BIOLOGICAL ASSESSMENT AND MONITORING OF SINGAPORE AQUATIC ENVIRONMENTS USING NGS TOOLS

BILGENUR BALOGLU

B. Sc. (Hons.), Istanbul Technical University

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2017

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

_____

Bilgenur Baloglu

24 AUGUST 2017

ACKNOWLEDGEMENTS

I would like to thank my Ph.D. advisor Prof. Rudolf Meier for supporting me during these past few years, for his motivation, and immense knowledge. You were instrumental in helping me crank out this thesis, all in one month. Thank you for your brilliant comments on several of my drafts. You really helped me get things into perspective. Your mentorship these past four years has been invaluable.

I also thank the members of my thesis advisory committee, Asst. Prof. Darren Yeo and Asst. Prof. Frank Rheindt for their very helpful feedback and encouragement. Your questions helped widen my research from various perspectives.

My sincere thanks also go to Asst. Prof. Roman Carrasco, for teaching the most useful lecture I took in NUS and for the stimulating discussions regarding R and the related analysis. You are the reason why I love R so much!

Thanks also to Prof. Peter Cranston for sharing his boundless chironomid wisdom with me.

Thank you NUS for funding my studies through SINGA scholarship, PUB for funding the chironomid project, members of TMSI for collecting all my chironomid samples, providing the environmental data, and for teaching me 100 ways to be patient.

Amrita Srivathsan, thank you for answering my bioinformatics related questions, your amazing time-saving scripts, assistance in editing of my manuscript, and the comforting conversations we had whenever I needed. You have been a very supportive colleague and a friend. Gowri Rajaratnam for checking up on me all those times and trying to make me feel better about PhD, physically and mentally. You've always been a supportive and caring friend, and your presence helped make the lab bearable.

Wing Hing Wong for your immense support ever since I first started in the lab and for all the hugs. You have been a big part of EvoLab. Yuchen Ang for teaching me Photoshop and how to do the morphological sorting. Darren for bearing with my random questions and for being such a great friend. Your absence in the student office and the lab is really felt. Maosheng for helping me collect my chironomids and letting me access to the museum collections at all the times I asked for. Jonathan for all of our awesome fish and midge related conversations and your suggestions on the data analysis. Honor, Wan Ting, Jake, Molly, you were the best midge team I could ever ask for! Thank you for all the PCRs, agarose gels, countless plates and making the lab a lot more fun. Theo, thank you so much for making the primer plates, all the lunch invitations and the takeaway food you brought me many times. I survived and felt cared thanks to you. Other members of EvoLab: Ywee Chieh for helping me get used to the lab, nice conversations and many lunches, and Sujatha for helping me edit my manuscript and your caring questions whenever we came across. Diego, Mindy, Kathy, Jinfa, Wan, you all made the lab such a great place.

A number of friends, my former flatmates Maedeh and Elham, my first two years in Singapore were so amazing thanks to you. Francesca for constantly inviting me for coffee breaks and the house parties, and listening to me during my several cycles of depression. You've been my support system in all this madness. Graham for the cover letter samples and our great conversations. Serap for being the amazing person you are, our Wednesday walks, and your constant support in all these years. Jonathan for being such a great listener and treating me all those dinners. Ahmet for helping me see the bright side of the things. Nesibe, Erhan, Mehmet, and Aysu for helping me keep physically active. Doğan Can for being my climbing partner on Fridays. Duygu, Betül, Begüm, Canan, Tuçe, and Marc, thanks for putting up with me despite the

strange level of communication in the last few years. Yıldıray Lise and Mustafa, thank you for all the postcards.

My forever encouraging and always enthusiastic grandma, thank you for always putting your trust in me and being there since I was a little kid. Special thanks go to my mom, for being the biggest support in my life since day one, for teaching me how to be strong whatever bad happens, for all your love. My late father, an agricultural engineer, even though you left us so early, you and the genetic traits I inherited from you may be why I am a biologist today. Thank you. My little sister Bestenur, thank you for always being there for me. My little brother Eşref, thank you for our humorous conversations. Seda teyze for the food packages you sent and the laptop you gave me without which I could not write my thesis. Thank you for being the perfect aunt since 1990. Ümit abi, for your invaluable support and our comforting phone calls. My dear cousin Bahar, for all the insect questions you have been asking me, and wanting to study insects. Gizem abla, you may be the only one in the family who understands me truly – and scientifically for that matter. Thank you. Meryem teyze, for your moral support whenever I needed and for all the food.

Zhewei for sticking with me all these years, for the letters, chocolates, motivational speeches, ideas. For making me remember whenever I am depressed that life is worth living, that there are many mountains to climb, and places to see. Thank you.

TABLE OF CONTENTS

SUMMARY

Non-biting midges (Chironomidae: Diptera) are an important component of freshwater ecosystems. However, most freshwater quality assessment or conservation biology studies rarely incorporate species-level information on midges. This is because traditional methods for sorting and identifying midges are too expensive. Here, I optimize, test, and use a new DNA barcoding technique that is based on Next Generation Sequencing (NGS). I use "NGS barcodes" for >30,000 individual specimens to demonstrate how NGS barcodes can improve analyzing the community structure of specimen- and species-rich invertebrate taxa. I first demonstrate that the midge fauna of a reservoir can be characterized by barcoding 500-1000 specimens (Chapter 2). I recommend that biomonitoring programs could cheaply gather data with only a small number of NGS-barcoded specimens or metabarcoded bulk samples. Next, I show how a new sequencing technique (MinION™) can be used for obtaining NGS barcodes within 24 hours (Chapter 3). I estimate that a single run of MinION™ can generate >100 barcodes and conclude that an estimate of species composition can be obtained 10 hours since sample handling. Lastly, I reveal that Singapore's biggest swamp forest remnant (Nee Soon Swamp Forest) maintains a rich and largely unique fauna (>400 chironomid species) that is resilient against the invasion of species from surrounding artificial reservoirs (Chapter 4). I show that the

chironomid occurrence in the swamp forest is driven by several physicochemical variables rather than the presence of or distance to the reservoirs. These findings suggest that even small or fragmented swamp forests can be suitable habitats for chironomids. This has an important conservation implication for many other swamp forests in Southeast Asia that are under threat. Overall, these studies expose the enormous power of NGS barcoding in ecological research, to study ecosystem health, biological diversity, and habitat conservation.

List of Figures

List of Tables

CHAPTER 1

General Introduction

When limited to strictly traditional methods, science is almost certainly guaranteed to fall far behind (Smith & Fisher, 2009).

1.1 A new voyage of biodiversity discovery

Imagine a future where any plant, animal, or microbe can be grouped into species and identified with relative ease. Instead of spending a lot of time on identification tasks and filing away specimens, ecologists and conservation biologists could then focus on biodiversity, morphology, and species interactions and gain a thorough understanding of species' role in the ecosystem. However, we are far from this vision. Instead, biologists struggle and spend much time on species-level sorting and identification. This is due to the fact that the earth is home to an estimated 2 million to 100 million known species (Vié *et al.* 2009). Around 2 million have been described so far and since 2006, ca. 18,000 species are described each year (Costello *et al.* 2013; Wheeler & Pennak, 2011). However, the discovery of new species and assessment of endangered ecosystems is difficult because the use of existing taxonomic tools is labor-intensive and slow and there is a shortage of resources and experts in taxonomy (Drew, 2011). In my dissertation, I explore Next Generation Sequencing (NGS)

barcoding tools for overcoming some of these challenges related to species identification and habitat assessment.

Throughout human history, biodiversity has been a source of inspiration and wonderment for humankind (Tilman, 2000). Our knowledge of the diversity of life is as old as humanity. Our hunter-gatherer ancestors explored the world and recorded what can be eaten, what is poisonous, what animals can be hunted or what animals pose a danger (Research Matters, 2017). In modern societies, other ecosystem services, such as clean air, fresh water, and shelter are more important, which are provided by biodiversity (Purvis & Hector, 2000; Tilman, 2000; Mace *et al.* 2012; Naeem *et al.* 2012; Sandifer *et al.* 2015; but also see Ridder, 2008). Ecosystem services provided by wild animals and plants are essential for our sustenance (Millennium Ecosystem Assessment, 2005; CBD UNEP, 2010). Biodiverse and green habitats also improve our psychological well-being (Kaplan, 2001; Fuller *et al.* 2007; Maller *et al.* 2009; Nisbet *et al.* 2011).

However, natural ecosystems are increasingly threatened by the explosive growth of human populations. Biodiversity loss at multiple levels ranging from species, over phylogenetic, and genetic, to functional diversity occurs at a rate much higher than outside of geologic mass extinction events (Smith & Fisher, 2009; Pereira *et al.* 2010; Naeem *et al.* 2012). The full extent of the problem is poorly understood, however, because the vast majority of species are unknown to humankind and the extinction risks for only a small proportion of the described species have been assessed. For example, as of 2008, only about 2.7% of the 1.8 million described species have been assessed by IUCN (Vié *et al.* 2009). Based on these limited data, a global multi-taxon meta-analysis suggests that the mean observed extinction risk by 2100 is 12.6% in

plants, 9.4% in invertebrates and 17.7% in vertebrates (Maclean & Wilson, 2011; Bellard *et al.* 2012).

Among natural ecosystems, islands and freshwater environments are particularly threatened. Freshwater ecosystems occupy less than 1% of the earth's surface (Dudgeon *et al.* 2006), yet support close to 12% (~126,000 species) of all described species on earth (Balian *et al.* 2008; Garcia-Moreno *et al.* 2014). Despite limited records, scientists estimate that at least 8-16% of all freshwater species on the planet have become extinct within the last century or are currently endangered (Strayer, 2006; Strayer & Dudgeon, 2010). One of the most imperiled freshwater habitats is swamp forests. In Southeast Asia, at least 45% of the original 27 million ha of peat swamp forests have been logged, and almost as much has been drained (Hooijer *et al.* 2006; Yule, 2010). In chapter 4, I study the chironomid midge fauna of the last remnant of freshwater swamp forest of Singapore and test whether it is resilient against faunal invasion from neighboring reservoirs.

1.2 Biodiversity assessment and biomonitoring

Given that urbanization and industrialization are destroying natural freshwater environments, could humans find a way to coexist with the natural world? How can we protect these natural ecosystems? How can we monitor their health? Water quality can be assessed using physical, chemical and biological characteristics. Physical and chemical analyses provide a snapshot of the water quality, i.e., they only reflect the quality a the time the sample was collected. This is not useful for a holistic assessment because the chemical and physical properties of water fluctuate with meteorological cycles and many relevant chemicals cannot be detected by conventional methods. On

the other hand, biological methods can assess the effects of the physicochemical variables on the organisms over longer periods and indirectly cover more parameters that are relevant to the quality of the environment (Bartram & Ballance, 1996). In chapter 2, I use and test "NGS barcoding" for obtaining species-level data for chironomid midges that I then use for bioassessment and biomonitoring.

The use of living organisms to assess changes in the environment, i.e., biomonitoring, is not new. In the 19th century, mine workers used canaries to detect noxious gas leaks in coal mines (Cairns & Pratt, 1993). In the 20th century, biomonitoring matured into a systematic approach based on collecting living specimens, performing taxonomic identification, and using inventories to assess the ecological health of a given site (Keck *et al.* 2017). Early use of stream and lake benthic organisms focused on detection of organic pollution. However, as anthropogenic impacts increased, the studies shifted towards monitoring the change in species richness and abundance (Johnson *et al.* 1993). Today, biomonitoring studies also start to employ the evaluation of environmental DNA (i.e., DNA molecules released into the environment from the skin, mucous, saliva, eggs, feces, urine, root, leaves, fruit, pollen, and rotting bodies, Taberlet *et al.* 2012). eDNA shows much promise for biomonitoring of freshwater environments. A good example is our study (Lim *et al.* 2016) of eDNA for two Singaporean reservoirs (Bedok and Pandan) which revealed evidence for hundreds of species (>500 animal signatures in a 1.2-pint glass of water). However, most of these genetic signals could not be identified to species. An exception was chironomid midges because my Ph.D. work had yielded identified NGS barcodes for the species in the reservoirs.

Biomonitoring programs are now routinely used in many countries such as the United States (Paulsen & Linthurst, 1994), the members of the European Union (Kallis & Butler, 2001), and in developing countries in Africa, Asia, Eastern Europe, and Latin America (Resh, 2007). These programs aim to understand the impact of environmental changes on species composition in a particular ecosystem over space and time (Hajibabaei *et al.* 2011), by using biological indicators (Fausch *et al.* 1990). A biological indicator is an organism which is associated with specific environmental signals such that its presence indicates the existence of those signals (Patton, 1987).

Indicator organisms are selected based on a number of criteria. First, the organisms need to be easy to sample. Second, they need to reproduce quickly, so that their response to changes in the environment is rapid. Third, indicator organisms should be relatively immobile such that their abundances reflect the health of the local ecosystem. Finally, they need to be easy to identify, i.e., rapidly, at low cost, and ideally with high taxonomic resolution (Hilty & Merenlender, 2000). Most vertebrate species (fish, Marshall *et al.* 1987; Hourigan *et al.* 1988; frogs, Hecnar & M'Closkey, 1996; birds, Bharucha & Gogte, 1990; Davis, 1989; foxes, Davis, 1989; bears, wolfs, and goats, Kiester & Eckhardt, 1994) are not particularly suitable for biomonitoring (Hilty & Merenlender, 2000), because species with large bodies are usually found in low densities (Blueweiss *et al.* 1978) and are more susceptible to local extinction (Shaffer, 1981). This is one of the reasons why invertebrates are more widely used in biomonitoring of freshwater habitats.

1.3 Invertebrates and their use in biomonitoring research

Benthic macroinvertebrates are widely used because they live in a variety of habitats, are abundant and diverse, are relatively immobile, and are responsive to environmental stresses (Nazarova *et al.* 2004; Morse *et al.* 2007; Roque *et al.* 2009; Morais *et al.* 2010; Fu *et al.* 2012). They also play critical roles in the ecosystem by acting as important food sources for higher order predators, such as fish and birds, and by contributing to the decomposition of organic matter (Margalef, 1983; Real *et al.* 2000; Murakami & Nakano, 2002). Despite their importance, benthic communities of tropical freshwater systems are rarely used for biomonitoring because the fauna is largely unknown (Lucca *et al.* 2010). Only a few invertebrate groups are well studied. This includes dragonflies (Odonata), caddisflies (Trichoptera), and mayflies (Ephemeroptera) (Clements *et al.* 2002; Ball *et al.* 2005; Rainbow *et al.* 2012) while many specimen- and species-rich taxa are either widely ignored or the specimens are not identified to species. This is due to the high cost and expertise required for obtaining species identifications (Hilty & Merenlender, 2000).

1.4 Invertebrates and their use in conservation research

Compared to mammals and birds, invertebrates are often neglected in conservation biology. Conservation projects disproportionately focus on saving populations of large vertebrates that are often on the brink of extinction but gather little information on the extinction threats to organisms that underpin much of the food chain. Such projects rely on the notion that ‘the whole ecosystem will be saved if indicator species are saved’ (Graul *et al.* 1976). This approach can lead to ill-informed conservation campaigns. Moreover, such a vertebrate-centric species approach

overemphasizes the protection of rare species, although such species are often of minor importance for overall ecosystem health. For example, many resources are spent on artificial breeding programs and human intervention (Hutto *et al.* 1987; Simberloff, 1998).

Instead of a species-centric approach, evaluating biological communities that contribute to ecosystem processes can provide a clearer picture of ecosystem health. Invertebrates act as primary consumers (i.e., the link between plants and the higher animals) and their relative abundances and distribution can reflect the environmental stresses acting on all animals above the food chain. Conservation campaigns should account for the contribution of the selected biological communities to ecosystem processes (such as energy exchange, nutrient cycling, herbivory, etc.) rather than emphasizing the protection of individual populations (i.e., community-centered vs. species-centered conservation). Maintaining stable ecosystem functions increases the chances of protecting endangered individuals as well as the many species yet to be described (Walker, 1992). In chapter 4 of the thesis, I use chironomid communities for the conservation of a swamp forest remnant.

