Testing the utility of DNA barcoding in South African Hemiptera: Using eThekwini species as a case study

By

Ashrenee Govender

BSc. (*Hons*) Genetics

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

Discipline of Genetics

School of Life Sciences

College of Agriculture, Engineering, and Science

University of KwaZulu-Natal

Pietermaritzburg

South Africa

UNIVERSITY OF
KWAZULU-NATAL
INYUVESI
YAKWAZULU-NATALI

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

Supervisor: Dr S. Willows-Munro

Signature:

Date: 16 January 2017

Supervisor: Prof M. Rouget

Signature:

Date: 16 January 2017

Preface

The research contained in this dissertation was completed by the candidate while based in the discipline of Genetics, School of Life Sciences of the College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa under the supervision of Dr S. Willows-Munro and Prof M. Rouget.

These studies represent original work by the candidate and have not otherwise been submitted in any form to another University. Where use has been made of the work by other authors it has been duly acknowledged in the text.

Supervisor: Dr S. Willows-Munro

Signature:

Date: 16 January 2017

Supervisor: Prof M. Rouget

Signature:

Date: 16 January 2017

College of Agriculture, Engineering, and Science

Declaration of Plagiarism

I, Ashrenee Govender, declare that:

(i) the research reported in this dissertation, except where otherwise indicated or acknowledged, is my original work;

(ii) this dissertation has not been submitted in full or in part for any degree or examination to any other university;

(iii) this dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;

(iv) this dissertation 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, their writing has been placed inside quotation marks, and referenced;

(v) where I have used material for which publications followed, I have indicated in detail my role in the work;

(vi) this dissertation is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included;

(vii) this dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the References sections.

Signature: Ashrenee Govender

Date: 16 January 2016

Declaration Plagiarism 22/05/08 FHDR Approved

Abstract

The eThekwini municipal region and surrounding areas (Durban, Kwazulu-Natal, South Africa) are situated within the globally important Maputaland-Pondoland-Albany biodiversity hotspot. The biodiversity present in this region is under significant pressure from urbanization and climate change. This highlights the need to provide tools that can assist in the discovery and identification of species at an accelerated pace, to create biodiversity inventories which can be used for appropriate conservation planning. The creation of species inventories is a difficult task, more especially for hyper-diverse groups such as terrestrial arthropods. These groups can be morphologically cryptic or difficult to identify using traditional morphology-based taxonomy. Therefore, molecular-based methods of species identification have been proposed to assist in traditional taxonomy. DNA barcoding has been suggested as a mechanism which enables biologists to "label" or "tag" species, using nucleotide variations in short sequences known as DNA barcodes. This study investigates the utility of DNA barcoding and the use of the mitochondrial cytochrome oxidase c subunit 1 (COI) marker to identify species of Hemiptera efficiently and accurately. This study presents a preliminary DNA barcode reference library for Hemiptera collected from 18 different localities within and around the eThekwini municipal region. To test the success of DNA barcoding and the COI marker, matches between morphospecies and barcode clusters (BINs) were analyzed and the presence of the DNA barcode gap in the data was examined. The DNA barcode gap is the gap between the intraspecific and interspecific genetic distances, the lack of the DNA barcode gap suggests that taxa cannot be reliably sorted into species based on the genetic data. Analyses revealed that DNA barcoding using the COI marker is a successful method of identifying Hemiptera species in this study. Thereafter, a case study was selected within the Buffelsdraai Landfill Site Community Reforestation Project, to test whether DNA barcoding could be used to assess the potential of Hemiptera as an indicator of ecological restoration success. The Hemiptera species composition and assembly were assessed by analyzing multiple diversity indices, ordination, UPGMA cluster analysis and phylogenetic analysis. Hemiptera was seen to be sensitive to changes in an ecosystem which make this order an effective environmental and biological indicator. With the help of DNA barcoding, specific families of Hemiptera were identified as habitat-specific and good biological indicators for future studies of ecological restoration and reforestation.

Acknowledgments

I would like to express my heartfelt gratitude to the following people and organizations for their support:

My supervisor, Dr Sandi Willows-Munro, for all your guidance, patience, constant support and for being my critical reader. I am truly grateful for all your help and encouragement throughout the years. Thank you for allowing me to grow and improve greatly as a scientist.

My co-supervisor, Prof Mathieu Rouget, for all your valuable input, and guidance. Thank you for providing me with funding and for allowing me an opportunity of a life time to attend my first international conference in Morocco. I am truly grateful for all your support.

The National Research Foundation (NRF), Durban Action Research Partnership (D'RAP), and Wildlands Conservation Trust for their financial support throughout the years.

Thank you to the former and present members of the conservation genetics laboratory- Courtnee', Sihle, Andrinajoro, Riel, Thina, Bright, Jimmy, Jade, Sophia, Isabella, Ian, Melanie and Martyn. Each one of you made the laboratory a better place, with a whole lot of laughter and fun times.

My best friends – Nolyn, thank you for all the love, patience, support and for always believing in me. Courtnee', thank you for all the craziness, laughter and motivation. Phylicia, even though you are so far away, thank you for always being there for me. Each one of you has played a vital role in my life and I truly love and appreciate you all.

To my amazing parents, Anand and Kubashni, thank you for providing the greatest opportunity for me to pursue my dreams. You both, together with my sibling, Melashen, have given me so much of love and support not only through my studies but throughout the years of me growing up. Thank you for always being my pillars of strength and for always believing in me. I love all three of you with all my heart.

List of conferences and symposia

2015

1) **SAEON GFW Drakensberg Global Change Monitoring Platform: Mini Symposium** (October 2015) – Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study.
2) **Symposium of Contemporary Conservation Practice** (November 2015) - Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study **– Prize won: (KZN Premier's Student Award) - 2nd Prize for the best student presentation.**
3) **Durban Action Research Partnership (D'RAP) Year-end Symposium** (December 2015) - Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study **– Prize won: 2nd Prize for the best student project presentation in the Community Reforestation Research Programme.**

2016

4) **Joint SAAB/SASSB Conference 2016** (January 2016) – Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study - **Prize won: 2nd Prize for the best oral presentation by an MSc student.**
5) **UKZN SLS Research Day** (May 2016) - Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study - **Prize won: 1st Prize for the best oral presentation in the venue.**
6) **3rd African Congress of Conservation Biology –** Morocco (September 2016) - Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study.
7) **Durban Action Research Partnership (D'RAP) Year-end Symposium** (December 2016) - Title: The Utility of DNA Barcoding in South African Hemiptera: Using eThekwini as a Case Study.

Table of Contents

List of Figures

List of Tables

Chapter One

Literature review

1.1 Urbanization and biodiversity

The rapid increase in the world's population has resulted in a vast movement of people from rural areas into urban areas (Grimm *et al.* 2008; Moorhead & Philpott 2013). Current estimates suggest that there are more than 6 billion people around the world living in urban areas (Magura *et al.* 2010; Moorhead & Philpott 2013). This shift is known as urbanization and is considered a form of environmental disturbance with important ecological and evolutionary consequences (Rebele 1994; Buczkowski & Richmond 2012). Urbanization is responsible for many land-use changes, such as habitat loss and fragmentation, that result in the alteration and modification of natural habitats (Rebele 1994; Miyashita *et al.* 1998; Gibbs & Stanton 2001; Buczkowski & Richmond 2012; Moorhead & Philpott 2013). Urban areas are densely populated and comprise of a highly diverse mosaic of land-use types which include residential, commercial, industrial, and infrastructural developments (Magura *et al.* 2010; Jones & Leather 2012). Interspersed between these highly transformed habitats are residual patches of natural habitats, which can are referred to as green spaces within cities (Magura *et al.* 2010; Jones & Leather 2012). The current increase in the human population indicates that urbanization is set to continue at an accelerated pace (Jones & Leather 2012). Therefore, understanding the impact of land-use changes on the environment and on the distribution and abundance of biodiversity in urban areas, specifically within the green spaces, is a research priority (Savard *et al.* 2000; Gibbs & Stanton 2001; Grimm *et al.* 2008; Buczkowski & Richmond 2012). It is also important to understand how green spaces within cities can offer refuge for species which would otherwise be displaced by urbanization.

Over the last decade cataloging and understanding the distribution of biodiversity has been at the forefront of worldwide conservation efforts (Savard *et al.* 2000). Biodiversity can be defined in many ways, but the most common definition is the variety of life that transcends all levels of life from genes to communities (Salwasser 1990; Savard *et al.* 2000). The definition of biodiversity also includes the genetic difference within and between species; greater species diversity ensures natural sustainability for all life forms (McKinney 2008).

The presence of high-quality biodiversity in urban areas have intrinsic value and provides essential ecosystem services to communities that live around the green spaces within the cities (McKinney 2008). These services include the purification of air and water, therefore, protecting biodiversity is of both local and global interest (McKinney 2008). Biodiversity comprises of two main attributes; richness and evenness. Richness is the number of species found in an area and is usually referred to as species richness. Evenness is the proportion of species or functional groups present in each area. A region with low evenness indicates that only a few species dominate the area.

Urbanization is responsible for the decrease in both species richness and evenness for most biotic communities (McKinney 2008). In contrast, some species find urban areas favorable for their survival while other species are forced to adapt resulting in a shift in community structure (McKinney 2008; Jones & Leather 2012). For example, urbanization may be responsible for the alteration of species composition by causing a shift from the presence of specialized species to generalist species that are better able to exploit urban environments (Grimm *et al.* 2008; McKinney 2008; Jones & Leather 2012). As urbanization increases, urban green spaces become important for the promotion of biodiversity (Savard *et al.* 2000; McKinney 2008; Jones & Leather 2012). Biodiversity and ecosystems should be valued and managed as part of the urban environment, therefore, green spaces should be accounted for and included in urban planning (Savard *et al.* 2000; Jones & Leather 2012).

1.1.1 Importance of species identification and discovery

The escalating extinction crisis shows the pressure that humanity is placing on the planet. The rapid extinction of species has been referred to as the "global biodiversity crisis" (Alroy 2002; Brooks *et al.* 2006; McKinney 2008; Telfer *et al.* 2015). According to the International Union for Conservation of Nature (IUCN), it is estimated that by mid-century around 30% - 50% of all species on earth will be heading towards extinction (Thomas *et al.* 2004). The current biodiversity crisis is one of the most severe in the history of multi-cellular life on earth (Alroy 2002; Brooks *et al.* 2006). There is great uncertainty as to the number of species at risk of extinction due to incomplete taxonomic coverage for some groups (Alroy 2002; Hebert *et al.* 2003; Meyer & Paulay 2005; Smith *et al.* 2005; Swartz *et al.* 2008). In particular, highly diverse groups that have many morphologically cryptic species which are poorly studied may be problematic (Alroy 2002). This highlights the need to provide tools that can assist in the discovery and identification of species at an accelerated pace, to create biodiversity inventories which can be used for appropriate conservation planning and to highlight species that are threatened by extinction (Savard *et al.* 2000; Alroy 2002; Hebert *et al.* 2003; Smith *et al.* 2005; Jinbo *et al.* 2011; Telfer *et al.* 2015).

Traditional morphology-based taxonomy has made progress in species identification and discovery, however, there is still work to be done. With the use of traditional taxonomy, studies show that over the past 250 years; only 1.2-1.8 million species have been identified and described out of an estimated 7-15 million species (Alroy 2002; Hebert *et al.* 2003; Meier *et al.* 2006; Swartz *et al.* 2008; Packer *et al.* 2009; Chapple & Ritchie 2013; Ajmal Ali *et al.* 2014). The low identification rate is a concern for conservationists and scientists as species may be extinct before identification (Alroy 2002; Smith *et al.* 2005; Swartz *et al.* 2008). This presents a problem for conservation planning and prioritization because species that have not been identified cannot be protected efficiently.

1.1.2 Taxonomy

Taxonomy also known as 'alpha taxonomy' is the backbone of species identification, classification, and discovery (Pires & Marinoni 2010; Ajmal Ali *et al.* 2014). It is the driving force behind the construction of the phylogenetic tree of life, the construction of biodiversity inventories and it helps provide baseline data for conservation and ecological studies (Wilson 2004; Pires & Marinoni 2010; Ajmal Ali *et al.* 2014). Without taxonomy, scientists cannot accurately report experimental results or access published information on species (Meier *et al.* 2006; Pires & Marinoni 2010). Traditional taxonomy sorts specimens and groups them into taxonomic units by using multiple morphological characteristics, similarity, and contiguity to delineate species (Ajmal Ali *et al.* 2014; Decraemer & Backeljau 2015). Species are usually named using the Linnaean nomenclature (Decraemer & Backeljau 2015). To accurately discover and identify species, special skills through extensive training and experience are required. As such, species identification is usually performed by expert taxonomists, and trained technicians (Hebert *et al.* 2003; Jinbo *et al.* 2011).

The greatest challenge placed on traditional morphology-based taxonomy is the slow rate at which taxonomists can identify and describe species (Hebert *et al.* 2003; Smith *et al.* 2005; Meier *et al.* 2006; Jinbo *et al.* 2011). Current taxonomists cannot cope with the overwhelming need to create species inventories and are faced with a lack of funding for basic taxonomic research resulting in the decline of taxonomic expertise (Swartz *et al.* 2008; Pires & Marinoni 2010). In addition, Hebert and colleagues (2003) highlighted four major limitations using traditional morphology-based taxonomy. First, phenotypic plasticity and variability in the morphological characters used for species identification may lead to the incorrect assignment of specimens to species (Hebert *et al.* 2003). Second, morphologically cryptic species may be overlooked (Hebert *et al.* 2003). Third, the lack of taxonomic keys to identify specimens at different life stages and different genders may lead to incorrect identification of species (Hebert *et al.* 2003; Swartz *et al.* 2008). Finally, high levels of expertise are required for traditional taxonomic assignments which are a problem given the current decline in taxonomists (Hebert *et al.* 2003). Due to the challenges and limitations of traditional taxonomy, alternative, and complementary approaches to identify species are needed (Lambert *et al.* 2005; Pires & Marinoni 2010; Jinbo *et al.* 2011).

1.1.3 Molecular-based taxonomy

Molecular-based methods of species identification have been proposed to assist in taxonomy (Lambert *et al.* 2005; Pires & Marinoni 2010; Jinbo *et al.* 2011). Molecular approaches exploit the diversity among DNA sequences to classify organisms, as there is expected to be less genetic diversity among taxa that belong to the same taxonomic grouping (Hebert *et al.* 2003). DNA sequences can be used to assess the diversity of life and the analysis of evolutionary relationships within and among groups of organisms (Stoeckle 2003; Hajibabaei *et al.* 2007).

The main justifications for using a DNA-based system is that sequences can be used for the identification of species and species boundaries in populations of morphologically similar organisms, thus accelerating cryptic species discovery (Hebert *et al.* 2003; Stoeckle 2003; Lambert *et al.* 2005; Jinbo *et al.* 2011; Nunes *et al.* 2014). Biotechnology and systematics are both rapidly progressing fields and are needed to play a practical role in taxonomy (Smith *et al.* 2005; Jinbo *et al.* 2011). These fields have contributed to the establishment of DNA barcoding as a complementary system to morphology-based species identification (Smith *et al.* 2005; Jinbo *et al.* 2011). DNA barcoding has been suggested as a mechanism which enables biologists to “label” or “tag” species, using nucleotide variations in short sequences known as DNA barcodes (Hebert *et al.* 2003).

1.2 DNA Barcoding

The beginning of the DNA barcoding era started in 2003 when researchers (Hebert and colleagues, 2003) at the University of Guelph in Ontario, Canada proposed DNA barcoding as a way of identifying species (Hebert *et al.* 2003; Meyer & Paulay 2005; Jinbo *et al.* 2011; Nunes *et al.* 2014). The advantages of DNA barcoding include being able to sort specimens to a species level in a relatively rapid, accurate, automated, and cost-effective manner (Hebert *et al.* 2003; Hebert & Gregory 2005; Meyer & Paulay 2005; Jinbo *et al.* 2011; Nunes *et al.* 2014). DNA barcoding is no replacement for traditional-morphology based taxonomy, however, it is used to aid the taxonomic workflow and complement traditional taxonomy, thus making the Linnaean taxonomic system more accessible to non-taxonomic experts (Hebert & Gregory 2005; Hajibabaei *et al.* 2007). DNA barcoding uses short, standardized gene regions as internal species tags (Hebert *et al.* 2003; Hebert & Gregory 2005; Kress & Erickson 2008; Luo *et al.* 2011; Chapple & Ritchie 2013).

In May 2004, a project known as The Consortium for the Barcode of Life (CBOL) was established with the main aim of developing a standard protocol for DNA barcoding and providing bioinformatic solutions to facilitate the storage, analysis, and curation of DNA barcode data (Hajibabaei *et al.* 2007; Ratnasingham & Hebert 2007; Pires & Marinoni 2010; Jinbo *et al.* 2011). This project has grown rapidly and now includes more than 170 organizations from 45 nations (Ratnasingham & Hebert 2007; Pires & Marinoni 2010; Jinbo *et al.* 2011). The Barcode of Life Data System (BOLD) is an online database which can provide an integrated bioinformatics platform and support all stages of the analytical pathway of the DNA barcoding process from specimen collection to the DNA-based identification of specimens to species (Ratnasingham & Hebert 2007). The success of this global barcoding initiative rests on the construction of a comprehensive DNA barcode reference library for all life on earth (Meyer & Paulay 2005; Ratnasingham & Hebert 2007; Virgilio *et al.* 2010; Jinbo *et al.* 2011; Nunes *et al.* 2014).

1.2.1 Selection of the mitochondrial cytochrome c oxidase 1 (COI) gene as the standard identification marker

In 1977, Carl Woese was the first researcher to investigate the difference between nucleotide sequences of a single gene to understand evolutionary relationships (Woese & Fox 1977; Hebert *et al.* 2004). He illustrated that analyzing sequence differences using a conserved gene such as ribosomal RNA (rRNA) could be used to infer phylogenetic relationships (Woese & Fox 1977; Hebert *et al.* 2004). Researchers then investigated the use of genes that evolve rapidly to differentiate among closely related organisms (Brown *et al.* 1979; Hebert *et al.* 2004). The results of this investigation showed that rapidly evolving genes can provide insight into both the genetic relationships among closely related species and the genetic variability within species (Brown *et al.* 1979; Hebert *et al.* 2004). Most studies that focus on the relationship among closely related species use markers from the mitochondrial genome because it has a higher rate of nucleotide substitution than the nuclear genome (Brown *et al.* 1979; Hebert *et al.* 2004).

For a gene region to be considered as a DNA barcode, the region must satisfy three criteria (Kress & Erickson 2008). First, the gene must contain a significant species-level genetic variability (Kress & Erickson 2008). Second, it must contain conserved flanking sites for universal PCR primers to be developed that can be used for wide taxonomic application (Kress & Erickson 2008). Last, the gene must have a short sequence length to facilitate easy DNA extraction and PCR amplification (Kress & Erickson 2008). The standardized DNA barcode used in most animal groups is a 658-base pair (bp) protein-coding region known as the mitochondrial cytochrome c oxidase 1 (COI) gene (Hebert *et al.* 2003; Hajibabaei *et al.* 2007; Ratnasingham & Hebert 2007; Kress & Erickson 2008; Luo *et al.* 2011; Nunes *et al.* 2014). This region is easily amplified and Sanger sequenced using universal primers (Folmer *et al.* 1994; Vrijenhoek 1994; Hebert *et al.* 2004; Nunes *et al.* 2014). The COI region is easily alignable as insertions and deletions are not common and this mitochondrial gene contains enough DNA mutations to provide resolution at the species level (Hebert *et al.* 2004; Barrett & Hebert 2005; Nunes *et al.* 2014).

Despite being successfully used in many animal groups such as Aves (Hebert *et al.* 2004), Lepidoptera (Hebert *et al.* 2003) and Araneae (Barrett & Hebert 2005), the use of a single marker to determine species has been criticized (Moritz & Cicero 2004). Although this gene fragment is often used in animal barcoding studies, it is not effective in plants and fungi (Hebert *et al.* 2003; Kress & Erickson 2008; Chase & Fay 2009; Hollingsworth 2011; Jinbo *et al.* 2011). Instead, two standard barcoding regions are used in plants; these regions are situated in the plastid (chloroplast) genome and include the rbcL and matK genes (Kress & Erickson 2008; Chase & Fay 2009; Hollingsworth 2011; Jinbo *et al.* 2011). In fungi, the internal transcribed spacers (ITS) of the nuclear ribosomal DNA region are used for DNA barcoding (Kress & Erickson 2008; Chase & Fay 2009; Hollingsworth 2011; Jinbo *et al.* 2011).

1.2.2 DNA Barcode Gap as a measure of identification accuracy

One of the main requirements of DNA barcoding is a database containing reference sequences from already described species that can be used as part of an automated identification system (Ratnasingham & Hebert 2007; Jinbo *et al.* 2011). The BOLD database is a very useful bioinformatic tool for biodiversity researchers, as it allows free access to a large amount of data of taxa collected from all around the world (Hajibabaei *et al.* 2007; Ratnasingham & Hebert 2007; Jinbo *et al.* 2011). Sorting of taxa into species using DNA barcoding is done using genetic distance thresholds (Barrett & Hebert 2005; Meyer & Paulay 2005; Wiemers & Fiedler 2007). When comparing an unknown sample to known samples there are one of two outcomes. First, if the genetic distance of the unknown sample is less than the threshold value of an existing sample, it is assumed that the two samples represent the same species (Meyer & Paulay 2005; Wiemers & Fiedler 2007). Second, if the genetic distance of the unknown sample is greater than the threshold value of existing samples, distinctness is indicated and it is then assumed that the unknown sample represents a new taxon (Meyer & Paulay 2005; Wiemers & Fiedler 2007).

An important consideration of DNA barcoding is identification accuracy (Meyer & Paulay 2005). Identification accuracy using threshold-based approaches relies on the level of overlap between intraspecific and interspecific variation across a selected phylogeny (Meyer & Paulay 2005). The existence of the 'DNA barcode gap' in a dataset, is a useful way of predicting whether DNA barcoding will be successful for the taxon of choice (Hebert *et al.* 2003). The DNA barcode gap is the separation between the mean intraspecific and interspecific genetic distance (Meyer & Paulay 2005; Wiemers & Fiedler 2007; Puillandre *et al.* 2012; Čandek & Kuntner 2015). The DNA barcode gap occurs when the interspecific genetic divergence is greater than the intraspecific genetic variation (Figure 1A) (Barrett & Hebert 2005; Wiemers & Fiedler 2007; Luo *et al.* 2011; Chapple & Ritchie 2013). This gap indicates that the selected marker used in DNA barcoding can reliably and successfully distinguish between species of the taxon of choice (Meyer & Paulay 2005; Jinbo *et al.* 2011; Čandek & Kuntner 2015). When the intraspecific and interspecific genetic distances overlap, the DNA barcode gap is absent and taxa cannot be reliably sorted into species based on the genetic data (Figure 1B) (Meyer & Paulay 2005; Jinbo *et al.* 2011; Puillandre *et al.* 2012; Chapple & Ritchie 2013).

