:ACB1476
Diptera|ETKD377-12|IphGr3Dipt1601|BOLD:ACA6854
Diptera|ETKD378-12|IphGr3Dipt1602|BOLD:ACA6854
Diptera|ETKD532-12|NewGGr2Dipt2803|BOLD:ACB1714
Diptera|ETKD531-12|NewGGr2Dipt2802|BOLD:ACB1552


Diptera|ETKD721-12|PalGr3Dipt17|BOLD:ACC2602
Agromyzidae|ETKD051-12|MSIGRDIPT1601|BOLD:ABW2515
Agromyzidae|ETKD050-12|MSIGRDIPT1501|BOLD:ABW2515
Diptera|ETKD919-13|GibaFo2Dipt1701|BOLD:ACH1750
Diptera|ETKD920-13|GibaFo2Dipt1702|BOLD:ACH1750
Diptera|ETKD376-12|IphGr3Dipt16|BOLD:ACA6469
Diptera|ETKD528-12|NewGGr2Dipt28|BOLD:ACA6469
Diptera|ETKD343-12|NorPGr1Dipt42|BOLD:ACA6468
Diptera|ETKD530-12|NewGGr2Dipt2801|BOLD:ACB1475
Diptera|ETKD257-12|PalFoDipt012|BOLD:ABX4386
Diptera|ETKD571-12|IphFo2Dipt36|BOLD:ACC3869
Diptera|ETKD573-12|IphFo2Dipt3602|BOLD:ACC3869
Diptera|ETKD086-12|MSIGRDIPT3302|BOLD:ABW2746
Diptera|ETKD087-12|MSIGRDIPT34|BOLD:ABW2746
Diptera|ETKD749-12|SprGr3Dipt31|BOLD:ACC4030
Diptera|ETKD572-12|IphFo2Dipt3601|BOLD:ACC3840
Diptera|ETKD835-13|KSGr1Dipt2001|BOLD:ACH1574
Diptera|ETKD456-12|IphFo2Dipt2002|BOLD:ACB1938
Diptera|ETKD357-12|IphGr3Dipt02|BOLD:ACA6796
Diptera|ETKD586-12|IphFo3Dipt09|BOLD:ACC3898
Diptera|ETKD230-12|MsiGrDipt52|BOLD:ABX4362
Diptera|ETKD582-12|IphFo3Dipt05|BOLD:ACC3801
Diptera|ETKD649-12|MsiFo1Dipt37|BOLD:ACC3427
Diptera|ETKD661-12|MsiFo1Dipt47|BOLD:ABX4380
Diptera|ETKD474-12|IphFo2Dipt3401|BOLD:ABX4380
Diptera|ETKD458-12|IphFo2Dipt2101|BOLD:ABX4380
Diptera|ETKD457-12|IphFo2Dipt21|BOLD:ABX4380
Diptera|ETKD459-12|IphFo2Dipt2102|BOLD:ABX4380
Diptera|ETKD263-12|PalFoDipt017|BOLD:ABX4380
Diptera|ETKD282-12|PalFoDipt01702|BOLD:ABX4380
Diptera|ETKD650-12|MsiFo1Dipt3701|BOLD:ABX4380
Diptera|ETKD664-12|MsiFo1Dipt49|BOLD:ABX4380
Diptera|ETKD420-12|NewGF02Dipt2902|BOLD:ACB1791
Diptera|ETKD454-12|IphFo2Dipt20|BOLD:ACB1791
Diptera|ETKD419-12|NewGF02Dipt2901|BOLD:ACB1791
Diptera|ETKD421-12|NewGF02Dipt2903|BOLD:ACB1791
Diptera|ETKD447-12|IphFo2Dipt14|BOLD:ACB1791
Diptera|ETKD455-12|IphFo2Dipt2001|BOLD:ACB1791
Diptera|ETKD418-12|NewGF02Dipt29|BOLD:ACB1791
Drosophilidae|ETKD387-12|NewGF02Dipt03|BOLD:ACB1791
Diptera|ETKD819-13|KSGr1Dipt12|BOLD:AAA1831
Drosophilidae|ETKD493-12|NewGGr2Dipt08|BOLD:AAA1831
Diptera|ETKD673-12|MsiFo1Dipt57|BOLD:ACC1790
Diptera|ETKD473-12|IphFo2Dipt34|BOLD:AAV6732
Diptera|ETKD666-12|MsiFo1Dipt51|BOLD:AAV6732
Diptera|ETKD274-12|PalFoDipt01701|BOLD:AAV6732
Diptera|ETKD667-12|MsiFo1Dipt5101|BOLD:AAV6732
Diptera|ETKD622-12|MsiFo1Dipt1401|BOLD:AAV6732
Drosophilidae|ETKD398-12|NewGF02Dipt1202|BOLD:AAV6732
Diptera|ETKD933-13|GibaFo2Dipt25|BOLD:ACH1506
Drosophilidae|ETKD396-12|NewGF02Dipt12|BOLD:AAV6733
Drosophilidae|ETKD397-12|NewGF02Dipt1201|BOLD:AAV6733
Ephydridae|ETKD073-12|MSIGRDIPT2601|BOLD:ABW3352
Diptera|ETKD035-12|MSIGRDIPT1005|BOLD:ABW2735
Diptera|ETKD756-12|UKZNGr1Dipt0602|BOLD:ABW2721
Diptera|ETKD235-12|PalGrDipt04|BOLD:ABW2721
Diptera|ETKD755-12|UKZNGr1Dipt0601|BOLD:ABW2721
Diptera|ETKD234-12|PalGrDipt03|BOLD:ABW2721
Diptera|ETKD758-12|UKZNGr1Dipt0604|BOLD:ABW2721
Diptera|ETKD034-12|MSIGRDIPT1004|BOLD:ABW2721
Platypezidae|ETKD312-12|NorPGr1Dipt24|BOLD:ABW2721
Diptera|ETKD412-12|NewGF02Dipt23|BOLD:ACB1936
Diptera|ETKD414-12|NewGF02Dipt25|BOLD:ACB1936
Diptera|ETKD541-12|NewGGr2Dipt33|BOLD:ACB1519
Diptera|ETKD015-12|MSIGRDIPT06|BOLD:ABW2748
Diptera|ETKD248-12|PalFoDipt06|BOLD:ABX4383
Diptera|ETKD440-12|IphFo2Dipt09|BOLD:ABX4383
Diptera|ETKD246-12|PalFoDipt04|BOLD:ABX4383
Diptera|ETKD614-12|MsiFo1Dipt07|BOLD:ABX4383
Diptera|ETKD627-12|MsiFo1Dipt18|BOLD:ACC3429
Diptera|ETKD787-13|HMGr1Dipt02|BOLD:ACH1713
Diptera|ETKD1009-13|GibaGr1Dipt05|BOLD:ACH1077
Diptera|ETKD853-13|DrumGr2Dipt13|BOLD:ACC3937
Diptera|ETKD757-12|UKZNGr1Dipt0603|BOLD:ACC3937
Diptera|ETKD1044-13|GibaFo2Dipt02|BOLD:ACH0911
Diptera|ETKD633-12|MsiFo1Dipt22|BOLD:AAG6788
Diptera|ETKD1032-13|GibaFo1Dipt08|BOLD:AAG6788
Diptera|ETKD613-12|MsiFo1Dipt06|BOLD:AAG6788
Diptera|ETKD788-13|HMGr1Dipt03|BOLD:AAG6788
Diptera|ETKD770-13|SpringGr1Dipt0402|BOLD:ACH1787
Diptera|ETKD754-12|UKZNGr1Dipt06|BOLD:ACC3791
Diptera|ETKD779-13|CHGr1Dipt04|BOLD:ACC3791
Diptera|ETKD903-13|GibaFo2Dipt08|BOLD:ACB1872
Diptera|ETKD443-12|IphFo2Dipt1101|BOLD:ACB1872
Diptera|ETKD601-12|IphGr1Dipt13|BOLD:ACB1872
Diptera|ETKD904-13|GibaFo2Dipt09|BOLD:ACH1695
Diptera|ETKD305-12|NorPGr1Dipt19|BOLD:ACA6779
Diptera|ETKD883-13|DrumGr3Dipt1101|BOLD:AAD7633
Diptera|ETKD884-13|DrumGr3Dipt1102|BOLD:AAD7633
Diptera|ETKD442-12|IphFo2Dipt11|BOLD:ACB1873
Diptera|ETKD970-13|CarrGr1Dipt04|BOLD:ABV1242
Diptera|ETKD364-12|IphGr3Dipt0602|BOLD:ABV1242
Diptera|ETKD605-12|IphGr1Dipt17|BOLD:ABV1242
Diptera|ETKD885-13|DrumGr3Dipt12|BOLD:ACH1827
Calliphoridae|ETKD481-12|NewGGr2Dipt03|BOLD:ACB1739
Diptera|ETKD483-12|NewGGr2Dipt0302|BOLD:ACB1739
Diptera|ETKD484-12|NewGGr2Dipt0303|BOLD:ACB1739