1.5 Chironomidae (Diptera) as indicator taxon in bioassessment

In the thesis, I develop, test, and use NGS tools for quantifying non-biting midge (Diptera: Chironomidae) communities. I prepared NGS barcodes for >30,000 chironomid specimens in an attempt to develop them into a species-level model invertebrate taxon for biomonitoring purposes. Chironomids are a common group of freshwater macroinvertebrates found in virtually all aquatic environments. Combined, they sometimes have more biomass and are more species-rich than all other common

macroinvertebrate groups (Marziali *et al.* 2010; Nicacio & Juen, 2015). In addition, many species have specific habitat requirements which allow them to serve as biological indicators.

The main challenge in studying chironomids is species delimitation (i.e., inferring the boundaries and numbers of species) and identification. Delimiting/grouping species of chironomid midges based on morphology is laborious, time-consuming and costly (Meier *et al.* 2006; Pfenninger *et al.* 2007; Friberg *et al.* 2011; Carew *et al.* 2013). In particular, cryptic species or immature life stages are notoriously difficult to classify (Thomsen & Willerslev, 2015). Moreover, taxonomic expertise is becoming increasingly scarce as specialists age and retire.

The difficulties and high-cost of morphology-based species delimitation are some of the reasons why the low taxonomic resolution is used in many biomonitoring studies. This is undesirable because it has been shown chironomids, that congeneric species in *Cricotopus*, *Polypedilum*, and *Tanytarsus* differ considerably with regard to their tolerance to heavy metals, pesticides, and nutrient-levels (Cranston, 2000; Riva-Murray *et al.* 2002). This means that if one were to use genus- or family-level identification incorrect or imprecise conclusions are drawn with regard to ecosystem health (Lenat & Resh, 2001; Metzeling *et al.* 2002; Greffard *et al.* 2011). I here use NGS barcodes which overcome the cost problem of species delimitation because it allows for rapid species-level sorting in the bioassessment project. During my Ph.D., I received hundreds of chironomid samples from the national water agency (PUB) of Singapore and often only had 2-3 days to sequence them. Yet, I was often able to match them to species because the sequences could be matched to local barcode

database for chironomid species. However, NGS barcodes often only delimit species. They rarely allow for identifying sequences/specimens to species (i.e., assign scientific names) given that all existing barcode databases are very incomplete. This is particularly so for tropical invertebrates.

1.6 The need for high-throughput (cheaper) processing of specimens

The use of molecular markers for species delimitation is not new (Kurtzman, 1994; Wilson, 1995; Avise, 2012). The term "DNA barcoding" was first coined in 2003 when a 658 bp fragment of Cytochrome Oxidase Subunit I (COI), a mitochondrial protein-coding gene, was used for identifying species based on DNA sequences (Hebert *et al.* 2003). Mitochondrial genes were considered desirable because they lack introns, were inherited maternally (Saccone *et al.* 1999), and easy to align due to the rarity of indel mutations. In addition, mitochondrial genes evolve 2–9 times faster than nuclear protein-coding genes for most metazoan animals (DeSalle *et al.* 1987; Monteiro & Pierce, 2001; Moriyama & Powell, 1997; Johnson *et al.* 2003), thus making them useful for tracing recent speciation events (Voigt *et al.* 2012). This is also the reason why I use mitochondrial DNA (mtDNA) throughout my thesis. My target gene was the metazoan barcoding gene COI.

Hebert *et al.* (2003) showed that COI is conserved within many species, yet usually variable between species in species pairs. This meant that many species could be identified based on COI. Hebert *et al.* (2003) also suggested the use of genetic distance as a standard method when analyzing barcode data. The implicit assumption was that the intraspecific (within species) divergences would be shorter than interspecific (between species) divergences (i.e., barcoding gap; Meyer & Paulay, 2005). Clustering thresholds proposed for COI sequences varied depending on the

organism studied. Several studies have used a 2 or 3% threshold (Hebert *et al.* 2003; Song *et al.* 2008; Strutzenberger *et al.* 2011; Ng'endo *et al.* 2013), while a range of 2-5% thresholds was shown to be stable in midges (Meier *et al.* 2015; Baloglu *et al.* unpublished).

One criticism of DNA barcoding is that the barcode for recently evolved species complexes and closely related taxa is often (nearly) identical (Will & Rubinoff, 2004; Meyer & Paulay, 2005; Meier *et al.* 2006). Indeed, little or no COI barcode divergence is expected for close relatives (Pentinsaari *et al.* 2017), as the divergences in COI are not a cause but a consequence of speciation (Kwong *et al.* 2012a); i.e., barcoding gaps evolve – often a long time after a speciation event. Another criticism of DNA barcoding is the lack of generally accepted clustering thresholds for grouping specimens by genetic distance (Meyer & Paulay, 2005; Stribling, 2006; Meier *et al.* 2006; 2008; Wiemers & Fiedler, 2007). However, there is no reason to expect that there is a universal barcoding threshold (Meier *et al.* 2006). Instead, thresholds are mainly guidelines for sorting specimen samples to species. My thesis will address the threshold problem by applying multiple thresholds to all my data. I investigate whether species estimates are stable across these thresholds. If not, I determine how many species and specimens are affected by threshold choices in order to test whether the thresholds affect community-level conclusions.

Overall, it is clear that DNA barcodes are not the solution to all problems and cannot replace alpha taxonomic research (i.e., describing species). I would argue that they are useful for sorting through large numbers of specimens, but ultimately still require scrutiny by experts in order to connect barcodes to the existing taxonomic

literature (Will, 2005). It is possible that in the future, 80-90% of all species can be identified with DNA barcodes, but taxonomy experts will still be needed for distinguishing closely related species and for pointing to those species that cannot be distinguished based on DNA barcodes.

Generating global barcode databases for identified species has been a failure, particularly for invertebrates. As of 2012, the majority of GenBank entries were unidentified to species (74%). For example, most Lepidoptera (78%) but also a large proportion of fish barcodes (34%) in the BOLD barcode database was not identified to species (Kwong *et al.* 2012b). One may conclude that I would not be able to use NGS barcodes in my thesis. However, delimiting species does not require species identifications. In my thesis, I obtain barcodes, cluster them at preset thresholds, and use the available barcode databases to identify at least some of the clusters to species. For the remaining clusters, I can still determine abundance and distributions because barcodes can be matched across space and time.

In my thesis, I optimize a fairly new way for obtaining DNA barcodes with NGS. Traditionally, obtaining DNA barcodes required genomic DNA extraction from individual specimens, DNA amplification, and Sanger sequencing (Sanger *et al.* 1977). Sanger sequencing can generate long reads (up to 1000 bases). However, it requires a relatively high concentration of DNA amplicon template and only allows for low sequence throughput (Shokralla *et al.* 2014). Here, I overcome the problems with Sanger sequencing via Next-Generation Sequencing (NGS). NGS technologies used to be divided into four major platforms: Roche 454, Ion Torrent, Illumina, and single molecule sequencing platforms such as Pacific Biosciences and Oxford

Nanopore (Goodwin *et al.* 2016). In the thesis, I only use Illumina platforms and Oxford Nanopore technology.

NGS barcoding in my thesis refers to tagged amplicon sequencing that consists of three steps: i) PCR amplification, where the COI amplicon for each specimen is individually labeled using a set of oligonucleotides that contain a known tag sequence, ii) PCR clean-up of pooled samples, and iii) Illumina sequencing (Meier *et al.* 2016). Unless otherwise stated, I amplify the specimens using direct PCR, thus avoiding the need for time-consuming and costly DNA extraction (Wong *et al.* 2014). NGS with Illumina system (i.e., MiSeq and HiSeq) has several advantages. Firstly, short-read barcodes can be produced at a fraction of the cost of traditional Sanger methods and with less effort (Metzker, 2010). Secondly, DNA sequence data from thousands of specimens can be read in parallel in a single sequencing run with amplicon sequencing (Meier *et al.* 2016). NGS platforms with optimized lab protocols can quickly tackle specimen- and species-rich invertebrate taxa for routine biomonitoring purposes and this is what I will present in three chapters of this thesis.

The drawback of NGS technologies such as Roche 454 (Shokralla *et al.* 2014) and Illumina (Shokralla *et al.* 2015b; Meier *et al.* 2016) is that sequencing run times are long, and barcoding is only cost-effective when thousands of specimens are simultaneously barcoded. A recently released portable sequencing device (Oxford Nanopore MinION™) alleviates one of these drawbacks (Loman & Watson, 2015) by providing real time sequencing. Data turnaround is quite fast (<24 hours) (Judge *et al.* 2015; Laver *et al.* 2015; Loose *et al.* 2016) and MinION™ is very small and promises portable and rapid sequencing, which would make bioassessment based on

macroinvertebrates much easier and faster (Quick *et al.* 2014; Leggett *et al.* 2015). Chapter 3 of the thesis will show how MinION™ can be applied in biomonitoring research. However, the new sequencing technology is not without its challenges. MinION™ has higher sequencing error rate than the Illumina system, making sequence alignment and data post-processing challenging. In chapter 3, I will discuss the nature of the errors and make recommendations for future experiments with this device.

1.7 Aims and outline of the thesis

The main goal of my thesis is to demonstrate how NGS barcoding can be used for analyzing the community of specimen- and species-rich invertebrate taxa. My thesis focuses on chironomid midges, but the techniques can also be applied to other taxa.

In the first data chapter (chapter 2), I investigate the chironomid community composition at Bedok Reservoir, an artificial aquatic habitat, where several recent nuisance midge outbreaks have occurred. I identify environmental parameters that are correlated with the outbreaks and provide recommendations for controlling these outbreaks. In addition, I test how many midges have to be barcoded in order to use chironomid midges for biomonitoring. I demonstrate that biomonitoring programs could gather the necessary data at low cost because only a small number of specimens have to be sequenced. Further manpower cost reductions may be possible through the metabarcoding of bulk samples. The DNA barcodes I generated for this chapter were also used for an eDNA study published in Royal Society Open Science, where I am a co-author.

In Chapter 3, I explore a new technique (MinION™) that can be used to make NGS barcoding faster. This manuscript is currently under review. I am a co-first author of this publication. I performed the field and bench work and contributed to the writing of the manuscript. We describe a pipeline for DNA barcoding using a MinION™ sequencer. The results suggest that >100 specimens can be multiplexed in a single run of the MinION™ sequencer and an estimate of species composition can be obtained 10 hours since sample handling.

In Chapter 4, I test whether adjacent natural and artificial habitats can maintain distinct midge faunas. To do this, I compare the community of a previously unstudied natural habitat, Nee Soon Swamp Forest, with the communities of three surrounding man-made reservoirs. I demonstrate that the swamp forest is home to a surprisingly species-rich chironomid community which is largely resilient and different from the midge communities in the adjacent reservoirs. I also analyze the spatial composition of the swamp forest community with respect to the several environmental parameters. Lastly, I discuss how these data will be essential for the conservation of the Nee Soon Swamp Forest in particular and Southeast Asian swamp forests in general.

CHAPTER 2[1]

Dissecting the causes of a nuisance outbreak of chironomid midges with NGS barcodes

2.1 Abstract

In this chapter, I used NGS barcoding to perform cost-effective biomonitoring of non-biting midges (Chironomidae: Diptera) from Bedok Reservoir in Singapore. Bedok Reservoir was the site of several recent swarming events by non-biting midges. NGS barcoding was used to identify 10,340 adults and 5,427 larval midges sampled bi-weekly for nearly a year from the reservoir. From molecular identification, midge communities were found to be composed of only 31 species (molecular operational taxonomic units, MOTUs). Dissolved oxygen levels at the center of the reservoir were associated with the nuisance outbreaks caused by *Tanytarsus oscillans,* a species of Chironomidae that was previously not known to be involved in mass swarming. Subsampling techniques were used to determine how many midge larvae and adults should be barcoded to establish a complete species profile that would be useful for biomonitoring. Sequences for 600-1000 midges obtained over a two-months period was sufficient for characterizing the species profile of chironomid communities from

[1] A version of this chapter is in prep as "**Baloglu, B.**, Clews, E., Cranston, P., Meier, R. (2017). Dissecting the causes of a nuisance outbreak of chironomid midges with NGS barcodes." I am the first author of this publication. I performed the bench work, data analysis and writing of the manuscript.

the tropical reservoir. 100-200 midges from one sampling event were sufficient for characterizing the most common species. My analyses suggest that biomonitoring programs could cheaply gather data with only a small number of NGS-barcoded specimens or metabarcoded bulk samples.

2.2 Introduction

Non-biting midges (Chironomidae: Diptera) are an important family of freshwater invertebrates. Chironomids possess higher species diversity than other macroinvertebrates (Marziali *et al.* 2010; Nicacio & Juen, 2015) and have specific habitat requirements that render them as useful ecological indicators. Chironomid swarms are also a health and economic nuisance in urban residential areas around Singapore (Lin & Quek, 2001; Cranston *et al.* 2013). Understanding the life cycle and structure of chironomid communities is essential for identifying the cause of the outbreaks. However, characterizing insect communities is difficult because conventional methods of chironomid species identification based on morphology (Pfenninger *et al.* 2007) are laborious and costly (Meier *et al.* 2006; Friberg *et al.* 2011; Carew *et al.* 2013). With traditional taxonomic techniques, regular monitoring of chironomids at the species-level is not feasible (Raunio *et al.* 2011). This is unfortunate because midge biodiversity information can significantly increase the accuracy of biomonitoring programs and help provide recommendations for preventing or mitigating insect swarming behavior (Nicacio & Juen, 2015).

An alternative to morphological identification is molecular identification using DNA barcodes. Identification based on a DNA sequence can be cheaper, faster, and yield better taxonomic resolution than traditional techniques (Stein *et al.* 2014; Wong *et al.* 2014; Meier *et al.* 2015). In monitoring chironomid outbreaks, DNA barcoding is ideal because the number of swarming insects is large, and swarms may consist of multiple species (Darby, 1962; Armitage *et al.* 2012). Although the use of molecular tools in species identification is not new (Ander *et al.* 2013; Lin *et al.* 2015), the study

of whole communities with presorting information is rare (i.e., for multiple sites, over many months, and life stages). Little is known about the correlation between environmental variables and species-level change in chironomid community structure.

Here I use an affordable NGS barcoding approach to study the community structure of chironomid midges in a tropical reservoir. The gene of choice for barcoding is the 313bp fragment of COI with which I identify nearly 16,000 specimens from one of Singapore's freshwater reservoirs with several recent (>2011) mass swarming events (Lin & Quek, 2011; Cranston *et al.* 2013). The reservoir covers an area of 880,000 m^2 and stores 12.8-million m^3 of water. Bedok Reservoir is also relatively deep with a mean depth of 9 m and a maximum depth of 18.2 m (Clews *et al.* 2014). To understand what is causing the outbreaks and locate the larvae responsible for the outbreaks, I study the temporal and spatial dynamics of chironomid community composition for several sites (five), dates (up to 26), sampling methods (three), and life stages (two) over a period of a year. With this experiment design, I can identify environmental parameters correlated with swarming and determine whether different sampling techniques for midges yield similar results.