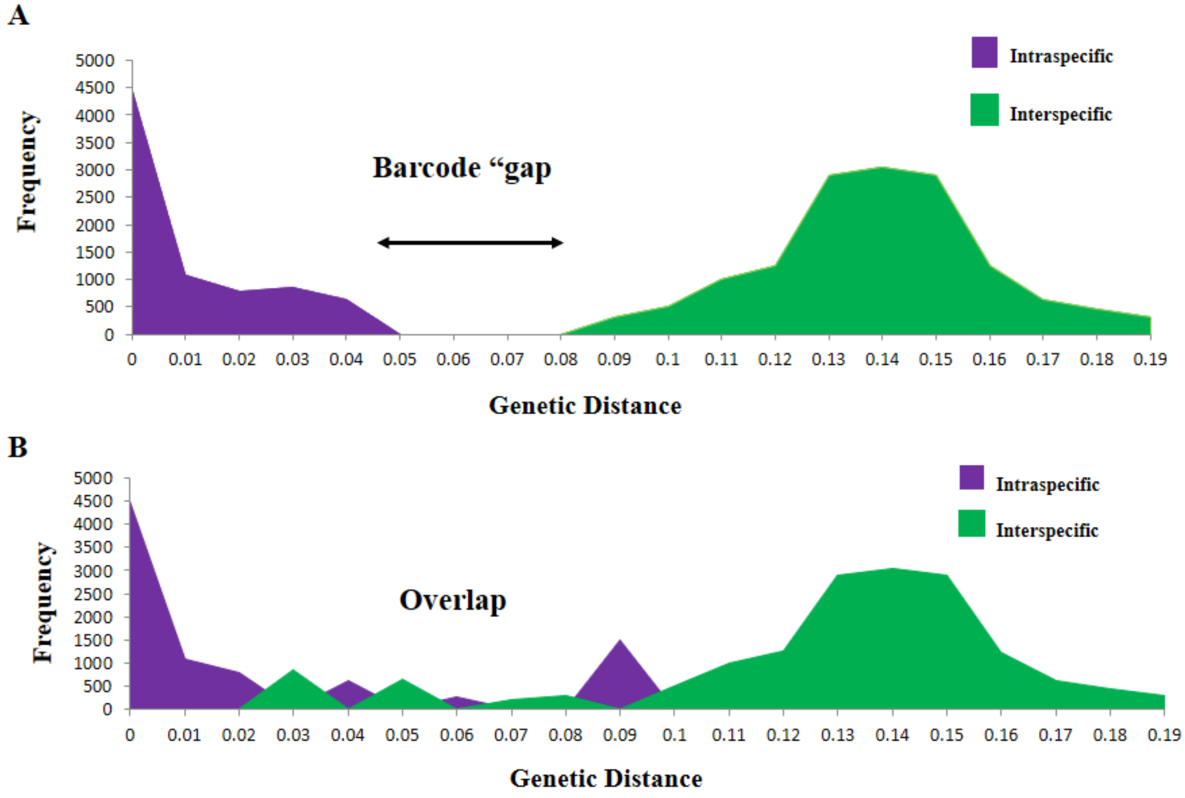

Figure 1.1: The DNA barcode gap, Figure adapted from Meyer and Paulay (2005). The distribution of intraspecific variation is shown in purple and the interspecific divergence is shown in green.

When the barcode gap is present (Figure 1 A) there is a distinct difference between intraspecific variation and interspecific divergence. When there is no DNA barcode gap (Figure 1 B) the intraspecific variance and interspecific genetic distance are continuous. Taxa with sequence divergence values which fall within the overlap cannot be sorted into species with confidence.

There are many reasons as to why there may be a lack of the DNA barcode gap in a given dataset. First, there might be elevated levels of genetic diversity below the species level (Jinbo *et al.* 2011). Second, the extent of the geographic sampling region could affect rates of intraspecific divergence (Jinbo *et al.* 2011; Chapple & Ritchie 2013). Third, the presence of closely related species and/or poorly understood taxonomy of selected groups (i.e. cryptic species) (Wiemers & Fiedler 2007; Jinbo *et al.* 2011; Chapple & Ritchie 2013).

Despite the importance of the DNA barcode gap in the accuracy of DNA barcoding, few studies have tested for the presence of the gap. The DNA barcode gap analysis that has previously been carried out on birds (Hebert *et al.* 2004) and arthropods (Barrett & Hebert 2005) showed the presence of a distinct barcode gap. The existence of this gap, however, has been put into question when the analysis was carried out on groups such as gastropods (Meyer & Paulay 2005) and Diptera (Meier *et al.* 2006), where there was no barcode gap present. Understanding what factors affect the success of barcoding is a vital component of any DNA barcoding study.

1.3 Importance of applying DNA barcoding to arthropods

The phylum Arthropoda includes insects, arachnids, myriapods, and crustaceans. Arthropods are regarded as the most species-rich animal phylum in most terrestrial ecosystems (Bolger *et al.* 2000; Wiemers & Fiedler 2007). This phylum is very useful to study the effects of climate change and urbanization, because of their great abundance, diversity, functional importance and because they can be found in a wide range of diverse ecosystems (Kremen *et al.* 1993; Bolger *et al.* 2000; Andersen *et al.* 2004; Buczkowski & Richmond 2012; Jones & Leather 2012).

Arthropods play important functional roles in ecosystem processes and as significant biological indicators of ecological change due to their short life cycles and sensitivity to habitat disturbances (Kremen *et al.* 1993; Bolger *et al.* 2000; Andersen *et al.* 2004; Buczkowski & Richmond 2012; Jones & Leather 2012). They are therefore able to respond to environmental changes more rapidly than vertebrates (Kremen *et al.* 1993). Despite the increasing issues surrounding the global biodiversity crisis and conservation planning, very little attention has been given to the creation of biodiversity inventories and monitoring of terrestrial arthropods in ecologically sensitive parts of the world (Kremen *et al.* 1993; Andersen *et al.* 2004; Smith *et al.* 2005).

To create biodiversity inventories for arthropods, routine taxonomic identification needs to be carried out. Two of the major problems surrounding the creation of biodiversity inventories for arthropods includes the species identification of this phylum being extremely time-consuming and the lack of taxonomists available to carry out routine taxonomic identifications for most groups

(Wiemers & Fiedler 2007). The percentage of undescribed species within arthropods is much higher than that of vertebrates, resulting in an increased need for improved identification tools to identify a large range of arthropod samples successfully and reliably (Wiemers & Fiedler 2007; Virgilio *et al.* 2010).

Since the beginning of the DNA barcoding era, there have been many studies that have been carried out on selected arthropod taxa. These studies include Diptera (Meier *et al.* 2006), Lepidoptera (Wiemers & Fiedler 2007; Sourakov & Zakharov 2011), Coleoptera (Greenstone *et al.* 2005), Araneae (Čandek & Kuntner 2015), and Formicidae (Smith *et al.* 2013). DNA barcoding is crucial for the identification of certain arthropods at different developmental stages (e.g. eggs, larvae, nymphs, or pupae) which are difficult or impossible to identify using traditional taxonomy (Virgilio *et al.* 2010). For example, DNA barcoding has been proven to be successful when trying to identify eggs and larvae of closely related carabids and spiders which are natural predators that are important for the biological control of agricultural pests (Greenstone *et al.* 2005; Virgilio *et al.* 2010). Furthermore, using DNA barcoding, researchers from the University of Florida and the University of Guelph were able to clarify taxonomic relationships within the Caribbean butterfly genus *Calisto* (Sourakov & Zakharov 2011). They described this genus as "often highly cryptic and confusing". These researchers sequenced barcodes linked to 31 separate taxa and their efforts resulted in a taxonomic revision of the genus now comprising of 34 species and 17 subspecies (Sourakov & Zakharov 2011).

1.3.1 Hemiptera

Hemiptera, known as true bugs, are one of the largest, most diverse groups of hemimetabolous insects (i.e. insects undergoing incomplete metamorphosis in which the young do not resemble the adult) (Park *et al.* 2011; Raupach *et al.* 2014). Currently, there are more than 42 000 described species worldwide, consisting of 5800 genera and 140 families (Henry 2009; Park *et al.* 2011; Raupach *et al.* 2014). The Hemiptera order is divided into three suborders: Hydrocorizae (aquatic bugs); Amphibicorizae (semi-aquatic or shore-inhabiting bugs) and Geocorizae (terrestrial bugs) (Wheeler 2001). This order is found worldwide in a variety of ecosystems and habitats, with exception of the deep sea and polar regions (Raupach *et al.* 2014).

Hemiptera are of high ecological and economic importance making their identification highly desirable (Henry 2009; Park *et al.* 2011; Raupach *et al.* 2014). They are known to play an important role in agriculture, they cause direct damage to plants by herbivory, they are indirectly involved in the transportation of diseases and are used as biological control agents (Wheeler 2001; Park *et al.* 2011; Raupach *et al.* 2014). However, the vast majority of Hemiptera are harmless insects and some of the predatory forms can be regarded as beneficial when they habitually prey on insect pests (Raupach *et al.* 2014).

Given the lack of Hemiptera taxonomic expertise in South Africa and the general difficulty in identifying species, especially throughout their various nymphal stages, traditional identification of this group has proven to be very difficult and time-consuming (Raupach *et al.* 2014). To alleviate this difficulty, DNA barcoding has been proposed as a complementary tool to morphological-based methods for the identification of Hemiptera (Raupach *et al.* 2014).

DNA barcoding of the South African Hemiptera will not only improve the representation of South African taxa in the BOLD, it will also allow for the assessment of the local biodiversity within the study sites (Valentini *et al.* 2008). In general, most of the biodiversity surveys (particularly DNA barcode studies) have focused on rural areas and protected areas; very rarely do they survey the biodiversity and species richness within an urban context (Helden & Leather 2004). The richness of diverse groups of species belonging to Hemiptera has not been previously surveyed (nature reserves, parks, and undeveloped land) in South Africa. Biodiversity data at a regional scale is therefore needed for accurate and appropriate conservation planning.

1.4 Aim of study

There are two main aims of this study:

1. To construct a preliminary reference library for Hemiptera collected from the eThekwini region, South Africa. Here I will assess the utility of the DNA barcode marker (COI) to differentiate among morphospecies of Hemiptera. One key consideration is the existence of the DNA barcode gap. If there is no significant differentiation between inter- and intraspecific genetic diversity then this will limit specimen identification success using DNA barcoding. This project will contribute towards the multi-taxon Urban Barcode project currently being run by the Willows-Munro research group at UKZN and through consultation with taxonomic experts the data from

the project will be added to the global Hemiptera DNA reference library available through Barcode of life database (BOLD).

2. To test the suitability of Hemiptera to monitor the success of ecological habitat restoration. The Buffelsdraai Landfill Site Community Reforestation Project will be examined as a case study. This portion of the study will involve the establishment of a reference library of Hemiptera collected from the Buffelsdraai site. Examining the data will allow for a suite of indicator species with high habitat specificity to be selected. This project will also aim to assess changes in species richness and diversity in reforested sites of different ages ranging from the years 2010 to 2015.

1.5 References

Ajmal Ali M., Gyulai G., Hidvégi N., Kerti B., Al Hemaid F.M.A., Pandey A.K. & Lee J. (2014) The changing epitome of species identification – DNA barcoding. *Saudi Journal of Biological Sciences* **21**, 204-231.

Alroy J. (2002) How many named species are valid? *Proceedings of the National Academy of Sciences* **99**, 3706-3711.

Andersen A.N., Fisher A., Hoffmann B.D., Read J.L. & Richards R.O.B. (2004) Use of terrestrial invertebrates for biodiversity monitoring in Australian rangelands, with particular reference to ants. *Austral Ecology* **29**, 87-92.

Barrett R.D.H. & Hebert P.D.N. (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* **83**, 481-491.

Bolger D.T., Suarez A.V., Crooks K.R., Morrison S.A. & Case T.J. (2000) Arthropods in urban habitat fragments in southern California: area, age, and edge effects. *Ecological Applications* **10**, 1230-1248.

Brooks T.M., Mittermeier R.A., da Fonseca G.A.B., Gerlach J., Hoffmann M., Lamoreux J.F., Mittermeier C.G., Pilgrim J.D. & Rodrigues A.S.L. (2006) Global biodiversity conservation priorities. *Science* **313**, 58-61.

Brown W.M., George M. & Wilson A.C. (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences* **76**, 1967-1971.

Buczkowski G. & Richmond D.S. (2012) The effect of urbanization on ant abundance and diversity: a temporal examination of factors affecting biodiversity. *PLoS One* **7**, e41729.

Čandek K. & Kuntner M. (2015) DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources* **15**, 268-277.

Chapple D.G. & Ritchie P.A. (2013) A retrospective approach to testing the DNA barcoding method. *PLoS One* **8**, e77882.

Chase M. & Fay M. (2009) Barcoding of plants and fungi. *Science* **325**, 682.

Decraemer W. & Backeljau T. (2015) Utility of classical α-taxonomy for biodiversity of aquatic nematodes. *Journal of Nematology* **47**, 1-10.

Folmer O., Black M., Hoeh W., Lutz R. & Vrijenhoek R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology Biotechnology* **3**, 294-199.

Gibbs J.P. & Stanton E.J. (2001) Habitat fragmentation and arthropod community change: carrion beetles, phoretic mites, and flies. *Ecological Applications* **11**, 79-85.

Greenstone M.H., Rowley D.L., Heimbach U., Lundgren J.G., Pfannenstiel R.S. & Rehner S.A. (2005) Barcoding generalist predators by polymerase chain reaction: carabids and spiders. *Molecular Ecology* **14**, 3247-3266.

Grimm N.B., Faeth S.H., Golubiewski N.E., Redman C.L., Wu J., Bai X. & Briggs J.M. (2008) Global change and the ecology of cities. *Science* **319**, 756-760.

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**, 167-172.

Hebert P. & Gregory R. (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology* **54**, 852-859.

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

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S. & Francis C.M. (2004) Identification of birds through DNA barcodes. *PLoS Biology* **2**, e312.

Helden A.J. & Leather S.R. (2004) Biodiversity on urban roundabouts—Hemiptera, management and the species–area relationship. *Basic and Applied Ecology* **5**, 367-377.

Henry T.J. (2009) Biodiversity of Heteroptera. In: Foottit R. & Adler P. (eds), *Insect Biodiversity*: *Science and society*. Washington, DC: Wiley-Blackwell. pp. 223-263.

Hollingsworth P.M. (2011) Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **108**, 19451-19452.

Jinbo U., Kato T. & Ito M. (2011) Current progress in DNA barcoding and future implications for entomology. *Entomological Science* **14**, 107-124.

Jones E. & Leather S. (2012) Invertebrates in urban areas: A review. *European Journal of Entomology* **109**, 463-478.

Kremen C., Colwell R.K., Erwin T.L., Murphy D.D., Noss R.F. & Sanjayan M.A. (1993) Terrestrial arthropod assemblages: Their use in conservation planning. *Conservation Biology* **7**, 796-808.

Kress W.J. & Erickson D.L. (2008) DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences* **105**, 2761-2762.

Lambert D.M., Baker A., Huynen L., Haddrath O., Hebert P.D.N. & Millar C.D. (2005) Is a large-scale DNA-based inventory of ancient life possible? *Journal of Heredity* **96**, 279-284.

Luo A., Zhang A., Ho S.Y., Xu W., Zhang Y., Shi W., Cameron S.L. & Zhu C. (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. *BMC Genomics* **12**, 1-13.

Magura T., Horváth R. & Tóthmérész B. (2010) Effects of urbanization on ground-dwelling spiders in forest patches, in Hungary. *Landscape Ecology* **25**, 621-629.

McKinney M.L. (2008) Effects of urbanization on species richness: A review of plants and animals. *Urban Ecosystems* **11**, 161-176.

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**, 715-728.

Meyer C.P. & Paulay G. (2005) DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology* **3**, e422.

Miyashita T., Shinkai A. & Chida T. (1998) The effects of forest fragmentation on web spider communities in urban areas. *Biological Conservation* **86**, 357-364.

Moorhead L.C. & Philpott S.M. (2013) Richness and composition of spiders in urban green spaces in Toledo, Ohio. *Journal of Arachnology* **41**, 356-363.

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

Nunes V.L., Mendes R., Marabuto E., Novais B.M., Hertach T., Quartau J.A., Seabra S.G., Paulo O.S. & Simões P.C. (2014) Conflicting patterns of DNA barcoding and taxonomy in the cicada genus Tettigettalna from southern Europe (Hemiptera: Cicadidae). *Molecular Ecology Resources* **14**, 27-38.

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.

Pires A.C. & Marinoni L. (2010) DNA barcoding and traditional taxonomy unified through Integrative Taxonomy: a view that challenges the debate questioning both methodologies. *Biota Neotropica* **10**, 339-346.

Puillandre N., Lambert A., Brouillet S. & Achaz G. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology* **21**, 1864-1877.

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

Raupach M.J., Hendrich L., Küchler S.M., Deister F., Morinière J. & 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**, e106940.

Rebele F. (1994) Urban ecology and special features of urban ecosystems. *Global Ecology and Biogeography Letters* **4**, 173-187.

Salwasser H. (1990) Conservation of diversity in forest ecosystem conserving biological diversity: A perspective on scope and approaches. *Forest Ecology and Management* **35**, 79-90.

Savard J.P.L., Clergeau P. & Mennechez G. (2000) Biodiversity concepts and urban ecosystems. *Landscape and Urban Planning* **48**, 131-142.

Smith M.A., Fisher B.L. & Hebert P.D. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **360**, 1825-1834.

Smith M.A., Hallwachs W., Janzen D.H. & Segura R.B. (2013) DNA Barcoding a collection of ants (Hymenoptera: Formicidae) from Isla Del Coco, Costa Rica. *Florida Entomologist* **96**, 1500-1507.

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.

Stoeckle M. (2003) Taxonomy, DNA, and the Bar Code of Life. *BioScience* **53**, 796-797.

Swartz E.R., Mwale M. & Hanner R. (2008) Review article: A role for barcoding in the study of African fish diversity and conservation. *South African Journal of Science* 104, 293-298.

Telfer A.C., Young M.R., Quinn J., Perez K., Sobel C.N., Sones J.E., Levesque-Beaudin V., Derbyshire R., Fernandez-Triana J., Rougerie R., Thevanayagam A., Boskovic A., Borisenko A.V., Cadel A., Brown A., Pages A., Castillo A.H., Nicolai A., Glenn Mockford B.M., Bukowski B., Wilson B., Trojahn B., Lacroix C.A., Brimblecombe C., Hay C., Ho C., Steinke C., Warne C.P., Garrido Cortes C., Engelking D., Wright D., Lijtmaer D.A., Gascoigne D., Hernandez Martich D., Morningstar D., Neumann D., Steinke D., Marco DeBruin D.D., Dobias D., Sears E., Richard E., Damstra E., Zakharov E.V., Laberge F., Collins G.E., Blagoev G.A., Grainge G., Ansell G., Meredith G., Hogg I., McKeown J., Topan J., Bracey J., Guenther J., Sills-Gilligan J., Addesi J., Persi J., Layton K.K., D'Souza K., Dorji K., Grundy K., Nghidinwa K., Ronnenberg K., Lee K.M., Xie L., Lu L., Penev L., Gonzalez M., Rosati M.E., Kekkonen M., Kuzmina M., Iskandar M., Mutanen M., Fatahi M., Pentinsaari M., Bauman M., Nikolova N., Ivanova N.V., Jones N., Weerasuriya N., Monkhouse N., Lavinia P.D., Jannetta P., Hanisch P.E., McMullin R.T., Ojeda Flores R., Mouttet R., Vender R., Labbee R.N., Forsyth R., Lauder R., Dickson R., Kroft R., Miller S.E., MacDonald S., Panthi S., Pedersen S., Sobek-Swant S., Naik S., Lipinskaya T., Eagalle T., Decaens T., Kosuth T., Braukmann T., Woodcock T., Roslin T., Zammit T., Campbell V., Dinca V., Peneva V., Hebert P.D. & deWaard J.R. (2015) Biodiversity inventories in high gear: DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve. *Biodiversity Data Journal* **3**, e6313.

Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J., Collingham Y.C., Erasmus B.F.N., de Siqueira M.F., Grainger A., Hannah L., Hughes L., Huntley B., van Jaarsveld A.S., Midgley G.F., Miles L., Ortega-Huerta M.A., Townsend Peterson A., Phillips O.L. & Williams S.E. (2004) Extinction risk from climate change. *Nature* **427**, 145-148.

Valentini A., Pompanon F. & Taberlet P. (2008) DNA barcoding for ecologists. *Trends in Ecology & Evolution* **24**, 110-117.

Virgilio M., Backeljau T., Nevado B. & De Meyer M. (2010) Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* **11**, 1-10.

Vrijenhoek R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology* **3**, 294-299.

Wheeler A.G. (2001) *Biology of the Plant Bugs (Hemiptera: Miridae): Pests, Predators, Opportunists*. New York: Cornell University Press.

Wiemers M. & Fiedler K. (2007) Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* **4**, 8.

Wilson E.O. (2004) Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**, 739.

Woese C.R. & Fox G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5088-5089.

Chapter Two

Construction of a preliminary reference library for Hemiptera collected from the eThekwini region and surrounding areas and the assessment of DNA barcoding using the COI marker

Abstract

Hemiptera are useful bio-indicators and constitute an important component of South African insect fauna. This order has been used in the biological control of invasive plants and plays an important role in agriculture. Identification of Hemiptera using traditional taxonomy can be challenging. Challenges include the separation of cryptic species and the delimitation of individuals from different developmental stages. DNA barcoding is a molecular-based system designed to provide a rapid, accurate and automatable method for species identification. A short, standardized gene region, the mitochondrial cytochrome c subunit 1 (COI), is used as a species tag. This study presents a preliminary DNA barcode reference library for Hemiptera collected from the eThekwini region and surrounding areas (Durban, KwaZulu-Natal, South Africa). This region is situated within the Maputaland-Pondoland-Albany biodiversity hotspot. In total, a reference library containing over 1456 mitochondrial COI barcodes representing a total of at least 357 barcode clusters collected from 18 localities were analyzed. This study tested for the presence of the "DNA barcode gap". This is the gap between the intraspecific and interspecific genetic distances – the larger the gap the more accurate species delimitation. This study highlights the utility of DNA barcoding and assesses the biodiversity within the eThekwini region for suitable conservation and management of this region.

2.1 Introduction

2.1.1 Biodiversity hotspots in South Africa and the eThekwini region

Rapid growth in the human population has had a negative impact on the natural ecosystem and has resulted in a substantial depletion of biodiversity (Grimm *et al.* 2008; Moorhead & Philpott 2013). Therefore, there is a crucial need to act, protect and conserve the biodiversity that is left on our planet during this global biodiversity crisis (Brooks *et al.* 2006; McKinney 2008). Questions have been raised as to which geographic regions are a priority to protect and maintain biodiversity (Reid 1998; Myers *et al.* 2000). A successful method to select important geographic regions is to identify biodiversity 'hotspots' (Myers *et al.* 2000). Biodiversity hotspots are regions, which are rich in species, contain many endemic and threatened species and are experiencing a grave loss of natural habitat (Reid 1998; Myers *et al.* 2000).

South Africa, located on the southernmost tip of the African continent, is recognized by its biodiversity hotspots of global importance (Rouget *et al.* 2004). These include the Maputaland-Pondoland-Albany region, the Cape Floristic region and the Succulent Karoo (Rouget *et al.* 2004; Zachos & Habel 2011). In South Africa, approximately 65 500 animal species have been described, with over 44 000 of these being insects (Scholtz 1999; Hamer 2013; Silva & Willows-Munro 2016). However, this represents less than half of the actual fauna richness in South Africa (Hamer 2013).

This study focuses on the globally important Maputaland-Pondoland-Albany biodiversity hotspot, with interest being placed on the eThekwini municipal region and surrounding areas. EThekwini is a metropolitan municipality situated in KwaZulu-Natal, South Africa. Although the municipality is situated within the globally-important Maputaland-Pondoland-Albany biodiversity hotspot, the region is under increasing pressure from urbanization (Croucamp 2009). The increase in urbanization is due to the city's rapid expansion to accommodate its increasing human population (> 3.5 million) (Govender 2014). Not only is it threatened by urbanization it is also vulnerable to the effects of the global climate change (Govender 2014). This urban area is highly developed with a 55% loss of natural habitat and is a critically endangered ecosystem with limited knowledge of the biodiversity present within the region.

To conserve the biodiversity in the city, the Durban Metropolitan Open Space System (D'MOSS) was developed to assist in town planning. D'MOSS is a network of green open spaces linking together areas of high biodiversity value such as nature reserves and undeveloped pieces of privately and municipality-managed land (Willows-Munro 2013; Govender 2014). The D'MOSS network consists of 74 500 ha of open spaces and is composed of a variety of habitat types, however, only 10% of the D'MOSS area is formally protected (Govender 2014). There are more than 2200 described plant species, 82 terrestrial mammal species, 520 bird species, 69 reptile species, 37 frog species and 25 endemic invertebrate groups found in the D'MOSS region (Govender 2014). This only represents a fraction of the biodiversity found within the region. D'MOSS was created to generate much needed knowledge of the biodiversity and to assist managers in the municipality in the decision making of prioritizing regions for biodiversity and conservation.