Diptera|ETKD854-13|DrumGr2Dipt1401|BOLD:ACA6756
Diptera|ETKD855-13|DrumGr2Dipt1402|BOLD:ACA6756
Muscidae|ETKD304-12|NorPGr1Dipt18|BOLD:ACA6756
Diptera|ETKD886-13|DrumGr3Dipt13|BOLD:AAU6684
Diptera|ETKD1015-13|GibaGr2Dipt04|BOLD:ABY1720
Diptera|ETKD475-12|IphFo2Dipt35|BOLD:ACB1986
Diptera|ETKD1037-13|GibaFo1Dipt1301|BOLD:ACA6833
Diptera|ETKD685-12|PalGr2Dipt01|BOLD:ACC3870
Diptera|ETKD1014-13|GibaGr2Dipt03|BOLD:ACH1080
Diptera|ETKD733-12|SprGr3Dipt05|BOLD:ACC3875
Diptera|ETKD861-13|DrumGr2Dipt1501|BOLD:ACH1504
Diptera|ETKD864-13|DrumGr2Dipt1602|BOLD:ACH1504
Diptera|ETKD901-13|GibaFo2Dipt06|BOLD:ACH1696
Diptera|ETKD988-13|DrumGr2Dipt0101|BOLD:AAY9761
Diptera|ETKD989-13|DrumGr2Dipt0102|BOLD:ACH1075
Diptera|ETKD862-13|DrumGr2Dipt1502|BOLD:ACH1830
Diptera|ETKD949-13|UKZNGr2Dipt0801|BOLD:ACH1830
Diptera|ETKD745-12|SprGr3Dipt25|BOLD:ACC4028
Diptera|ETKD902-13|GibaFo2Dipt07|BOLD:ACH1688
Diptera|ETKD543-12|NewGGr2Dipt35|BOLD:ACB1576
Diptera|ETKD307-12|NorPGr1Dipt21|BOLD:ACA6778
Diptera|ETKD694-12|PalGr2Dipt05|BOLD:ACC3790
Diptera|ETKD696-12|PalGr2Dipt07|BOLD:ACC3790
Diptera|ETKD753-12|SprGr3Dipt36|BOLD:ACC3790
Diptera|ETKD968-13|CarrGr1Dipt02|BOLD:ACH1025
Diptera|ETKD778-13|CHGr1Dipt03|BOLD:ACH1789
Diptera|ETKD408-12|NewGF02Dipt19|BOLD:ACB1984
Diptera|ETKD482-12|NewGGr2Dipt0301|BOLD:ACL1020
Diptera|ETKD907-13|GibaFo2Dipt1003|BOLD:ACC3868
Diptera|ETKD702-12|PalGr3Dipt04|BOLD:ACC3868
Diptera|ETKD630-12|MsiFo1Dipt20|BOLD:ABX4375
Diptera|ETKD616-12|MsiFo1Dipt09|BOLD:ABX4375
Diptera|ETKD637-12|MsiFo1Dipt25|BOLD:ABX4375
Diptera|ETKD905-13|GibaFo2Dipt1001|BOLD:ABX4375
Diptera|ETKD615-12|MsiFo1Dipt08|BOLD:ABX4375
Diptera|ETKD247-12|PalFoDipt0501|BOLD:ABX4375
Diptera|ETKD599-12|IphGr1Dipt11|BOLD:ABX4375
Diptera|ETKD648-12|MsiFo1Dipt36|BOLD:ABX4375
Diptera|ETKD638-12|MsiFo1Dipt26|BOLD:ACC2607
Diptera|ETKD872-13|DrumGr3Dipt0201|BOLD:ACC2607
Diptera|ETKD624-12|MsiFo1Dipt16|BOLD:ACC2607
Diptera|ETKD906-13|GibaFo2Dipt1002|BOLD:ACC2607
Diptera|ETKD619-12|MsiFo1Dipt12|BOLD:ACC2607
Diptera|ETKD644-12|MsiFo1Dipt32|BOLD:ACC2607
Diptera|ETKD646-12|MsiFo1Dipt34|BOLD:ACC2607
Diptera|ETKD647-12|MsiFo1Dipt35|BOLD:ACC2607
Diptera|ETKD873-13|DrumGr3Dipt0202|BOLD:ACC2607
Diptera|ETKD640-12|MsiFo1Dipt28|BOLD:ACC2607
Diptera|ETKD774-13|CHGr1Dipt0101|BOLD:ACC2607
Diptera|ETKD874-13|DrumGr3Dipt0203|BOLD:ACC2607
Diptera|ETKD775-13|CHGr1Dipt0102|BOLD:ACC2607
Diptera|ETKD890-13|DrumGr3Dipt1601|BOLD:ABW3763
Muscidae|ETKD306-12|NorPGr1Dipt20|BOLD:ABW3763
Diptera|ETKD863-13|DrumGr2Dipt1601|BOLD:ABW3763
Diptera|ETKD865-13|DrumGr2Dipt1603|BOLD:ABW3763
Diptera|ETKD891-13|DrumGr3Dipt1602|BOLD:ABW3763
Diptera|ETKD893-13|DrumGr3Dipt1604|BOLD:ABW3763
Diptera|ETKD894-13|DrumGr3Dipt1605|BOLD:ABW3763
Diptera|ETKD969-13|CarrGr1Dipt03|BOLD:ABW3763
Diptera|ETKD192-12|MsiGrDipt0303|BOLD:ABW3763
Diptera|ETKD196-12|MsiGrDipt141|BOLD:ABW3763
Diptera|ETKD645-12|MsiFo1Dipt33|BOLD:ABW3763
Diptera|ETKD837-13|KSGr1Dipt2003|BOLD:ABW3763
Diptera|ETKD838-13|KSGr1Dipt2004|BOLD:ABW3763
Diptera|ETKD839-13|KSGr1Dipt2005|BOLD:ABW3763
Muscidae|ETKD005-12|MSIGRDIPT0304|BOLD:ABW3763
Muscidae|ETKD006-12|MSIGRDIPT0305|BOLD:ABW3763
Diptera|ETKD892-13|DrumGr3Dipt1603|BOLD:ABW3763
Diptera|ETKD836-13|KSGr1Dipt2002|BOLD:ABW3763
Muscidae|ETKD308-12|NorPGr1Dipt22|BOLD:ABW3763
Diptera|ETKD866-13|DrumGr2Dipt1604|BOLD:ACH1503
Diptera|ETKD416-12|NewGF02Dipt27|BOLD:ACB1939
Diptera|ETKD896-13|DrumGr3Dipt1801|BOLD:ACD4493
Diptera|ETKD897-13|DrumGr3Dipt1802|BOLD:ACD4493
Diptera|ETKD400-12|NewGF02Dipt14|BOLD:ACB1874
Diptera|ETKD889-13|DrumGr3Dipt15|BOLD:ACC3953
Diptera|ETKD699-12|PalGr3Dipt02|BOLD:ACC3953
Diptera|ETKD700-12|PalGr3Dipt0201|BOLD:ACC3953
Diptera|ETKD776-13|CHGr1Dipt0201|BOLD:ACF4607
Diptera|ETKD777-13|CHGr1Dipt0202|BOLD:ACF4607
Diptera|ETKD950-13|UKZNGr2Dipt0802|BOLD:ACC3866
Diptera|ETKD678-12|PalGr1Dipt20|BOLD:ACC3866
Diptera|ETKD595-12|IphGr1Dipt06|BOLD:ACB1526
Diptera|ETKD520-12|NewGGr2Dipt2404|BOLD:ACB1526
Diptera|ETKD518-12|NewGGr2Dipt2402|BOLD:ACB1526
Diptera|ETKD516-12|NewGGr2Dipt24|BOLD:ACB1526
Diptera|ETKD489-12|NewGGr2Dipt0404|BOLD:ACB1526
Diptera|ETKD488-12|NewGGr2Dipt0403|BOLD:ACB1526
Diptera|ETKD487-12|NewGGr2Dipt0402|BOLD:ACB1526
Diptera|ETKD486-12|NewGGr2Dipt0401|BOLD:ACB1526
Diptera|ETKD485-12|NewGGr2Dipt04|BOLD:ACB1526
Diptera|ETKD689-12|PalGr2Dipt03|BOLD:ACB1526
Diptera|ETKD706-12|PalGr3Dipt0701|BOLD:ACB1526
Diptera|ETKD707-12|PalGr3Dipt0702|BOLD:ACB1526
Diptera|ETKD708-12|PalGr3Dipt0703|BOLD:ACB1526