2.3 Materials and Methods

2.3.1 Sampling sites and protocols

Emergence trap and sediment grab. The sampling of chironomid specimens and the collection of environmental data were carried out by Tropical Marine Science Institute (TMSI). Four sites were sampled within Bedok reservoir. Three of the sites were at the edge of the reservoir: WBA — 1°20'43.6"N 103°55'21.5"E, FAD — 1°20'35.6"N 103°55'47.7"E, FDA — 1°20'25.7"N 103°55'19.4"E. One of the sites was at the center, ~750m from reservoir edge— 1°20'34.7"N 103°55'30.9"E. Three sample replicates were collected from each of the four sites. The center and the edge sites differ from each other. At the edge of the reservoir, there are relatively few breeding sites for chironomids due to the steep and rocky slope. Most of the nutrient substrate is present in the reservoir center. Hence, larvae were sampled in two locations (center and only one edge habitat: FDA), while adults were sampled in four of the sites. Environmental parameters such as Chlorophyll *a*, dissolved oxygen, temperature, pH, conductivity and turbidity were only recorded for the center. Samples were collected on a bi-weekly basis between 2013 and 2014. Sediment grab sampling was used to capture larvae, and emergence traps were used to capture adults.

Colonizer. A cage sampler was deployed at 1.2 m water depth for four weeks and allowed for colonization of the sampler by invertebrates (Loke *et al.* 2010). Three replicates of the colonizer sampler were collected from only one site (Colonizer — 1°20'42.9"N 103°55'13.8"E) between 2013 and 2014, every 1-2 months. All samples were pre-sorted into morphotypes. The specimens were preserved in 70% ethanol for adults and 99% ethanol for larvae.

2.3.2 DirectPCR and high-throughput sequencing

I obtained DNA barcodes for each specimen using the direct polymerase chain reaction ("directPCR") protocol described in Wong *et al.* (2014). Sample-specific amplicon sequencing was carried out using unique combinations of tagged primers. This allowed for a trace-back of sequences to specimens as well as sequencing of thousands of amplicons in small numbers of MiSeq runs and libraries. Initially, I performed the PCR amplification using a few of the whole specimens with primer pair LCO 1490/HCO 2198 (Folmer *et al.* 1994). Initial success rates were as low as reported in the literature (Wong *et al.* 2014). To improve the success rates, I reduced the amount of specimen tissue. The tissue amount for medium- and large-sized adults was 2.5–3.4 mm and >3.5 mm, respectively. For all three size classes of larvae, 3.0–4.4 mm, 4.5–6.4 mm and 6.5–8.0 mm was used, respectively. Degenerate metazoan primers [COI; mlCO1intF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' (Leray *et al.* 2013) and jgHCO2198: 5'-TAIACYTCIGGRTGICCRAARAAYCA-3' (Geller *et al.* 2013)] were used for the new PCR reaction conditions. Each reaction used its own combination of 9-bp primer tags generated by the "Barcode Generator" as described by Meier *et al.* (2015). Unique combinations of the nine nucleotides long 5' overhangs were used to identify each specimen from a PCR reaction. PCR products were pooled and sent for library preparation. Within each library, every specimen was tagged by the unique forward and reverse primer tag combinations. NGS barcoding of specimens (n=15,767) was carried out on multiple MiSeq 2 X 300 cycle runs.

2.3.4 Bioinformatics

The bioinformatic pipeline used for processing the MiSeq data is described in Meier *et al.* (2015); in short, the paired-end reads were merged with PEAR (Zhang *et al.* 2013) and then demultiplexed and assigned to each specimen using the unique combination of 9bp tags (perfect match required) and primer sequence (<=2 bp). All reads <100 bp in length were discarded. The dominant read was identified, and all sequences that were identical but were length variants of this read were merged. The number of total reads, merged reads, the ratio between the dominant and second-most dominant read was recorded. A specimen was considered successfully barcoded if it met the following criteria: 1) >50x read coverage; 2) >10x read coverage for dominant read and 3) dominant read 5x times more common than the second-most dominant read. All retained reads were aligned against the local database of chironomids using the MegaBLAST algorithm (Zhang *et al.* 2000). MegaBLAST searches were performed with a minimum similarity of 97%.

2.3.5 MOTU delimitation and morphological identifications

MOTU delimitation. Objective Clustering was used to delimit sequences into putative species units, or molecular operational taxonomic units (MOTUs), at 2-5 % using uncorrected pairwise distances (Srivathsan & Meier, 2012). Previously it was shown that this range of thresholds does not affect the main conclusions (Meier *et al.* 2015). Automatic Barcode Gap Discovery (ABGD) was also conducted for MOTU delimitation (Puillandre *et al.* 2012). As opposed to Objective Clustering, ABGD uses a series of prior intraspecific divergences to infer from the dataset a one-sided confidence limit for interspecific divergence (Puillandre *et al.* 2012). The fasta file of

aligned sequences was analyzed using the *p*-distance in the ABGD online species delineation tool (wwwabi.snv.jussieu.fr/public/abgd/) with following parameters (p=0.005, P=0.1, n=20, s=0.1). Some of the MOTUs were further identified to species with a barcode database that was established based on morphology.

Barcode database based on morphology. I had an initial barcode database that was based on specimens that were identified with morphology. Identical haplotypes were generated for the obtained 12,622 COI sequences and the previously morphologically verified sequences (morphoIDs) using Objective Clustering at 0% in SpeciesIdentifier (TaxonDNA 1.6.2; Meier *et al.* 2006). To ensure that the obtained sequences were morphologically verified, the haplotype database was clustered at 1%, a threshold used for BOLD identifications. Sequences that were more than 1% apart from the morphoIDs were detected and considered unidentified.

2.3.6 Statistical analyses

Community analyses. I evaluated the sampling sufficiency using rarefaction curves generated with the iNEXT package (Hsieh *et al.* 2016) in R v2.0.12 (Team R, 2017). Community comparisons were visualized using three-dimensional non-metric multidimensional scaling (NMDS) ordination based on Bray-Curtis dissimilarities, with metaMDS, as implemented in *vegan* 2.4–3 package (Oksanen *et al.* 2017) in R and only the samples with more than 30 individuals were included. Multiple stage community comparisons were evaluated using the Morisita index (Chao *et al.* 2016) of SpadeR package in R. Sample sizes were corrected when comparing diversity between samples of different size. Mantel tests were used to determine the relationship between rarefied communities from (i) different sites, (ii) different

sampling intervals (i.e., every two weeks, every four weeks etc.), (iii) different life stages (adult or larvae), and (iv) different sampling techniques, using the ade4 package (Dray & Dufour, 2007) in R. The significance of the relationships were assessed using 999 randomizations. Richness estimates were documented for various sampling intervals (2, 4, 6, 8, and 10 weeks) using EstimateS (Colwell, 2013). As earlier analysis indicated no significant difference between the community compositions collected at different collection intervals and the number of sites used in this comparison was small (n = 3), I combined the edge community datasets.

Environmental parameters. I applied a linear mixed effect (lme) model to investigate the relationship between environmental parameters and species (MOTU) abundances, using the lme4 package (Bates *et al.* 2012). The collinearity between the explanatory variables was assessed using the vif function of the package car in R. The model considered time as a random effect to account for interdependence among sampling weeks. Null models were constructed containing only the random effects. ANOVA was used to compare the final lme model with the null model.

2.4 Results

2.4.1 Sequence data analysis

I obtained COI sequences for 12,622 specimens (80% success rate) from ca. 16,000 processed specimens. In total, 13,3 million reads were acquired from multiple shared MiSeq runs, which implies that one MiSeq 2 X 300 cycle run (15 million reads) would have been sufficient for the entire experiment. The overall cost per specimen is 0.29 USD (Lab cost: 0.16 USD/specimen, Meier *et al.* 2015; Miseq 2X300 PE and library cost: 0.13USD/specimen, according to NYU Genome Technology Center).

2.4.2 Structure of chironomid diversity

The previously carried out integrative taxonomy project on the midges of Singapore's reservoirs yielded barcodes for most species. Therefore, only 7% of the 227 haplotypes I obtained in this study do not have identifications because they differed by >1% from all identified barcodes. These haplotypes belong to rare species (0.9% of the total number of specimens). If the identification threshold for barcodes is increased to 2-5% genetic thresholds, even fewer MOTUs lack identification (0.3% - 0.008% of the specimens). Clustering the 12,622 COI barcodes at 2 - 5% using Objective Clustering generated a total of 34 – 29 MOTUs. The number of MOTUs defined by the ABGD (Appendix 1, Fig. S1) was similarly stable (29 - 32) at different a priori threshold values. Here, clusters obtained with a 3% threshold were used (Ball *et al.* 2005). Five MOTUs were singletons, three doubletons, and another four were rare (2< specimen counts <10; see Table 2.1).

Table 2.1: Taxonomic composition and abundance of the species of midges collected in Bedok Reservoir. The number of specimens (n) obtained per species is provided for each field site.

	Field sites				
Species	**CENTER**	**FAD**	**FDA**	**WBA**	**COLONIZER**
	n	**n**	**n**	**n**	**n**
Ablabesmyia 2segpalp	5	6	46	54	1
Ablabesmyia typeTMSI	3	23	150	58	29
Chironomini genus indet	0	0	2	0	0
Chironomus circumdatus Kieffer, 1916	2	0	2	0	0
Chironomus kiiensis Tokunaga, 1936	1	0	0	0	0
Cladopelma sp.	1	0	0	0	0
Cladotanytarsus sp.	3	174	547	106	0
Cladotanytarsus sp2.	1	31	85	21	0
Cladotanytarsus sp4.	1	0	9	1	0
Cladotanytarsus spGC34.	0	0	6	0	0
Cryptochironomus fulvus Johannsen, 1905	0	0	11	0	0
Dicrotendipes flexus Johannsen, 1932	0	0	4	0	0
Dicrotendipes pelechoris	0	1	4	1	1
Microchironomus tener Kieffer, 1918	0	0	2	0	0
Nanocladius sp.	0	35	29	30	2
Parachironomus sp.	0	71	10	8	1
Paratanytarsus sp.	0	0	1	0	0
Polypedilum cf.griseoguttatum	4	0	0	0	0
Polypedilum griseoguttatum Kieffer, 1921	93	14	16	1	0
Polypedilum leei Freeman, 1961	0	972	843	1191	109
Polypedilum leei2	0	147	87	276	18
Polypedilum masudai Tokunaga, 1938	0	0	1	0	0
Polypedilum nodosum Johannsen, 1932	2	140	260	509	57
Polypedilum nubifer Skuse, 1889	3	1	28	1	0

Polypedilum quasinubifer	12	18	2116	74	5
Procladius choreus Meigen, 1804	72	34	3	1	0
Procladius sp4.	0	0	1	0	0
Tanytarsus formosanus Kieffer, 1912	16	12	47	8	6
Tanytarsus infundibulus	0	0	2	0	0
Tanytarsus oscillans Johannsen, 1932	3342	173	57	74	0
Tanytarsus ovatus Johannsen, 1932	63	18	73	38	5
Total number of individuals	3624	1870	4442	2452	234
Total number of species	17	17	28	18	11

For adults, 6 of the 24 species (25%) were found at all four sites. For larvae, 13 out of 24 (54.2%) were shared when sediment grab samples were compared, and 6 of the 27 (22.2%) were shared between the three sampling sites (Center, FDA, and Colonizer). With regard to abundant species, species accumulation curves reached saturation for colonizer, center and the edge sites (Fig. 2.1a). However, for overall species richness, including the rare species, data for more specimens may be needed for some sites and methods (e.g., Colonizer; Fig. 2.1b). All specimens identified to MOTUs were deposited in the Lee Kong Chian Natural History Museum (LKCNHM) of the National University of Singapore (NUS).

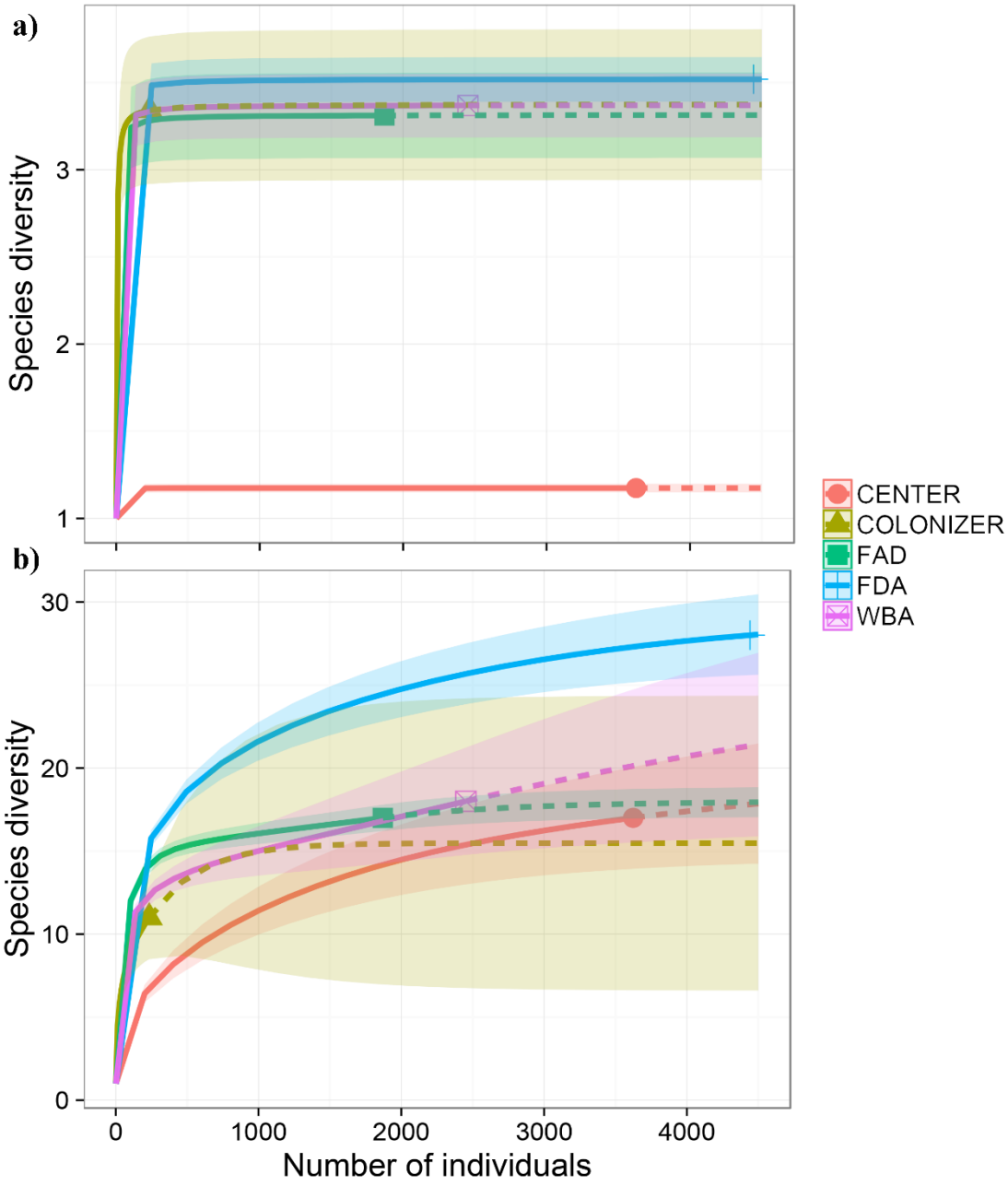

Figure 2.1: Individual-based rarefaction (solid line) and extrapolation (dashed line) for a) abundant and b) all species of chironomid communities, based on data from all samples from five sites at Bedok Reservoir. The 95% confidence intervals (shaded areas) were obtained by a bootstrap method based on 200 replications. FAD, Forest Adventure; FDA, Floating Deck A; WBA, Wakeboard. Data are given in Table 2.1.

2.4.3 Adult community at the edge and center sites

Based on 6173 adult specimens collected at the three edge sites (FAD, FDA, and WBA), 23 adult species were identified. 15 species were found after one sampling occasion at the three edge sites, 17 species after three sampling dates, and 19 species after five sampling dates, or 65, 74, and 83% of cumulative chironomid species respectively (Table 2.2).