2.1.2 Species inventories in South Africa

While South Africa contains major globally important hotspots, our country is not exempt from concerns over the conservation of biodiversity. The rapid expansion of urbanization within the eThekwini region highlights the importance of creating species inventories for the biodiversity hotspots within South Africa. The creation of species inventories helps to protect and conserve the biodiversity that is being placed under increasing anthropogenic pressure (Savard *et al.* 2000; Alroy 2002; Hebert *et al.* 2003; Smith *et al.* 2005). Knowledge of what species exist within South Africa is important for biodiversity surveillance, the detection of invasive taxa and potentially uncovering new or cryptic species (Hebert *et al.* 2003; Jinbo *et al.* 2011; Krishna Krishnamurthy & Francis 2012). To conserve this biodiversity, we need to take note of several factors such as the distribution and habitat association of species, species richness and diversity in each area, and what endemic species occur within each area (Krishna Krishnamurthy & Francis 2012).

When trying to create a comprehensive species inventory difficulty is experienced in hyper-diverse groups such as terrestrial arthropods (Alroy 2002). The reason for this is the presence of closely related species, cryptic species, and individuals at different developmental stages. In creating these species inventories, South Africa has made very slow progress in describing all the species because we do not have the taxonomic capacity due to a shortage of expert taxonomists (Krishna Krishnamurthy & Francis 2012). However, attempting to catalogue and monitor biodiversity even at a small regional scale is a daunting task (Krishna Krishnamurthy & Francis 2012). Therefore, molecular techniques have been proposed to complement traditional taxonomy, one of the techniques gaining popularity is that of DNA barcoding (Hebert *et al.* 2003; Jinbo *et al.* 2011).

DNA barcoding can be used for biodiversity assessments at both regional and global scales, as well as enabling the discovery of new species by providing a rapid, accurate and automatable way of species identification and discovery (Hebert *et al.* 2003; Jinbo *et al.* 2011). The short, standardized gene region, the mitochondrial cytochrome c subunit 1 (COI), is used as a species tag for many animal taxa. DNA barcoding can also be used to create species inventories by enabling the creation of reference libraries in which molecular data is coupled with reliable morphological taxonomic knowledge (Cesari *et al.* 2013). A reliable reference library contains reference sequences of voucher specimens that have previously been verified by expert taxonomists (Cesari *et al.* 2013). DNA barcoding can be considered as a complementary tool to traditional taxonomy for the creation of reference libraries for key taxa (Cesari *et al.* 2013).

This study will focus on the biodiversity present within the eThekwini region. Creating a reference library of the biodiversity within this region is important to enable the assessment of biodiversity and assist with proper town planning. This study will focus on creating a reference library of Hemiptera for the different open spaces incorporated into D'MOSS. If the data generated by this study is coupled with climatic data generated by the municipality, this genetic data has a potential to provide a valuable environmental monitoring tool to understand the impact of urbanization on the biodiversity within the eThekwini region (Willows-Munro 2013). Hemiptera was selected as the study species as they are found within a wide range of ecosystems and can be used as environmental indicators. Hemiptera are useful biological indicators and constitute an important component of South African insect fauna. Currently, there are more than 42 000 species of Hemiptera described worldwide (Raupach *et al.* 2014). Hemiptera are well represented in the BOLD database, however, a very few of the represented species come from South Africa. Currently there is a total of 267 511 barcode records present on BOLD, however, only 180 521 records have been published. These records form 13862 barcode index number (BINs), with specimens from 145 countries deposited by 133 institutions. Of these published records, only 72 506 barcodes have species names, representing 6 915 species. Moreover, only 3 414 published records were collected from South Africa. Identification of Hemiptera using traditional morphology-based identification is time consuming and in most cases, technically difficult. For example, the family Miridae is challenging to taxonomists as it is one of the most species rich families of Hemiptera (Coeur d'acier *et al.* 2014; Raupach *et al.* 2014).

2.1.3 DNA barcoding as an identification tool

DNA barcoding has been incorporated into studies which aim to understand and assess aspects that are related to ecology, evolution, conservation, and biodiversity (Hebert *et al.* 2003; Krishna Krishnamurthy & Francis 2012; Raupach *et al.* 2014). The success of species delimitations through barcoding is based on the taxonomic coverage and accuracy of the DNA barcoding reference library available. Nearly all organisms carry a version of the COI gene in their mitochondrion (Hebert *et al.* 2003; Jinbo *et al.* 2011). In animals, this gene is used as a standard identification marker which allows unknown individuals to be assigned to species and enables the discovery of potentially new species (Hebert *et al.* 2003; Tavares & Baker 2008; Jinbo *et al.* 2011; Krishna Krishnamurthy & Francis 2012). DNA barcoding has, however, been criticized by some members of the scientific community (Rubinoff *et al.* 2006; Tavares & Baker 2008; Krishna Krishnamurthy & Francis 2012; Raupach *et al.* 2014). Critics suggest that taxonomic uncertainty, interspecific hybridization, phylogeographic scales of sampling and insufficient intraspecific sampling in favour of greater taxonomic coverage decreases the efficacy of DNA barcoding (Tavares & Baker 2008; Baker *et al.* 2009; Bergsten *et al.* 2012; Krishna Krishnamurthy & Francis 2012; Chapple & Ritchie 2013). A major limiting factor of DNA barcoding is the scale of sampling, which can cause intraspecific distances to vary greatly (Bergsten *et al.* 2012; Čandek & Kuntner 2015). Genetic divergence of individuals of the same species can be higher than anticipated because of different geographic locations and below species-level phylogeographic structuring (Bergsten *et al.* 2012; Čandek & Kuntner 2015).

In addition to the above-mentioned limitations, critics of DNA barcoding claim that the use of a single gene to carry out routine species identification has its own limitations (Rubinoff *et al.* 2006; Tavares & Baker 2008; Baker *et al.* 2009; Krishna Krishnamurthy & Francis 2012). The suggested limitation regarding DNA barcoding is how accurately this tool can identify novel species and its ability to assign them within a wider taxonomic context (Rubinoff *et al.* 2006; Tavares & Baker 2008; Baker *et al.* 2009; Krishna Krishnamurthy & Francis 2012). It is suggested that the COI mitochondrial gene as a single identification marker for species identification is not adequate due to the following suggested limitations; heteroplasmy, compounding evolutionary processes, reduced effective population size and introgression, maternal inheritance, inconsistent mutation, and recombination (Rubinoff *et al.* 2006; Tavares & Baker 2008; Krishna Krishnamurthy & Francis 2012).

2.1.4 The DNA barcode gap as a measure of identification accuracy

One key consideration in the field of DNA barcoding is the existence of the DNA “barcode gap”. The “barcode gap” can be defined as the difference between interspecific and intraspecific genetic distances within a group of organisms (Barrett & Hebert 2005; Meyer & Paulay 2005). DNA barcoding relies on the assumption that intraspecific COI variation is lower than the interspecific variability (Barrett & Hebert 2005; Meyer & Paulay 2005; Wiemers & Fiedler 2007). The separation between intraspecific and interspecific species divergence is known as the DNA barcode gap (Barrett & Hebert 2005; Meyer & Paulay 2005; Wiemers & Fiedler 2007). If there is no significant differentiation or an overlap between intraspecific and interspecific genetic divergence then this will limit species identification success using DNA barcoding and the COI gene as a single identification marker (Meyer & Paulay 2005; Wiemers & Fiedler 2007). However, very few studies using barcoding actually check for the presence of the DNA barcode gap. Previous studies on birds and arthropods have supported the presence of the DNA barcode gap (Hebert *et al.* 2004; Barrett & Hebert 2005; Hajibabaei *et al.* 2007), while other studies have not (Meyer & Paulay 2005; Meier *et al.* 2006; Wiemers & Fiedler 2007).

The DNA barcode gap analysis relies on the correct estimation of interspecific and intraspecific genetic distances. These genetic distances are generated through the use of a substitution model to generate pairwise distances between species and among individuals belonging to the same species. Most DNA barcoding studies carry out the barcode gap analysis using a single, relatively simple, model of nucleotide substitution known as the Kimura-2-parameter (K2P) model (Kimura 1980). The original species identification method proposed by Hebert et al. (2003) introduced the construction of neighbor-joining (NJ) trees based on K2P divergence. It was suggested that because most barcoding datasets consist of closely related species, the K2P model is sufficient to calculate genetic distances (Hebert *et al.* 2003; Barley & Thomson 2016). This assumption could be problematic given that the K2P model is rarely the most appropriate model for the barcode data examined (Srivathsan and Meier 2012). The use of an incorrect model can result in the estimation of inaccurate genetic distances, causing inaccurate results to be obtained. Therefore, model selection procedures should be carried out to select a model that best represents mutational processes of the data, while minimising the loss of preditive ability through overparameterisation (Sullivan & Joyce 2005).

2.1.5 Aims of study

Despite the promise of DNA barcoding, only a handful of studies throughout the world have included DNA barcoding of Hemiptera (Virgilio *et al.* 2010; Park *et al.* 2011; Coeur d'acier *et al.* 2014; Raupach *et al.* 2014; Wang *et al.* 2015). This study aims to test the utility of DNA barcoding in South African Hemiptera using the mitochondrial COI gene to identify species and create a barcode reference library of Hemiptera collected from an important biodiversity hotspot threatened by urbanization (eThekwini and surrounding regions). This study has three key objectives. First, I will test for correlation between the number of morphospecies and genetic clusters of Hemiptera (which are also known Barcode Index Numbers; BINs) collected from each of the selected localities in this study. Second, to test the success of using DNA barcoding in South African Hemiptera, I will test for the presence of the DNA barcoding gap in the eThekwini Hemiptera data, in addition to this objective, I will examine the impact of different nucleotide substitution models on the position and presence of the DNA barcoding gap. Finally, using phylogenetic methods I will test if the mitochondrial COI marker can provide resolution for associations above the species level (i.e. among genera and families) in Hemiptera. Although this study will only include taxa collected from a small regional scale, the conclusion drawn will be applicable to future studies of Hemiptera in South Africa.

2.2 Materials and methods

2.2.1 Study area

The eThekwini municipal area (the city of Durban and surrounding areas) covers a land area of 2297 km^2 which makes up 1.4% of KwaZulu-Natal (Govender 2014). South Africa contains a total of nine globally important terrestrial biomes, the eThekwini region contains four of these biomes, namely: thicket, savanna, forest and grassland (Govender 2014). The region consists of eight broad vegetation types, namely: Eastern Valley Bushveld, Mangroves, Ngongoni Veld, Northern Coastal Forest, Scarp Forest, KwaZulu-Natal Coastal Belt, KwaZulu-Natal Hinterland Thornveld and the KwaZulu-Natal Sandstone Sourveld. In this study, we will focus on the KwaZulu-Natal Sandstone Sourveld region (Figure 2.1). Most of the sites selected for this study took place within the eThekwini municipal region, however, some sites were selected around the eThekwini region.

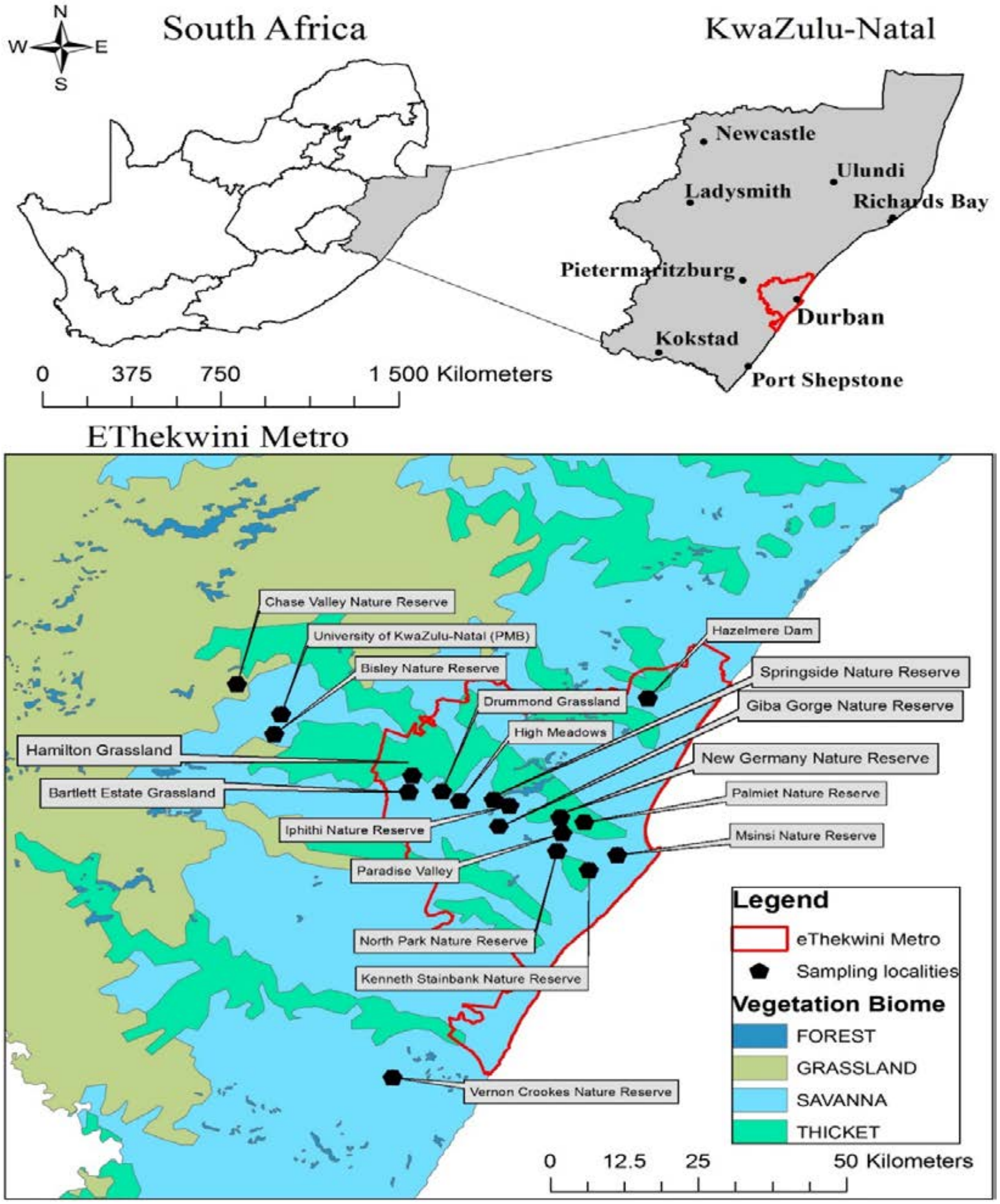

Figure 2.1: Map illustrating the 18 sampling localities in and around the eThekwini region (KwaZulu-Natal, South Africa). The extent of the eThekwini municipality area is outlined in red.

2.2.2 Sampling of specimens

All analyzed Hemiptera for this study were collected between 2011 and 2015 using a combination of sweep netting, tree beating, and active searching. A total of 1726 specimens were collected from 18 open spaces within and around the eThekwini municipality (Figure 2.1). Many of these localities form part of the D'MOSS system. These open spaces include protected areas/nature reserves (Chase Valley Nature Reserve, Bisley Nature Reserve Springside Nature Reserve, Iphithi Nature Reserve, Giba Gorge Nature Reserve, New Germany Nature Reserve, Palmiet Nature Reserve, Msinsi Nature Reserve, North Park Nature Reserve, Kenneth Stainbank Nature Reserve, Paradise Valley, Vernon Crookes Nature Reserve and Hazelmere Dam), privately managed areas (University of KwaZulu-Natal (Pietermaritzburg campus) and Drummond) and undeveloped areas (Hamilton Grassland, Bartlett Estate Grassland and High Meadows) within the sampling region. At each open space, Hemiptera was collected from all vegetation types present. All specimens were stored in ethanol (100%) at -20 degrees Celsius.

2.2.3 Sorting of specimens into morphospecies

Specimens were sorted into the lowest taxonomic level possible using morphological keys available in the taxonomic literature (Schuh & Slater 1995; Maw *et al.* 2000; Cassis *et al.* 2002). Morphological features such as body size, shape, colour patterns and wing structure were taken into consideration (Shen *et al.* 2013). Where possible, five individuals of each morphospecies per sampling locality were selected for DNA barcoding. Multiple individuals of each morphospecies were included to incorporate spatially-correlated variation within species (Bergsten *et al.* 2012). Each specimen was given a unique code to assist with identification and data tracking. Each code is made up four characteristics, the location from where the specimen was collected, the type of vegetation present at the location, and a unique specimen number. Each specimen selected for DNA analysis was photographed using a USB Digital Microscope 2.0 - Moticam 2300. Specimens were deposited at the University of KwaZulu-Natal (Pietermaritzburg campus). This repository is recognized by the Barcode of Life project as an official storage facility for specimens that have been barcoded. Successful DNA barcodes in this study will be uploaded to BOLD.

2.2.4 Genomic DNA extractions and Polymerase Chain Reaction (PCR) amplification

A single leg from ethanol-fixed specimens was removed and the total genomic DNA was extracted from 536 specimens using the ZR Tissue and Insect MicroPrep^{TM} kit (ZYMO research), following the manufacturer's instructions. The kit uses bashing beads to lyse the tissue and the extracted DNA was purified using the fast-spin column. The concentrations and quality of the extracted genomic DNA were quantified using a NanoDrop Spectrophotometer 2000 (Thermo Scientific).

The mitochondrial COI gene was amplified using the universal primers LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TAAACTTCAGGGTGACCAAAAAATCA-3′) originally developed by Folmer et al. (1994). The PCRs were carried out using a 25µl reaction volume made up of: genomic DNA (40.5 ng/µl, on average), DreamTaq DNA Polymerase (5µ/1µl; Fermentas, South Africa), DreamTaq buffer (10X, containing $MgCl_{2,}$ 20mM; Fermentas, South Africa), dNTP mix (10µM), forward and reverse primer (10µM each), sterile nuclease free water, additional $MgCl_2$ (25µM) and Bovine Serum Alumin (BSA) (1mg.m^{-1}) was added. Negative controls were carried out to test for contamination of reagents. Thermocycler amplification was carried out using the following conditions: initial denaturation at 95 ºC for 3 minutes, 10 cycles (denaturation at 95ºC for 30 seconds, annealing at 42ºC for 45 seconds, extension at 72º for 1 minute 25 seconds), 25 cycles (denaturation at 95ºC for 30 seconds, annealing at 45ºC for 45 seconds, extension at 72º for 1 minute 25 seconds) and final extension at 72ºC for 10 minutes.

The PCR products were visualized on a 0.8% (w/v) TBE agarose gel stained with Ethidium bromide (0.2mg). A 100-bp molecular weight marker (Solis Biodyne) was used to estimate the size of the PCR products. The expected product size was 658-bp. Agarose gels were viewed under ultra violet light using the MiniBis Pro gel capture instrument (Bio-Imagining System) and the PCR products were sized using a standard curve. Only products of expected size (658-bp) were sent for sequencing.

2.2.5 Sequencing and data processing

Successfully amplified PCR products were sent to the Central Analytical Facility at Stellenbosch University, South Africa for sequencing. Both forward and reverse sequences were generated. The electropherogram of each PCR amplicon was examined for quality using the BioEdit 7.0.9.0 sequence alignment editor (Hall 1999). Barcode compliant sequences (>500-bp of length, no stop codons, misidentifications, or contamination) generated in the present study (n =517) were aligned together with barcode complaint sequence data for eThekwini Hemiptera downloaded from BOLD (n = 939) using ClustalX 2.1 (Larkin *et al.* 2007). Summary statistics such as the number of variable characters (V), number of parsimony informative characters (Pi) and the average nucleotide compositions were estimated for the data using MEGA 6.0 (Tamura *et al.* 2007).

2.2.6 Phylogenetic tree construction and database construction

Using the COI sequence data for the eThekwini Hemiptera project available on BOLD, a neighbor-joining (NJ) tree was constructed using the K2P substitution model. This tree was created using the sequence analysis function within the BOLD interface (Ratnasingham & Hebert 2007). Specimens were assigned to Barcode Index Numbers (BINs) or barcode clusters on the tree using the clustering algorithm implemented within BOLD. These BINs represent operational taxonomic units (OTUs) which are synonymous with species (Blaxter *et al.* 2005). The use of BINs is useful when taxonomic information is lacking as the clustering algorithm used by BOLD allows specimens to be assigned to supposed BINs using the available sequence data (Ratnasingham & Hebert 2007). MEGA 6.0 was then used to construct a NJ tree using both the COI sequence data downloaded from BOLD and the COI sequence data generated in the present study. The tree was constructed using the K2P substitution model which is the standard model used on BOLD, (Hebert *et al.* 2004). This model was selected so that both the NJ trees constructed in BOLD and MEGA were comparable. Barcode Index Numbers were then assigned to barcode clusters on the newly created NJ tree.

Each COI sequence on the NJ tree was blasted against the global BOLD reference library. Sequence similarity to the top search hit was recorded. Taxa with similarity values greater than a 95% were considered already present in BOLD. Conditional genus-level or family-level identification was allocated to taxa with sequence similarity values less than 95%.

2.2.7 Estimating species diversity and abundance

A database of the Hemiptera specimens was constructed using Microsoft Office Excel 2010. The database served as a personal reference library and included the following information for each specimen: BINs, the unique specimen name, morphospecies allocation, taxonomy (order, family and where possible, genus and species name), sampling locality, vegetation type and the sequence similarity score from the top hit in BOLD. This information was then used to compare the number of morphospecies and BINs at the different sampling localities. A chi-squared test was performed to check for any significant difference between the number of morphospecies and number of BINs recorded for each of the sampling localities.

Species diversity analyses were carried out for Hemiptera at each of the sampling localities using the program Past 3.44 (Hammer *et al.* 2001) on the dataset where individuals were assigned to BINs. To link sample heterogeneity and estimates of species richness and abundance, both within and between each locality, species diversity was calculated using four most widely accepted diversity indices, namely, Margalef's diversity index (*d*) (Margalef 1958), Fisher's α (α) (Fisher *et al.* 1943), Shannons diversity index (*H'*) (Shannon & Weaver 1963) and Simpson's diversity index (*D*) (Simpson 1949). Margalef's diversity index was calculated using the total number of species and the total number of individuals present in each sampling area to highlight the most species-rich sites (Margalef 1958). Fisher α diversity index is a parametric index of diversity that assumes the abundance of species follows the log series of distribution, this index describes how individuals sampled are divided among species in each sampling site (Fisher *et al.* 1943). Shannons index takes into account the number of individuals as well as the number taxa at each sampling site (Shannon & Weaver 1963). Simpson's index measures the evenness of a community from 0 to 1 at in each sampling site (Simpson 1949).

2.2.8 DNA Barcode gap analysis

Intraspecific and interspecific genetic distances were calculated using two datasets, namely, (i) individuals assigned to BINs (number of BINs = 357) and (ii) individuals assigned to morphospecies (number of morphospecies = 256). The best-fit nucleotide substitution models were selected using the program jModelTest 2 (Darriba *et al.* 2012) along with the Akaike Information Criterion (AIC) (Akaike 1974). In both cases, the model of best-fit for the data was the general time reversible model (GTR) with both gamma distribution (G) and proportion of invariable sites (I). To test what effect model choice had on barcode gap estimation, both the K2P nucleotide substitution model and the GTR + I + G nucleotide substitution model were used to generate pairwise genetic distance matrices using MEGA and RAxML 8.0 (Stamatakis 2014).

The intraspecific and interspecific distances were plotted using Microsoft Excel 2010. The maximum intraspecific distance was subtracted from the minimum interspecific distance to check for the presence or absence of a barcoding gap (Meier *et al.* 2006).

Further statistical analyses for the DNA barcode gap was carried out in the R statistical software (http://www.r-project.org). The Jeffries-Matusita Distance (J-M) statistic was used to test if the intraspecific and interspecific genetic distances are separable. The J-M distance is widely used as a reparability criterion and is used to assess the potential of band pairs to discriminate between two different regional classes (Trigg & Flasse 2001). The J-M separability criterion considers the distance between class means and the distribution of values from the mean (Dabboor *et al.* 2014). The J-M distance is asymptotic to 1.414 and as such, a value of 1.414 or greater suggests that two regional classes are statistically separable (Trigg & Flasse 2001).