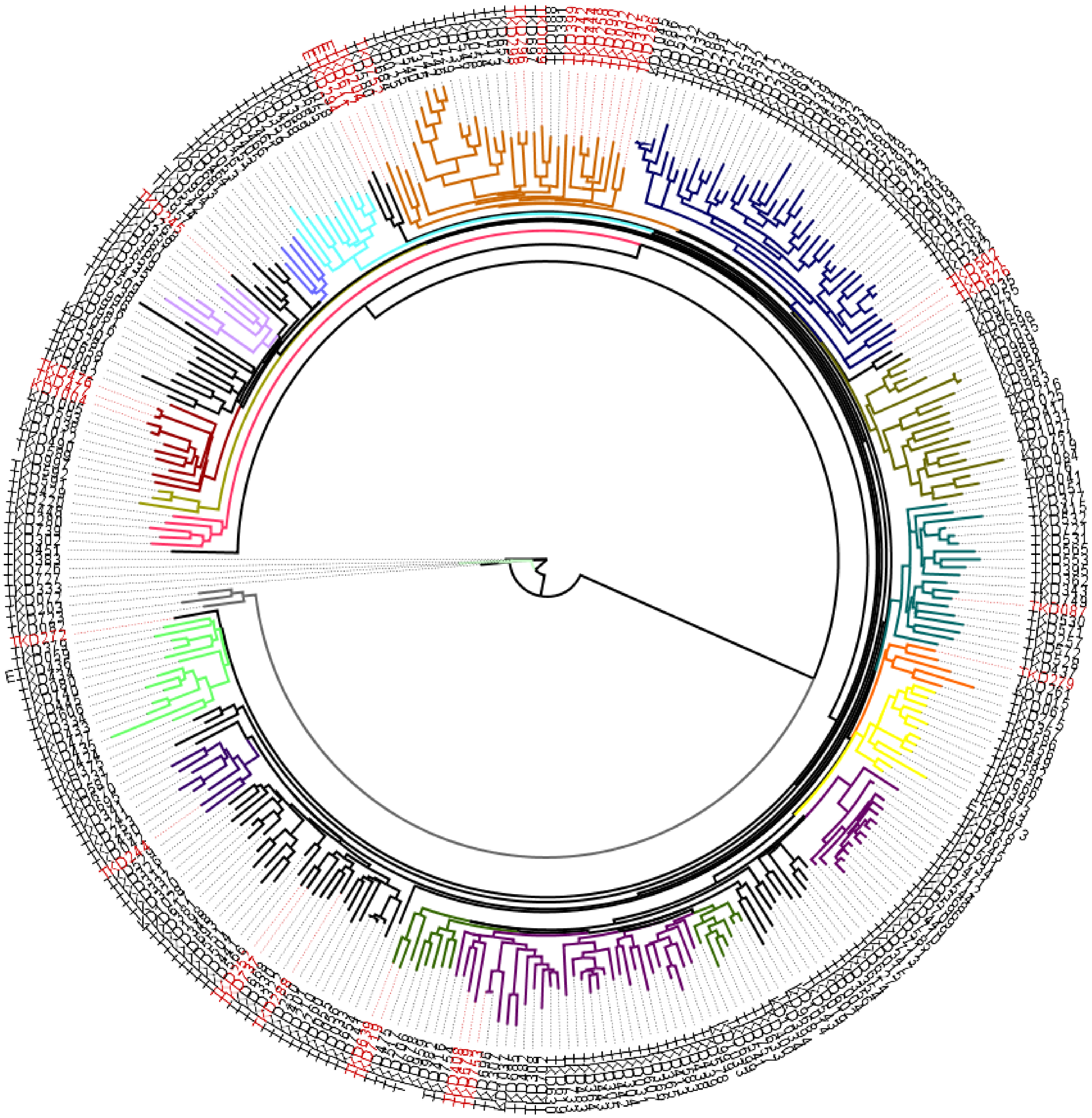

Diptera|ETKD959-13|BNRGr1Dipt08|BOLD:ACH1024
Diptera|ETKD191-12|MsiGrDipt022|BOLD:ABX4366
Diptera|ETKD1001-13|BEGr1Dipt03|BOLD:ABX4366
Diptera|ETKD517-12|NewGGr2Dipt2401|BOLD:ABX4366
Diptera|ETKD982-13|DrumGr1Dipt05|BOLD:ACH1031
Diptera|ETKD1002-13|BEGr1Dipt04|BOLD:ACH1031
Diptera|ETKD786-13|HMGr1Dipt01|BOLD:ACH1031
Diptera|ETKD879-13|DrumGr3Dipt07|BOLD:ACH1829
Diptera|ETKD607-12|MsiFo1Dipt01|BOLD:ACC3425
Diptera|ETKD608-12|MsiFo1Dipt0101|BOLD:ACC3425
Diptera|ETKD031-12|MSIGRDIPT1001|BOLD:ABW2745
Diptera|ETKD578-12|IphFo3Dipt02|BOLD:ACC4031
Diptera|ETKD579-12|IphFo3Dipt0201|BOLD:ACC4031
Diptera|ETKD583-12|IphFo3Dipt06|BOLD:ACC3800
Diptera|ETKD831-13|KSGr1Dipt1902|BOLD:ACH1798
Diptera|ETKD833-13|KSGr1Dipt1904|BOLD:ACH1798
Diptera|ETKD830-13|KSGr1Dipt1901|BOLD:ACH1798
Diptera|ETKD832-13|KSGr1Dipt1903|BOLD:ACH1798
Diptera|ETKD887-13|DrumGr3Dipt1401|BOLD:ACH1798
Diptera|ETKD888-13|DrumGr3Dipt1402|BOLD:ACH1798
Diptera|ETKD834-13|KSGr1Dipt1905|BOLD:ACH1798
Diptera|ETKD365-12|IphGr3Dipt0603|BOLD:ACA6560
Diptera|ETKD366-12|IphGr3Dipt07|BOLD:ACA6514
Diptera|ETKD413-12|NewGF02Dipt24|BOLD:ACB1856
Diptera|ETKD438-12|IphFo2Dipt07|BOLD:ACB1857
Diptera|ETKD439-12|IphFo2Dipt08|BOLD:AAG6781
Diptera|ETKD1034-13|GibaFo1Dipt10|BOLD:ACH0903
Diptera|ETKD394-12|NewGF02Dipt10|BOLD:ACB1793
Diptera|ETKD999-13|BEGr1Dipt0201|BOLD:AAA6020
Diptera|ETKD1000-13|BEGr1Dipt0202|BOLD:AAA6020
Diptera|ETKD823-13|KSGr1Dipt1501|BOLD:AAX3121
Diptera|ETKD825-13|KSGr1Dipt1503|BOLD:AAX3121
Diptera|ETKD1033-13|GibaFo1Dipt09|BOLD:ACH0904
Diptera|ETKD824-13|KSGr1Dipt1502|BOLD:ACH1650
Diptera|ETKD909-13|GibaFo2Dipt1102|BOLD:ACB1869
Diptera|ETKD908-13|GibaFo2Dipt1101|BOLD:ACB1869
Diptera|ETKD471-12|IphFo2Dipt32|BOLD:ACB1869
Diptera|ETKD501-12|NewGGr2Dipt12|BOLD:ACB1500
Diptera|ETKD822-13|KSGr1Dipt1402|BOLD:ACH1820
Diptera|ETKD368-12|IphGr3Dipt09|BOLD:ACA6467
Diptera|ETKD544-12|NewGGr2Dipt36|BOLD:ACB1458
Diptera|ETKD545-12|NewGGr2Dipt3601|BOLD:ACB1458
Diptera|ETKD596-12|IphGr1Dipt07|BOLD:ACC2097
Diptera|ETKD709-12|PalGr3Dipt08|BOLD:ACC3966
Diptera|ETKD449-12|IphFo2Dipt16|BOLD:ACB1870
Diptera|ETKD698-12|PalGr3Dip09|BOLD:ACC3967
Diptera|ETKD710-12|PalGr3Dipt10|BOLD:ACC3967
Diptera|ETKD585-12|IphFo3Dipt08|BOLD:ACC3802
Diptera|ETKD711-12|PalGr3Dipt1001|BOLD:ACC3965
Diptera|ETKD712-12|PalGr3Dipt1002|BOLD:ACC1791
Diptera|ETKD910-13|GibaFo2Dipt1201|BOLD:ACH1825
Diptera|ETKD734-12|SprGr3Dipt06|BOLD:ACC3987
Diptera|ETKD877-13|DrumGr3Dipt0601|BOLD:ACH1505
Diptera|ETKD820-13|KSGr1Dipt13|BOLD:ACH1505
Diptera|ETKD821-13|KSGr1Dipt1401|BOLD:ACH1505
Diptera|ETKD194-12|MsiGrDipt041|BOLD:AAG6787
Diptera|ETKD713-12|PalGr3Dipt1003|BOLD:AAG6787
Muscidae|ETKD008-12|MSIGRDIPT0404|BOLD:AAG6787
Muscidae|ETKD007-12|MSIGRDIPT0403|BOLD:AAG6787
Diptera|ETKD878-13|DrumGr3Dipt0602|BOLD:AAG6787
Diptera|ETKD193-12|MsiGrDipt0302|BOLD:AAG6787
Muscidae|ETKD009-12|MSIGRDIPT0405|BOLD:AAG6787
Diptera|ETKD912-13|GibaFo2Dipt1203|BOLD:ACH1722
Diptera|ETKD913-13|GibaFo2Dipt1204|BOLD:ACH1555
Diptera|ETKD911-13|GibaFo2Dipt1202|BOLD:ACG2170
Diptera|ETKD914-13|GibaFo2Dipt1205|BOLD:ACG2170
Diptera|ETKD929-13|GibaFo2Dipt22|BOLD:ACH1508
Phoridae|ETKD347-12|NorPGr1Dipt4501|BOLD:ACA6605
Diptera|ETKD921-13|GibaFo2Dipt1703|BOLD:ACH1832
Diptera|ETKD922-13|GibaFo2Dipt1704|BOLD:ACH1832
Diptera|ETKD679-12|PalGr1Dipt21|BOLD:ACC3872
Diptera|ETKD682-12|PalGr1Dipt24|BOLD:ACC3872
Diptera|ETKD683-12|PalGr1Dipt25|BOLD:ACC3872
Diptera|ETKD600-12|IphGr1Dipt12|BOLD:ACC2605
Diptera|ETKD602-12|IphGr1Dipt14|BOLD:ACC2605
Anthomyiidae|ETKD019-12|MSIGRDIPT0704|BOLD:AAZ4294
Anthomyiidae|ETKD020-12|MSIGRDIPT0705|BOLD:AAZ4294
Anthomyiidae|ETKD018-12|MSIGRDIPT0703|BOLD:AAZ4294
Anthomyiidae|ETKD017-12|MSIGRDIPT0702|BOLD:AAZ4294
Anthomyiidae|ETKD091-12|MSIGRDIPT3801|BOLD:AAZ4294
Anthomyiidae|ETKD092-12|MSIGRDIPT3802|BOLD:AAZ4294
Anthomyiidae|ETKD093-12|MSIGRDIPT3803|BOLD:AAZ4294
Anthomyiidae|ETKD016-12|MSIGRDIPT0701|BOLD:AAZ4294
Diptera|ETKD840-13|KSGr1Dipt21|BOLD:AAZ4294
Diptera|ETKD639-12|MsiFo1Dipt27|BOLD:ACC1788
Diptera|ETKD719-12|PalGr3Dipt15|BOLD:ACA6755
Muscidae|ETKD350-12|NorPGr1Dipt4701|BOLD:ACA6755
Diptera|ETKD356-12|IphGr3Dipt01|BOLD:ACA6859
Diptera|ETKD899-13|GibaFo2Dipt0501|BOLD:ACH1689
Diptera|ETKD900-13|GibaFo2Dipt0502|BOLD:ACH1689
Diptera|ETKD705-12|PalGr3Dipt07|BOLD:ACC3867
Sarcophagidae|ETKD004-12|MSIGRDIPT0301|BOLD:ABW3546
Diptera|ETKD931-13|GibaFo2Dipt2302|BOLD:ABX4382
Diptera|ETKD266-12|PalFoDipt019|BOLD:ABX4382
Diptera|ETKD930-13|GibaFo2Dipt2301|BOLD:ABX4382
Diptera|ETKD681-12|PalGr1Dipt23|BOLD:ABX4382
Diptera|ETKD509-12|NewGGr2Dipt20|BOLD:ACB1608
Diptera|ETKD826-13|KSGr1Dipt1601|BOLD:ACH1573