Table 2.2: Accumulating percentage of species richness (±standard error, SE) collected at edge sites and the center at each sampling event (every 2 weeks). Diversity indices (±SE) provided for each individual sampling date. The analysis was carried out using EstimateS.

	Edge sites				Center			
Total sampling events	No. of individuals	Species richness (%)	Shannon's index	Simpson's index	No. of individuals	Species richness (%)	Shannon's index	Simpson's index
1	359	0.65 (±0.05)	1.51	0.64	656	0.62 (±0.09)	0.20	0.08
2	259	0.72 (±0.05)	1.81	0.72	616	0.68 (±0.10)	0.27	0.10
3	239	0.76 (±0.05)	2.02	0.82	119	0.72 (±0.11)	0.12	0.05
4	263	0.79 (±0.05)	1.87	0.77	216	0.77 (±0.11)	0.20	0.08
5	287	0.82 (±0.05)	1.67	0.72	162	0.80 (±0.11)	0.39	0.20
6	160	0.84 (±0.05)	1.36	0.56	384	0.84 (±0.11)	0.12	0.04
7	63	0.85 (±0.05)	1.35	0.56	5	0.87 (±0.11)	0.67	0.48
8	100	0.87 (±0.05)	1.24	0.55	80	0.90 (±0.10)	0.73	0.38
9	71	0.88 (±0.05)	1.95	0.76	51	0.93 (±0.09)	0.29	0.11
10	134	0.90 (±0.05)	1.78	0.74	196	0.95 (±0.07)	0.20	0.09
11	89	0.91 (±0.05)	1.89	0.81	39	0.98 (±0.06)	0.74	0.38
12	254	0.92 (±0.04)	1.60	0.69	194	1 (±0)	0.03	0.01
13	196	0.93 (±0.04)	1.60	0.68				
14	195	0.94 (±0.04)	1.26	0.55				
15	218	0.94 (±0.04)	1.62	0.75				
16	6	0.95 (±0.04)	1.01	0.61				
17	99	0.96 (±0.04)	1.66	0.76				
18	253	0.96 (±0.03)	1.44	0.68				
19	106	0.97 (±0.03)	2.02	0.75				
20	296	0.97 (±0.03)	1.50	0.67				
21	272	0.98 (±0.03)	1.43	0.64				
22	219	0.98 (±0.02)	1.71	0.74				
23	200	0.99 (±0.02)	1.56	0.68				
24	491	0.99 (±0.02)	1.47	0.63				

25	804	1 (±0.01)	1.59	0.66	
26	540	1 (±0)	1.81	0.78	

Across time, 2 - 4 dominant species each constituted >10% of all specimens while the remaining 13 - 16 species represented <5% of all specimens at the edge sites. Throughout the study period, there was considerable variation in the relative abundances of the dominant midge species. Only two species, *Polypedilum leei*, and *Cladotanytarsus sp.* were detected during most sampling events and at most edge sites (Fig. 2.2). The specimens' community were very similar for the edge communities across different sampling intervals, even when the number of weeks between sampling events increased (Appendix 1, Tables S1-2). Specifically, a high correlation was present between the communities of FAD and WBA (0.85 km apart), followed by that of FDA and WBA (0.56 km apart).

Unlike the edge sites, 99% of the adult specimens were collected during the first 12 sampling events at the center of the reservoir where only 7 species were identified. Likewise, for the center site, 4 species were observed after one sampling occasion, 5 species after three sampling dates and 6 species after five sampling dates, or 60, 70 and 85% of cumulative chironomid species. Only one abundant species was detected at the center throughout the sampling period, and this was *Tanytarsus oscillans* (Table 2.1), a species that has so far only produced nuisance outbreaks in Singapore.

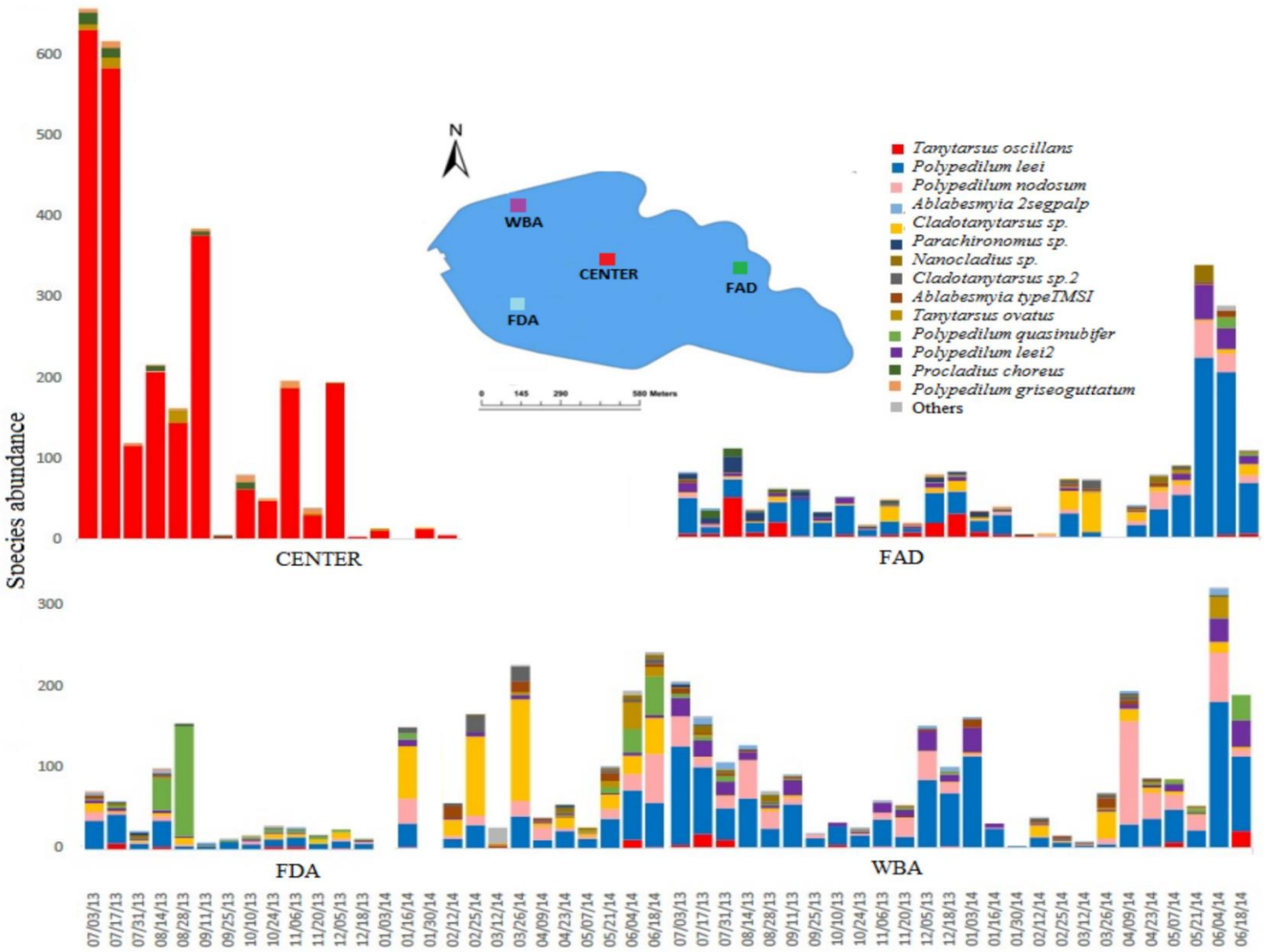

Figure 2.2: Abundances of chironomid midge species in four sites of Bedok Reservoir from July 2013 to June 2014, using the adult dataset. FAD, Forest Adventure; FDA, Floating Deck A; WBA, Wakeboard. Note the maximum values on the y axis differ for the center and the other sites.

2.4.4 Spatial variation in MOTU richness and environmental parameters

Edge habitats collectively differed from the center (Appendix 1, Table S2) as is also evidenced from non-metric multidimensional scaling (NMDS; Fig. 2.3). Of the environmental parameters recorded, dissolved oxygen measurements explained much of the variance in the center (P <0.05), which is in line with the known high oxygen demand of *Tanytarsus larvae*. The center was characterized by a high relative abundance of *Tanytarsus oscillans*, accounting for 95% and 84% of the adult and larval chironomids, respectively. At edge sites, the relative abundance of this species was 1% for larvae and ranged from 3 to 9% for adults. The bottom of the reservoir center site is sixteen meters deep and was expected to have low levels of oxygen as was typical of tropical lakes (Townsend, 1999; Ambasht & Ambasht, 2012). However, the high numbers of *T. oscillans* at the center site may have been due to artificial oxygen availability. Indeed, oxygen concentrations measured at the bottom of the reservoir were higher than normal and significantly affected the abundances of *Tanytarsus oscillans* larvae **(**χ^2= 7.84, P = 0.01), as well as the adults **(**χ^2= 5.75, P = 0.02) (Table 2.3). The oxygen concentrations did not influence the abundances for either of the life stages at other depths of the reservoir (Table 2.3).

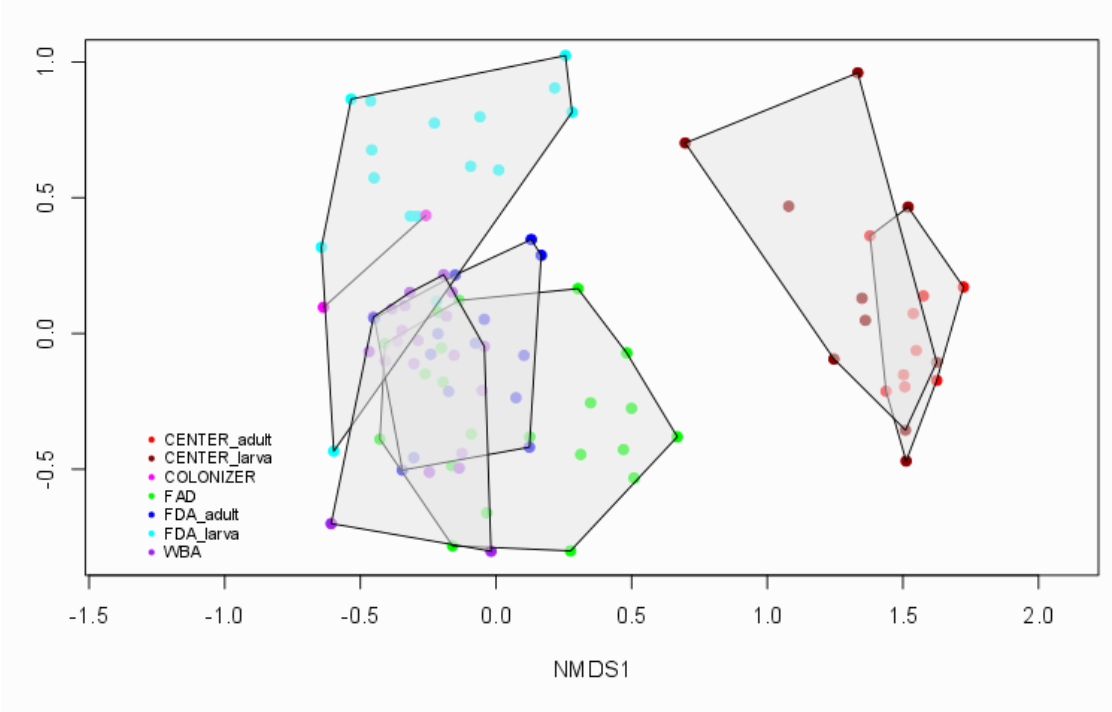

Figure 2.3: NMDS ordination of Bray-Curtis similarities in MOTU composition between the sampling dates, based on abundance dataset for all five sites and two life stages. The points represent the samples collected biweekly from the corresponding sites. FAD, Forest Adventure; FDA, Floating Deck A; WBA, Wakeboard.

Table 2.3: Model comparisons and *P* values. A total of eight models were compared with null models. ChiSquare and p values are shown, with significant values in bold. Due to collinearity between explanatory variables (data not shown), models 1 and 5 were used.

Tanytarsus oscillans life stage	Explanatory Variable (Oxygen measurements at different depths)	Anova $\chi 2$ (null model vs final model)	*P*
Adult	Model1: bottom	**5.75**	0.02*
	Model2: middle	0.83	0.36
	Model3: surface	0.18	0.67
	Model4: bottom + middle + surface	**9.28**	0.03*
Larvae	Model5: bottom	**7.84**	0.01*
	Model6: middle	0.01	0.93
	Model7: surface	0.17	0.68
	Model8: bottom + middle + surface	**10.01**	0.02*

2.4.5 Chironomid community at different sampling intervals

To determine sampling requirements, I compared the community compositions obtained using different sampling intervals and sample sizes. If the sampling interval and number of specimens collected was reduced to allow for a more time- and cost-effective bioassessment, would community composition change significantly? If communities were sampled every 2, 4, and 6 weeks for a maximum sampling duration of 48 weeks (12 months), all sampling strategies captured the maximum number of observed species (n = 23; Table 2.4). Using 8- and 10-weeks sampling intervals for the same duration, at most 74% of the species were detected in a sample of 800 individuals. The most common 10 species made up 95% of the total number of captured individuals. Thus, species richness at different sampling intervals was also analyzed for the most common species. Table 2.4 shows that all sampling intervals from 2 to 10 weeks resulted in a maximum number of species captured. This table is quite revealing in several ways. First, at every sampling interval, subsamples containing 300-400 individuals for two subsequent sampling events is sufficient to obtain more than 90% of the most common species. Secondly, the total sampling duration can be as short as 14 weeks (3.5 months) to characterize the most common species in the community, as opposed to the 48 weeks needed for characterizing the whole community. These results indicate that environmental assessment plans focused on measuring the relative abundances of the most common species can be time- and cost-effective to gauge ecosystem health.

Table 2.4: Number of sampling events, individuals (subsamples) and percentage of observed species richness at different sampling intervals for a community of a) all species, b) most common species. The analysis was carried out using EstimateS.

	All species			10 most common species		
Sampling interval	No. of sampling events	No. of individuals	Species richness	No. of sampling events	No. of individuals	Species richness
2 weeks	3	600	0.69	1	192	0.825
	7	1400	0.8	2	384	0.953
	14	2817	0.9	3	577	0.985
	24	4829	1	7	1346	1
4 weeks	3	550	0.71	1	174	0.84
	5	916	0.81	2	348	0.96
	8	1466	0.9	3	522	0.987
	12	2199	1	4	1045	1
6 weeks	2	336	0.7	1	158	0.79
	3	505	0.78	2	317	0.95
	5	841	0.9	3	475	0.98
	8	1346	1	4	951	1
8 weeks	2	428	0.63	1	206	0.833
	4	856	0.71	2	413	0.97
	6	1284	0.74	3	825	1
10 weeks	2	354	0.6	1	172	0.78
	5	886	0.7	2	344	0.97
	-	-	-	3	515	1

2.4.6 Midge community at different life stage and sampling methods

Life stages. In my study, both the adults and larvae were sampled. I analyzed the similarity between adult and larval communities using four different ways, namely Bray-Curtis, Jaccard, Chao and Shannon indices. The Mantel randomization tests between the adult and larval chironomid communities of the center revealed a significant similarity with the Bray-Curtis dissimilarity index, binary Jaccard and Chao tests (Pearson r = 0.41; 0.39; and 0.35 respectively, permutation $P < 0.05$ for all three tests). Furthermore, the effective number of species (ENS) using the Shannon index (Jost, 2006) was found to be 1.31 in adult, and 2.13 in the larval community at the center. On the other hand, the Mantel randomization tests between the adult and larval chironomid communities of site FDA, also shown in Fig. 2.3, revealed a nonsignificant correlation with Bray-Curtis dissimilarity index, binary Jaccard and Chao tests (Pearson r = 0.09; 0.05; and 0.12 respectively, permutation $P > 0.05$ for all three tests). Furthermore, an effective number of species (ENS) using Shannon index (Jost, 2006) was found as 7.98 in adult, and 3.29 in the larval community at the edge site.