2.2.9 Phylogenetic Analysis

In addition to the construction of a NJ tree, two additional phylogenetic approaches were taken to determine if the COI marker can provide resolution above the species level. Maximum likelihood and Bayesian inference analyses were performed on a truncated dataset containing only one representative for each BIN cluster. In this case, a dataset including 357 sequences was analyzed. The maximum likelihood analysis and Bayesian inference are both model-based approaches that can implement more sophisticated substitution models, therefore the GTR + I + G was used in both these analyses.

RAxML 8.0 was used to conduct the maximum likelihood analysis. The tree search method was ML + thorough bootstrap implementing 1000 bootstrap replicates. This analysis generated both the most likely tree and a bootstrap tree file, which contained the 1000 bootstrap trees. A consensus tree was created from the bootstrap file using Phylip 3.69 (Felsenstein 2005). Thereafter, trees were viewed and modified using the program Figtree 1.3.1 (Rambaut 2009).

MrBayes 3.1.2 (Ronquist & Huelsenbeck 2003) was used to conduct the Bayesian inference. This analysis was performed using two independent runs each consisting of four parallel chains. The MCMC chains were run for 30 million generations with trees sampled every 5000th generation. Once each run was completed, convergence of the MCM chains from each run was assessed in Tracer 1.5 (Rambaut & Drummond 2007). Convergence was obtained when effective sampling size (ESS) values were all above 200. The first 6 million trees were removed as burn-in.

The resulting trees were used to create a 50% majority rule consensus tree using Phylip 3.69. Branch support values were provided as posterior probabilities (Bayesian analysis) and bootstrap values (maximum likelihood analysis). The resulting maximum likelihood and Bayesian topologies were compared by analyzing significantly supported nodes (bootstrap ≥ 75% and posterior probabilities ≥ 0.95). Since the maximum likelihood and Bayesian trees agreed on one topology, I combined the posterior probabilities values with the bootstrap values and annotated them onto the most likely tree from the maximum likelihood analysis.

2.3 Results

2.3.1 Data description

In the present study, DNA was extracted from 536 specimens. The COI gene region was successfully amplified and sequenced from 517 specimens. These 517 barcodes met the barcode-compliance criteria stipulated by the Consortium for DNA Barcoding (>500-bp of length, no stop codons, misidentifications, or contamination). These sequences were added to the larger Hemiptera data already available on BOLD from the eThekwini region (939 specimens). The sequences were easily aligned as there were no insertions, deletions, or stop codons. The final alignment was 615 base pairs in length (465 variable sites of which 426 were parsimony informative). The Hemiptera database used in this study consisted of 1456 specimens, which were clustered into 357 BINs and 256 morphospecies. The average base composition for the full dataset was 35.8% thymine (T), 17.6% cytosine (C), 31.4% adenine (A) and 15.2% guanine (G). The nucleotide compositions of all the sequences were heavily biased toward A and T nucleotides. This is not unusual as insects are known to have an AT-rich nucleotide composition in COI sequences (Raupach et al, 2014).

2.3.2 Species richness and diversity

The number of specimens (n = 1456), morphospecies (n = 256) and BINs (n = 357) per sampling locality were plotted (Figure 2.2) and used to estimate the species richness and diversity within the eThekwini region. Generally, the number of BINs and morphospecies observed from each locality were similar. However, a chi-square test of coherence between the number of BINs and morphospecies demonstrated a significant difference in Hamilton grassland, Drummond, Springside Nature Reserve, Palmiet Nature Reserve, and North Park Nature Reserve ($p < 0.05$). The variation in the number of morphospecies and BINs recorded from these sites could suggest cryptic speciation. For example, Palmiet has more BINs that morphospecies, which suggests that there are a larger number of distinct genetic lineages in comparison to the number of morphospecies. In contrast, it is observed that there is an over estimation of species based on morphology when compared to BINs in Paradise Valley, which suggests the presence of species which are very morphologically variable or specimens which are at different developmental stages at this locality.

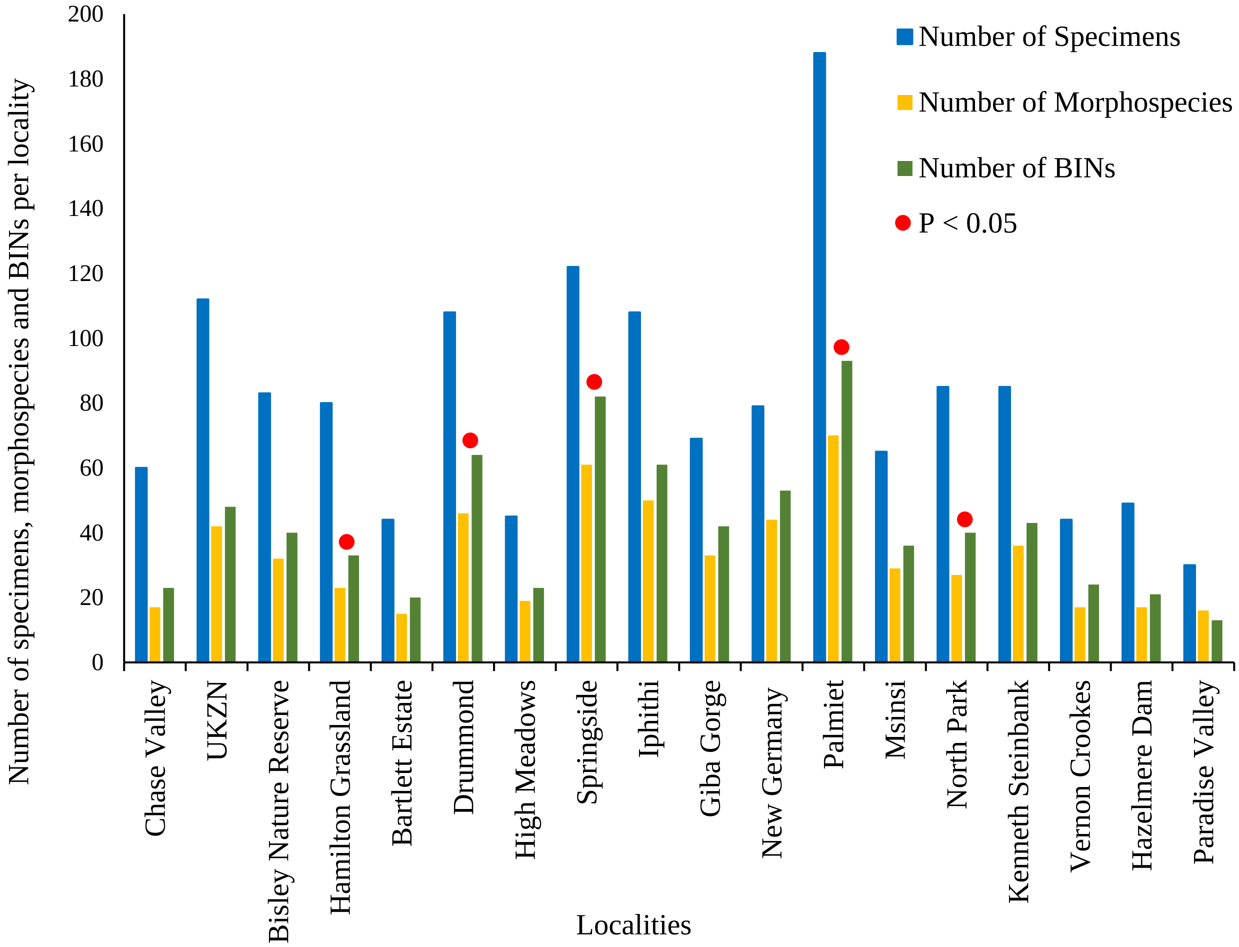

Figure 2.2: Bar graph representing the distribution of specimens, morphospecies and barcode clusters (BINs) amongst the different sampling localities included in the present study. Localities with a significant difference between the number of BINs and the number of morphospecies are highlighted by a red dot.

Species richness, abundance and diversity were calculated and used to assess the different localities (Table 2.1). The highest species richness value was observed at Palmiet Nature Reserve (S = 79), followed by Springside Nature Reserve (S = 70) and Drummond (S = 55). The Margalef's index values obtained showed that species diversity among sampling sites varied between *d* = 3.46 and 14.39 and the Fisher α diversity index values obtained ranged between α = 8.16 and α = 69.29. The highest species diversity and α values were observed in Springside Nature Reserve (*d* = 14.39, α = 69.29), Palmiet Nature Reserve (*d* = 14.34, α = 47.69), and Drummond (*d* = 11.56, α = 45.44). In contrast, the lowest species diversity was observed in Paradise Valley (*d* = 3.46, α = 8.46) and

Hazelmere Dam ($d = 4.13$, $\alpha = 9.39$). These trends were further supported by the Shannons index (H') and Simpsons index (D) values. Both these indices ranked the community of Springside nature reserve (H = 4.01, D = 0.98), Palmiet Nature Reserve (H = 3.93, D = 0.97), and Drummond (H = 3.78, D = 0.97) as the most diverse. Whereas, Paradise Valley (H = 2.33, D = 0.89) and Hazelmere Dam (H = 2.58, D = 0.90) were ranked the least diverse. Furthermore, the high Simpson's index values ($D > 0.20$) at all the localities suggest that the distribution of species appeared to be unevenly matched or balanced. This suggests that most species are not present in similar abundance throughout the localities, but rather that there are one or more species dominating the different sites.

In general, when assessing molecular data and statistically supported data, UKZN (Pietermaritzburg campus), Drummond, Springside Nature Reserve, Iphithi Nature Reserve, and Palmiet Nature Reserve have the highest number of BINs, indicating high species richness and abundance in these regions. In contrast, Bartlett Estate, High Meadows, Vernon Crookes, Hazelmere Dam, and Paradise Valley have the lowest number of BINs. These regions are classified as the least diverse regions, even though they were sampled extensively. This could mean that there is a true lack of diversity in these regions. The overall structure, vegetation, and diversity of the different ecological niches could have affected the variation in species diversity across the 18 localities.

Table 2.1 Univariate models of biodiversity estimation indices for each locality. S represents the total number of BINs present (richness), N represents the total number of individuals per a locality (abundance), *d* represents species diversity calculated from the Margalef's diversity index, α represents Fisher's α, indicating that the division of species per sampling locality, H' represents Shannons index which estimates biodiversity richness and diversity and D represents Simpsons index which quantifies the biodiversity present at each locality.

Sites	**S**	**N**	***d***	**α**	**H'**	**D**
Chase Valley Nature Reserve	19	60	4.40	9.58	2.68	0.92
University of KwaZulu-Natal (PMB)	48	112	9.96	31.82	3.53	0.96
Bisley Nature Reserve	34	79	7.55	22.64	3.22	0.95
Hamilton Grassland	27	80	5.93	14.33	2.81	0.91
Bartlett Estate	20	47	4.94	13.16	2.80	0.93
Drummond	55	107	11.56	45.44	3.78	0.97
High Meadows	21	45	5.25	15.33	2.83	0.93
Springside Nature Reserve	70	121	14.39	69.29	4.01	0.98
Iphithi Nature Reserve	54	108	11.32	42.98	3.64	0.96
Giba Gorge	37	70	8.47	31.80	3.39	0.96
New Germany	44	79	9.84	40.93	3.52	0.96
Palmiet Nature Reserve	79	187	14.34	47.69	3.93	0.97
Msinsi Nature Reserve	31	64	7.21	23.68	3.10	0.94
North Park Nature Reserve	32	58	6.98	18.66	3.25	0.95
Kenneth Steinbank Nature Reserve	38	84	8.35	26.74	3.41	0.96
Vernon Crookes	19	44	4.76	12.0	2.74	0.92
Hazelmere Dam	17	48	4.13	9.39	2.58	0.90
Paradise Valley	13	32	3.46	8.16	2.33	0.89

2.3.3 DNA Barcode gap analysis

The presence of the DNA barcode gap was tested on the Hemiptera individuals assigned to morphospecies and DNA barcode clusters (BINs). Two different nucleotide substitution models (K2P and GTR+I+G) were used.

Using the COI morphospecies dataset (n = 256), the frequency distribution of the intraspecific and interspecific pairwise genetic distances was computed using K2P (Figure 2.3A) and GTR+I+G (Figure 2.3B) nucleotide substitution models. The frequency of the K2P pairwise distance values was distributed between the range of 0.00 to 0.4 whereas the frequency for the GTR+I+G pairwise distance values was distributed between the range of 0.00 to 1.20. The K2P pairwise distance values between species (range 0.00 to 0.4) were greater than within species (range 0.00 to 0.14) and the GTR+I+G pairwise distance values between species (range 0.00 to 1.20) were greater than within species (range 0.00 to 0.27). There is an overlap observed between the intraspecific and interspecific classes using both the K2P and GTR+I+G nucleotide substitution models on the COI morphospecies dataset (refer to the graph inserts in Figures 2.3A and 2.3B). The overlap using the K2P pairwise distance values occurred between 0.00 to 0.14, however here was a shift in the overlap using the GTR+I+G pairwise distance values, where the overlap occurred between 0.00 to 0.24. The GTR+I+G nucleotide substitution model showed a much larger overlap than the K2P model. The presence of the overlap observed between the intraspecific and interspecific classes using the morphospecies dataset is common, it suggests that the identification of specimens within this overlap region was inaccurate. Species pairs found within the overlap region have low divergence values and include some whose taxonomic status as distinct species is debatable. Some of the families occurring within this overlap cannot be well differentiated phenotypically due to close morphological similarities. These families include Cercopidae, Cydnidae, Rhopalidae, Coreidae, Miridae, Pyrrhocoridae, Cicadellidae, Reduviidae, Rhyparochromidae, Pentatomidae, Fulgoridae, Flatidae, Thyreocoridae and Nabidae. Therefore, one must proceed with caution when dealing with the above-mentioned families.

Using the COI BINs dataset (n = 357), the frequency distribution of the intraspecific and interspecific genetic distances was computed using K2P (Figure 2.3C) and GTR+I+G (Figure 2.3D) substitution models. The frequency of the K2P pairwise distance values was distributed between the range of 0.00 to 0.4 whereas the frequency for the GTR+I+G pairwise distance values was distributed between the range of 0.00 to 1.20. The K2P pairwise distance values between species (range 0.07 to 0.4) were greater than within species (range 0.00 to 0.1) and the GTR+I+G pairwise distance values between species (range 0.08 to 1.20) were greater than within species (range 0.00 to 0.17). There is an overlap observed between the intraspecific and interspecific classes using both the K2P and GTR+I+G nucleotide substitution models on the COI BINs dataset (refer to the graph inserts in Figures 2.3C and 2.3D). The overlap using the K2P pairwise distance values occurred between 0.07 to 0.1, however here was a shift in the overlap using the GTR+I+G pairwise distance values, where the overlap occurred between 0.09 to 0.17. The GTR+I+G nucleotide substitution model showed a much larger overlap than the K2P model. The overlap observed between the intraspecific and interspecific classes indicates possible misidentification of taxa or the presence of cryptic species within the BINs dataset. Unlike the morphospecies dataset, the overlap observed when using BINs was due to one specific taxon. The intraspecific distance values within the species *Spilostethu pandurus* created the overlap with the interspecific K2P and GTR+I+G pairwise divergence. If this taxon was removed from the dataset, a clear barcode gap would be observed. Therefore, one must proceed with caution when dealing with the above-mentioned taxon.

When comparing the use of the morphospecies dataset (Figure 2.3 A and 2.3B) to the BINs dataset (Figure 2.3C and 2.3D) with regards to the different substitution models it is observed that there is a greater overlap using the morphospecies dataset (K2P ranged between 0.00 to 0.14, and GTR+I+G ranged between 0.00 to 0.24) then BINs dataset (K2P ranged between 0.07 to 0.1, and GTR+I+G ranged between 0.09 to 0.17). There is far less misidentification of taxa present within the overlap of intraspecific and interspecific divergence when using the BINs dataset as compared to the morphospecies dataset.

The Jeffries-Matusita (J-M) distance was calculated for the different datasets (morphospecies and BINs) using the different substitution models (K2P and GTR+I+G) to statistically analysis the overlap between the intraspecific and interspecific classes. The Jeffries-Matusita distance was calculated to be 1.72 for the morphospecies dataset using K2P, 1.67 for the morphospecies dataset using GTR+I+G, 1.97 for the BINs dataset using K2P, and 1.76 for the BINs dataset using GTR+I+G. These values were all greater than 1.414 and is very close to 2, suggesting that the intraspecific and interspecific classes are separable. This indicates that the degree of overlap observed in all the cases above is not statistically significant and there is, in fact, a barcode gap. The observed results in Figures 2.3A, 2.3B, 2.3C and 2.3D showed that the COI region is a suitable marker in the mitochondrial protein coding genes as a universal DNA barcode for the use in Hemiptera species delimitation within the eThekwini region.

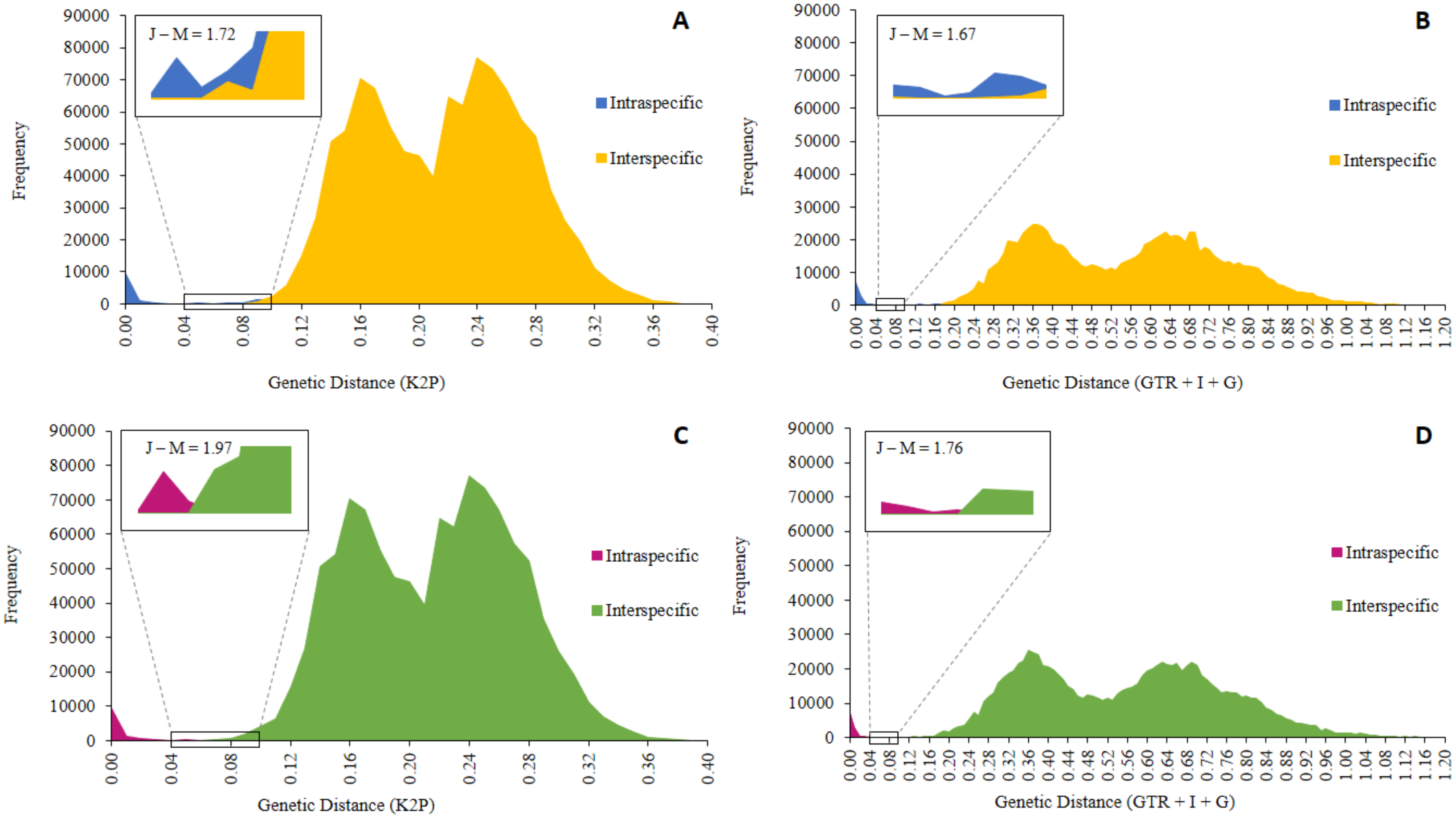

Figure 2.3: Frequency distributions of pairwise distances (K2P and GTR + I + G) based on the COI gene of 1456 individuals sorted into 256 morphospecies and 357 BINs of Hemiptera from the eThekwini region. (**A**) Computed using K2P on the morphospecies dataset (**B**) computed using GTR + I + G on the morphospecies dataset (**C**) computed using K2P on the BINs dataset (**D**) computed using GTR + I + G on the BINs dataset. A degree of overlap is observed between the intraspecific and interspecific genetic divergence (graph inserts). Genetic distances are placed on the x-axis and the frequencies are placed on the y-axis.

2.3.4 Phylogenetic analysis

The phylogenetic trees constructed using the maximum likelihood approach and the Bayesian Inference recovered very similar topologies. There were no instances of conflict i.e. competing hypotheses supported by >75 % bootstrap support or > 0.95 posterior probability. As such all branch support values are shown on the most likely tree (Figure 2.7). All values above 75% bootstrap (BS) and 0.95 posterior probabilities (PP) were considered significantly well supported and highlighted on the phylogeny. This phylogeny was midpoint rooted as no outgroup taxa was included. The phylogeny represents 22 families and 181 genera. The mutation rate of the COI marker is appropriate to resolve species-level associations but our data suggests that this molecule may also be useful at higher taxonomic levels such as family and genus levels.

There is evidence of monophyly in the following families (n = 14), however, in most of these families the branch support values were not very high (<75 % BS or < 0.95 PP): Cicadellidae (BS = 46.3, PP = 1.00), Tingidae (BS = 64.4, PP = 1.00), Anthocoridae, Miridae (BS = 17.5, PP = 0.76), Reduviidae (BS = 26.4, PP = 0.98), Berytidae, Pyrrhocoridae (BS = 44.5, PP = 0.99), Pentatomidae (BS = 10.1, PP = 0.97), Coreidae (BS = 0.6, PP = 0.13), Psyllidae (BS = 50.8, PP = 0.07), Aphrophoridae (BS = 87.7, PP = 0.99), Achilidae, Tropiduchidae and Delphacidae (BS = 59.3, PP = 0.97). There is evidence of paraphyly in the following families (N = 4): Lygaeidae (BS = 4.5, PP = 0.80), Rhyparochromidae (BS = 2, PP = 0.27), Tettigometridae (BS = 100, PP = 0.95), and Eurybrachidae (BS = 68.2, PP = 0.96). There is evidence of polyphyly in the following families (n = 4): Fulgoridae (BS = 18.3, PP = 0.39), Cixiidae (BS = 18.9, PP = 0.45), Lophopidae, and Flatidae (BS = 18.7, PP = 0.68). Overall it is observed that a majority of the families in the COI Hemiptera dataset are monophyletic suggesting that the COI marker can resolve Hemiptera families relatively well. Most of the branches present within each family had well supported bootstrap and/or posterior probabilities.

The phylogenetic tree (Figure 2.4) was further analyzed to the genus level to see how well the COI marker could resolve at this taxonomic level. There was evidence of monophyly in 128 of the genera, paraphyly is 31 of the genera, and polyphyly in 22 of the genera. Most of the genera were well supported, with majority of them being monophyletic, this suggests that the COI marker can resolve the different genera of Hemiptera relatively well.