Diptera|ETKD994-13|DrumGr2Dipt04|BOLD:ACH1074
Diptera|ETKD1008-13|GibaGr1Dipt04|BOLD:ACH1074
Diptera|ETKD1003-13|GibaGr1Dipt0101|BOLD:ACH1078
Diptera|ETKD1007-13|GibaGr1Dipt03|BOLD:ACH1078
Diptera|ETKD1010-13|GibaGr2Dipt0101|BOLD:ACH1078
Diptera|ETKD1011-13|GibaGr2Dipt0102|BOLD:ACH1078
Diptera|ETKD850-13|DrumGr2Dipt10|BOLD:ACH1502
Diptera|ETKD981-13|DrumGr1Dipt04|BOLD:ACH0910
Diptera|ETKD956-13|BNRGr1Dipt05|BOLD:ACH1029
Diptera|ETKD944-13|UKZNGr2Dipt03|BOLD:ACH1613
Diptera|ETKD895-13|DrumGr3Dipt17|BOLD:ACH1797
Diptera|ETKD996-13|DrumGr3Dipt04|BOLD:ACH1073
Diptera|ETKD259-12|PalFoDipt014|BOLD:ABX4387
Diptera|ETKD262-12|PalFoDipt016|BOLD:ABX4387
Diptera|ETKD429-12|IphFo2Dipt01|BOLD:ABA8397
Diptera|ETKD772-13|SpringGr1Dipt06|BOLD:ACH1791
Diptera|ETKD773-13|SpringGr1Dipt07|BOLD:ACH1790
Asilidae|ETKD476-12|NewGGr2Dipt01|BOLD:ACB1477
Asilidae|ETKD477-12|NewGGr2Dipt0101|BOLD:ACB1604
Diptera|ETKD783-13|CHGr1Dipt08|BOLD:ACB1604
Diptera|ETKD860-13|DrumGr1Dipt13|BOLD:ACH1100
Diptera|ETKD987-13|DrumGr1Dipt08|BOLD:ACH1100
Diptera|ETKD1004-13|GibaGr1Dipt0102|BOLD:ACH1264
Diptera|ETKD1005-13|GibaGr1Dipt0103|BOLD:ACH1265
Diptera|ETKD1006-13|GibaGr1Dipt02|BOLD:ACH1265
Diptera|ETKD1012-13|GibaGr2Dipt0201|BOLD:ACH1076
Diptera|ETKD1013-13|GibaGr2Dipt0202|BOLD:ACH1076
Diptera|ETKD898-13|DrumGr3Dipt19|BOLD:ACH0905
Diptera|ETKD1038-13|GibaFo1Dipt1302|BOLD:ACH0905
Asilidae|ETKD490-12|NewGGr2Dipt05|BOLD:ACB1662
Diptera|ETKD497-12|NewGGr2Dipt1002|BOLD:ACB1662
Diptera|ETKD506-12|NewGGr2Dipt17|BOLD:ACB1662
Diptera|ETKD593-12|IphGr1Dipt05|BOLD:ACC3431
Diptera|ETKD594-12|IphGr1Dipt0501|BOLD:ACC3431
Diptera|ETKD597-12|IphGr1Dipt08|BOLD:ACC3431
Diptera|ETKD827-13|KSGr1Dipt1602|BOLD:ACH1572
Diptera|ETKD940-13|UKZNGr2Dipt0102|BOLD:ACH1079
Diptera|ETKD939-13|UKZNGr2Dipt0101|BOLD:ACH1079
Diptera|ETKD941-13|UKZNGr2Dipt0103|BOLD:ACH1079
Diptera|ETKD997-13|BEGr1Dipt0101|BOLD:ACH1079
Diptera|ETKD998-13|BEGr1Dipt0102|BOLD:ACH1079
Diptera|ETKD847-13|DrumGr2Dipt0801|BOLD:ACH1079
Diptera|ETKD848-13|DrumGr2Dipt0802|BOLD:ACH1079
Diptera|ETKD958-13|BNRGr1Dipt07|BOLD:ACB1525
Diptera|ETKD492-12|NewGGr2Dipt07|BOLD:ACB1525
Diptera|ETKD1016-13|GibaGr2Dipt05|BOLD:ACH0907
Diptera|ETKD232-12|PalGrDipt01|BOLD:ABX4389
Diptera|ETKD589-12|IphGr1Dipt01|BOLD:ABX4389
Diptera|ETKD590-12|IphGr1Dipt02|BOLD:ABX4389
Dolichopodidae|ETKD053-12|MSIGRDIPT1603|BOLD:ABW3660
Diptera|ETKD245-12|PalFoDipt0303|BOLD:ABX4394
Diptera|ETKD281-12|PalFoDipt028|BOLD:ABX4373
Diptera|ETKD380-12|UKZNGr1Dipt02|BOLD:ACA6855
Diptera|ETKD498-12|NewGGr2Dipt1003|BOLD:ACA6855
Diptera|ETKD507-12|NewGGr2Dipt18|BOLD:ACA6855
Diptera|ETKD641-12|MsiFo1Dipt29|BOLD:ACC2606
Diptera|ETKD656-12|MsiFo1Dipt43|BOLD:ACC2606
Diptera|ETKD653-12|MsiFo1Dipt40|BOLD:ACC2606
Diptera|ETKD658-12|MsiFo1Dipt45|BOLD:ACC2606
Diptera|ETKD620-12|MsiFo1Dipt13|BOLD:ABW2402
Stratiomyidae|ETKD001-12|MSIGRDIPT0201|BOLD:ABW2402
Stratiomyidae|ETKD002-12|MSIGRDIPT0203|BOLD:ABW2402
Diptera|ETKD448-12|IphFo2Dipt15|BOLD:ACB1806
Diptera|ETKD302-12|NorPGr1Dipt16|BOLD:ACA6781
Diptera|ETKD675-12|PalGr1Dipt1701|BOLD:ACA6781
Diptera|ETKD280-12|PalFoDipt027|BOLD:ABX4374
Diptera|ETKD739-12|SprGr3Dipt14|BOLD:ACC3985
Asilidae|ETKD059-12|MSIGRDIPT19|BOLD:ABW2466
Diptera|ETKD207-12|MsiGrDipt40|BOLD:ABW2478
Simuliidae|ETKD080-12|MSIGRDIPT30|BOLD:ABW2478
Diptera|ETKD268-12|PalFoDipt020|BOLD:ABX4368
Diptera|ETKD360-12|IphGr3Dipt05|BOLD:ACA6795
Diptera|ETKD361-12|IphGr3Dipt0501|BOLD:ACA6795
Chironomidae|ETKD081-12|MSIGRDIPT3101|BOLD:ABW4210
Chironomidae|ETKD083-12|MSIGRDIPT3103|BOLD:ABW4210
Diptera|ETKD359-12|IphGr3Dipt04|BOLD:ACA6863
Diptera|ETKH008-12|PalmGr3Hemi06|
Diptera|ETKD932-13|GibaFo2Dipt24|BOLD:ACH1507
Diptera|ETKD957-13|BNRGr1Dipt06|BOLD:ACH1030
Diptera|ETKD514-12|NewGGr2Dipt23|BOLD:ACB1465
Diptera|ETKD515-12|NewGGr2Dipt2301|BOLD:ACB1465
Diptera|ETKD784-13|CHGr1Dipt09|BOLD:ACB1465
Therevidae|ETKD542-12|NewGGr2Dipt34|BOLD:ACB1490
Diptera|ETKD204-12|MsiGrDipt3903|BOLD:ABX4354
Diptera|ETKD740-12|SprGr3Dipt19|BOLD:ACB2071
Diptera|ETKD472-12|IphFo2Dipt33|BOLD:ACB2071
Tipulidae|ETKD431-12|IphFo2Dipt0201|BOLD:ACB2071
Tipulidae|ETKD432-12|IphFo2Dipt0202|BOLD:ACB2071
Tipulidae|ETKD430-12|IphFo2Dipt02|BOLD:ACB2071
Diptera|ETKD272-12|PalFoDipt022|BOLD:ABX4369
Diptera|ETKD285-12|PalFoDipt02201|BOLD:ABX4369
Diptera|ETKD405-12|NewGF02Dipt17|BOLD:ACB1861
Diptera|ETKD461-12|IphFo2Dipt23|BOLD:ACB1774
Diptera|ETKD466-12|IphFo2Dipt28|BOLD:ACB1859
Diptera|ETKH086-12|IphiGr3Hemi17|BOLD:ACB1859
Diptera|ETKD453-12|IphFo2Dipt19|BOLD:ACB1777
Diptera|ETKH047-12|PalmGr1Hemi03|BOLD:ACK4876
Diptera|ETKD389-12|NewGF02Dipt05|BOLD:ACB1862
Diptera|ETKD948-13|UKZNGr2Dipt07|BOLD:ACH1612