Sampling methods. To test if communities estimated using different sampling strategies would differ, I pooled the larval samples from site FDA in month-long intervals and compared against colonizer samples which were collected on three different dates. The communities were negatively associated and not significantly correlated (September 2013: $R = -0.21$; $p = 0.95$; January 2014: $R = -0.21$; $p = 0.71$ and April 2014: $R = -0.09$; $p = 0.67$). Finally, the effective number of species (ENS) using Shannon index (Jost, 2006) was found to be 4.71 in FDA, and 4.53 in colonizer samples.

2.5 Discussion

2.5.1 Dissecting a mass swarming event using NGS barcoding

Although Chironomidae is prevalent in aquatic ecosystems, especially in tropical habitats (Coffman & de la Rosa, 1998), few people study their species composition. The lack of interest is due to challenge in species identification and the associated high cost. Here, NGS barcoding of chironomids can overcome this challenge even for large bioassessment campaigns. For my study, a one-year sampling campaign across 5 sites and 26 sampling dates cost approximately 0.29 USD/specimen.

One criticism of DNA barcoding is that the species units obtained with barcodes are not stable. I here used Objective Clustering and ABGD methods to estimate the number of MOTUs. These estimates were stable (ABGD: 29 - 32; Objective Clustering: 29 - 34). The minor differences between the two approaches were due to splitting or lumping of three MOTUs. One of these MOTUs is nearly 5% distant from the sequences that are being lumped with it during the ABGD procedure (*Polypedilum cf. griseoguttatum*, a rare species at Bedok Reservoir, <10 specimens). All the sequences that belong to the other MOTU (*Tanytarsus ovatus*) are less than 3% different from each other, but ABGD still splits them into two MOTUs. Finally, another MOTU that causes a discrepancy between the two approaches is *Cladotanytarsus sp.*, which ABGD lumps with *Cladotanytarsus sp.2*. If I were to use 4% genetic threshold instead of 3% during my analysis, these two species would be considered as one. This discrepancy is minor because only 1% of all specimens in the dataset were affected. Overall, the method for determining MOTUs is of secondary importance to understanding the overall community structure.

A study conducted in 2013 identified *Tanytarsus oscillans* as the nuisance midge in Bedok Reservoir (Cranston *et al.* 2013). However, we knew little about midge community structure and the cause of the outbreak in Bedok Reservoir. Here, NGS barcoding provided the MOTU/species level resolution that is needed for distinguishing the larvae of the nuisance species (*Tanytarsus oscillans*) from larvae of closely related species and enabled the study of whole midge communities from Bedok Reservoir. NGS barcoding revealed that chironomid communities were less evenly distributed in the center due to the dominance of *T. oscillans* species than at the three edge sites (see Fig. 2.2). The center of this particular reservoir is rich in sediments that could be inhabited by many larvae and is an attractive feeding site. However, low oxygen levels previously prevented larval colonization. Oxygen concentration is known to be a major community structuring factor in chironomids (Thienemann, 1921; Brundin, 1949). Many species of the tribe Chironomini are known to adapt well to changes in oxygen content (Pinder, 1995). *Tanytarsus sp*. of the tribe Tanytarsini, in particular, are reported as indicative of better-oxygenated standing water environments (Walshe, 1947; Thienemann, 1913 as cited in Esteves, 1988; Heinis & Davids, 1993; Takahashi *et al.* 2008). The high abundance of *T. oscillans* species observed in my study in the reservoir center suggests that dissolved oxygen concentrations are the driving force causing nuisance outbreaks in Bedok Reservoir. Indeed, dissolved oxygen levels at the bottom of the reservoir correlated with both adult and larvae *T. oscillans* abundances. However, I did not detect any significant association between *T. oscillans* abundances and oxygen levels at other depths. My findings suggest that nuisance outbreaks only occur when the center has favorable growth conditions for *Tanytarsus* larvae.

In contrast with the center, few *T. oscillans* were collected from edge sites (Fig. 2.3). The most common species observed across the three edge sites were minor or absent in the center, namely *Polypedilum leei* and *Cladotanytarsus sp.* (Fig. 2.2). The disparity between the center and edge communities could primarily be due to the habitat differences, e.g. the edge sites are steep and rocky, while the center has most of the sediments and growth substrate. Furthermore, the three edge communities display significant similarity despite the subtle differences observed among them (Appendix 1, Tables S1-2). This finding has further implications for rapid and cost-effective bioassessment.

2.5.2 Rapid bioassessment: How often should midges be sampled?

Ideally, the best estimate of a community is based on all the specimens. Species richness, with rare species being its largest component, was traditionally the focus of many ecological theories, such as island biogeography (MacArthur, 1972), disturbance theory (Connell 1978; Robinson & Minshall 1986) and biodiversity conservation (Peet 1974; May 1988; Baltanas 1992). However, exclusion of rare species may not bias the outcome in ecological bioassessment (Arscott *et al.* 2006). Rapid bioassessment can also be carried out using only the abundance of the most common species.

Total species richness. In my study, a sampling duration of one year and sampling intervals of 2, 4, or 6 weeks captured the entire species profile of the reservoir edge sites. My findings indicate that a routine biomonitoring using chironomids can be as infrequent as 6 weeks and still reveal the whole community structure. 70% of the total community was observed after two sampling intervals (4 weeks) for the reservoir

edge sites, and three sampling intervals (6 weeks) for the reservoir center site (Table 2.2). Frequent sampling of chironomids can be beneficial for understanding local and regional species diversity in temperate regions (Raunio & Muotka, 2005; Ekrem *et al.* 2010), but is not necessary for tropical environments where the environment is less seasonal. Sampling sites can also be more coarse-grained geographically while revealing the same level of information. The three reservoir edge sites used in this study did not have significant differences in community composition, yet the sites were 1 km apart from each other. Overall, I suggest sampling fewer sites and increasing the distance between the sites. Sampling midges across one or two sites at 10-week intervals could be a suitable long-term biomonitoring regimen for tropical reservoirs. If resources are scarce, sampling every 2 weeks for only 4-6 weeks can also be sufficient for capturing a snapshot of the chironomid community (~70%).

Most common species. Bioassessment can be even cheaper and time-effective if the target is to characterize the most common species. Marchant (2002) stated that common species provide the most obvious signals in environmental degradation, and bioassessment should incorporate the common or abundant species to interpret the effects of habitat disturbance. Based on my subsampling analysis, a total sampling duration of 3.5 months with a 2-week sampling interval was sufficient for capturing the abundance of the most common species. Such a campaign would take only 30% of the resources of the original study. For example, a single sampling event (>150 individuals) is sufficient to obtain 80% of the common species. Tropical freshwater habitats are more biodiverse than their temperate counterparts (Lake *et al.* 1994). This has also been the case for chironomid midges (Cranston *et al.* 1997; Coffman & de la

Rosa, 1998). Thus, even fewer resources can be sufficient for characterizing the chironomid species profile of temperate freshwater reservoirs.

2.5.3 Rapid bioassessment: Which life stage should be sampled?

A major cost of bioassessment programs is the sampling strategy. Larval sampling methods, such as the sediment grab and colonizers used in my study, require more labor, time, and money than adult sampling with emergence traps (Grant, 2002). Larvae are small and difficult to separate from sediment (Guardiola *et al.* 2016). Sediment sampling is also difficult in rocky habitats, where the bottom substrate comprises of granite boulders, rocks, pebbles and coarse sand (Loke *et al.* 2010).

Here, I tested whether the larval communities obtained with two different sampling techniques, sediment grab (FDA) and colonizer, was similar. I found that the larval communities were uncorrelated. Colonizer samples had a sample coverage of >95%. However, the comparison was limited only to three sampling events due to high experimental failure rates. The lack of correlation could be an artifact of the small sample size. Despite the lack of correlation between sediment grab and colonizer samples, the colonizer samples seemed to cluster tightly with the sediment grab samples (Fig. 2.3). In my study, sediment grab samples were more species-rich than the colonizer samples, indicating a possibility of nested communities. Colonizer communities could be a subset of sediment grab communities.

Furthermore, I compared the adult and larval communities collected from the center, and I found that their community composition was similar. The NMDS ordination of Chironomidae larval and adult communities indicated a relatively large

spatial overlap for the center reservoir site. I expected this correlation, given that the center community was imbalanced with the influence of the dominant species. On the other hand, in the site FDA, the adult community obtained with emergence trapping was richer in species composition and significantly different from the larval community. Here, the lack of correlation between larval and adult samples could be due to an increase in chironomid mobility during the pupal life stage. During the larval stage, chironomids mostly live in the sediment where they hatch from the egg. During the pupae stage, chironomids become adrift in water and rise to the water surface in preparation for their emergence as an aerial organism (Armitage *et al.* 2012). Thus, the adult sample composition could be more species-rich because pupae from adjacent communities also drift into the emergence traps. Hence, emergence trapping may represent a sampling from a wider ecological area. If ecosystem spatial resolution is necessary, sediment grabs should be used because sediments are not likely to contain organisms from a spatially distant ecosystem. For bioassessment programs interested in finding the common species over a large area, emergence trapping of adult chironomids should be sufficient.

2.6 Conclusion

In my study, I used NGS barcoding to obtain species-level taxonomic identification of larval and adult chironomids in a cost- and time-effective manner. Spatiotemporal analysis of chironomid community structure at Bedok Reservoir revealed the cause of the midge outbreak. Increased oxygen levels in the reservoir sediment caused an increase in abundance of the nuisance midge species *Tanytarsus oscillans*. A cost-effective regimen for future biomonitoring at Bedok Reservoir can rely on small samples of the whole community: Sampling adults midges from one or two sites over a period of 4-6 weeks will provide enough information for community bioassessment. For my study, adult midges were easier to handle, and the relative abundances of the most common species already provided sufficient information about the health of the ecosystem. Ultimately, the sample target, regimen, and capture method will depend on the research question. However, here I suggest that the small numbers of NGS-barcoded specimens can be sufficient to characterize whole or most of the chironomid community. This finding has important implications for cost-effective specimen- or metabarcoding-based environmental biomonitoring of chironomids, as they are well-known bioindicators throughout the world. With cheap DNA barcodes, similar studies can be carried out any place on earth. Future studies with NGS barcoding will provide us with a better understanding of how species community compositions are affected by changing environmental conditions.

CHAPTER 3[2]

Towards quick and cheap barcoding in the field: A specimen-based MinION barcode pipeline

3.1 Abstract

DNA barcodes are an important tool for species discovery and the identification of pests, exotic species, pathogens, vectors, and detecting fraud in the food industry. The existing methods for generating barcodes require a well-equipped molecular laboratory and can be time-consuming and expensive. This is unfortunate because many potential users lack access to such facilities and place a premium on obtaining DNA barcodes rapidly. We here test whether reliable barcodes can be generated using the recently introduced Oxford Nanopore MinION™ sequencer. We produce 50 tagged COI amplicons (313 bp) for 50 specimens of non-biting midges (Chironomidae: Diptera) before pooling and sequencing them with MinION™ and Illumina MiSeq. We then develop a bioinformatics pipeline that accommodates the high base call error rates of MinION™ and assess the MinION™ ON barcodes against the MiSeq barcodes. MinION™ recovers all 50 specimen barcodes at 19-609X coverage with few mismatches (1 bp across 50 barcodes). Indel error rates are higher (4-10 bp), but >98% are concentrated in homopolymeric regions. Resampling at different depths suggests that 10X coverage per specimen can yield a fairly

[2] A version of this chapter has been submitted as "Srivathsan, A., **Baloglu, B.**, Bertrand, D., Boey, E. J. H., Koh, J. Y., Nragarajan, N., Meier, R. (2017). Towards quick and cheap barcoding in the field: A specimen-based MinION™ barcode pipeline." I am a co-first author of this publication. I performed the field and bench work and contributed to the writing of the manuscript.

accurate barcode (mismatch rate = 0.09%). We recover 40/50 specimen barcodes within 2.5 hours and can estimate species composition within 1-1.5 hours. We estimate that a single run of MinION™ can generate >100 barcodes and conclude that MinION™ already out-competes other barcoding pipelines with regard to instrumentation needs and matches Sanger sequencing with regards to cost and speed. Despite containing errors, MinION™ barcodes are likely to be accurate enough for most identification needs.

3.2 Introduction

DNA barcodes are widely used for species identification, but existing pipelines for generating barcodes are not optimized for speed and efficiency when there is a need for barcoding 50-100 specimens. Yet, this is the number of samples that is commonly in need of barcoding in order to identify pests, pathogens, vectors, illegally traded species, and verifying food ingredients (Ander *et al.* 2013; Ball & Armstrong, 2006; Gonçalves *et al.* 2015; Shokralla *et al.* 2015a; Tsui *et al.* 2011). Currently, most barcodes are still obtained with Sanger sequencing which requires access to a well-equipped molecular laboratory including an ABI sequencer. Unfortunately, Sanger sequencing is fairly slow and is costly in terms of consumables and manpower. The literature is quite vague about the cost (Meier 2008), but high throughput facilities like the Canadian Centre for DNA Barcoding charges C$2,200 per plate (http://ccdb.ca/pricing/) which translates to ca. USD 17/specimen. Barcoding protocols based on next-generation-sequencing technologies such as Illumina (Meier *et al.* 2016; Shokralla *et al.* 2015b) and Roche 454 (Shokralla *et al.* 2014) have been described but they also require expensive equipment, the sequencing run times tend to be long, and/or the barcodes are only cost-effective when large numbers of specimens are barcoded simultaneously. What is arguably still missing is a barcoding pipeline that is quick and cost-effective and yet only requires minimal equipment. Such a pipeline would be welcome news for the kind of small and time-sensitive identification tasks that are common in academia, industry, and government.

Oxford Nanopore Technologies (ONT) MinION^{TM} sequencer, first introduced in 2014, paints a desirable picture: it is a palm-sized sequencing device with a USB3.0-

interface and has an initial hardware cost of at most $900 USD. MinION™ has several advantages, such as easy library preparation protocols and real time data generation. These features made MinION™ appealing when a rapid response is required as in the case of medical diagnostics and forensics (Børsting & Morling 2015; Greninger *et al.* 2015; Hoenen *et al.* 2016; Kalianski *et al.* 2015; Quick *et al.* 2015; Quick *et al.* 2016) and when data need to be generated closer to the field (Kalianski *et al.* 2015; Mikheyev & Tin, 2014). Less desirable has been its high error rates, when compared with standard procedures that yield per base accuracy of <90% (Hargreaves & Mulley 2015; Ip *et al.* 2015; Mikheyev & Tin, 2014; Sović *et al.* 2016). To this end, developments have been made, both molecular and analytical, that enable sample characterization with reasonable accuracy (Jain *et al.* 2015; Li *et al.* 2016; Loman *et al.* 2015). Such advancements have led to, for example, the reconstruction of *Escherichia coli* genome with 99.5% nucleotide identity (Loman *et al.* 2015); identifying bacteria and characterizing microbiomes (Benítez-Páez *et al.* 2016; Li *et al.* 2016; Shin *et al.* 2016), DNA fingerprinting for identifying anonymous human DNA (Zaaijer *et al.* 2016) and more recently for possibilities of sequencing DNA in space under zero-gravity conditions (Cesare 2015).