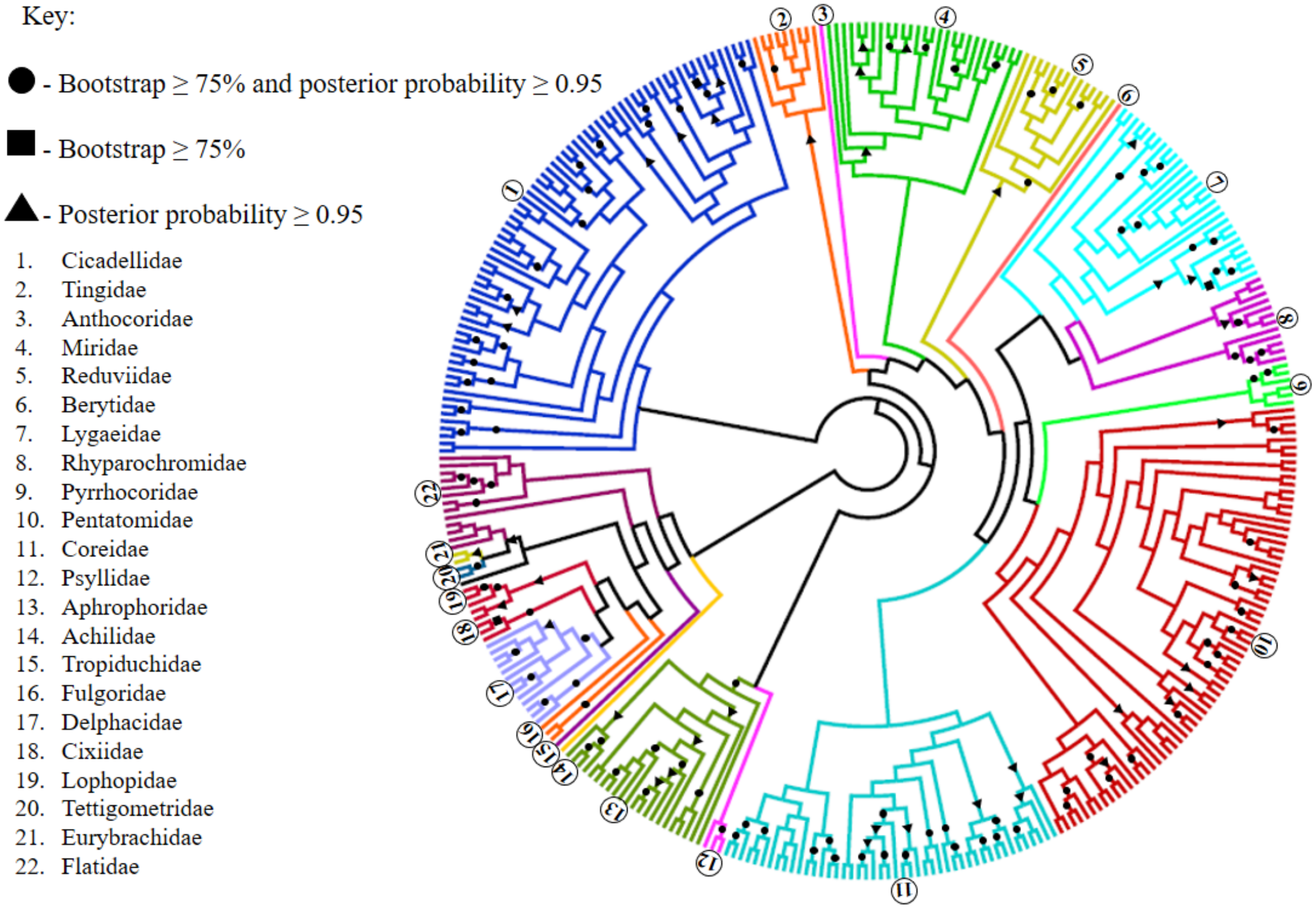

Figure 2.4: Most likely tree of 357 representative (BINs) Hemiptera specimens using their COI sequence data. The branches of the tree are colour coded by family. Only significantly supported branch support values (Bootstrap value ≥ 75%, Posterior probability value ≥ 0.95) are shown.

2.4 Discussion

Since 2003 DNA barcoding has gained momentum as an automatable, accurate and rapid method of species identification (Hebert & Gregory 2005). DNA barcoding has the ability to identify specimens to previously described species or families, by comparing an unknown query barcode sequence to barcodes of taxonomically identified species in the BOLD database. DNA barcoding can be successfully applied when a good reference library is created and the data conform to two main criteria. First, there needs to be a match between the number of species recognized by traditional morphological techniques and genetic barcode clusters (BINs). Second, there needs to be minimal overlap between intraspecific and interspecific genetic divergences. If these assumptions are met, then DNA barcoding offers a great number of advantages and is a complementary tool for the traditional morphological identification in biodiversity studies.

Developing a complete reference library for any region is a formidable task. In this study, I focused on a small regional area as a case study. The main aim of this chapter was to test for the utility of DNA barcoding to accurately identify Hemiptera species collected from the eThekwini region and surrounding areas. This was done by statistically comparing the number of morphospecies collected at each locality to the number of BINs recovered by the DNA data. I also tested for the presence of the barcode gap in the data and looked at what effect model choice has on the barcode gap. The conclusions drawn from this study are, however, applicable to the construction of reference libraries for larger geographic areas and for other taxa.

2.4.1 DNA barcoding compliments traditional morphological-based techniques

There will always be possible flaws during species identification when dealing with either the traditional morphology-based method or DNA barcoding. DNA technology can never replace morphological analysis but this study supports the use of DNA barcoding as a complementary tool to traditional morphological species identification. An integrative approach combining the strengths of these two methods would be the most comprehensive way to achieve a more sound, reliable, and efficient species identification system.

In this study, specimens were separated into 256 morphospecies and subsequent molecular testing and phylogenetic tree construction suggested the presence of at least 357 unique genetic barcode clusters (BINs). The barcode data suggests that the variation in the number of morphospecies and BINs were a result of cryptic speciation and the presence of Hemiptera at different developmental stages (i.e. immature and adult forms of Hemiptera). Cryptic speciation presents a challenge when attempting to quantifying biodiversity. The estimation of species richness and endemism is crucial to identifying habitats which qualify for conservation. The discovery of cryptic speciation has been accelerated by the ease of obtaining DNA sequences from organisms. Investigating whether cryptic speciation is more common in a certain geographic area, habitat, biome, or vegetation type could enhance our knowledge and speed up conservation efforts (Bickford et al. 2006).

In this study, the DNA barcode data provides evidence for possible cryptic speciation in Hemiptera. An example of possible cryptic speciation which would have been overlooked by a novice taxonomist can be seen in the family Aphrophoridae (Figure 2.5), where the species *Clovia albomarginata* and *Aphrophora parallella.* These two taxa look morphologically very similar but belonged to very different genetic lineages. An example of Hemiptera at different developmental stages was seen in the family Pentatomidae (Figure 2.6). In this case the immature form of *Palomena prasina* looks very different from the adult morph. This study highlights the fact that DNA barcoding is useful when cryptic species are involved and morphological identification is difficult (Gregory 2005). This is particularly important when diversity studies are being conducted by researchers, which are not expert taxonomists as was the case in this study.

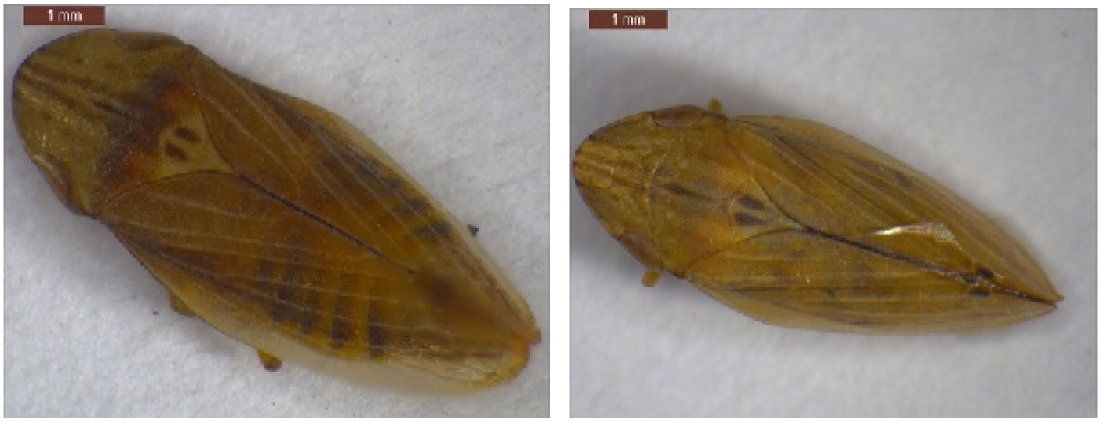

Figure 2.5: The representation of cryptic speciation in the family Aphrophoridae. The individual on the right is *Clovia albomarginata* and the individual on the left is *Aphrophora parallella.*

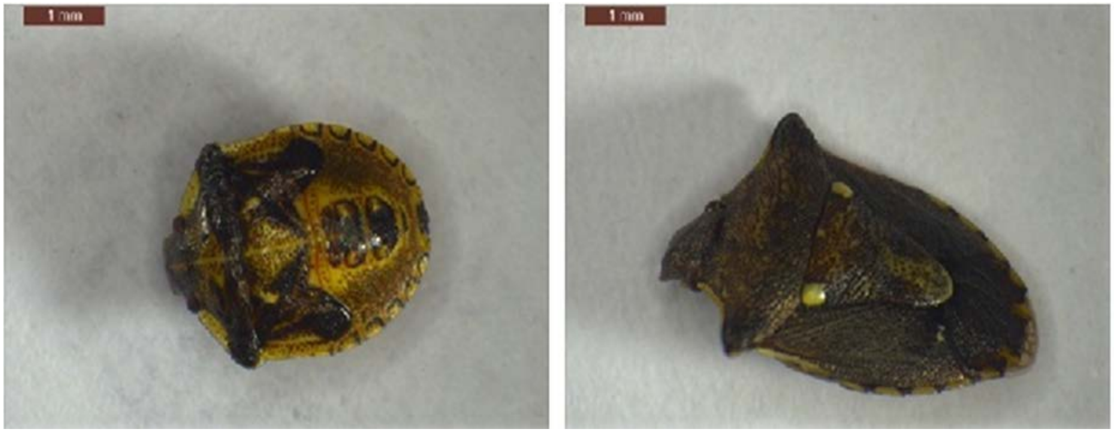

Figure 2.6: The representation of Pentatomidae *Palomena prasina* during their different life cycles. The individual on the right is the immature form and the individual on the left is the adult form.

2.4.2 Assessing the diversity of the eThekwini South African Hemiptera

In this study the species diversity and species richness was assessed at a small regional scale, a total of 1456 specimens were collected from the eThekwini region and surrounding areas and comprised of an estimated 357 distinct genetic lineages which may represent species and 22 families. In comparison, several other studies which focused on much larger geographic scales found comparable or less Hemiptera diversity (Liu *et al.* 2013; Foottit *et al.* 2014; Raupach *et al.* 2014; Tembe *et al.* 2014). For example, a study conducted in Germany surveyed most of the country and collected a total of 1742 specimens which comprised of 457 species and 39 families (Raupach *et al.* 2014). A study conducted in Canada collected 1482 Hemiptera specimens from the Canadian National Collection of Insects center, these specimens comprised of 471 species and 147 genera (Foottit *et al.* 2014). Moreover, a study conducted in the Western Ghats of India collected 80 specimens comprising of 43 species and 35 genera (Tembe *et al.* 2014). Finally, an Asian study collected 214 Hemiptera specimens across the South-east of Asia, which made up 42 species and 9 genera (Liu *et al.* 2013). The high levels of diversity seen in our preliminary data set highlight the remarkable diversity within the eThekwini region. This however, places emphasis on the fact that this region needs to be conserved and effectively managed in order to protect the valuable biodiversity found within this region. Continued urbanization may threaten this biodiversity and any further developments within the region should consider incorporating green spaces and biodiversity corridors in town planning.

2.4.3 The DNA barcode gap

This study was carried out on a small regional scale, this is an important factor to consider as many species are expected to be closely related. As such, this study is an excellent case study to test the utility of DNA barcoding and its success. This study provides evidence that a COI-based identification system is effective for the identification of Hemiptera species within the eThekwini region. These results suggest that this system will also be applicable to the rest of South Africa. The results obtained from the gap analysis was consistent in terms of the distinct overlap between the intraspecific and interspecific pairwise distances, despite assigning individuals to either morphospecies or barcode clusters (BINs) and the use of different nucleotide substitution models (K2P versus GTR+I+G). However, there was a greater overlap seen when using morphospecies then when using BINs, together with a distinct shift in the overlap when using the different nucleotide substitution models. This observation suggests that the use of BINs is more reliable then use of morphology-based identification when testing for the presence of the DNA barcode gap. Further statistical analysis (J-M distance test) suggests that despite some overlap between the intraspecific and interspecific genetic divergence of Hemiptera, the two classes are separable. This provides strong evidence for the existence of the DNA barcode gap in our data as the overlap was seen in a very insignificant portion of our data. Species found in the overlap should be flagged and caution should be taken when working with those species.

The barcode gap found in this study supports the findings of previous studies done on Hemiptera (Park *et al.* 2011; Raupach *et al.* 2014; Tembe *et al.* 2014). Park et al (2011) and Raupach et al (2014) both demonstrated the success of DNA barcoding in the species identification of Hemiptera and Tembe et al (2014), reported that a distinct barcode gap was observed in the intraspecific and interspecific genetic divergence seen at higher taxonomic levels of Hemiptera. Throughout the literature of DNA barcoding and the DNA barcode gap analysis, an important observation was made, this observation was that the range in which the barcode gap occurs differs across different datasets. It is, therefore, important that each study that used DNA barcoding include the barcode gap analysis to assess the success of the COI marker to reliably identify species of their study organisms.

2.4.4 Phylogenetic analysis of the COI marker

The mitochondrial COI gene has a mutational rate appropriate for the resolution of associations at the species level. One of the aims of this study was to determine if this single marker system could also provide information on higher-level relationships (genus- or family-level associations) within South African Hemiptera. In this study, using individuals from the 22 families and 181 genera of Hemiptera, the phylogeny-based analysis (maximum likelihood and Bayesian inference) indicated that the COI sequences are phylogenetically informative enough to provide information at both family and generic level. Most of the families and genera were well supported with a majority of them being monophyletic, suggesting the reliability of the COI marker at higher taxonomic levels.

2.5 Conclusion

The increase in urbanization within and around the eThekwini region increases the risk of the loss of biodiversity and in turn increasing the risk of species extinctions. The data generated by the present study will contribute towards D'MOSS and facilitate conservation planning. In this study, it was shown that DNA barcoding using the selected DNA barcode marker (COI) can distinguish between the different barcode clusters and morphospecies allowing for species identification. It is recommended that DNA barcodes should be used in combination with morphology-based methods. The results from this study warrants further investigation using approaches such as phytogeography or landscape genetics. The barcode data could then be used to assess the diversity of different species in an ecosystem and could shed light on genetic diversity of taxa below the species level. Using patterns of genetic variation gene flow along urban green corridors can be assessed and quantified. Connectivity network can be constructed for eThekwini which can be used to model the movement and spatial distribution of biodiversity in the region. Stakeholder's can use this information to inform the public and implement conservation plans. This, however, was a preliminary study, additional sampling, and sequencing to ensure the presence of barcode gap at a larger scale of sampling is required before full reference library will be available for use.

2.6. References

Akaike H. (1974) A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716-723.

Alroy J. (2002) How many named species are valid? *Proceedings of the National Academy of Sciences* **99**, 3706-3711.

Baker A.J., Tavares E.S. & Elbourne R.F. (2009) Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Molecular Ecology Resources* **9** , 257-268.

Barley A.J. & Thomson R.C. (2016) Assessing the performance of DNA barcoding using posterior predictive simulations. *Molecular Ecology* **25**, 1944-1957.

Barrett R.D.H. & Hebert P.D.N. (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* **83**, 481-491.

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. *Systematic Biology* **61**, 851-869.

Bickford D., Lohman D.J., Sodhi N.S., Ng P.K.L., Meier R., Winker K., Ingram K.K. & Das I. (2006) Cryptic species as a window on diveristy and conservation. *Trends in Ecology and Evolution* **22**, 148-155.

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

Brooks T.M., Mittermeier R.A., da Fonseca G.A.B., Gerlach J., Hoffmann M., Lamoreux J.F., Mittermeier C.G., Pilgrim J.D. & Rodrigues A.S.L. (2006) Global biodiversity conservation priorities. *Science* **313**, 58-61.

Čandek K. & Kuntner M. (2015) DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources* **15**, 268-277.

Cassis G., Gross G.F. & Study A.B.R. (2002) *Hemiptera*. Australia: CSIRO Publishing.

Cesari M., Guidetti R., Rebecchi L., Giovannini I. & Bertolani R. (2013) A DNA barcoding approach in the study of tardigrades. *Journal of Limnology* **72**, 182-198.

Chapple D.G. & Ritchie P.A. (2013) A retrospective approach to testing the DNA barcoding method. *PLoS One* **8**, e77882.

Coeur d’acier A., Cruaud A., Artige E., Genson G., Clamens A.L., Pierre E., Hudaverdian S., Simon J.C., Jousselin E. & Rasplus J.Y. (2014) DNA barcoding and the associated PhylAphidB@se website for the identification of European aphids (Insecta: Hemiptera: Aphididae). *PLoS One* **9**, e97620.

Croucamp A. (2009) Report: Our Biodiverse City. Department of Environmental Management. Durban: eThewkwini Municipality.

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* **35**, 6859-6873.

Darriba D., Taboada G., Doallo R. & Posada D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772.

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

Fisher R., A C. & Williams C. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**, 42-58.

Folmer O., Black M., Hoeh W., Lutz R. & Vrijenhoek R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**, 294-299.

Foottit R.G., Maw E. & Hebert P.D.N. (2014) DNA barcodes for Nearctic Auchenorrhyncha (Insecta: Hemiptera). *PLoS One* **9**, e101385.

Govender N. (2014) Durban: State of biodiversity report 2013/2014. Durban: eThewkwini Municipality.

Gregory T.R. (2005) DNA barcoding does not compete with taxonomy. *Nature* **434**, 1067.

Grimm N.B., Faeth S.H., Golubiewski N.E., Redman C.L., Wu J., Bai X. & Briggs J.M. (2008) Global change and the ecology of cities. *Science* **319**, 756-760.

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**, 167-172.

Hall T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98.

Hamer M. (2013) A national strategy for zoological taxonomy (2013-2020). Pretoria: South African National Biodiversity Institute.

Hammer R., Harper D. & Ryan P. (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4**, 9.

Hebert P. & Gregory R. (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology* **54**, 852-859.

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

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S. & Francis C.M. (2004) Identification of birds through DNA barcodes. *PLoS Biology* **2**, e312.

Jinbo U., Kato T. & Ito M. (2011) Current progress in DNA barcoding and future implications for entomology. *Entomological Science* **14**, 107-124.

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

Krishna Krishnamurthy P. & Francis R.A. (2012) A critical review on the utility of DNA barcoding in biodiversity conservation. *Biodiversity and Conservation* **21**, 1901-1919.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. & Higgins D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23**, 2947-2948.

Liu Q.H., Jiang L.Y. & Qiao G.X. (2013) DNA barcoding of Greenideinae (Hemiptera : Aphididae) with resolving taxonomy problems. *Invertebrate Systematics* **27**, 428-438.

Margalef R. (1958) Information theory in ecology. *General Systems* **3**, 36-71.

Maw H.E.L., Foottit R.G., Hamilton K.G.A. & Scudder G.G.E. (2000) *Checklist of the Hemiptera of Canada and Alaska*. Canada: NRC Press.

McKinney M.L. (2008) Effects of urbanization on species richness: A review of plants and animals. *Urban Ecosystems* **11**, 161-176.

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**, 715-728.

Meyer C.P. & Paulay G. (2005) DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology* **3**, e422.

Moorhead L.C. & Philpott S.M. (2013) Richness and composition of spiders in urban green spaces in Toledo, Ohio. *Journal of Arachnology* **41**, 356-363.

Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A.B. & Kent J. (2000) Biodiversity hotspots for conservation priorities. *Nature* **403**, 853-858.

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.

R Development Core Team. (2010) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for statistical computing. Retrieved from http://www.R-project.org. Date accessed: 06 July 2016

Rambaut A. (2009) FigTree v.1.3.1 Computer program and documentation. Retrieved from http:tree.bio.ed.ac.ukz/software. Date accessed: 26 May 2016

Rambaut A. & Drummond A.J. (2007) Tracer v. 1.5 Computer program and documentation. Retrieved from at http://beast.bio.ed.ac.uk/Tracer. Date accessed: 26 May 2016

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

Raupach M.J., Hendrich L., Küchler S.M., Deister F., Morinière J. & 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**, e106940.

Reid W.V. (1998) Biodiversity hotspots. *Trends in Ecology & Evolution* **13**, 275-280.

Ronquist F. & Huelsenbeck J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.

Rouget M., Reyers B., Jonas Z., Desmet P., Drivers A., Maze K., Egoh B. & Cowling R. (2004) South African National Spatial Biodiversity Assessment 2004: Technical Report. Volume 1: Terrestrial Component. South Africa: South African National Biodiversity Institute.

Rubinoff D., Cameron S. & Will K. (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity* **97**, 581-594.

Savard J.P.L., Clergeau P. & Mennechez G. (2000) Biodiversity concepts and urban ecosystems. *Landscape and Urban Planning* **48**, 131-142.

Scholtz C.H. (1999) Review of insect systematics research in South Africa. *Transactions of the Royal Society of South Africa* **54**, 53-63.

Schuh R.T. & Slater J.A. (1995) *True bugs of the world (Hemiptera: Heteroptera): classification and natural history*. New York: Cornell university press.

Shannon C. & Weaver W. (1963) *The Mathematical Theory of Communication*. USA: The University of Illinois Press.

Shen Y.Y., Chen X. & Murphy R.W. (2013) Assessing DNA barcoding as a tool for species identification and data quality control. *PLoS One* **8**, e57125.

Silva J.M.D. & Willows-Munro S. (2016) A review of over a decade of DNA barcoding in South Africa: A faunal perspective. *African Zoology* **51**, 1-12.

Simpson E. (1949) Measurement of diversity. *Nature* **163**, 688.

Smith M.A., Fisher B.L. & Hebert P.D. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B Biological Sciences* **360**, 1825-1834.

Srivathsan A., & Meier R. (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcode literature. *Cladistics* **28**, 190-194.

Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313.

Sullivan J. & Joyce P. (2005) Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **36**, 445-466.

Tamura K., Dudley J., Nei M. & Kumar S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.

Tavares E.S. & Baker A.J. (2008) Single mitochondrial gene barcodes reliably identify sister-species in diverse clades of birds. *BMC Evolutionary Biology* **8**, 81.

Tembe S., Shouche Y. & Ghate H.V. (2014) DNA barcoding of Pentatomomorpha bugs (Hemiptera: Heteroptera) from Western Ghats of India. *Meta Gene* **2**, 737-745.

Trigg S. & Flasse S. (2001) An evaluation of different bi-spectral spaces for discriminating burned shrub-savannah. *International Journal of Remote Sensing* **22**, 2641-2647.

Virgilio M., Backeljau T., Nevado B. & De Meyer M. (2010) Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* **11**, 1-10.

Wang X.B., Deng J., Zhang J.T., Zhou Q.S., Zhang Y.Z. & Wu S.A. (2015) DNA barcoding of common soft scales (Hemiptera: Coccoidea: Coccidae) in China. *Bulletin of Entomological Research* **105**, 545-554.

Wiemers M. & Fiedler K. (2007) Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* **4**, 8.

Willows-Munro S. (2013) eThekwini urban barcoding project: DNA barcoding explores invertebrate diversity in an urban setting. *Barcode Bulletin* **2**, 18.

Zachos F.E. & Habel J.C. (2011) *Biodiversity hotspots: distribution and protection of conservation priority areas*. Germany: Springer Science & Business Media.

Chapter Three

The utility of DNA barcoding as a practical tool to assess the success of ecological restoration using Hemiptera as a biological indicator

Abstract

DNA barcoding has been used in a wide range of studies to understand and assess aspects that are related to ecology, evolution, conservation, and biogeography. In addition, it is useful in many different ecological applications such as species monitoring, and ecological habitat restoration. Invertebrates are excellent biological indicators as they can be used to assess changes in species diversity or community assemblage in the context of restoration ecology. Understanding trends in species composition and assemblage of key invertebrate groups can provide important insight into the condition of, or changes in the environment. In this study, DNA barcoding is used to assess the potential of Hemiptera as an indicator of restoration success for the Buffelsdraai Landfill Site Community Reforestation Project. A total of 393 specimens were collected from sites reforested at distinct phases in the years 2010, 2012 and 2015 and reference sites (forest and grassland) selected within the buffer zone of the Buffelsdraai regional landfill site. The Hemiptera species composition and assembly were assessed by analyzing multiple diversity indices, ordination, UPGMA cluster analysis and phylogenetic analysis. A significant difference was uncovered for Hemiptera species composition among the different reference sites as well as between 2015, 2012 and 2010 reforested sites. This study highlights the utility of DNA barcoding as a tool to monitor the success and progress of the reforestation being carried out in the buffer zone of the Buffelsdraai regional landfill site by further analyzing the use of Hemiptera as a suitable biological indicator.