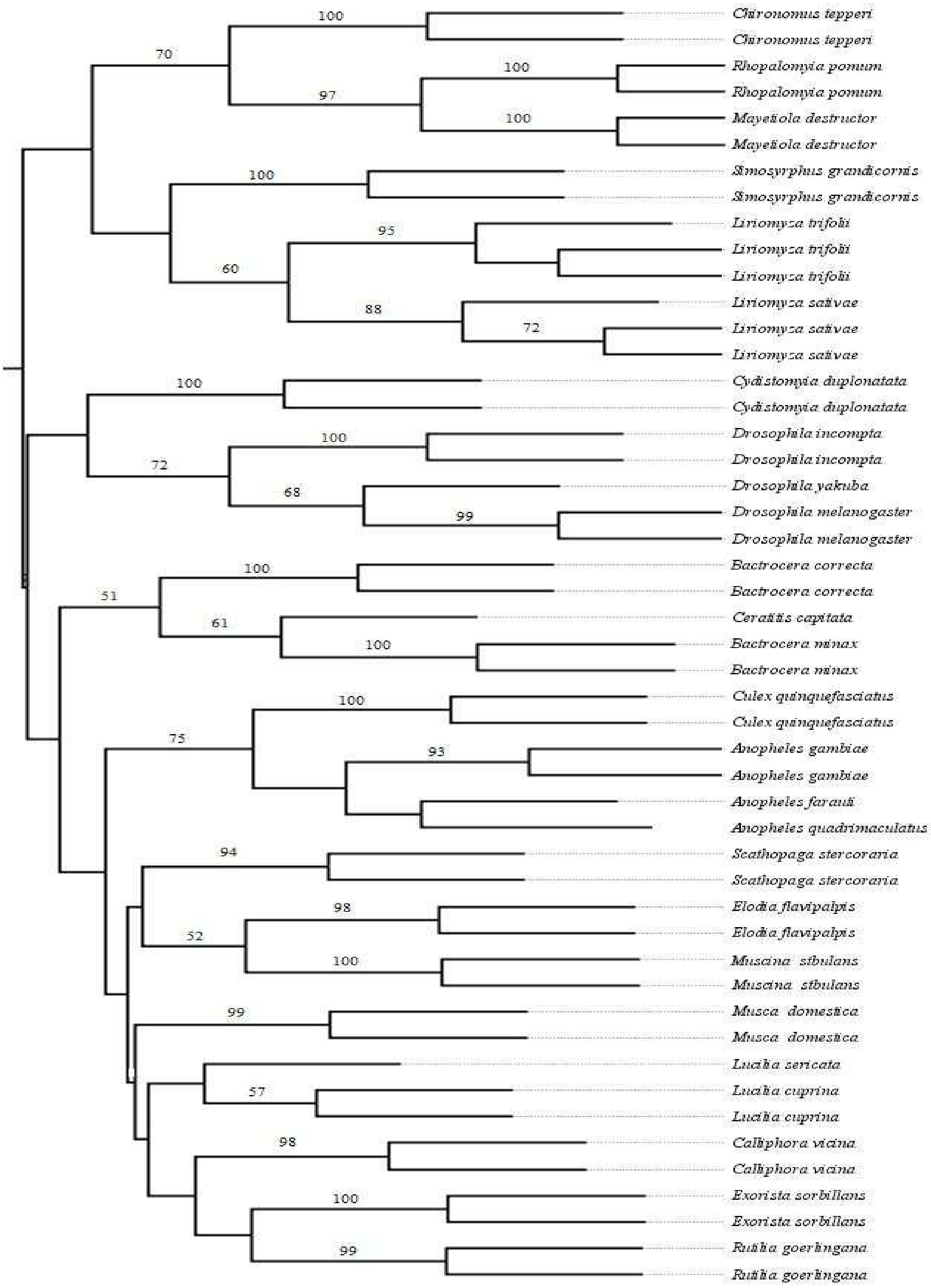


Diptera|ETKD284-12|PalFoDipt0904|BOLD:ABX4355
Diptera|ETKD283-12|PalFoDipt0903|BOLD:ABX4356
Diptera|ETKD222-12|MsiGrDipt4802|BOLD:ABX4359
Diptera|ETKD223-12|MsiGrDipt4803|BOLD:ABX4359
Diptera|ETKD465-12|IphFo2Dipt27|BOLD:ABX4359
Diptera|ETKD655-12|MsiFo1Dipt42|BOLD:ABX4359
Diptera|ETKD224-12|MsiGrDipt4804|BOLD:ABW4194
Chloropidae|ETKD078-12|MSIGRDIPT2803|BOLD:ABW4194
Diptera|ETKD221-12|MsiGrDipt4801|BOLD:ABW4194
Phoridae|ETKD338-12|NorPGr1Dipt37|BOLD:ABW4194
Diptera|ETKD468-12|IphFo2Dipt30|BOLD:ACB1860
Diptera|ETKD469-12|IphFo2Dipt3001|BOLD:ACB1860
Diptera|ETKD581-12|IphFo3Dipt04|BOLD:ACB1860
Diptera|ETKD229-12|MsiGrDipt5102|BOLD:ABX4388
Diptera|ETKD556-12|PalGr1Dipt0301|BOLD:ABX4363
Muscidae|ETKD353-12|NorPGr1Dipt50|BOLD:ABX4363
Diptera|ETKD346-12|NorPGr1Dipt45|BOLD:ABX4363
Diptera|ETKD228-12|MsiGrDipt5101|BOLD:ABX4363
Muscidae|ETKD351-12|NorPGr1Dipt48|BOLD:ABX4363
Phoridae|ETKD352-12|NorPGr1Dipt49|BOLD:ABX4363
Diptera|ETKD446-12|IphFo2Dipt13|BOLD:ACB1858
Diptera|ETKD464-12|IphFo2Dipt26|BOLD:ACB1858
Diptera|ETKD587-12|IphFo3Dipt11|BOLD:ACC3804
Phoridae|ETKD354-12|NorPGr1Dipt51|BOLD:ACA6644
Diptera|ETKD789-13|HMGr1Dipt04|BOLD:ACH1709
Diptera|ETKD495-12|NewGGr2Dipt10|BOLD:ACB1550
Diptera|ETKD504-12|NewGGr2Dipt15|BOLD:ACB1550
Diptera|ETKD551-12|NewGGr2Dipt39|BOLD:ACB1642
Diptera|ETKD236-12|PalGrDipt05|BOLD:ABX4391
Diptera|ETKD676-12|PalGr1Dipt1702|BOLD:ABX4391
Diptera|ETKD691-12|PalGr2Dipt0401|BOLD:ABX4391
Diptera|ETKD690-12|PalGr2Dipt04|BOLD:ABX4391
Diptera|ETKD692-12|PalGr2Dipt0402|BOLD:ABX4391
Diptera|ETKD693-12|PalGr2Dipt0403|BOLD:ABX4391
Diptera|ETKD198-12|MsiGrDipt021|BOLD:ABX4355
Diptera|ETKD238-12|PalGrDipt07|BOLD:ABX4392
Phoridae|ETKD348-12|NorPGr1Dipt46|BOLD:ABX4392
Diptera|ETKD231-12|MsiGrDipt53|BOLD:ABX4361
Diptera|ETKH375-12|NPGr1Hemi16|BOLD:ACB2151
Diptera|ETKD623-12|MsiFo1Dipt15|BOLD:AAW7902
Diptera|ETKD652-12|MsiFo1Dipt39|BOLD:AAW7902
Diptera|ETKD945-13|UKZNGr2Dipt04|BOLD:ACH1614
Diptera|ETKD591-12|IphGr1Dipt03|BOLD:ACC2096
Diptera|ETKD592-12|IphGr1Dipt04|BOLD:ABV1114
Diptera|ETKD771-13|SpringGr1Dipt05|BOLD:ABV1114
Orthoptera|ETKII1554-13|NPGr1Orth05|BOLD:ACK3198
Orthoptera|ETKII1555-13|NPGr1Orth0501|BOLD:ACK3198
Diptera|ETKD220-12|MsiGrDipt4702|BOLD:ABX4360
Diptera|ETKD785-13|CHGr1Dipt10|BOLD:ACH1792
Diptera|ETKD203-12|MsiGrDipt2502|BOLD:ABX4353
Diptera|ETKD723-12|SprFo3Dipt01|BOLD:ACL1210
Diptera|ETKH496-12|MsiFo1Hemi0504|BOLD:ACC4228
Diptera|ETKH036-12|MsiFo3Hemi04|BOLD:ACA6844
Sciaridae|ETKD082-12|MSIGRDIPT3102|BOLD:ABW3432
Diptera|ETKD436-12|IphFo2Dipt05|BOLD:ACB1778
Tipulidae|ETKD434-12|IphFo2Dipt0301|BOLD:ACB1845
Diptera|ETKD876-13|DrumGr3Dipt05|BOLD:ABV1132
Diptera|ETKD211-12|MsiGrDipt43|BOLD:ABX4351
Diptera|ETKD379-12|UKZNGr1Dipt01|BOLD:AAN4393
Diptera|ETKD294-12|NorPGr1Dipt09|BOLD:ACA6556
Diptera|ETKD672-12|MsiFo1Dipt56|BOLD:ACC1789
Culicidae|ETKD435-12|IphFo2Dipt04|BOLD:ACB2024
Diptera|ETKD423-12|NewGF02Dipt31|BOLD:ACB1804
Diptera|ETKD335-12|NorPGr1Dipt34|BOLD:AAW3995
Diptera|ETKD659-12|MsiFo1Dipt46|BOLD:ABV3572
Diptera|ETKD660-12|MsiFo1Dipt4601|BOLD:ABV3572
Diptera|ETKD424-12|NewGF02Dipt32|BOLD:AAZ3622
Culicidae|ETKD334-12|NorPGr1Dipt33|BOLD:AAA4210
Diptera|ETKD445-12|IphFo2Dipt12|BOLD:ACB1789
Diptera|ETKD946-13|UKZNGr2Dipt05|BOLD:AAE2099
Diptera|ETKD606-12|IphGr1Dipt18|BOLD:ACC3426
Diptera|ETKD759-12|VERCFO1Dipt01|BOLD:ACC3426
Diptera|ETKD760-12|VERCFO1Dipt0101|BOLD:ACC3426
Diptera|ETKD427-12|NewGF02Dipt3302|BOLD:ACB1807
Diptera|ETKD426-12|NewGF02Dipt3301|BOLD:ACB1807
Diptera|ETKD428-12|NewGF02Dipt3303|BOLD:ACB1807
Tipulidae|ETKD433-12|IphFo2Dipt03|BOLD:ACB1846
Diptera|ETKD064-12|MSIGRDIPT2205|BOLD:ABW2722
Diptera|ETKD062-12|MSIGRDIPT2203|BOLD:ABW2722
Diptera|ETKD061-12|MSIGRDIPT2201|BOLD:ABW2722
Diptera|ETKD063-12|MSIGRDIPT2204|BOLD:ABW2722
Diptera|ETKD199-12|MsiGrDipt2102|BOLD:ABW2722
Diptera|ETKD875-13|DrumGr3Dipt03|BOLD:ACC3986
Diptera|ETKD738-12|SprGr3Dipt12|BOLD:ACC3986
Diptera|ETKD451-12|IphFo2Dipt1701|BOLD:ACB1773
Diptera|ETKH217-12|IphiFo2Hemi01|BOLD:ACK6982
Diptera|ETKH252-12|IphiFo2Hemi0101|BOLD:ACK6982
Coleoptera|ETKII084-12|MSIGRCOL3302|
Hemiptera|ETKC076-12|PalmGr2Col13|BOLD:ABV2583
Hemiptera|ETKD470-12|IphFo2Dipt31|BOLD:ACB3037
Hemiptera|ETKD628-12|MsiFo1Dipt19|BOLD:ACK6996
Hemiptera|ETKD618-12|MsiFo1Dipt11|BOLD:ACK6996
Hemiptera|ETKD629-12|MsiFo1Dipt1901|BOLD:ACK6996
Diptera|ETKD510-12|NewGGr2Dipt21|BOLD:ACB1501
Psocoptera|ETKD463-12|IphFo2Dipt25|BOLD:ACK2744
Lepidoptera|ETKD411-12|NewGF02Dipt22|BOLD:ACK2772