In the present study, we aimed to develop a straightforward procedure for specimen based DNA barcoding by sequencing COI amplicon from several specimens rapidly and with minimal technical expertise. Amplicon sequencing using MinION™ is not new and has been done in the context of 16S sequencing for characterizing microbiomes and obtaining species identities (Benítez-Páez *et al.* 2016; Shin *et al.* 2016). This "metabarcoding" approach using error-prone MinION™ data has till date relied on the mapping of reads onto reference database sequences. Despite its utility to

characterize bulk samples, it remains unclear how novel Molecular Operational Taxonomic Units (or MOTUs) can be identified from MinION^{TM} data when reference DNA barcodes are not available. Secondly, bulk sequencing does not allow association of a DNA barcode to a physical specimen (Meier *et al.* 2016). On the other hand, DNA barcoding as done by traditionally by Sanger sequencing has a wide appeal for any application where specimen level information is required and in the absence of reference databases good quality DNA barcodes can be grouped into novel MOTUs.

To use a MinION^{TM} sequencer for DNA barcoding we need 1) a multiplexing strategy allowing samples from several specimens to be sequenced in one sequencing run and later on discriminated, 2) straightforward procedures for bench work and 3) an analytical pipeline that demultiplexes data and determines DNA barcodes. For this, we adapt our recently described DNA barcoding procedure using Illumina platforms which allows us to sequence thousands of specimens (Meier *et al.* 2016). Here the first step is to use indexed or tagged primers to amplify DNA from specimens. The amplification procedure is simplified by 'directPCR' where PCR was done using specimen tissues with high success rates (Wong *et al.* 2014). The tagged amplicons are pooled together, purified, library-prepped and sequenced. When compared with Sanger sequencing, this approach reduces the tedious treatment of PCR products separately through purification, cycle-sequencing, and clean-up. When compared bulk sequencing (metabarcoding), it retains both intact specimens and the sequence to specimen association for further morphological work.

Although the general framework can be used for MinION™ sequencing, the downstream bioinformatics procedures developed for Illumina technologies cannot be applied due to the error rates of Nanopore sequencing. Here we develop a pipeline for this and test if we can obtain accurate DNA barcodes for species identification. We use the case of the non-biting midges (Diptera: Chironomidae) which emerge from eutrophic aquatic environments and can cause nuisance in urban areas as they increase in numbers rapidly (Cranston *et al.* 2013; Baloglu *et al.* unpublished). Due to morphological similarities between species, identification of these midges is time-consuming and requires taxonomic expertise (Epler, 2001). Timely identifications using DNA barcoding would be useful, and here we test if we can obtain these using MinION™ sequencer. We compare the results from MinION™ to DNA barcodes obtained from the same specimens using Illumina MiSeq. The pipeline developed in the study recovers all the barcodes sent for sequencing with high accuracy rapidly, with >80% of them recovered within first 2.5 hours of sequencing.

3.3 Materials and Methods

3.3.1 Sampling

I collected adult chironomid midges from one freshwater habitat (Upper Seletar reservoir— 1°24'01.3"N 103°48'21.9"E) in Singapore via kick net sampling and preserved in 70% ethanol. I then pre-sorted the specimens into morphotypes maximize the species diversity in the sample and selected 55 specimens representing multiple morphotypes for PCR amplification.

3.3.2 PCR amplification and sequencing

I amplified DNA barcodes for 55 specimens using the direct polymerase chain reactions (directPCR) protocol (Wong *et al.* 2014). I used a whole specimen for small midges (anterior to posterior 1.5–2.4 mm), two legs for medium- (2.5 to 3.4 mm) and large-sized (>3.5 mm) midges. I conducted PCRs in 50 uL reaction volume containing 5 uL of BioReady rTaq 10x Buffer, 3.75 uL of 2 mM dNTP mixture, 0.625 uL of BioReady rTaq DNA polymerase, 5 uL of 10 uM forward and reverse primers. I amplified a short 313 bp of the COI barcode using the degenerate metazoan primers [COI; mlCO1intF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' (Leray *et al.* 2013) and jgHCO2198: 5'-TAIACYTCIGGRTGICCRAARAAYCA-3' (Geller *et al.* 2013)]. Each primer was tagged with 9 bp sequence for the sequence to specimen association (Meier *et al.* 2016), and it was ensured that both forward and reverse tags were unique to the specimen. Cycling conditions were: Initial denaturation at 95 °C for 2 min, 35 cycles of denaturation at 94 °C for 30 s, annealing at 47 °C for 1 min, and extension at 72 °C for 1 min in cycles 1–34 or for 5 min in cycle 35. After PCR, I

checked the amplification results with electrophoresis on a 1% agarose gel stained with GelRed (Biotium Inc.). I then pooled together and purified the successfully amplified products with SureCleanTM (Bioline). I did an additional clean-up using 0.2% Sera-Mag Carboxylate-Modified Beads (GE Healthcare Life Sciences) in 18% PEG-8000 (polyethylene-glycol) solution at 5:4 DNA to Beads+PEG solution ratio. Purified products were sent for library preparation and sequencing using MiSeq and the MinION^{TM} sequencer. For MiSeq sequencing, the libraries were prepared using TruSeq Nano DNA library preparation kit, and 300 bp paired end reads were obtained.

Amplified product concentration was determined using Qubit fluorometer 2.0 (Thermo Fisher Scientific). One microgram of amplified product was used for MinION^{TM} library preparation using the NSK007 library preparation kit according to the manufacturer's protocol. Briefly, amplified product was end-repaired using NEBNext Ultra II End-Repair/ dA-tailing Module (New England Biolabs) at 20°C for 5 min and 65°C for 5 min. The end-repaired product was cleaned up with 1x AMPure XP beads (Beckman Coulter) and eluted in 31 µl of nuclease-free water. Adapter ligation was carried out using the NEB Blunt / TA Ligase Master Mix (New England Biolabs) together with the adapter mix and HPA from the NSK007 library kit. The ligation reaction was incubated at room temperature for 10 min. Subsequently, HPT was added, and ligation reaction was further incubated at room temperature for 10 min. Adapted DNA was purified using washed MyOne C1 beads. Adapted library was then loaded on an R9 flow cell and sequenced using the NC_48Hr_Sequencing_Run_FLO_MIN104.py workflow on MinKNOW. Total library preparation time was estimated as seventy minutes. The library was split into

two loads and loaded 24 hrs apart. Reads were base called using Metrichor (RNN SQK007 1.99) software with 2D Basecalling for SQK-NSK007. The fastq file was generated using poretools (version v0.5.1-17, option --type 2D).

3.3.3 Bioinformatics

The Illumina data were analyzed as described in Meier *et al.* (2016). However, for the MinION™ data, a new bioinformatics pipeline had to be developed that is described in Fig. 3.1. It only requires the FASTA file with the MinION™ reads, the demultiplexing file for the samples, and the installation of Python, *glsearch36*, and MAFFT v7. The script executes the following steps:

1) Primer identification and removal (step 1 in Fig. 3.1). To identify COI sequences without a reference and demultiplex them to specimen bins, we first identified the primer and then retrieved the barcode and tag sequences at the 3' and 5' ends of the primer respectively (Fig. 3.1). Due to its error rates, a search of exact primer sequences in MinION™ reads yielded few matches. Hence primers were aligned to the reads generated from MinION™ using *glsearch36* (Pearson, 1990) which allow for an alignment that is global to the query and local for the reference sequence. A generous e-value threshold (10,000) was used given the short length of the query primer sequence. This e-value ensures that many alignments of a primer sequence are found. Given that there are several degenerate nucleotides in the primers, different possible primer sequences were tested until no new matches were found. If the primer was found in the forward orientation, it was considered if it was in the first half of the MinION™ read and similarly if it was found in the reverse

orientation, it was only considered if it was identified in the second half of the read. All matches with >5 gaps were excluded. If multiple primer matches were found in the same sequence, only the one with the best identity was retained. This procedure was applied to both forward and reverse primers. For those reads where primer(s) were identified, we retrieved the "tag" and "barcode" information. The "barcode" sequence was identified as the sequence between the two primers if both primers were identified for a particular sequence, or as the 300 bp following the 3' end of the primer if only a single primer was identified. If a barcode sequence was <200 bp, the read was discarded. The "tag" sequence was identified as the 9 bp preceding the 5' end of the primer.

2) Demultiplexing (step 2 in Fig. 3.1): The tags for the barcodes were used to bin the barcodes for each specimen. All barcodes were binned either had perfect tag matches to the tags used during PCR or matches that were 1 bp away from the reference tags. The amplicons for each specimen have a unique forward and unique reverse primer tag and identifying perfect matches is straightforward as these require exact string matches. To identify close tag matches, we created a 1-bp "mutant" set for each tag, i.e., a set of tag sequences that are 1-bp away from original tag sequence (either by a substitution, insertion, or deletion). The sequences were then again demultiplexed using this "mutant tag" set. If at any point conflicting signal was obtained between forward or reverse tags, the sequence was discarded.

3) Alignment and consensus calling (step 3 in Fig. 3.1): Here, we aligned all barcodes in the specimen-specific bins and then called a consensus sequence that constitutes the MinION™ barcode for this specimen. We first merged identical demultiplexed barcodes and retained the count information. Merged reads were aligned using MAFFT v7 (Katoh & Standley, 2013) where several parameters (--auto,--ensi, --globalpair,--genafpair and multiple gap opening penalties) were tested and a gap opening penalty of 0 (--op 0) was found to be suitable. This allowed for a large number of gaps in the alignment given that we found a large number of indel errors in the MinION™ reads. Note that the large number of gaps can later be corrected by calling a consensus barcode. A majority rule consensus from the alignment was determined, and for consensus calling, gaps were treated as a fifth character and any position lacking a base-call with > 0.5 ratio was considered ambiguous ("N"). All positions called as a gap in the consensus sequence were excluded. After this step, the consensus sequence was determined which was further trimmed to exclude 10 bp on both ends of the barcode. This step was desirable because preliminary results indicated large numbers of errors at the barcode edges.

4) Validation (step 4 in Fig. 3.1): This is an optional step that is only feasible if the correct barcodes for the specimens are known. Here, we compared the MinION™ barcode with a known barcode (a MiSeq barcode in our experiment). The MinION™ barcode was aligned to the Illumina barcodes using MAFFT v7 –op 0 (Katoh & Standley, 2013), and we calculated the (a) number of mismatches (b) number of gaps introduced (in and outside of homopolymer regions) and the (c) number of ambiguous bases in the

MinION™ barcode. The results for all 50 barcoded specimens are shown in Fig. 3.1.

The sets of barcodes obtained using MinION™ and Illumina platforms were aligned using MAFFT v7 (Katoh & Standley, 2013) under default parameters. The number of species was determined using 1) Automated Barcode Gap Discovery (ABGD, (Puillandre *et al.* 2012)) and SpeciesIdentifier (Meier *et al.* 2006), using *p*-distances at various thresholds (Srivathsan & Meier, 2012). For ABGD, the initial pairwise distance matrix was obtained using MEGA6 (Tamura *et al.* 2013), and parameters described by Ratnasingham & Hebert (2013) were used, except the maximum prior intraspecific divergence value, which we set at 0.2. For these MOTU delimitations, we treated gaps as missing data.

3.3.4 Effect of coverage on accuracy of barcode

To determine the coverage required to obtain an accurate barcode, we resampled the MinION™ reads for each specimen with >100 sequences. The reads were resampled at 10-100X coverage (10,000 iterations) and the same consensus calling procedure that was applied to the full dataset was used to determine the barcodes for the subsets. The accuracy of the consensus barcodes was determined by comparing them to the corresponding Illumina barcode using the same criteria described above (number of mismatches, gaps, ambiguous bases). For every specimen, the values across various iterations were averaged.

3.3.5 Assessing error rate biases in homopolymeric regions

Initial examination of the DNA barcodes revealed that consensus barcodes contained a significant number of indel errors and a visual examination suggested that they were concentrated in homopolymeric stretches of COI. We thus analyzed the extent to which homopolymers contributed to indel errors using a custom script: here MAFFT v7 was used for pairwise alignment between a MinION™ barcode and corresponding specimen Illumina barcode. Wherever a gap was observed (either insertion or deletion), we determined if a homopolymer was present at the site and measured the length of the homopolymer on the Illumina barcode. This was done separately for A/T and G/C homopolymers. We define homopolymers as stretches of identical nucleotides that are at least 3 bp long, but we also report the results for dinucleotides AA/TT and GG/CC.

3.3.6 Effect of run time on sample characterization

Lastly, we assessed the relationship between MinION™ run time and the number of reliable barcodes obtained. For each time point, specimen-specific datasets were created by calling consensus barcodes as described previously. These datasets could be generated using read time information.

3.4 Results

The success rate for directPCR from 55 midge specimens was 90.9% with 50/55 specimens yielding a PCR product. The products were sent for sequencing using both Illumina MiSeq and Oxford Nanopore MinION^{TM} sequencer. The processing of the MiSeq reads followed Meier *et al.* (2016), and the sequences represented three MOTUs according to ABGD and Objective Clustering [between *p*-distances 0.7-13.6% (ABGD) and 1.8-15% (Objective Clustering)]. The three MOTUs were represented by 25, 21, and 4 specimens, respectively. When I compared the MOTUs to the existing midge barcode database, I found that the rare MOTU represented a species whose barcode was previously unknown.

The same pool of PCR products yielded 14,772 reads with Oxford Nanopore MinION^{TM}. Of these 8,683 (58.8%) could be demultiplexed unambiguously and assigned to specimens (Fig. 3.1). A consensus barcode could be determined for all 50 specimens with coverage >10X. After trimming 10 bp at the edges, we obtained the same 3 stable MOTUs that were obtained with Illumina barcodes [between *p*-distances 0.7-13.6% (ABGD) and Objective Clustering: 1.8-15.1% (Objective Clustering)]. Comparison of the MinION^{TM} barcodes and Illumina barcodes revealed only one mismatch in the barcodes for the 50 specimens (Fig. 3.1, Table). However, each MinION^{TM} consensus barcode retained errors; mostly in the form of insertions and deletions that were 4-10 bp in length and we also observed up to three ambiguous bases (“N”; see Fig. 3.1).