3.1 Introduction

3.1.1 The use of DNA barcoding as a practical tool in conservation biology

Many biological fields such as ecology, evolution, biogeography, and conservation biology utilizes molecular analysis to understand species diversity, composition, and assembly (Fine & Kembel 2011; Harmon-Threatt & Ackerly 2013; Willows-Munro & Schoeman 2015). In recent years, the use of DNA barcoding as a molecular tool to identify species, has proved to be an appealing tool to help resolve taxonomic ambiguity, to enhance biodiversity inventories, and to support many applications that require species identification in conservation biology (Hebert *et al.* 2004; Rubinoff *et al.* 2006; Valentini *et al.* 2008). A combination of ecological processes and evolutionary changes influence community composition at both local and regional scales, therefore DNA barcoding has the potential to become an integral tool to understand these changes (Kress & Erickson 2008; Valentini *et al.* 2008; Fine & Kembel 2011; Harmon-Threatt & Ackerly 2013). In addition, DNA barcoding can be used in ecological applications such as species monitoring, identifying morphologically indistinguishable life stages, and monitoring of ecological change (Kress & Erickson 2008; Valentini *et al.* 2008).

Despite the utility of DNA barcoding in ecological applications, there are a limited number of studies that have utilized DNA barcode data at a practical level to monitor the progress of ecological restoration. Rather, a focus has been placed on the choice of the barcoding region and technical aspects of generating useful DNA sequence data (Kress & Erickson 2008). This study aims to test the practical potential for using DNA barcoding in conservation biology and restoration ecology.

3.1.2 Ecological restoration

The increase in urbanization has caused an increase in the loss and degradation of natural habitats and biodiversity, there is thus a need for ecologists and conservation biologists to pay attention to urban restoration (Wallington *et al.* 2005; Cabin *et al.* 2010; Verdú *et al.* 2012). Ecological restoration is the process of assisting in the recovery of an ecosystem that has been disturbed, degraded, and destroyed through human involvement such as deforestation and urbanization (Ruiz-Jaen & Mitchell Aide 2005; Clewell & Aronson 2007; Wortley *et al.* 2013). Ecological restoration projects differ in their objectives and goals; however, the main goals are usually to enhance, mitigate, and re-establish ecosystems that are resilient, self-sustaining and can recover from anthropogenic disturbances (Ruiz-Jaen & Mitchell Aide 2005; Wortley *et al.* 2013; Galimberti *et al.* 2016). Restoration can accomplish these goals by creating environments which have a diverse species composition, which demonstrates a complex community structure, are indigenous to the area being restored and that contains species which belong to multiple functional groups (Clewell & Aronson 2007; Galimberti *et al.* 2016). Restoration plays a key role in the conservation of biodiversity and by understanding the processes of restoration, important information on community and evolutionary ecology can be obtained (Verdú *et al.* 2012).

A key consideration in ecological restoration is the ability to monitor the progress and success of restoration efforts (Ruiz-Jaen & Mitchell Aide 2005). Monitoring of restoration is critical for the development of good land-use practices which would allow for adaptive management and maintenance of restored sites (Ruiz-Jaen & Mitchell Aide 2005; Clewell & Aronson 2007; Wortley *et al.* 2013). The Society for Ecological Restoration (SER) Primer is a driving force behind ecological restoration and contributes basic guidelines towards restoration planning together with a list of nine key attributes for successful restoration (Ruiz-Jaen & Mitchell Aide 2005; Wortley *et al.* 2013). These attributes can be narrowed down to three main factors, namely, vegetation structure, species diversity and abundance, and ecological processes (Ruiz-Jaen & Mitchell Aide 2005). This study focuses on how vegetation structure affects species diversity and abundance.

Another key aspect of restoration monitoring is the inclusion of information from reference sites. Reference sites are valuable sources of information that demonstrate the intended path of restoration projects and are an essential component to the evaluation of restoration success (Ruiz-Jaen & Mitchell Aide 2005; Clewell & Aronson 2007). Reference sites should preferably occur within the same landscape or close to the restoration site, as they need to share key ecological and geographic features to serve as a template for species composition and community assembly (Ruiz-Jaen & Mitchell Aide 2005; Clewell & Aronson 2007). Having more than one reference site in a restoration study is advised as variation often occurs among different sites (Ruiz-Jaen & Mitchell Aide 2005).

3.1.3 The use of biological indicators to monitor the progress and success of ecological restoration

The biodiversity present in restored plots are one of the most important factors to consider when assessing restoration success. Functional self-sustaining ecosystems vary in species composition, diversity, and stratification (Moir *et al.* 2005; Orabi *et al.* 2010; Pander & Geist 2013). The high levels of species diversity and abundance are associated with biodiverse environments which contribute to ecosystems resilience (Orabi *et al.* 2010; Pander & Geist 2013). Accurately sampling and monitoring all species in an environment is a difficult task. Instead suitable, and effective biological indicators need to be selected to evaluate restoration success (Anderson *et al.* 2011; Pander & Geist 2013). Good indicators should display a narrow ecological range, reflect the abiotic or biotic state of an environment, possess rapid response to environmental change, have a wide distribution, and must be inexpensive to sample (Bellinger & Sigee 2010; Anderson *et al.* 2011). In addition, an important aspect to consider when selecting biological indicators in ecological restoration is to understand the ecological relationship between the selected indicator and the wider community structure of the restoration site (Orabi *et al.* 2010; Anderson *et al.* 2011).

Invertebrates are well established as good biological indicators as they are very small, often mobile, have relatively short generation times and are sensitive to changes in an ecosystem (Anderson *et al.* 2011). They can also be used to monitor species diversity, community assembly and species composition (Orabi *et al.* 2010). Many taxonomic groups such as Aves (Galimberti *et al.* 2016), Coleoptera (Parmenter & Macmahon 1987) and Orthoptera (Parmenter *et al.* 1991) have been studied and used to monitor the success of ecological restoration, however, very few studies have focused on the use of Hemiptera as a biological indicator.

The order Hemiptera is one of the largest, most diverse groups of hemimetabolous insects (Park *et al.* 2011; Raupach *et al.* 2014). They are known to play a functional role in agricultural ecosystems and are used as biological control agents (Wheeler 2001; Park *et al.* 2011; Raupach *et al.* 2014). Members of this order are known to reflect biophysical changes in the environment as they are sensitive to habitat fragmentation and environmental changes (Orabi *et al.* 2010). In addition, Hemiptera may be suited to studying and understanding the progress of restoration as the group includes both specialist and generalist plant feeders, making them sensitive to changes in the quality, structure, and species composition of the plants they feed on (Prestidge & McNeill 1983; Hartley *et al.* 2003; Orabi *et al.* 2010).

Despite the dominance of Hemiptera in many ecosystems and their advantages as biological indicators, there is very limited knowledge and information on them regarding their use in ecological restoration and reforestation. This present study is unique in that restoration studies have seldom used genetic approaches to understand how species composition and assemblage are affected in restored environments with the changes in vegetation types. This study focuses on the use of DNA barcoding as a tool to monitor the success and progress of ecological restoration and reforestation. Hemiptera will also be assessed as suitable biological indicator taxa.

3.1.4 Buffelsdraai reforestation and restoration ecology case study

The Buffelsdraai Landfill Site Community Reforestation Project was initiated in 2008 by the eThekwini Municipality in partnership with the Wildlands Conservation Trust and Durban Solid Waste Management Department (Douwes *et al.* 2015a; Douwes *et al.* 2015b). This project started to help alleviate the impact of climate change and to help offset a percentage of the greenhouse gas emissions generated by hosting the FIFA 2010 World Cup^{TM} in Durban, KwaZulu-Natal (Douwes *et al.* 2015a; Douwes *et al.* 2015b; Diederichs & Roberts 2016). However, the aims of this community reforestation programme are three-fold: 1) to uplift the surrounding community, 2) to improve biodiversity, and 3) to contribute to climate change adaptation and mitigation.

This project was started within the 757-hectare buffer zone of the municipality's Buffelsdraai regional landfill site. The land that the landfill site and the buffer zone is situated on was previously old agricultural land, planted with sugarcane, or land with limited productive capacity, or infested by invasive alien plants (Douwes *et al.* 2015a; Douwes *et al.* 2015b; Diederichs & Roberts 2016). Of the 757-hectare buffer zone, 580-hectares has been set aside for the reforestation project. This restoration project aims to increase the native biodiversity present in an area that was previously under sugarcane production and to transform the buffer zone into a native forest (Douwes *et al.* 2015a; Douwes *et al.* 2015b; Diederichs & Roberts 2016). Initial indigenous tree planting started in 2009, this was then followed by more intensive planting from 2010 onwards (Douwes *et al.*

2015a; Douwes *et al.* 2015b; Diederichs & Roberts 2016). As of January 2015, a total of 595 476 trees were planted in the buffer zone, these trees include 51 locally indigenous species, the most common being *Acacia natalitia*, *Erythrina lysistemon* and *Bridelia micrantha* (Douwes *et al.* 2015a).

3.1.5 Aims

This study aims to test whether South African Hemiptera can be used as biological indicators to track the progress of ecological habitat restoration in the region. This will be carried out using Hemiptera specimens collected from the Buffelsdraai Regional Landfill Site's buffer zone. Hemiptera specimens were collected from the restoration sites which were reforested at distinct phases (2015, 2012 and 2010). Species diversity, richness, and composition of Hemiptera will be analysed. The Hemiptera collected from the selected sites will be compared to the native grassland and forest sites also present within the Buffelsdraai Regional Landfill Site's buffer zone. DNA barcoding will be used to identify Hemiptera species and to further analyse the use of Hemiptera as a suitable biological indicator taxa.

3.2 Materials and methods

3.2.1 The study area

This study took place at the Buffelsdraai Regional Landfill Site, north of Durban, in the KwaZulu-Natal province, South Africa (Figure 3.1). All landfill sites are required to have a buffer zone between the active landfill and the adjacent communities. The Buffelsdraai buffer zone is a minimum of 800 m wide and 787 hectares in extent which shields the neighbouring communities (Buffelsdraai and Osindisweni) from the impacts of the landfill site (Douwes *et al.* 2015b). The study areas encompass different types of vegetation and land use types. These include woodlands, sugarcane, riparian forest, transitional weed, fallow lands, thicket, indigenous forest (non-riparian), grassland, maintained areas, rural settlements, infrastructure and bareland (Douwes *et al.* 2015a; Douwes *et al.* 2015b; Diederichs & Roberts 2016). A large portion of the 757-hectare buffer zone was under sugarcane cultivation, until 2009 when the eThekwini municipality decided to rehabilitate the sugarcane, transitional weed and fallow lands to their original forested state

(Douwes *et al.* 2015b). The Buffelsdraai regional landfill site’s buffer zone were divided up into management blocks which were reforested in distinct phases. Indigenous tree planting started in 2009, which was followed by more intensive planting in 2010, 2012 and 2015 (Douwes *et al.* 2015a; Douwes *et al.* 2015b; Diederichs & Roberts 2016).

3.2.2 Study design and site selection

Hemiptera samples were collected from 20 plots from the 8 different sites which were selected from the Buffelsdraai regional landfill site’s buffer zone (Table 3.1, Figure 3.1). Samples were collected from three management blocks, which had been planted with indigenous trees in 2015, 2012, and 2010. The 2015 reforested site was predominantly covered in sugarcane with scattered tree saplings (less than 1 m in height), the 2012 site had a mixture of grassland and scattered trees (2 – 3 m height in average) and the 2010 site was predominately covered in a juvenile forest (small trees up to 5 m with a tree canopy cover between 50 - 75%) vegetation type. Samples were also collected from reference sites which had native forest and grassland. Each of the reforested and reference sites were sampled four times to get a true reflection of the biodiversity within the sites. The reforested areas were sampled at two wet sites and two dry sites. Sites that were classified as wet were found close to or alongside the drainage lines, whereas sites that were classified as dry were found further away from the drainage lines.

Table 3.1: Sites selected for this study. The table indicates the number of plots each site was sampled at, the age of reforestation at each site and the predominant vegetation type occurring at each site.

Site code	Restored/ reference	No. of plots sampled	Age class (years)	Vegetation type
2015 wet	Restored/ reference	2	1	Sugarcane + scattered trees
2015 dry	Restored/ reference	2	1	Sugarcane + scattered trees
2012 wet	Restored	2	4	Grassland + scattered trees
2012 dry	Restored	2	4	Grassland + scattered trees
2010 wet	Restored	2	6	Juvenile forest
2010 dry	Restored	2	6	Juvenile forest
Forest	Reference	4	-	Established forest
Grassland	Reference	4	-	Grassland

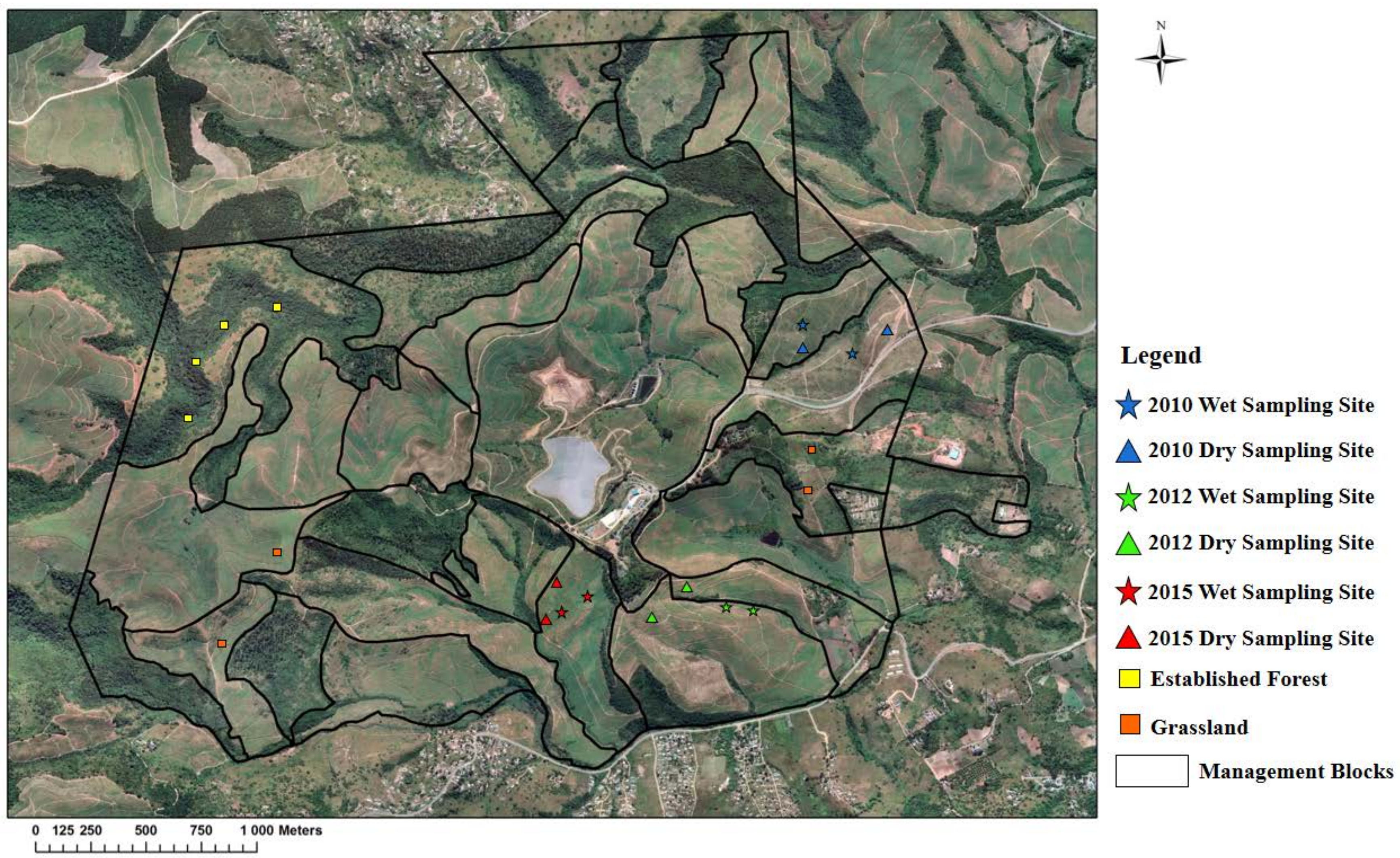

Figure 3.1: Map illustrating the 20 plots from the 8 different sites selected for sampling within the Buffelsdraai regional landfill site's buffer zone (KwaZulu-Natal, South Africa).

3.2.3 Sampling procedure

During October 2015 Hemiptera specimens were collected from the different sites using a combination of sweep netting, tree beating and active searching. A total of 393 specimens were collected. All specimens were stored in ethanol (100%) at -20 degrees Celsius.

3.2.4 Sorting of specimens into morphospecies

The collected Hemiptera specimens were identified to the lowest taxonomic level possible through the use of morphological keys (Schuh & Slater 1995; Maw *et al.* 2000; Cassis *et al.* 2002). Morphological features such as body size, shape, colour patterns and wing structure were taken into consideration (Shen *et al.* 2013). Each specimen was given a unique code to assist with identification and data tracking. Each code is made up three characteristics, the location from where the specimen was collected, the type of vegetation present at the location, and a unique specimen number. All the specimen was photographed using a USB Digital Microscope 2.0 - Moticam 2300. Specimens were deposited at the University of KwaZulu-Natal (Pietermaritzburg campus). This repository is recognized by the Barcode of Life project as an official storage facility for specimens that have been barcoded. Successful DNA barcodes in this study will be uploaded to BOLD.

3.2.5 Genomic DNA extraction and Polymerase Chain Reaction (PCR) amplification

Where possible, five individuals of each morphospecies per sampling site were selected for DNA barcoding analysis. Multiple individuals of each morphospecies were included to incorporate spatially-correlated variation within species (Bergsten *et al.* 2012). A single leg from ethanol-fixed specimens was removed and the total genomic DNA was extracted from 132 specimens using the ZR Tissue and Insect MicroPrep^{TM} kit (ZYMO research), following the manufacturer's instructions. The concentrations and quality of the extracted genomic DNA was quantified using a NanoDrop Spectrophotometer 2000 (Thermo Scientific).

The mitochondrial COI gene was amplified using the universal primers LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TAAACTTCAGGGTGACCAAAAAATCA-3′) originally developed by Folmer et al. (1994). The PCR and PCR product visualization was carried out as described in chapter two.

3.2.6 Sequencing, data processing, phylogenetic tree construction and database construction

Successfully amplified PCR products were sent to the Central Analytical Facility at Stellenbosch University for sequencing. Both forward and reverse sequences were generated. The electropherogram of each PCR amplicon was examined for quality in BioEdit 7.0.9.0 sequence alignment editor (Hall 1999). Barcode compliant sequences (>500-bp of length, no stop codons, misidentifications, or contamination) generated in the present study (n = 132) were aligned using ClustalX 2.1 (Larkin *et al.* 2007). Summary statistics such as the number of variable characters (V), number of parsimony informative characters (Pi) and the average nucleotide compositions were estimated for the data using MEGA 6.0 (Tamura *et al.* 2007).

MEGA 6.0 was used to construct a neighbor-joining (NJ) tree, thereafter, Barcode Index Numbers (BINs) were assigned to barcode clusters present on the NJ tree. Each COI sequence on the NJ tree was blasted against the global BOLD reference library. Sequence similarity to the top search hit was recorded. Taxa with similarity values greater than a 95% were considered already present in BOLD. Conditional genus-level or family-level identification was allocated to taxa with sequence similarity values less than 95%.

3.2.7 Estimating species diversity and abundance

A database of the Hemiptera specimens collected from Buffelsdraai was constructed using Microsoft Office Excel 2010. The database served as a personal reference library and included the following information for each specimen: BIN, the unique specimen name, morphospecies allocation, taxonomy (order, family and where possible, genus and species name), sampling site, vegetation type and sequence similarity score from the top hit in BOLD. This information was then used to compare the total number of specimens collected from the different restored (wet and dry) and reference sites, in addition this information was used to compare the number of morphospecies and BINs at the different sampling sites. A chi-squared test was done to check for any significant difference between the number of morphospecies and number of BINs recorded from each sampling site.

3.2.8 Diversity indices

Species diversity analyses were carried out for Hemiptera at each of the sampling sites using the program Past 3.44 (Hammer *et al.* 2001) on the full barcode dataset (BINs dataset) containing 132 sequences. These analyses were carried out on the dataset containing BINs to link sample heterogeneity and estimates of species richness and abundance, both within and between each locality, species diversity was calculated using four most widely accepted diversity indices, namely, Margalef's diversity index (*d*) (Margalef 1958), Fisher's α (α) (Fisher *et al.* 1943), Shannons diversity index (*H'*) (Shannon & Weaver 1963) and Simpson's diversity index (*D*) (Simpson 1949) as illustrated in chapter two.

3.2.9 Ordination

Non-metric multidimensional scaling (NMDS) analysis based on Bray-Curtis dissimilarity matrix (Bray & Curtis 1957) was carried out in Past 3.44 to visualize differences between sampling points in ordination space, on the BINs dataset. In this study, Hemiptera species composition and assemblage were considered. When using 2-dimentional ordination, the stress value associated with this analysis is affected by the quality of the data used. The general rule to analysis the stress value output from the ordination analysis is as follows: stress ≤0.05 gives an excellent representation, with no prospect of misinterpretation of the data, stress ≤0.1 represents a good ordination with no real risk of misinterpretation, stress ≤0.2 may still give a potentially useful ordination, but cross-checks with other techniques are recommended (Clarke & Warwick 1994).

The significance of the NMDS clusters was tested by using permutational MANOVA (PERMANOVA) which was carried out in Past 3.44. PERMANOVA is a multivariate analysis of variance technique that is used for abundance data, where significance is based on permutation of the dissimilarity matrix (Anderson 2001). PERMANOVA was conducted using the Bray-Curtis dissimilarity matrix to determine the significance between the different sampling site points in the ordination plot of the Buffelsdraai data.

3.2.10 UPGMA clustering analysis

In addition, the clustering analysis using unweighted pair-group average (UPGMA) was carried out to cross-check the reliability of the NMDS ordination. UPGMA was conducted using the Bray-Curtis dissimilarity matrix in Past 3.44 on the BINs dataset, clusters were joined based on the mean of the relative abundance of specimens for each locality.

3.2.11 Phylogenetic Analysis

In addition to the construction of a neighbor joining (NJ) tree, an additional maximum likelihood phylogenetic approach was taken to assess the taxonomic coverage of Hemiptera across the different sampling sites as well as test whether there were indicator species that were associated with a specific habitat. Maximum likelihood analysis was performed on the BINs dataset. The best-fit nucleotide substitution model was selected using the Akaike Information Criterion (AIC) (Akaike 1974) in the program jModelTest 2 (Darriba *et al.* 2012). The model of best-fit for the data was the general time reversible model (GTR) with both gamma distribution (G) and proportion of invariable sites (I). This model was used in all subsequent maximum likelihood analysis.

RAxML 8.0 (Stamatakis 2014) was used to construct the maximum likelihood analysis. The tree search method was ML + thorough bootstrap implementing 1000 bootstrap replicates. This analysis generated both the most likely tree and a bootstrap tree file, which contained the 1000 bootstrap trees. A consensus tree was created from the bootstrap file using Phylip 3.69 (Felsenstein 2005). Thereafter, trees were viewed and modified using the program Figtree 1.3.1 (Rambaut 2009).

3.3 Results

3.3.1 Data description

The COI gene region was successfully amplified and sequenced from all 132 specimens, these specimens were all barcode-compliant (>500-bp of length, no stop codons, misidentifications, or contamination). The 132 sequences were easily alignable as there were no insertions, deletions, or stop codons. The final alignment was 634 base pairs in length (381 variable sites of which 347 were parsimony informative). The Hemiptera database used for this study consisted of 132 specimens, which were clustered into 119 BINs and 117 morphospecies. The average base composition for the full dataset was 35.5% thymine (T), 18.2% cytosine (C), 31.1% adenine (A) and 15.2% guanine (G). The nucleotide compositions of all the sequences were heavily biased toward A and T nucleotides which are common in insect mitochondrial genomes (Raupach *et al.* 2014).