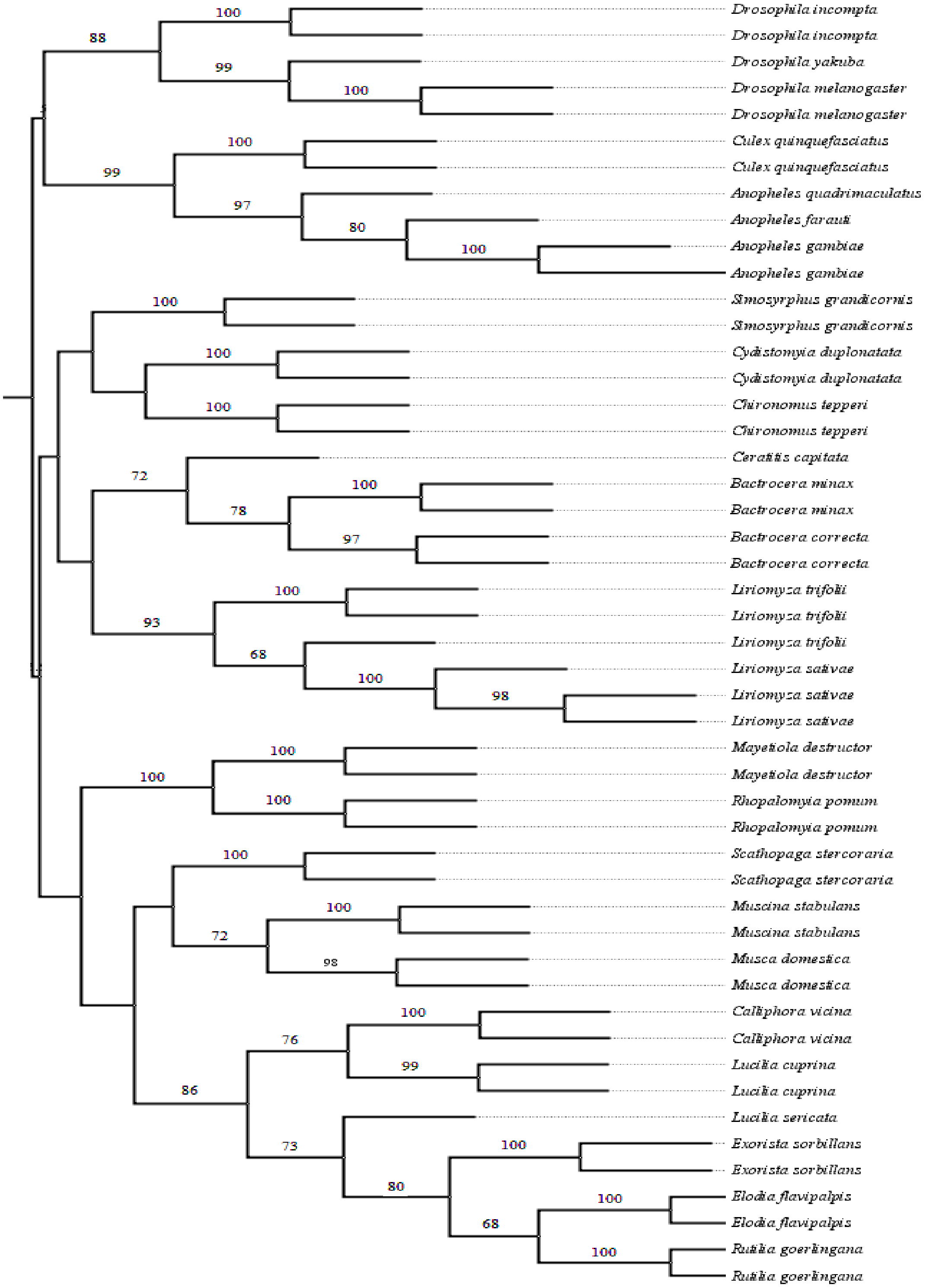
**Appendix 2.2** Annotated phylogenetic tree of one individual per BIN, the major families are colour coded on the tree. The annotations on the tree are Diptera ID numbers assigned in BOLD. The red coloured taxa had sequence similarities >95%.



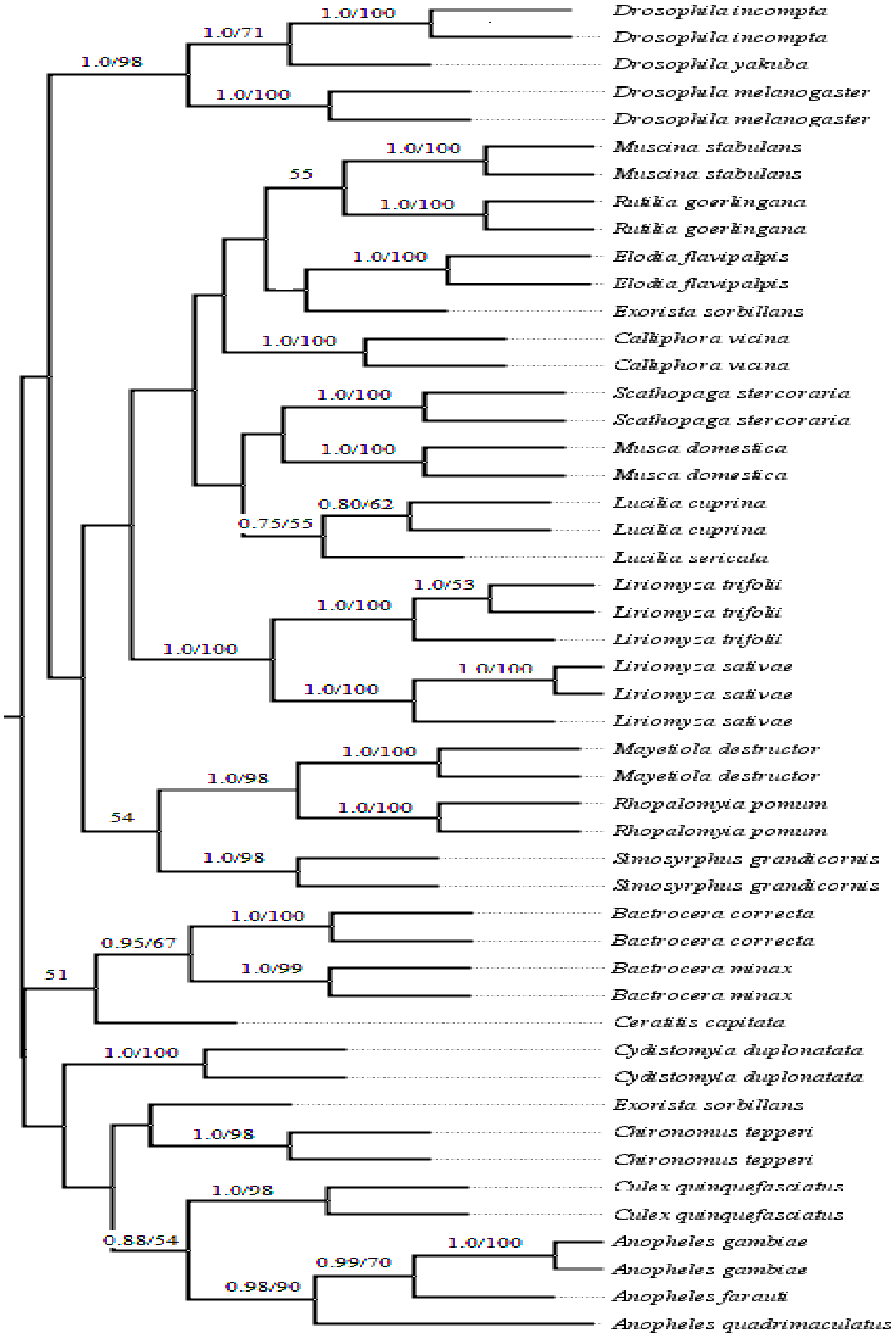
**Appendix 4.1** Maximum likelihood tree of the *ATP8* gene. Only bootstrap values above 50% are shown**.**