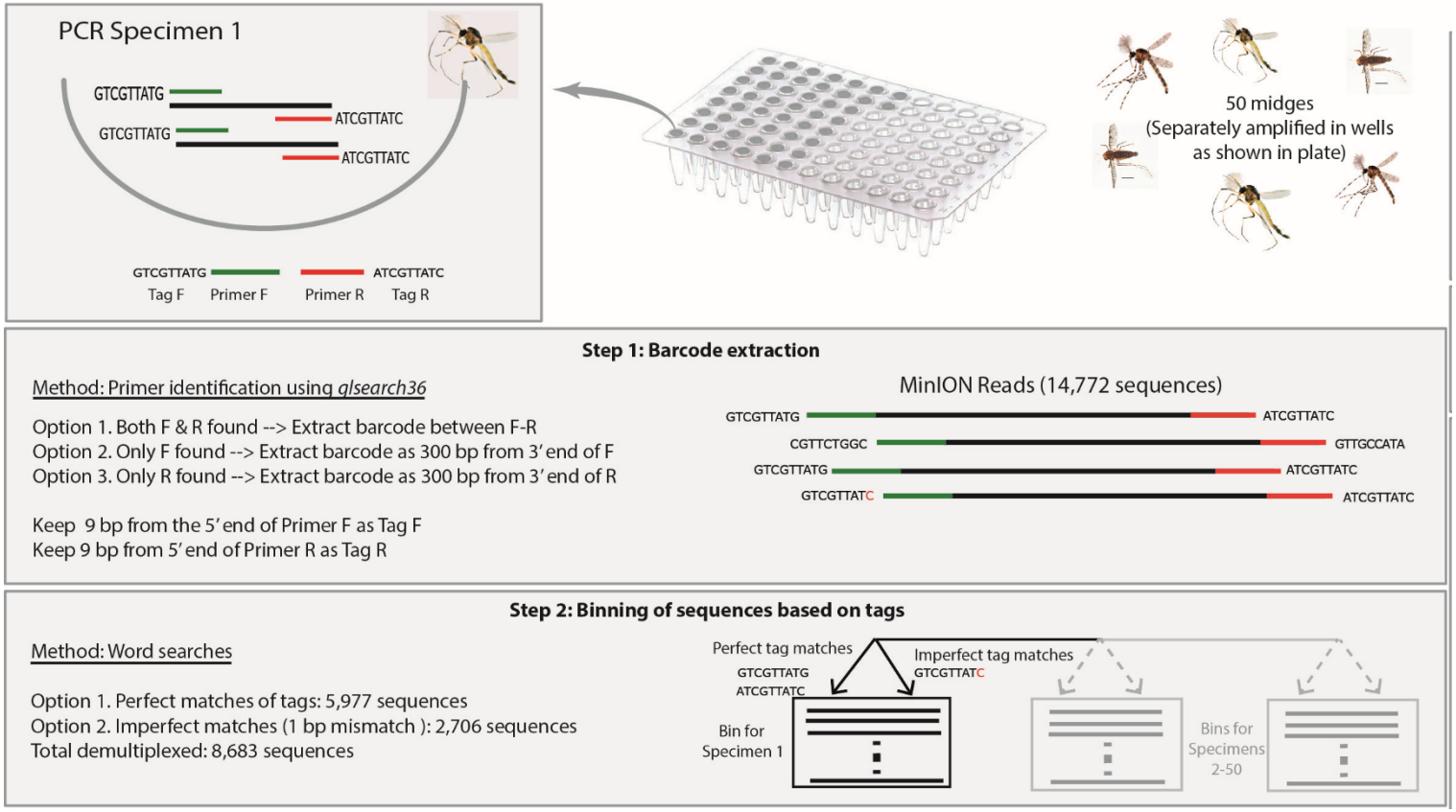
PCR Specimen 1
GTCGTTATG
ATCGTTATC
GTCGTTATG
ATCGTTATC
GTCGTTATG
ATCGTTATC
Tag F
Primer F
Primer R
Tag R
50 midges
(Separately amplified in wells
as shown in plate)
Step 1: Barcode extraction
Method: Primer identification using glsearch36
Option 1. Both F & R found --> Extract barcode between F-R
Option 2. Only F found --> Extract barcode as 300 bp from 3' end of F
Option 3. Only R found --> Extract barcode as 300 bp from 3' end of R
Keep 9 bp from the 5' end of Primer F as Tag F
Keep 9 bp from 5' end of Primer R as Tag R
MinION Reads (14,772 sequences)
GTCGTTATG
ATCGTTATC
CGTTCTGGC
GTTGCCATA
GTCGTTATG
ATCGTTATC
GTCGTTATC
ATCGTTATC
Step 2: Binning of sequences based on tags
Method: Word searches
Option 1. Perfect matches of tags: 5,977 sequences
Option 2. Imperfect matches (1 bp mismatch): 2,706 sequences
Total demultiplexed: 8,683 sequences
Perfect tag matches
GTCGTTATG
ATCGTTATC
Imperfect tag matches
GTCGTTATC
Bin for
Specimen 1
Bins for
Specimens
2-50

Figure 3.1: Graphical summary of DNA barcoding pipeline using MinION™ sequencer.

3.4.1 Effect of coverage on accuracy of barcode

Even at low coverage (10X), substitution errors are largely eliminated (median substitution error rate was 0.09%) and it further reduced to 0.02% at 30x and 0.01% at 50x; Fig. 3.2). Indel error rates declined with coverage from a median of 3.08% (10X) to 2.48% (30X), but a further increase in coverage did not improve the quality of the barcode. Lastly, the median percentage of "N" bases, i.e., ambiguities, declined from 2.01% to 0.62% and 0.40% when coverage was increased from 10X to 30X and 50X.

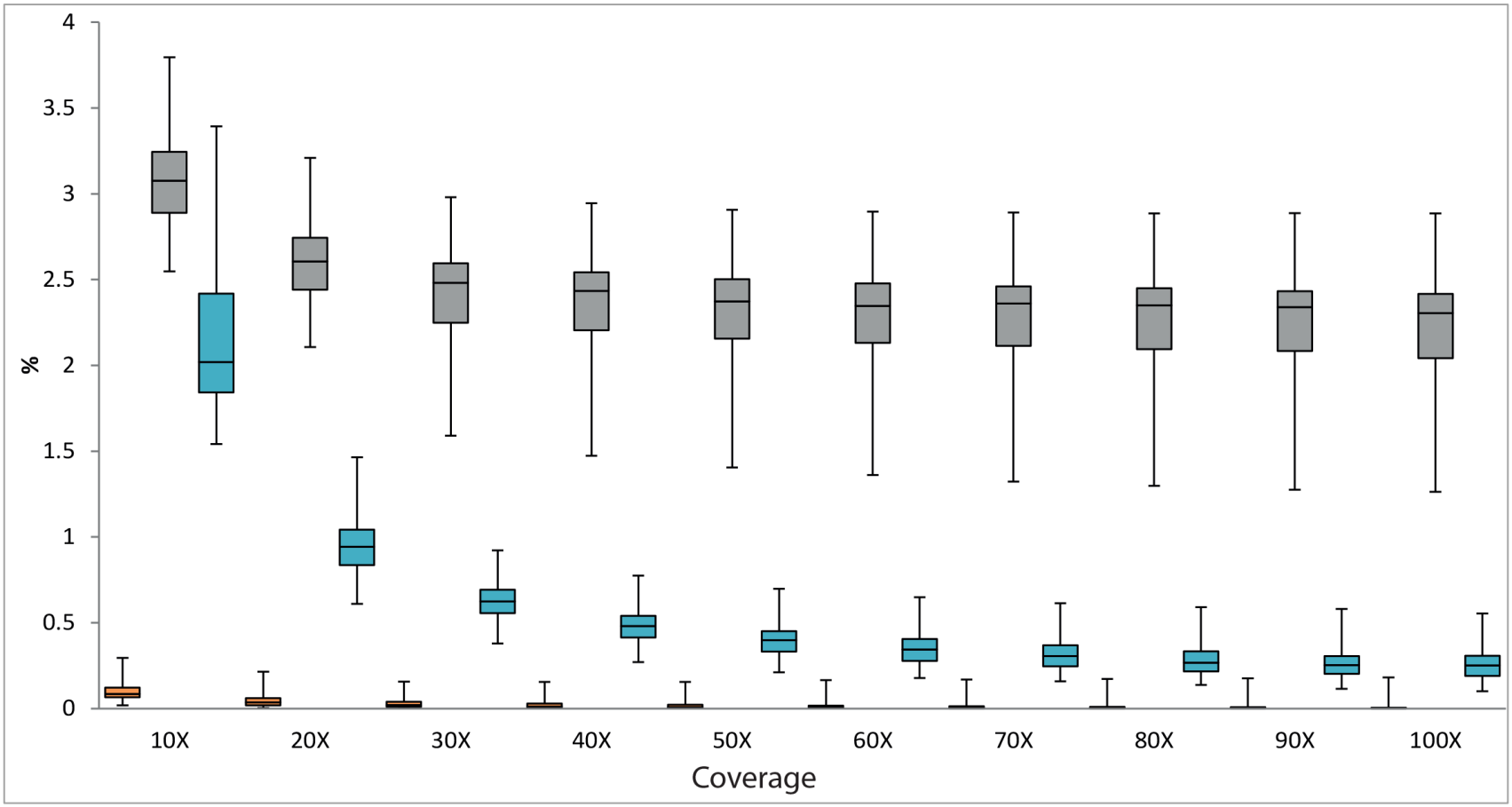

Figure 3.2: Box plots representing effect of coverage on barcode accuracy (orange: mismatch errors, gray: indel errors and blue: ambiguities, i.e., "N" bases). Resampling at various depths reveals a reduction in errors as coverage is increased.

3.4.2 Assessing error rate biases in homopolymeric regions

Most indel errors were found in regions of homopolymers. For instance, of the total of 324 indels introduced in the barcodes, only 5 were outside of homopolymer regions. When A/T and G/C homopolymer stretches were separately examined, we found that A/T homopolymer tracts were prone to insertions such that >90% of the homopolymer tracts of A/T that were >5 bp long had at least one insertion in the MinION^{TM} barcode (Table 3.1). On the other hand, deletions were largely limited to C/G homopolymers, where shorter homopolymers that were 3-4 bp in length contained errors, but at a lower frequency (50-60%).

Table 3.1: Effect of length of homopolymeric stretches in COI on indel errors

Length of homopolymer	**A/T**			**G/C**		
	Occurrence	**Insertion**	**Deletion**	**Occurrence**	**Insertion**	**Deletion**
1*	3922	0	1	3103	1	0
2	1461	0	0	706	1	4
3	420	0	0	137	0	73
4	204	0	0	51	0	27
5	4	1	0	0	NA	NA
6	84	81	0	0	NA	NA
7	12	10	0	1	0	1

3.4.3 Effect of run time on sample characterization

MinION™ allows for real-time sequencing and sequences can be analyzed at any point in time. We assessed how much data is needed and how many specimens are recovered at 10, 20 and 30X barcode coverage. The demultiplexed data from various time points can be used to generate the consensus barcode, and we find that within the first 12 hours of sequencing nearly all specimens are demultiplexed (10X: 50/50, 20X: 48/49 and 30X: 47/48, Fig. 3.3a). Moreover, 80% of the barcodes can be obtained within 2.5 hours at 10X coverage, 4 hours at 20X and 8 hours at 30X coverage. In this run, specimens from all three species are represented after 39 minutes at 10X coverage, 66 minutes at 20X coverage and 101 minutes at 30X coverage. If species compositions are measured at different time points, we find that reasonable compositional estimates could be obtained using a 10X coverage criterion after sequencing for 1-1.5 hours, while ~2.5 hours are required using a 20X criterion and 3-3.5 hours are required using a 30X criterion (Fig. 3.3b).

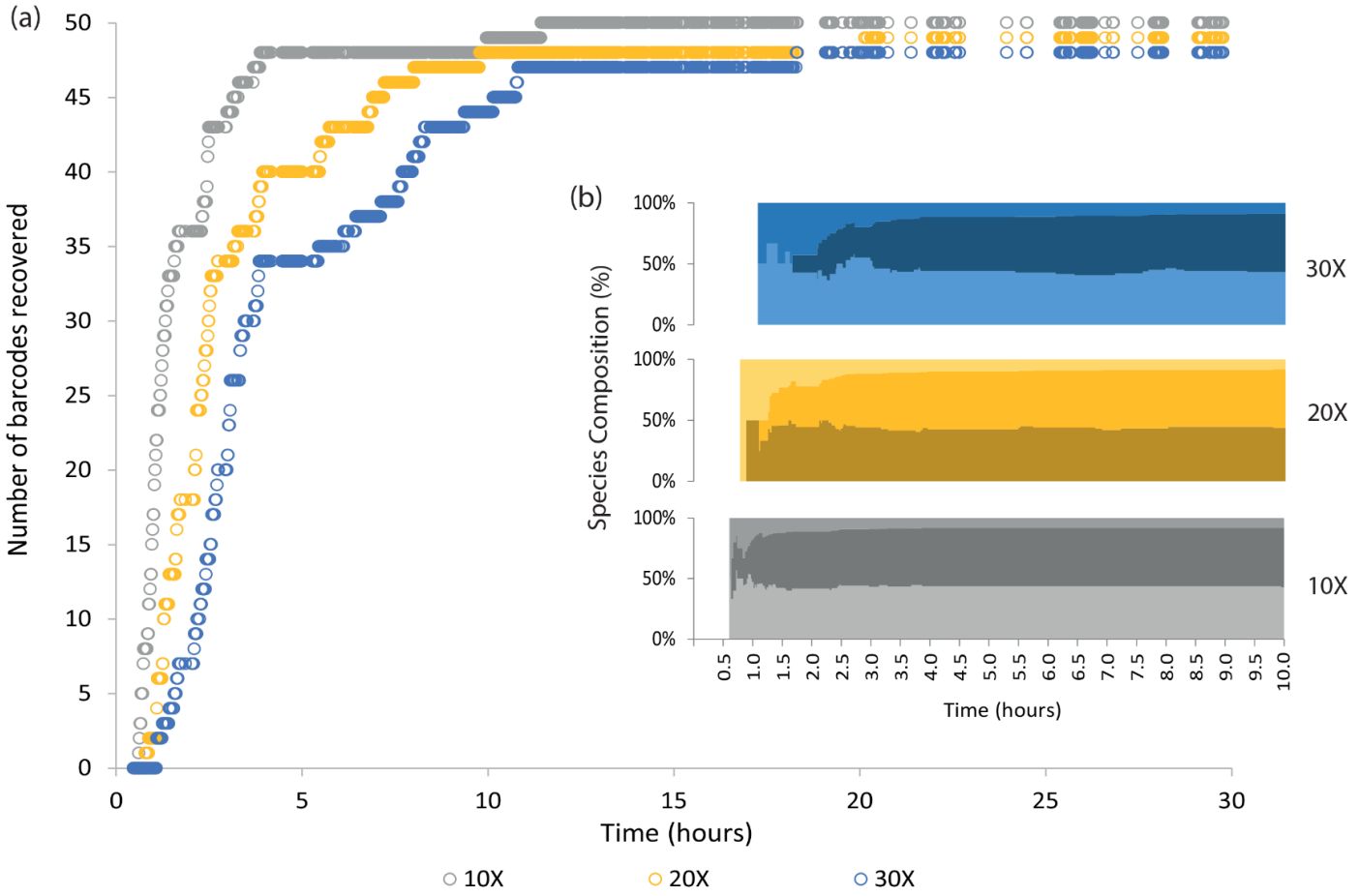

Figure 3.3: (a) Number of barcodes generated over time at 10X (grey), 20X (yellow) and 30X (blue) coverage.
(b) Species compositions estimated till 10 hours, color coding is identical to (a) and shades represent individual species.

3.5 Discussion

We here establish a pipeline for the *de novo* generation of DNA barcodes using the Oxford Nanopore MinION™ sequencer. We find that within 39 minutes of sequencing all three species are recovered, relative species abundances can be estimated within 1.5 hours, barcodes for 80% of the specimens are obtained within 2.5 hours, and 12 hours of sequencing yields barcodes for all specimens. Overall, the MinION™ pipeline can generate reliable DNA barcodes quickly, at low cost, and in a laboratory that does not require expensive equipment. These are attractive properties for identifying pests, pathogens, vectors, illegally traded species, and verifying food ingredients. We would argue that currently the MinION™ already out-competes other barcoding pipelines with regard to instrumentation needs and it at least matches Sanger sequencing with regard to speed and cost. However, raw MinION™ reads suffer from lower sequence quality. This is a concern, but we would argue that MinION™ barcodes are sufficiently accurate for most identification purposes.

Speed. All barcoding procedures require gene amplification and the cleanup of PCR products. For Sanger sequencing this cleanup is specimen-specific while all NGS-based techniques (including MinION™) can save time by cleaning up pooled products; i.e., Sanger sequencing requires more time unless liquid handling robots are used. Following clean-up, the library preparation for the MinION™ requires <1.5 hours while Sanger sequencing requires cycle sequencing and another round of cleanup for the individual products. By using liquid handling robots, the latter can apparently be accomplished in 35 minutes for short barcodes (Ivanova *et al.* 2009), but most protocols require ~2.5 hours. Capillary sequencing of short Sanger barcodes

requires another 45 minutes while the MinION™ sequencer provides species profiles in 1-1.5 hours of sequencing and specimen profiles after 2.5 hours (80% of specimens). So, overall MinION™ and Sanger sequencing require similar amounts of time. The remaining NGS barcoding protocols (Illumina: Meier *et al.* 2016; Shokralla *et al.* 2015b; Roche 454: Shokralla *et al.* 2014) are considerably slower because library preparation protocols and sequencing are more time-consuming.