3.3.2 Species richness and diversity

To estimate the species richness, abundance and diversity, the number of specimens for the dry sites (n = 99), wet sites (n = 127), overall total specimens (n = 393) and morphospecies (n = 307) per sampling site were plotted (Figure 3.2A) using the total number of specimens collected from the Buffelsdraai landfill sites' buffer zone. In general, there is an increase in the number of specimens in the wet sites as compared to the dry sites. This suggests that the moisture content of the different sites directly influences the distribution of specimens. Further, the trend shown in Figure 3.2 suggests that the number of specimens and morphospecies are directly proportional to the age of the reforested sites and the vegetation types of the different sites. As the vegetation type at the sites shifts from sugarcane and to native grasslands and forests, the species richness and abundance values increase, highlighting the success of the ecological restoration. The specimens collected from the forest and grassland reference sites are used to represent the natural undisturbed environments within Buffelsdraai landfill sites' buffer zone. There is evidence of an elevated number of specimens in the 2010 reforested site as compared to the forest reference site, this is due to three confounding factors. First, the samples collected from the forest reference site could be underrepresented as it is difficult to sample tall trees and the dense vegetation type which is predominate in the forest reference site. Second, there is a greater diversity of ecological niches with a mixed vegetation type present in the 2010 reforested site as compared to the forest reference site which only has one type of vegetation structure. The increased number of vegetation types in the 2010 reforested site could have affected the elevated number of specimens collected from this site as compared to the forest reference site. Third, there could be a high number of introduced species in the 2010 reforested site rather than natural species. Examples of introduced Hemiptera species in the 2010 reforested site include *Clavigralla horrida* (family: Coreidae), *Zicrona caerulea* (family : Pentatomidae), and *Spilostethu* s*pandurus* (family : Lygaeidae) which are native to Costa Rica and are known to be invasive species all around the world (Kaufman *et al.* 1929).

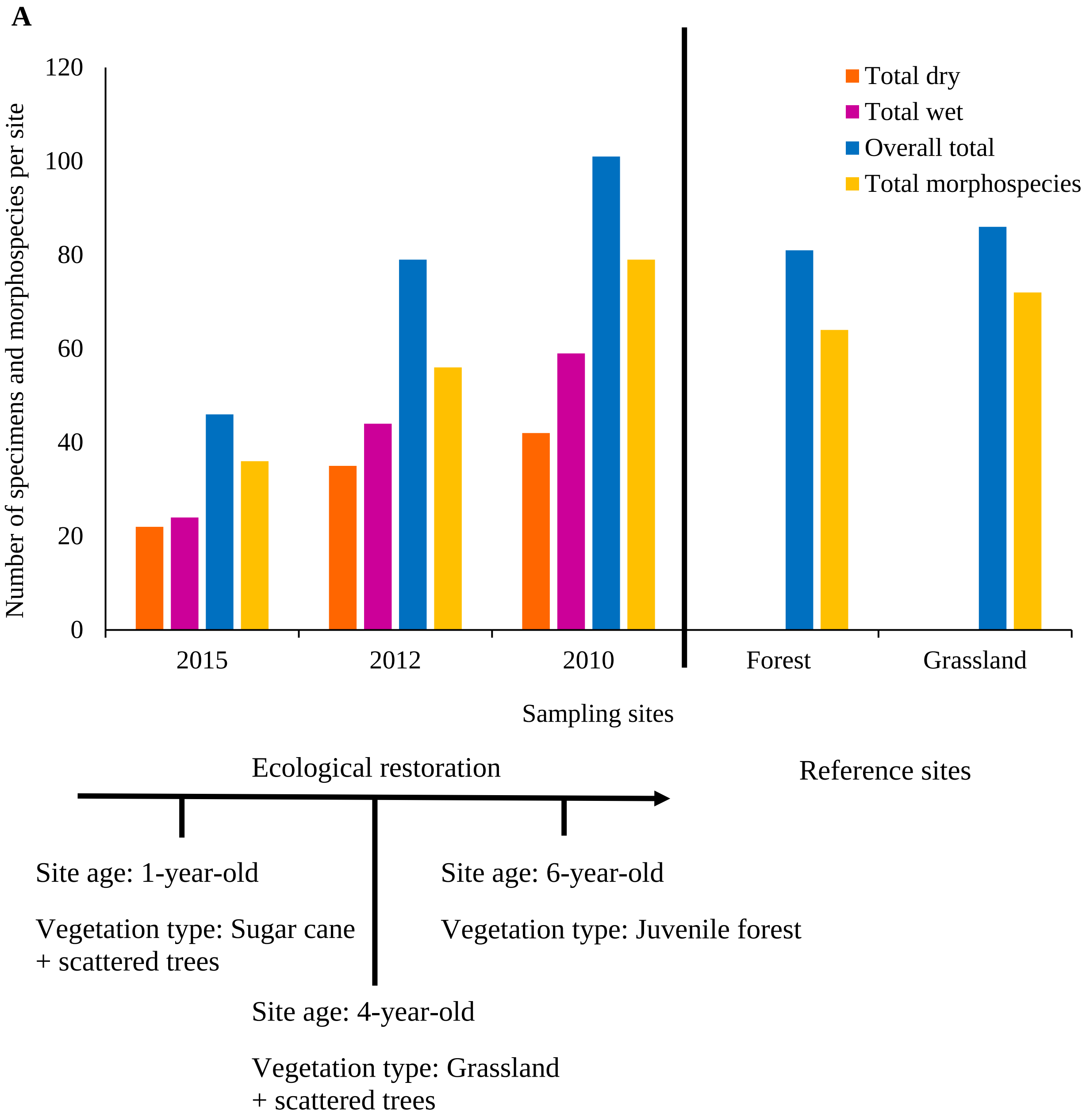

Figure 3.2: Bar graph representing the distribution of specimens and morphospecies amongst the different sampling sites included in the present study. Each of the test sites (2015, 2012, and 2010) were divided into wet and dry. The number of specimens per wet (orange) and dry (pink) were plotted together with the total number of specimens for both the sites (blue) and morphospecies (yellow).

3.3.3 The correlation between BINs and morphospecies

Specimens collected from the different sites within the Buffelsdraai landfill sites' buffer zone were sorted into morphospecies, thereafter, a maximum of 5 representatives of each morphospecies was selected for DNA analyses. The number of specimens (n = 132), morphospecies (n = 119) and BINs (n = 117) per sampling site were plotted (Figure 3.3) and used to assess the correlation between BINs and morphospecies observed. Generally, the number of BINs and morphospecies observed from each site were similar. In support of this observation, a chi-square test of coherence between BINs and morphospecies demonstrated no significant difference in any of the sites ($p > 0.05$). Both the 2012 and 2010 reforested sites have an increased number of BINs as compared to morphospecies, this variation in the number of morphospecies and BINs recorded from these sites could suggest cryptic speciation. In contrast, the grassland reference site is observed to have an over estimation of species based on morphology when compared to BINs, which suggests the presence of Hemiptera specimens which are morphologically variable or specimens which are at different developmental stages. The presence of cryptic species and species at different developmental stages highlight the importance of utilizing DNA barcoding as an identification tool in this study. When selecting a maximum of 5 morphospecies per site, it is observed that the reforested plots had a high species diversity when to compared both the forest and grassland reference sites. This is due to the three confounding factors mentioned in the previous section (3.3.2 Species richness and diversity).

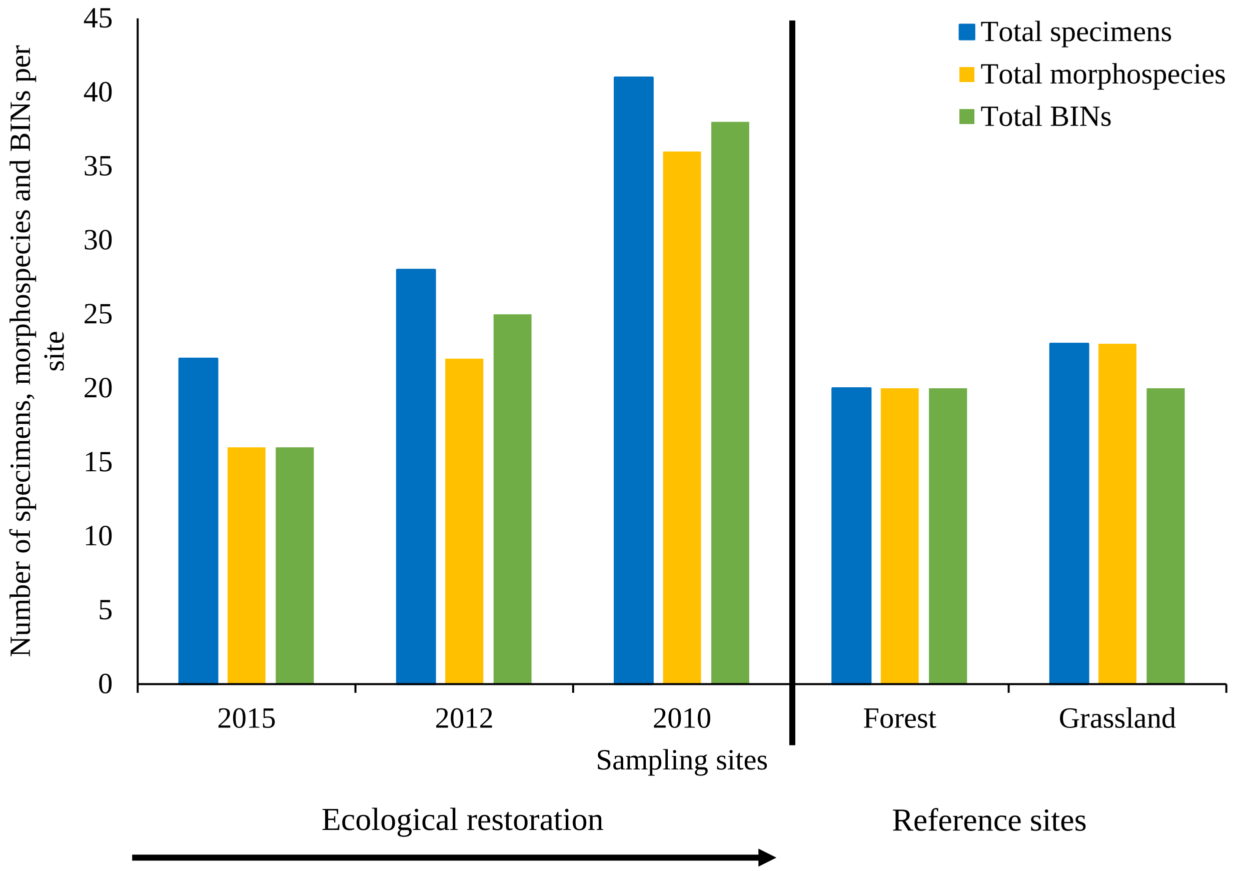

Figure 3.3: Bar graph representing the distribution of specimens, morphospecies and BINs amongst the different sampling sites included in the present study.

3.3.4 Diversity indices

Species richness, abundance and diversity were calculated and used to assess the different reforested sites (2015, 2012, and 2010) and reference sites (forest and grassland) in the Buffelsdraai landfill sites' buffer zone on the BINs dataset (n = 132) (Table 3.2). Diversity indices for each of the reforested sites were calculated for both the wet sites and the dry sites, in addition, the samples from both the wet and dry sites were pooled and the diversity indices were calculated overall for each of the reforested sites. Species richness (S) among the different sites varied significantly, with the 2010 reforested site having the highest species richness (S = 20 dry sites and 21 wet sites), followed by the 2012 reforested site (S = 10 dry sites and 18 wet sites) and the

2015 reforested site (S = 12 dry sites and 8 wet sites). In the reference sites, it was observed that the grassland sites had a higher species richness (S = 23) then the forest sites (S = 20). When comparing the overall reforested sites to the reference sites, it is observed that the 2010 reforested sites had a higher species richness (S = 36) then both the grassland and the forest reference sites and that the 2012 reforested site had a higher species richness (S = 21) then the forest reference site. The 2015 reforested sites had a much lower species richness (S = 17) then both reference sites. This suggests that the shift from the sugarcane vegetation type to the forest vegetation type has a positive impact on the species richness.

The Margalef's index values ranged between *d* = 2.38 and 5.32 and the Fisher α diversity index values ranged between α = 5.21 and α = 18.53. The highest species diversity and α values was observed in the 2010 reforested sites (*d* = 5.30 dry sites and 5.32 wet sites, α = 16.21 dry sites, 18.53 wet sites) followed by the 2012 reforested sites (*d* = 2.87 dry sites and 4.86 wet sites, α = 6.73 dry sites and 16.21 wet sites), while the 2015 reforested site had the lowest species diversity (*d* = 3.88 dry sites and 2.38 wet sites, α = 18.17 dry sites and 5.12 wet sites). The grassland sites had a higher diversity α (*d* = 5.14, α = 11.98) then the forest sites (*d* = 4.57, α = 9.99). A total of 39 morphospecies were present only in the reforested sites, 15 morphospecies were present only in the reference sites and 24 morphospecies that were present in both the reforested sites and reference sites. As the reforested sites get older, the species diversity within the sites increases.

Most of these trends were further supported by the Shannons and Simpsons indices in terms of each sites' abundance, richness, and diversity. However, both these indices ranked the Hemiptera community structure of the 2015 reforested sites (D = 0.89 dry sites and 0.81 wet sites, H = 2.51 dry sites, 1.93 wet sites) to be more diverse than 2012 reforested sites (D = 0.79 dry sites and 0.87 wet sites, H = 1.84 dry sites, 2.34 wet sites). This is unsurprising as the 2015 site has a well-established sugarcane vegetation type as compared to the intermediate changing phase from sugarcane into forest vegetation in the 2012 reforested site.

When comparing the dry and wet sites within the reforested sites, the wet sites shows a slightly higher richness in the 2010 and 2012 reforested sites than the dry sites. However, the dry sites of the 2015 reforested sites had a higher species richness and diversity (D = 0.89, H = 2.51) then the wet sites (D = 0.81, H = 1.93) suggesting that the species found within the 2015 reforested sites have adapted to the dry environment of the sugarcane and can survive within this vegetation type. The overall structure, vegetation, and diversity of the different ecological niches could have affected the variation in species diversity across the different reforested and reference sites.

Table 3.2 Univariate models of biodiversity estimation indices for each sampling site based on BINs. S represents the total number of species present (richness), N represents the total number of individuals per a locality (abundance), *d* represents species diversity calculated from the Margalef's diversity index, α represents Fisher's α, indicating that the division of species per sampling locality, H' represents Shannons index which estimates biodiversity richness and diversity and D represents Simpsons index which quantifies the biodiversity present at each locality.

	2015			**2012**			**2010**				
	Dry	Wet	Overall	Dry	Wet	Overall	Dry	Wet	Overall	Forest	Grassland
S	12	8	17	10	18	21	20	21	36	20	23
N	17	19	36	23	33	56	36	43	79	65	72
d	3.88	2.38	4.97	2.87	4.86	4.47	5.30	5.32	8.01	4.57	5.14
α	18.17	5.21	12.59	6.73	16.21	12.20	16.21	18.53	25.55	9.99	11.68
D	0.89	0.81	0.91	0.79	0.87	0.86	0.87	0.90	0.92	0.86	0.91
H	2.51	1.93	2.64	1.84	2.34	2.53	2.55	2.66	3.03	2.44	2.72

3.3.5 Ordination

The non-metric multidimensional scaling (NMDS) ordination (Figure 3.4, Stress value = 0.1794) and UPGMA clustering analysis (Figure 3.5) was performed using Bray-Curtis dissimilarity matrix between the different sites. The points of the ordination graph and the branches on the cladogram were colour coded to differentiate between the different sampling sites. In Figures 3.4 and 3.5 there is a clear grouping of species collected from the different vegetation types, namely, the grassland (maroon), forest (pink) and the 2015 reforested sites which are still predominantly covered in sugarcane (green). PERMANOVA confirmed a significant difference in species composition among these three vegetation types (Table 3.2, $p < 0.05$).

The NMDS ordination illustrates that the species composition from each of the reforested sites and references sites grouped together (Figure 3.4). A tight grouping was observed between the species found in the forest reference sites and the species found in the 2010 and 2012 reforested sites. These sites were found not to be significantly different from each other in terms of species composition (Table 3.3, $p > 0.05$). However, these sites are significantly different from the 2015 reforested sites (Table 3.3, $p < 0.05$). These groupings highlight the strong influence that restoration efforts have made in the 2010 and 2012 reforested sites. In the 2015 reforested sites, species for the dry sites grouped together and species for the wet sites were grouped together highlighting a distinct difference between the species composition between each of these sites of different moisture content (Figures 3.4 and 3.5).

The results obtained from Figures 3.4, 3.5 and Table 3.3 suggest that the species composition within the older reforested sites (2010 and 2012) have shifted towards a similar composition as the forest reference sites, indicating the effectiveness of the reforestation efforts being carried out in Buffelsdraai Landfill Site Community Reforestation Project.

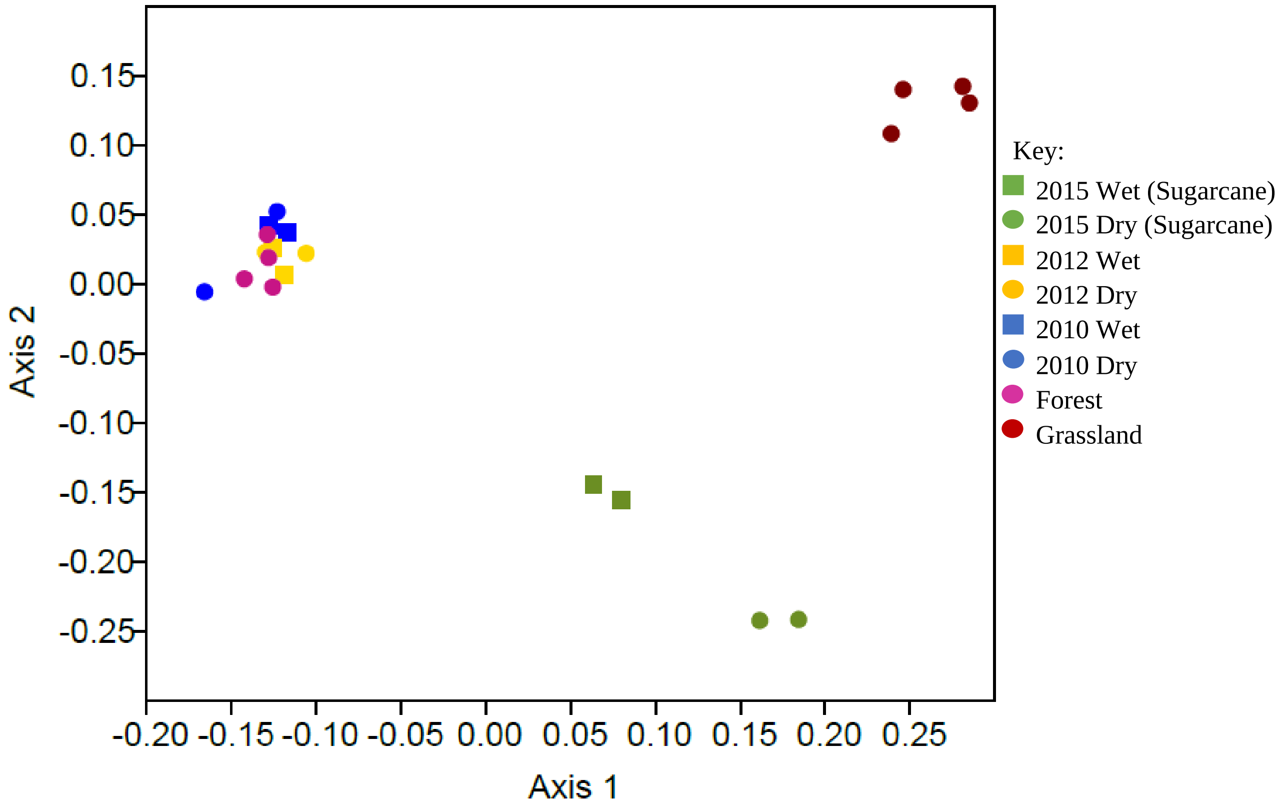

Figure 3.4: Non-metric multidimensional scaling (nMDS) plots of the first 3 dimensions based on Bray-Curtis dissimilarities for the community composition of Hemiptera collected from 2015 (green), 2012 (yellow), 2010 (blue), and the forest (pink) and grassland (maroon) reference sites. The stress value is 0.1794.

Table 3.3: P-values derived from PERMANOVA pairwise comparisons of the different sites community composition within the Buffelsdraai landfill sites' buffer zone using the Bray-Curtis dissimilarities values. The values highlight in bold are statistically significant (P < 0.05).

	2010	2012	2015	Forest	Grassland
2010	-	0.40	**0.030**	0.95	**0.028**
2012	0.40	-	**0.032**	0.18	**0.027**
2015	**0.030**	**0.032**	-	**0.032**	**0.028**
Forest	0.95	0.18	**0.032**	-	**0.027**
Grassland	**0.028**	**0.027**	**0.028**	**0.027**	-

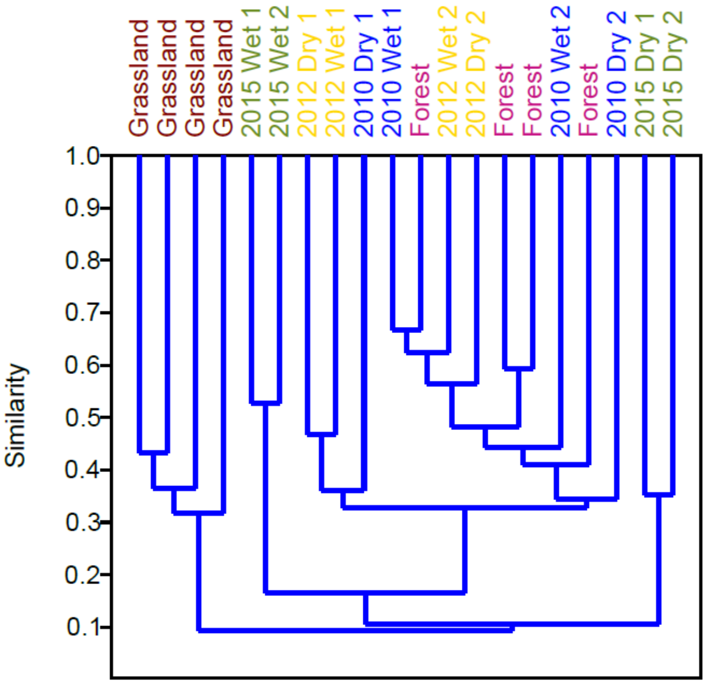

Figure 3.5: UPGMA cluster analysis of the Hemiptera communities for the different sampling sites in Buffelsdraai. The UPGMA clustering was based on Bray-Curtis dissimilarities matrix using the mean of the relative abundance of specimens for each locality.

3.3.6 Phylogenetic analysis

A total of 132 DNA sequences were available for phylogenetic analysis. The maximum likelihood (ML) tree clustered similar or putatively related individuals together (Figure 3.6). All values above 75% bootstrap (BS) were considered significantly supported and highlighted on the phylogeny. This phylogeny was midpoint rooted as no outgroup was included. The phylogeny represents 13 families and 48 genera. The mutation rate of the COI marker is appropriate to resolve species-level associations but our data suggests that this molecule may also be useful at higher taxonomic levels. Ten of the thirteen families present in the dataset were monophyletic and three of the families were either polyphyletic or paraphyletic. There is evidence of monophyly in the following families, however, in most of these families the branch support values were not very high supported (<75 % BS): Pentatomidae (BS = 56.4), Lygaeidae (BS = 30.2), Reduviidae (BS = 21.5), Coreidae (BS = 5.6), Cicadellidae (BS = 64.4), Dictyopharidae (BS = 53.4), Eurybrachidae (BS = 99.9), Aphrophoridae (BS = 99.5), Tingidae, and Pyrrhocoridae (BS = 67.9). There is evidence of paraphyly in Fulgoridae and Cixiidae and polyphyly in Lophopidae. Overall it is observed that majority of the families in the COI BINs dataset are monophyletic suggesting that the COI marker can resolve Hemiptera families relatively well. Thirty-six of the branches present within each family had well-supported bootstrap values.