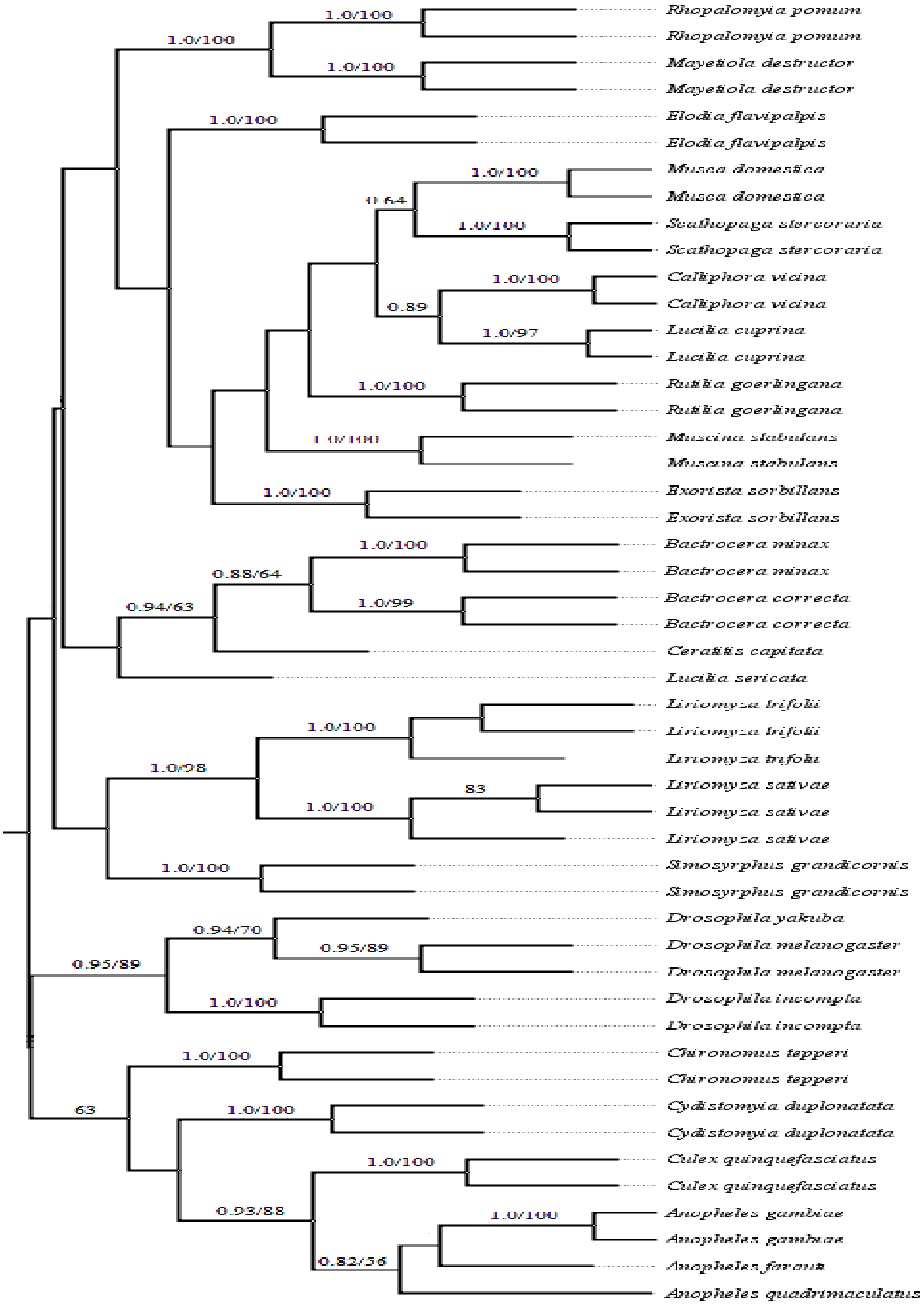
**Appendix 4.2** Maximum likelihood tree of the *Cytb* gene. Only bootstrap values above 50% are shown.



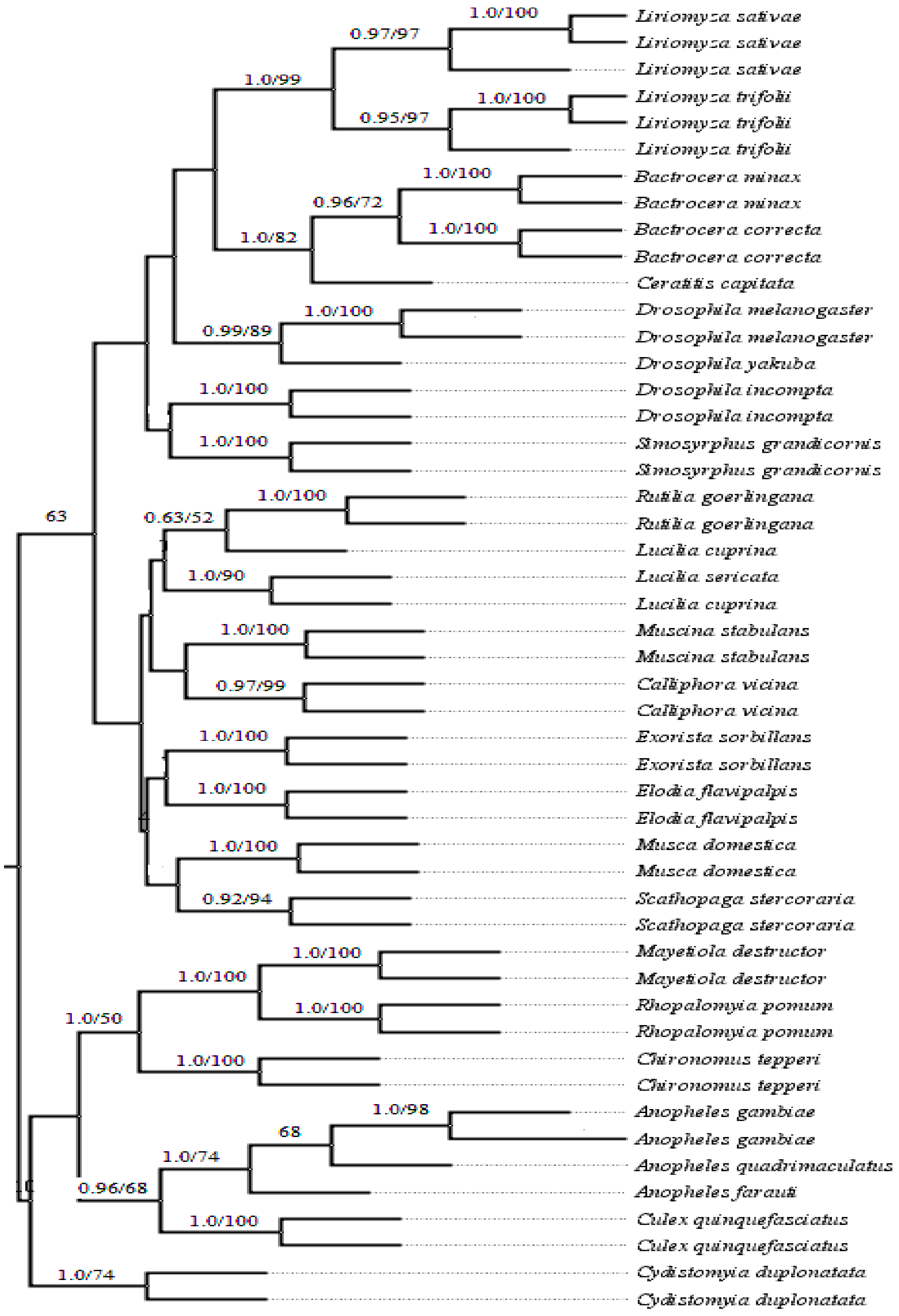
**Appendix 4.3** Maximum likelihood tree of the *COIII* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.



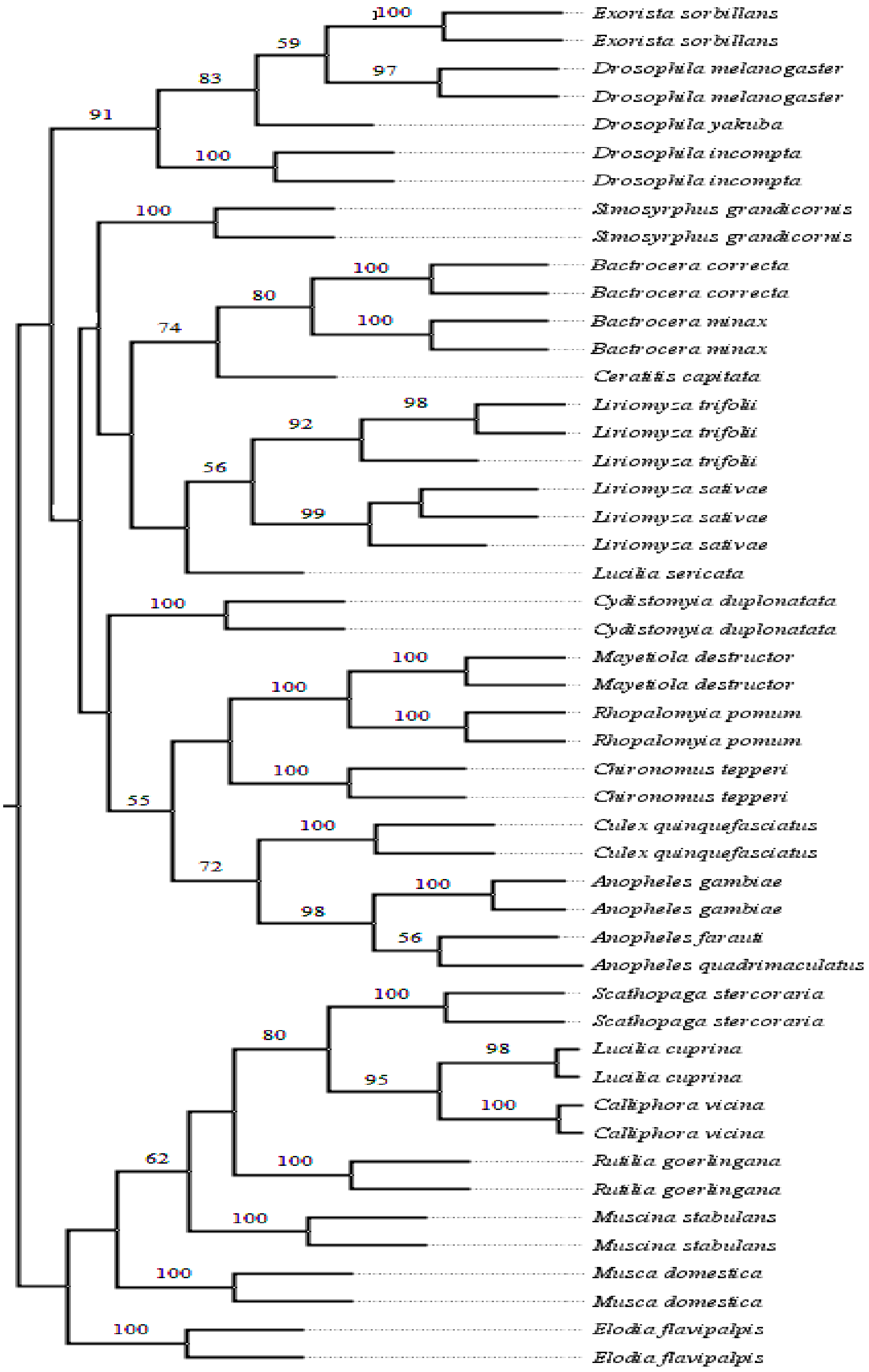
**Appendix 4.4** Maximum likelihood tree of the *NAD1* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown



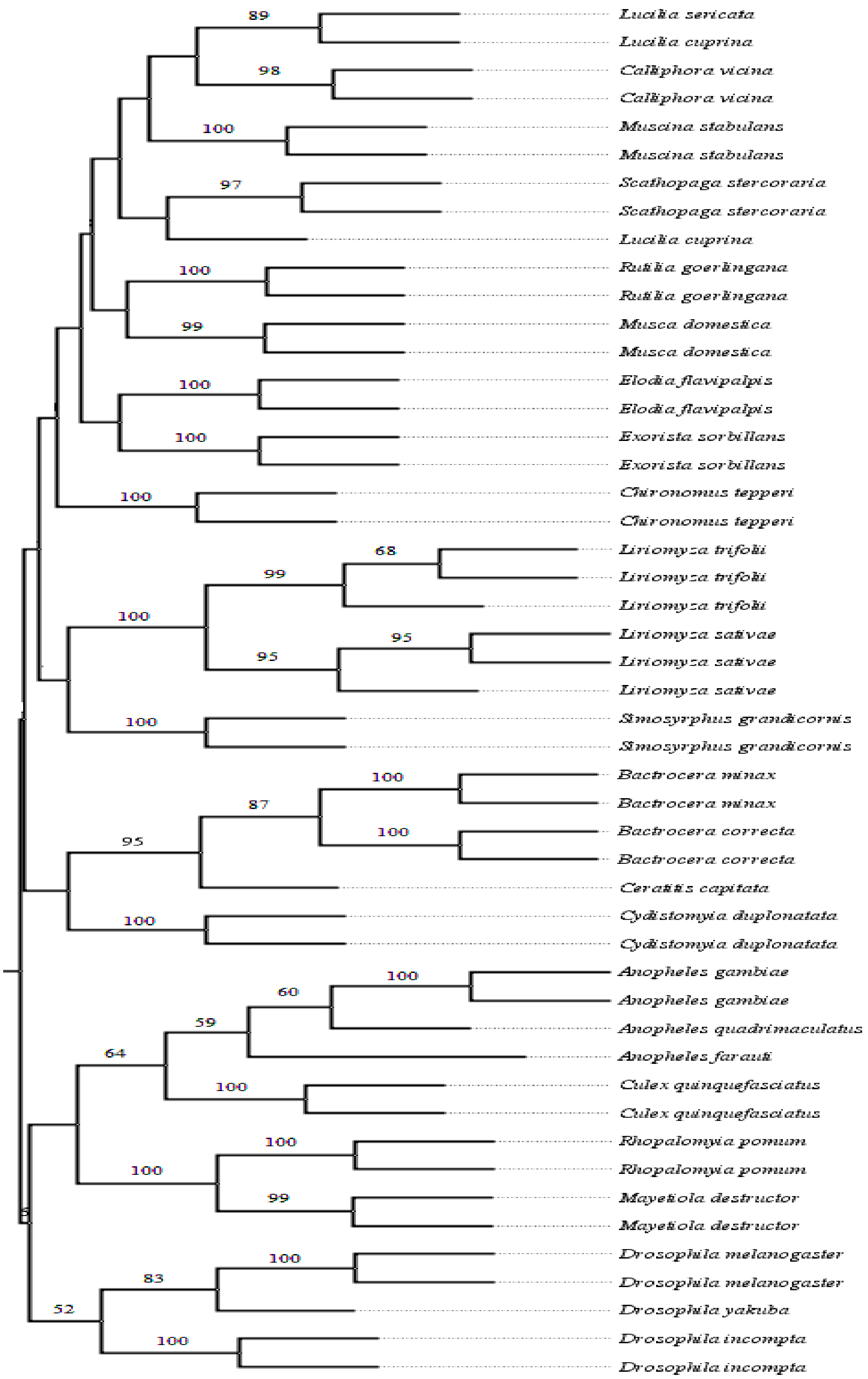
**Appendix 4.5** Maximum likelihood tree of the *NAD3* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.



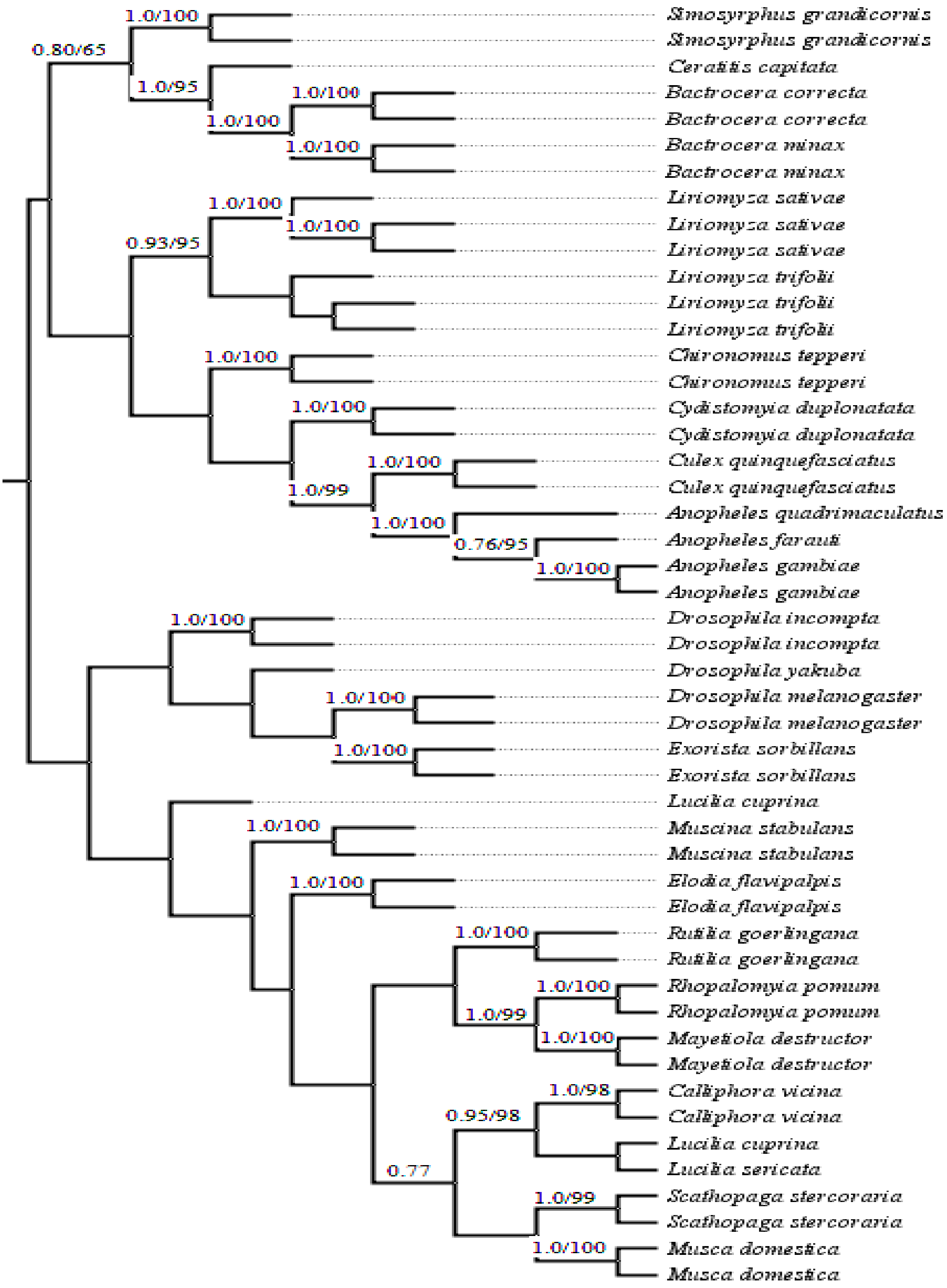
**Appendix 4.6** Maximum likelihood tree of the *NAD4* gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.



**Appendix 4.7** Maximum likelihood tree of the *NAD*5 gene. Only bootstrap values above 50% are shown.



**Appendix 4.8** Maximum likelihood tree of the *NAD*6 gene. Only bootstrap values above 50% and Bayesian posterior probabilities above 0.5 are shown.


1.0/100
0.80/65
1.0/95
1.0/100
1.0/100
1.0/100
1.0/100
1.0/100
0.93/95
1.0/100
1.0/100
1.0/100
1.0/99
1.0/100
0.76/95
1.0/100
1.0/100
1.0/100
1.0/100
1.0/100
1.0/100
1.0/100
1.0/99
1.0/100
1.0/98
0.95/98
0.77
1.0/99
1.0/100
Simosyrphus grandicornis
Simosyrphus grandicornis
Ceratitis capitata
Bactrocera correcta
Bactrocera correcta
Bactrocera minax
Bactrocera minax
Liriomyza sativae
Liriomyza sativae
Liriomyza sativae
Liriomyza trifolii
Liriomyza trifolii
Liriomyza trifolii
Chironomus tepperi
Chironomus tepperi
Cyclistomyia duplonatata
Cyclistomyia duplonatata
Culex quinquefasciatus
Culex quinquefasciatus
Anopheles quadrimaculatus
Anopheles farauti
Anopheles gambiae
Anopheles gambiae
Drosophila incompta
Drosophila incompta
Drosophila yakuba
Drosophila melanogaster
Drosophila melanogaster
Exorista sorbillans
Exorista sorbillans
Lucilia cuprina
Muscina stabulans
Muscina stabulans
Elodia flavipalpis
Elodia flavipalpis
Rutilia goerlingana
Rutilia goerlingana
Rhopalomyia pomum
Rhopalomyia pomum
Mayetiola destructor
Mayetiola destructor
Calliphora vicina
Calliphora vicina
Lucilia cuprina
Lucilia sericata
Scathopaga stercoraria
Scathopaga stercoraria
Musca domestica
Musca domestica