The ML tree was constructed to determine if there were any differences in the taxonomic coverage of the different Hemiptera families among the reforested sites (2010, 2012, and 2015) and reference sites (forest and grassland). Fulgoridae, Cixiidae, Dictyopharidae, Aphrophoridae, and Tingidae are underrepresented throughout the ML tree and conclusions as to whether they are a good biological indicator species are unclear. Further sampling of these families need to be done to test their habitat specificity. Pentatomidae, Lygaeidae, Cicadellidae, and Pyrrhocoridae are not good indicator species as these families are well represented throughout the different reforested and reference sites. This indicates that these families are not habitat specific and cannot be used as biological indicator species for reforestation. In contrast, Reduviidae, Coreidae, Lophopidae, and Eurybrachidae are good biological indicator species as they are not well represented throughout the different reforested and reference sites. These families are either absent (Reduviidae and Lophopidae) or underrepresented (Coreidae and Eurybrachidae) within the 2015 reforested site which is predominately covered in sugarcane. These families are habitat specific to either grassland or forest vegetation types making them good biological indicators for this study. Coreidae and Eurybrachidae are forest-specific families as they are predominately found within the 2012 reforested site, 2010 reforested site and the forest reference site, making them good biological indicators for assessing the progress of reforestation.

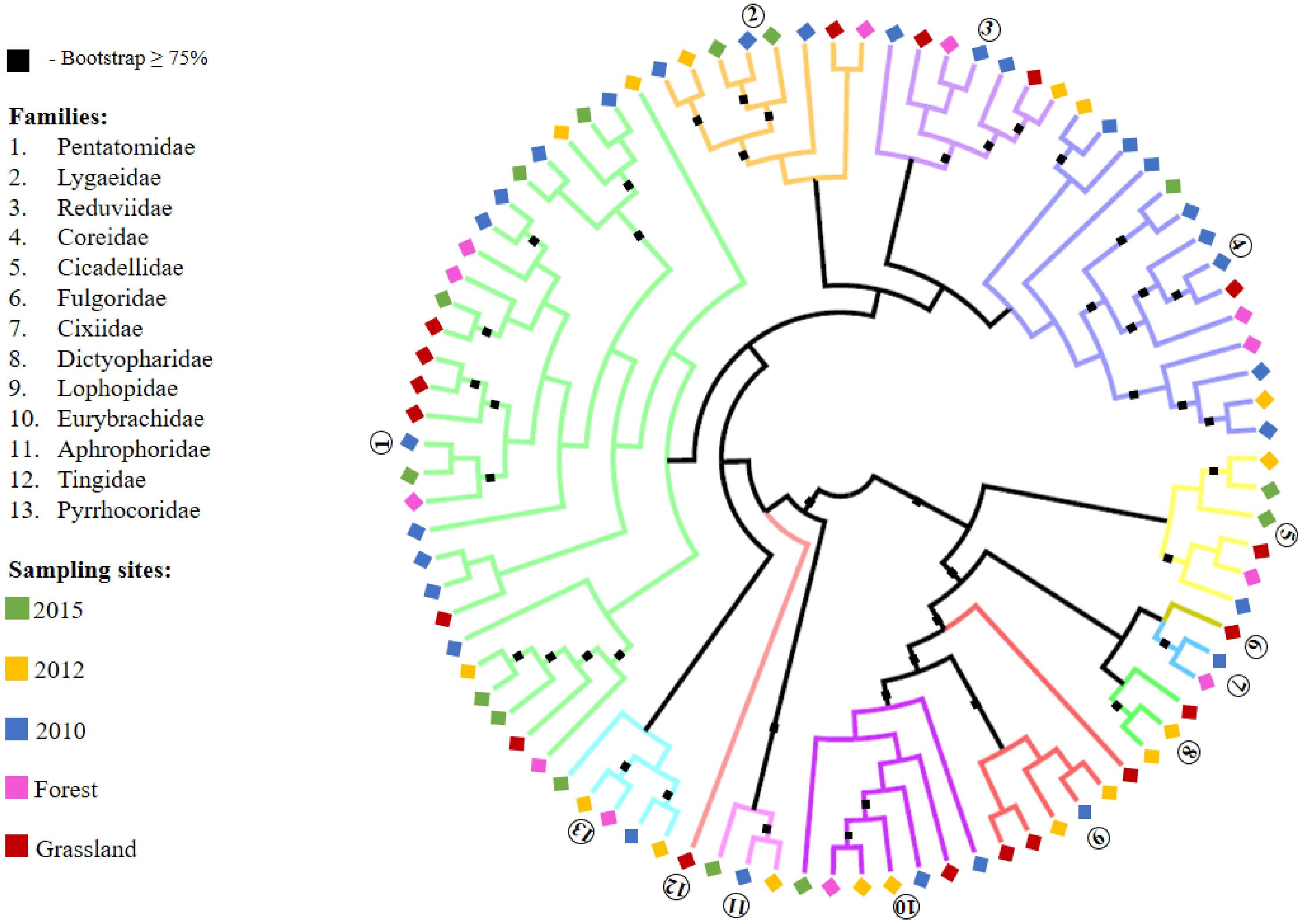

Figure 3.6: Most likely tree of 117 representative (BINs) Hemiptera specimens using their COI sequence data. The branches of the tree are colour coded by family. Only significantly supported branch support values (bootstrap value > 75%) are shown. Sampling sites are colour coded by the diamond shapes.

3.4 Discussion

Ecological habitat restoration is a process whereby a degraded ecosystem is altered and revived to a state in which it becomes resilient to environmental change and disturbances (Clewell & Aronson 2007). The aim of this study was to determine if Hemiptera could be a good indicator taxon for tracking the progress of restoration. What makes this study unique is that DNA barcoding was used for reliable species identification of cryptic species and species at different developmental stages. It was also useful to determine what families of Hemiptera are useful biological indicator species.

3.4.1 How does species richness and diversity of Hemiptera respond during ecological habitat restoration?

In this study, species richness, abundance, and diversity was linked to ages of the reforested patches (2010, 2012, and 2015), vegetation type (forest and grassland) and moisture content (wet versus dry). This supports the findings of previous studies using other invertebrate taxa such as Coleoptera (Parmenter & Macmahon 1987) and Orthoptera (Parmenter *et al.* 1991), which have shown an increase in species richness and diversity during ecological reforestation. The 2015 reforested site is seen to have a lower number of species richness, diversity and abundance as compared to the 2012 and 2010 reforested sites. This notion suggests that the plants that arthropods are attracted to are grassier, and are more densely populated vegetation types which are shifting towards forest plant cover and away from the sugarcane vegetation type. On a whole, the results which indicated a shift in Hemiptera species assemblage from the sugarcane vegetation type to the forest vegetation type are consistent with those of Moir et al. (2005). Moir et al. (2005) found similar Hemiptera species richness values between the restored sites and the native forest. This further supports the data in this study which suggests that Hemiptera species richness, abundance and diversity is affected by the different vegetation types.

3.4.2 How does species composition of Hemiptera change during ecological habitat restoration?

Over time, as the reforested habitat matures, there is a distinctive change in the species composition. The results obtained from this study are consistent with other studies that have selected Hemiptera as useful biological indicators to track restoration success and changes (Moir *et al.* 2005; Orabi *et al.* 2010). Despite examining the reforestation project within the Buffelsdraai Landfill Site' buffer zone over such a short time, the ordination, UPGMA clustering and phylogenetic analyses highlight a significant change in the species composition as the reforested sites shift from a predominantly sugarcane vegetation type towards the forest vegetation type. There is a clear clustering of the Hemiptera species composition and assembly between the 2012 and 2010 reforested sites and the forest reference sites. This highlights the success of the reforestation project currently being carried out within the Buffelsdraai Landfill Site Community Reforestation Project.

There is also evidence of clear clustering of the species composition and assemblage of Hemiptera in the different vegetation types (forest, grassland, and sugarcane), suggesting that the vegetation structure directly influences species composition. This correlates to findings from other studied which reveal that both plant species composition and vegetation structure can strongly effect Hemipteran species composition and assemblage (Sanderson *et al.* 1995; Hartley *et al.* 2003; Moir *et al.* 2005; Orabi *et al.* 2010).

3.4.3 Is the use of DNA barcoding and Hemiptera as suitable biological indicators to track the success of ecological restoration useful?

In this study, several Hemiptera species were specific to the different vegetation types (forest, grassland, and sugarcane). This provides evidence for the potential of this group to be biological indicators. Hemiptera are seen to be sensitive to changes in an ecosystem which makes this order an effective environmental and biological indicator. With the help of DNA barcoding, Reduviidae, Coreidae, Lophopidae, and Eurybrachidae were identified as good biological indicators for this study. These families were habitat specific to either grassland or forest enabling them to be good biological indicator species to understand and assess the progress of ecological reforestation. The results from this study remain consistent with other studies done on Hemiptera which highlight the use of this order as an effective biological indicator in ecological habitat restoration and their response to ecosystem changes and disturbances (Moir *et al.* 2005; Orabi *et al.* 2010). The use of Hemiptera in conjuncture with DNA barcoding is a cost-effective way to study changes in ecological habitat restoration, as DNA barcoding helps to rapidly identify large numbers of Hemiptera and overcome any taxonomic impediment. DNA barcoding in this study was proven to be an efficient and effective tool for the identification of cryptic species and Hemiptera at different developmental stages which were impossible to identify or to associate with the corresponding adult's forms. The results obtained from this identification tool allowed for meaningful comparisons to be made between the different reforested and reference sites within the Buffelsdraai landfill sites' buffer zone.

3.5 Conclusion

This study highlights the importance of multivariate analysis, biological indicator species-based approaches and molecular approaches to monitor the success of ecological habitat restoration. In this study, it is seen that Hemiptera can be used as a cost effective biological indicator as it has the potential to provide a high information content. The distinctive clustering of the species composition between the forest reference site and the 2012 and 2010 reforested sites suggest that the ecological reforestation project being carried out at Buffelsdraai Landfill Site Community Reforestation Project is successful.

3.6 References

Akaike H. (1974) A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716-723.

Anderson A., McCormack S., Helden A., Sheridan H., Kinsella A. & Purvis G. (2011) The potential of parasitoid Hymenoptera as bioindicators of arthropod diversity in agricultural grasslands. *Journal of Applied Ecology* **48**, 382-390.

Anderson M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46.

Bellinger E.G. & Sigee D.C. (2010) *Freshwater Algae: Identification and Use as Bioindicators.* United Kingdom: John Wiley & Sons.

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. *Systematic Biology* **61**, 851-869.

Bray J.R. & Curtis J.T. (1957) Àn ordination of upland forest communities of southern Wisconsin. *Ècological Monographs* **27**, 325-349.

Cabin R.J., Clewell A., Ingram M., McDonald T. & Temperton V. (2010) Bridging restoration science and practice: Results and analysis of a survey from the 2009 Society for Ecological Restoration International Meeting. *Restoration Ecology* **18**, 783-788.

Cassis G., Gross G.F. & Study A.B.R. (2002) *Hemiptera*. Australia: CSIRO Publishing.

Clarke K.R. & Warwick R.M. (1994) *Change in Marine Communities: An Approach to Statistical Analysis and Interpretation*. United Kingdom: Plymouth marine laboratory, Natural environment research council.

Clewell A. & Aronson J. (2007) *Ecological Restoration: Principles, Values, and Structure of an Emerging Profession.* Washington, DC: Island Press.

Darriba D., Taboada G., Doallo R. & Posada D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772.

Diederichs N. & Roberts D. (2016) Climate protection in mega-event greening: the 2010 FIFA™ World Cup and COP17/CMP7 experiences in Durban, South Africa. *Climate and Development* **8**, 376-384.

Douwes E., Rouget M., Diederichs N., O'Donoghue S., Roy K.E. & Roberts D. (2015a) Buffelsdraai Landfill Site Community Reforestation Project. In: *XIV World Forestry Congress. FAO (Food and Agriculture Organization of the United Nations)*. Durban, South Africa.

Douwes E., Roy K.E., Diederichs N., Mavundla K. & Roberts D. (2015b) *The Buffelsdraai Landfill Site Community Reforestation Project: Leading the way in community ecosystem-based adaptation to climate change.* South Africa: eThekwini Municipality.

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

Fine P.V.A. & Kembel S.W. (2011) Phylogenetic community structure and phylogenetic turnover across space and edaphic gradients in western Amazonian tree communities. *Ecography* **34**, 552-565.

Fisher R., A C. & Williams C. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**, 42-58.

Folmer O., Black M., Hoeh W., Lutz R. & Vrijenhoek R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**, 294-299.

Galimberti A., Spinelli S., Bruno A., Mezzasalma V., De Mattia F., Cortis P. & Labra M. (2016) Evaluating the efficacy of restoration plantings through DNA barcoding of frugivorous bird diets. *Conservation Biology* **30**, 763-773.

Hall T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98.

Hammer R., Harper D. & Ryan P. (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4**, 9.

Harmon-Threatt A.N. & Ackerly D.D. (2013) Filtering across spatial scales: Phylogeny, biogeography and community structure in bumble bees. *PLoS One* **8**, e60446.

Hartley S.E., Gardner S.M. & Mitchell R.J. (2003) Indirect effects of grazing and nutrient addition on the hemipteran community of heather moorlands. *Journal of Applied Ecology* **40**, 793-803.

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S. & Francis C.M. (2004) Identification of birds through DNA barcodes. *PLoS Biology* **2**, e312.

Kaufman E.E., Scott G.A., Nielsen N.I., Schiller C.M. & Blair R.E. (1929) *Callifornia Crop Report*. California: California State Print.

Kress W.J. & Erickson D.L. (2008) DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences* **105**, 2761-2762.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. & Higgins D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23**, 2947-2948.

Margalef R. (1958) Information theory in ecology. *General Systems* **3**, 36-71.

Maw H.E.L., Foottit R.G., Hamilton K.G.A. & Scudder G.G.E. (2000) *Checklist of the Hemiptera of Canada and Alaska*. Canada: NRC Press.

Moir M.L., Brennan K.E.C., Koch J.M., Majer J.D. & Fletcher M.J. (2005) Restoration of a forest ecosystem: The effects of vegetation and dispersal capabilities on the reassembly of plant-dwelling arthropods. *Forest Ecology and Management* **217**, 294-306.

Orabi G., Moir M.L. & Majer J.D. (2010) Assessing the success of mine restoration using Hemiptera as indicators. *Australian Journal of Zoology* **58**, 243-249.

Pander J. & Geist J. (2013) Ecological indicators for stream restoration success. *Ecological Indicators* **30**, 106-118.

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.

Parmenter R.R. & Macmahon J.A. (1987) Early successional patterns of arthropod recolonization on reclaimed strip mines in Southwestern Wyoming: The ground-dwelling beetle fauna (Coleoptera). *Environmental Entomology* **16**, 168-177.

Parmenter R.R., Macmahon J.A. & Gilbert C.A.B. (1991) Early successional patterns of arthropod recolonization on reclaimed Wyoming Strip Mines: The grasshoppers (Orthoptera: Acrididae) and allied faunas (Orthoptera: Gryllacrididae, Tettigoniidae). *Environmental Entomology* **20**, 135-142.

Prestidge R.A. & McNeill S. (1983) Auchenorrhyncha-host plant interactions: leafhoppers and grasses. *Ecological Entomology* **8**, 331-339.

Rambaut A. (2009) FigTree v.1.3.1 Computer program and documentation. Retrieved from: http:tree.bio.ed.ac.ukz/software. Date accessed: 26 May 2016

Raupach M.J., Hendrich L., Küchler S.M., Deister F., Morinière J. & 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**, e106940.

Rubinoff D., Cameron S. & Will K. (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity* **97**, 581-594.

Ruiz-Jaen M.C. & Mitchell Aide T. (2005) Restoration success: How is it being measured? *Restoration Ecology* **13**, 569-577.

Sanderson R.A., Rushton S.P., Cherrill A.J. & Byrne J.P. (1995) Soil, vegetation and space: An analysis of their effects on the invertebrate communities of a Moorland in North-East England. *Journal of Applied Ecology* **32**, 506-518.

Schuh R.T. & Slater J.A. (1995) *True bugs of the world (Hemiptera: Heteroptera): classification and natural history*. New York: Cornell university press.

Shannon C. & Weaver W. (1963) *The Mathematical Theory of Communication*. USA: The University of Illinois Press.

Shen Y.Y., Chen X. & Murphy R.W. (2013) Assessing DNA barcoding as a tool for species identification and data quality control. *PLoS One* **8**, e57125.

Simpson E. (1949) Measurement of diversity. *Nature* **163**, 688.

Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313.

Tamura K., Dudley J., Nei M. & Kumar S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.

Valentini A., Pompanon F. & Taberlet P. (2008) DNA barcoding for ecologists. *Trends in Ecology & Evolution* **24**, 110-117.

Verdú M., Gómez-Aparicio L. & Valiente-Banuet A. (2012) Phylogenetic relatedness as a tool in restoration ecology: a meta-analysis. *Proceedings of the Royal Society B: Biological Sciences* **279**, 1761-1767.

Wallington T.J., Hobbs R.J. & Moore S.A. (2005) Implications of current ecological thinking for biodiversity conservation: A review of the Salient issues. *Ecology and Society* **10**.

Wheeler A.G. (2001) *Biology of the Plant Bugs (Hemiptera: Miridae): Pests, Predators, Opportunists*. New York: Cornell University Press.

Willows-Munro S. & Schoeman M.C. (2015) Influence of killing method on Lepidoptera DNA barcode recovery. *Molecular Ecology Resources* **15**, 613-618.

Wortley L., Hero J.M. & Howes M. (2013) Evaluating ecological restoration success: A review of the literature. *Restoration Ecology* **21**, 537-543.

Chapter Four

General discussion

When DNA barcoding was first proposed 13 years ago, as a means of identifying species based on DNA sequences, very few biologists anticipated the impact DNA barcoding would have on the field of ecology, evolution, conservation, and biogeography. During my MSc study, I explored the utility of DNA barcoding and provided a way of testing the accuracy of using a single mitochondrial COI gene as a standardized marker to identify Hemiptera. In addition, I examined the Buffelsdraai Landfill Site Community Reforestation Project as a case study to further understand the utility of DNA barcoding in restoration ecology and the use of Hemiptera as a biological indicator species.

In **chapter one**, I highlighted the negative impacts of urbanization on biodiversity and the importance of species identification to create biodiversity inventories. I provided an overall review of DNA barcoding and its usefulness in assisting with species identification. DNA barcoding is used all around the world, however, most of the studies which involve this technique focus on taxa found in Europe and America, with very few studies focusing on the African continent. DNA barcoding can be used as a valuable tool for African systematics due to the current taxonomic impediment and lack of taxonomic knowledge on Africa taxa. This study focused on South African Hemiptera, more specifically Hemiptera collected from the eThekwini region and surrounding areas.

In **chapter two**, I demonstrated the utility of DNA barcoding and the mitochondrial COI gene to accurately identify Hemiptera species. A preliminary DNA barcode reference library was created for the Hemiptera collected from the eThekwini region and surrounding areas. This reference library currently consists of 1456 specimens from 357 putative species and 256 morphospecies, sampled from 18 geographic localities within and around the eThekwini region. Using molecular data and statistically supported data, high species richness and diversity were recorded in the University of KwaZulu-Natal (Pietermaritzburg campus), Drummond, Springside Nature Reserve, Iphithi Nature Reserve, and Palmiet Nature Reserve. In contrast, the lowest species richness and diversity was recorded in Bartlett Estate, High Meadows, Vernon Crookes, Hazelmere Dam, and Paradise Valley. The preliminary reference library was also used to compare the matches between morphospecies and BINs. There is strong coherence seen between the number of BINs and morphospecies recorded at most localities, suggesting that the DNA barcode data can be used as a reliable tool for rapidly identifying species. However, there was a mismatch of BINs and morphospecies recorded at some localities, these mismatches are due to either cryptic speciation or the presence of Hemiptera at different developmental stages. This suggests that these two methods should be used as complementary tools to one another to achieve more accurate results. In this chapter, I also tested for the presence of the DNA barcode gap, which was computed by the K2P and GTR+I+G nucleotide substitution models of evolution on datasets containing COI sequences sorted into morphospecies and genetic clusters (BINs). A degree of overlap was seen between the intraspecific and interspecific classes on both data sets using the different models of evolution. However, the use of the Jeffries-Matusita (J-M) distance indicated that there was a statistically significant barcode gap. This suggests that DNA barcoding is effective in Hemiptera species identification. Through phylogenetic analysis, the mitochondrial COI gene was observed to resolve above the species level providing support for Hemiptera families and genera, which further highlights the effectiveness of this marker when working with Hemiptera.

In **chapter three** I turned my attention to the Buffelsdraai Landfill Site Community Reforestation Project. The aim of this study was to determine whether Hemiptera can be used as biological indicators to track the progress of restoration. This was done by comparing species diversity, richness, and composition of Hemiptera species collected from sites which were reforested at distinct phases in the years 2015, 2012 and 2010. The Hemiptera from these sites were compared to native grassland and forest sites. DNA barcoding was used to identify Hemiptera species. A total of 393 specimens were collected from 20 different sites within the Buffelsdraai region. Once the 393 specimens were sorted into morphospecies, a maximum of 5 individuals were selected for DNA analysis. The DNA barcode reference library for this case study currently consists of 132 specimens from 119 putative species and 117 morphospecies. The morphospecies data and genetic data generated was used to assess the Hemiptera species richness and diversity across the different reforested sites (2015, 2012 and 2010) and the reference sites (grassland and forest) within the Buffelsdraai landfill sites' buffer zone. The highest species richness and diversity was recorded in the 2010 reforested site followed by the 2012 and 2015 reforested sites. When comparing the reforested sites to the reference sites, the 2010 reforested sites had a higher species richness and diversity then both the grassland and forest reference site. This was due to three confounding factors, however, one of these factors was due to the presence of introduced Hemiptera species in the 2010 reforested site. These introduced species included *Clavigralla horrida* (family Coreidae), *Zicrona caerulea* (family Pentatomidae), and *Spilostethus pandurus* (family Lygaeidae), which are native to Costa Rica and are known to be invasive species all around the world (Kaufman *et al.* 1929). This study assessed reforestation over a period of 6 years (2010 – 2015). Despite this short period of time, the results obtained from the ordination, UPGMA clustering and phylogenetic analyses displayed a significant clustering of species composition of the 2012 and 2010 reforested sites with the forest reference site. This is due to the shift from the sugarcane vegetation type to a forest vegetation type within the reforested plots. This highlights the success of the reforestation project being carried out within the Buffelsdraai Landfill Site Community Reforestation Project. This data also suggests that Hemiptera are reliable biological indicators to track the success of ecological restoration, as they are seen to be sensitive to environmental changes and disturbances. With the help of DNA barcoding four Hemiptera families namely; Reduviidae, Coreidae, Lophopidae, and Eurybrachidae were identified as good biological indicators for this study. This case study highlighted the importance of multivariate analysis, biological indicator species based approaches and molecular techniques to monitor the success of ecological restoration.

Overall, this study supports DNA barcoding as an exceptional tool and some of its advantages have been mentioned throughout this thesis. DNA barcoding can be used to perform large-scale screening of biodiversity and allow for a detailed investigation at different spatial scales. Furthermore, when DNA barcoding is coupled with traditional morphology-based taxonomy it has the potential to provide insight into biodiversity and taxonomic studies. The techniques and methods of analyses used in this study could be applied to building DNA barcode libraries for other organisms, not only in the eThekwini region but for the rest of South Africa. Thus, contributing to the barcode of life initiative to catalogue all of Earth's biodiversity. Only history will tell if DNA barcoding succeeds in advancing and improving research on biodiversity and in fostering close collaborations between barcoders and taxonomists to help overcome the current taxonomic impediment.

4.2 References

Kaufman E.E., Scott G.A., Nielsen N.I., Schiller C.M. & Blair R.E. (1929) *California Crop Report*. California: California State Print.