Cost. The cost of a MinION™ barcode in our experiment (50 specimens) was approximately USD 17/specimen (cost for flow cell; 675 USD, reagent costs: 170 USD). However, our bioinformatic analysis revealed that the quality of barcodes did not improve markedly beyond 30x coverage; i.e., several of our samples had overly luxurious coverage of >100X. If 100 samples had been pooled in this run and similar distribution of reads had been obtained (19-609X), we predict that 98% of the specimens would have been recovered at 10X coverage, 94% at 20X and 90% at 30X coverage. This implies that the number of samples that can be multiplexed in one MinION™ flowcell can be increased by at least two-fold; i.e., the cost of a barcode would drop to USD 8.5 per specimen. This is cheaper than the cost of Sanger barcodes (USD 15/specimen: http://ccdb.ca/pricing/), but considerably more expensive than the NGS barcodes obtained with Illumina sequencing (<1 USD: Meier *et al.* 2016). However, the low cost for Illumina barcodes can only be achieved when thousands of specimens are multiplexed. Note also that based on the experience with new NGS techniques, it is likely that the cost of MinION™ runs will decline rapidly so that the cost of MinION™ barcodes will become even more competitive.

Instrumentation. This is where the MinION™ out-competes all other techniques. The instrumentation needs at the PCR and clean-up stage are similar across all pipelines (pipettes, thermocycler) but the MinION™ sequencer has significant advantages in the subsequent procedures. The library preparation procedure requires the following equipment in addition to pipettes: a magnetic rack, a heating block, and a rotator. We believe that the latter two instruments are not likely to be essential, given that the heating can be done at body temperature (37ºC) and a biologist can physically rotate the tubes. Moreover, the sequencing instrument is considerably cheaper and smaller than ABI capillary, Illumina, Ion Torrent, or Roche 454 sequencers. These differences are not trivial because they indirectly affect the speed of sequencing: there are usually waiting times for getting access to expensive equipment because they have to be fully utilized to be cost-effective. On the other hand, most laboratories could afford multiple MinION™ sequencers (USD 1000) and leave one idle for urgent identification tasks. The low cost and small size make MinION™ also very suitable for establishing improvised laboratories under difficult conditions. Here, the MinION™ could complement the small thermocycler that is suitable for work in the field (Marx, 2015). Indeed, the total equipment cost for such a laboratory can now be <USD 3000 and technologies like MinION or the smartphone friendly SmidgION get biologists closer to the vision of being able to obtain sequences in the field.

Barcode quality. This remains the biggest drawback of MinION™ barcodes. The main concern is not their length (313 bp) because MinION™ is particularly suitable for long-range barcoding so that full-length COI barcodes (658 bp or even longer) could be obtained. Instead, it is currently unrealistic to expect MinION™ barcodes to

be error free. Our error-correction pipeline eliminates many but not all sequencing errors. The main problem is indels that are concentrated in homopolymer regions (Fig. 3.1, Table) while there are only a few base mismatches (error rate = 0.007%). Fortunately, the error rates decline as coverage increases. For example, at the 10X coverage, the consensus MinION^{TM} barcode is at most 0.3% away from the corresponding Illumina sequence (Fig. 3.2) while this is almost halved as coverage increases to ~30X. However, even at high coverage, all specimen barcodes have errors when compared to Illumina barcodes (Fig. 3.1). These errors are fortunately systematic and concentrated in homopolymeric regions of COI (see also Ashton *et al.* 2015; Jain *et al.* 2015; Loman *et al.* 2015) and the nature of the error can be predicted because it depends on the base composition. A/T rich regions are prone to insertions while G/C rich regions are prone to deletions (Table 3.1). Given that the Metazoa barcode (COI) is a protein-encoding gene, most errors can be identified and circumvented by treating gaps as missing. A second approach could involve masking of homopolymeric regions of >=3bp length in alignments. However, these errors nevertheless reduce the utility of MinION^{TM}, and it is currently not possible to generate high-quality barcodes that would be suitable for addition to reference databases.

3.6 Future improvements

Our bioinformatics pipeline is quite straightforward and only requires the input of the read file obtained from the MinION^{TM} and a demultiplexing file specifying the specimen-specific tags. The pipeline is also very fast and returns the set of consensus barcodes in <3 minutes using a standard laptop computer (using <1GB RAM, two cores). However, this pipeline could be further optimized. Currently, only ca. 60% of

the reads are demultiplexed because the amplicon tag information is ambiguous on the remaining 40%. To demultiplex more reads, an iterative process could be used. The reads with ambiguous tags could be mapped onto the consensus barcodes obtained during the first pass through the data. After mapping, additional tags could be identified that have more than one sequencing error. The corresponding reads could then be added to the bins for the respective specimens. This would increase coverage or allow for the barcoding of an even larger number of specimens.

Changes in lab procedures could similarly improve the performance of MinION™ for DNA barcoding. For example, longer tags could be used to increase the proportion of reads that can be demultiplexed. Another obvious target for improvement is the library preparation step. The current study uses "2D" reads that were generated using a commercially available library preparation kit from ONT. 2D workflow makes use of a hairpin adapter to read each molecule in both forward and complementary direction thus yielding a consensus of a higher quality (Risse *et al.* 2015). This was shown to reduce error rates from >20% to ~8% (Ip *et al.* 2015; Jain *et al.* 2016; Sović *et al.* 2016). However, error-rates can be further reduced by using "INC-seq" which uses circularized template DNA and rolling circle amplification to generate high-quality consensus reads (>97% accuracy) (Li *et al.* 2016). Once suitable for labs and facilities with limited technical expertise, these improvements may yield MinION™ barcodes that are of similar quality as Sanger and Illumina barcodes.

CHAPTER 4[3]

NGS barcoding reveals high resilience of a species-rich chironomid fauna (Diptera) against invasion from adjacent freshwater reservoirs

4.1 Abstract

Macroinvertebrates such as non-biting midges (Chironomidae: Diptera) are an important component of freshwater ecosystems. However, they are often neglected in conservation research because invertebrate species richness is difficult and expensive to quantify with traditional morphological methods. Here, I use a newly developed cost-effective method for barcoding ("NGS barcodes") to test the resilience of the midge fauna of Singapore's last remaining swamp forest (Nee Soon Swamp Forest) against invasions from three surrounding human-made reservoirs. The study of species diversity, species turnover, and resilience is based on >14,000 individually-barcoded specimens that are also used to understand how environmental, spatial, and temporal variables may shape the swamp forest community. I find that the swamp forest maintains a rich and largely unique fauna with 259 observed species (estimated diversity>400) in a small area (90 ha) that is <20% the combined size of the reservoirs. All reservoirs combined contained only 37 species (estimated diversity:

[3] A version of this chapter is in prep as "**Baloglu, B.**, Clews, E., Meier, R. (2017). NGS barcoding reveals high resilience of a species-rich chironomid fauna (Diptera) against invasion from adjacent freshwater reservoirs." I am the first author of this publication. I performed the bench work, data analysis and writing of the manuscript.

87). The resilience of the swamp forest appears high because only 7 of the 259 species are shared and the shared species are not particularly abundant (3% all specimens) despite the proximity of these two habitats. 5 of these shared species were found in higher abundances in the reservoirs. Within the swamp forest, redundancy analysis revealed that dissolved oxygen levels, pH, stream depth, latitude, and the sampling year were significant factors influencing chironomid community structure, but these environmental factors explained only ~16% of the variance. Analysis with LME showed that the total species richness decreased with conductivity. My study documents that natural habitats can be very resilient against the invasion of species from neighboring urban environments. I also document how NGS barcodes can be used to integrate diverse and specimen-rich invertebrate taxa into habitat assessments.

4.2 Introduction

Freshwater ecosystems are under threat worldwide from habitat destruction, pollution, and climate change and the global freshwater biodiversity is declining much more rapidly than the diversity of many stressed terrestrial ecosystems (1-8% species loss per decade; Ricciardi & Rasmussen, 1999). Such loss of freshwater biodiversity affects food webs, nutrient cycling, climate, air quality, and water supply (Gleick, 1993; Vaughn, 2010); i.e., the losses impact many important ecosystem services and affect large geographic areas beyond the boundaries of the affected freshwater ecosystems (Holmlund & Hammer, 1999). One problem with monitoring the health of freshwater systems is the lack of efficient and rapid assessment tools for species-rich invertebrates (Raunio *et al.* 2011). Much of the assessment work is driven by cost considerations and relies on comparatively species-poor taxa such as selected macroinvertebrate groups and fish (Resh, 2008). More ubiquitous and species-rich invertebrate taxa that often constitute more of the biomass are either omitted or studied at low taxonomic resolution (e.g., genus, family).

Aquatic invertebrates are essential for the health of ecosystems because they occupy many niches and are known to display rapid community responses to water quality and habitat changes (Hynes, 1960). Because aquatic invertebrates are largely immobile, they are also particularly suitable for assessing the water quality in a particular sampling location (Reynoldson & Meltcafe-Smith, 1992). Moreover, they tend to have shorter generation times and thus tend to respond more quickly to change (e.g., recover faster from disturbance) thus enabling a more rapid assessment of environments than long-lived species (Clarke, 1993; Resh 2008). Among aquatic

invertebrates, non-biting midges (Chironomidae: Diptera) are a particularly important indicator taxon because these midges are found in most freshwater habitats worldwide (Pinder, 1986; Pinder, 1995), are particularly species-rich (sometimes having more species than all other insect species in an aquatic environment combined: Heino & Paasivirta, 2008), and have high abundance. Chironomids are also an important food source for higher-order predators, such as odonates, fish, and birds, and act as important decomposers of organic matter (Armitage, 1995; Jones & Grey, 2004; Nicacio & Juen, 2015). High abundance, species richness and important ecosystem roles should thus make chironomids particularly attractive bioindicators for assessing the health of freshwater habitats. However, reliable sorting/identification to species using traditional techniques is so expensive that in most studies chironomids are either only identified to genus/subfamilies, or they are altogether neglected in bioassessment and conservation studies (Raunio *et al.* 2011). Habitat assessment studies instead focus on plants, vertebrates, and those invertebrate groups that are easier to identify (Fattorini, 2011).

The cost of midge identification via morphology is high because it usually requires the dissection and study of specimens that are mounted on microscopic slides (Epler, 2001; Carew *et al.* 2007; Cranston *et al.* 2013) and this can require 15–20 min per specimen (Wong *et al.* 2014) even if the sorting and identification is done by an experienced taxonomist. An additional complication is that usually larval midges are collected while the species names and much of the identification literature is for adults. As a result, species-level chironomid data is rarely used in conservation and bioassessment studies although such species-level information is highly desirable because different chironomid species vary in their sensitivity to habitat and environmental parameters (Pettigrove & Hoffmann, 2005; Marziali *et al.* 2010; Carew

et al. 2011; Nicacio & Juen, 2015). For instance, congeners in *Cricotopus*, *Polypedilum*, and *Tanytarsus* have been shown to differ considerably with regard to their tolerance to heavy metals, pesticides, and nutrient-levels (Cranston, 2000; Riva-Murray *et al.* 2002). This means that identification to genus or subfamily comes with considerable information loss for habitat assessment.

Due to the problems with species-level sorting, little is known about species turnover of chironomid communities (Delettre & Morvan, 2000) in tropical environments, and even less is known about the environmental variables that structure communities (Cranston *et al.* 1997; Helson *et al.* 2006). More information is available for temperate lake systems (e.g., Brodersen & Lindegaard, 1999; Wazbinski & Quinlan, 2013; Tarkowska-Kukuryk & Mieczan, 2014) while the response of chironomids to physicochemical variables in the tropics is poorly understood. Yet, it is likely to differ from those in temperate regions given that the streams in the tropics are known to receive more intense rainfall and have higher and more stable water temperature than their temperate counterparts (Boulton *et al.* 2008). Moreover, tropical lakes of moderate to great depth show permanent water stratification (Lewis, 1996). Based on the limited information that is available, chironomid species appear to be more widely distributed in tropical streams compared to their temperate counterparts (Cranston *et al.* 1997; Coffman & de la Rosa, 1998).

Here I use Next Generation Sequencing (NGS) for overcoming the species-sorting impediment in chironomid midges (adults and larvae) to study the resilience of a swamp forest midge community against invasion from neighboring reservoirs. NGS barcodes can be used for fast and large-scale species sorting with apparently little

compromise with regard to accuracy (Wong *et al.* 2014; Meier *et al.* 2016) given that DNA barcoding are capable of distinguishing most species of Chironomidae (>80-90% congruence; Sharley *et al.*, 2004; Carew *et al.*, 2007; Taenzler *et al.* 2012; Montagna *et al.* 2016). DNA barcodes have thus been successfully used for revealing the community patterns of taxonomically complex chironomid taxa (Carew *et al.* 2005; Ekrem *et al.* 2007; Sinclair & Gresens, 2008; Stur & Ekrem, 2011; Silva *et al.* 2013). However, one should keep in mind that such barcodes are likely to underestimate the species diversity of recently diverged species and overestimate species diversity for those species with diverging allopatric populations (Burns *et al.* 2007; Ward, 2009; Will & Rubinoff, 2004; Meyer & Paulay, 2005; Meier *et al.* 2006).

Currently, the biggest obstacle to the large-scale use of DNA barcoding for assessing specimen-rich invertebrate communities is the high cost of obtaining barcodes via Sanger sequencing. Very high barcode cost per specimen (normally, 16-34 USD; 8-17 USD, if submitted to International Barcode of Life Project, Meier *et al.* 2016) limits our ability to use Sanger sequencing for obtaining barcodes for thousands of specimens (Stein *et al.* 2014; Shokralla *et al.* 2015b). However, the cost-related problems can be addressed by using NGS, with NGS barcode costing as little as ~0.29 USD per specimen depending on the techniques used (Meier *et al.* 2016; Baloglu *et al.* unpublished).

I here use NGS barcodes for >14,000 chironomids to study the species richness and turnover between adjacent urban (reservoirs) and a wild habitat in Singapore (swamp forest). Chironomids living in the artificial reservoirs have been regularly collected as part of freshwater quality monitoring, and the reservoirs involved (Upper

Peirce, Lower Peirce, Upper Seletar Reservoir) are known have similar environmental conditions (Low, 2010) due to similar historical backgrounds and water flow. Less than 1 km away is Nee Soon Swamp Forest with a size of less than 20% of the reservoirs (90 ha). The plant and vertebrate species in the swamp forest had been previously studied, but its chironomid fauna is unknown. This swamp forest is the last remnant of its kind in Singapore, and it provides an ideal opportunity for studying the resilience of a Southeast Asian swamp forest against the anthropogenic influences by adjacent reservoirs.

There are nearly 5,000 described species of chironomids (Cranston & Martin, 1989), and the first aim of my study is to quantify the species diversity of the chironomid fauna in a swamp forest remnant using NGS barcoding applied to a large sample. The second aim is to compare the chironomid fauna of adjacent urban and wild habitats. With increasing urbanization, an increasingly important challenge is the replacement of native species with non-native species that often have less specific habitat-requirements and are thus more widespread. This replacement can lead to more homogeneous biotic communities by diminishing the faunal distinctions among regions (Blair, 2001). As shown for some taxa in urban-gradient studies (plants, Kuehn & Klotz, 2006; ants, Holway & Suarez, 2006; Roura-Pascual *et al.* 2010; birds, Blair & Johnson, 2008), native species are being replaced with non-native species upon the invasion of natural habitats. However, there is little data for invertebrates in general and even fewer data for chironomids in particular. Much of the midge research focuses on nuisance species while their impacts on the adjacent native fauna are largely ignored (Armitage, 1995; Haenel & Chown, 1998; Jacobsen & Perry, 2007; Failla *et al.* 2015). I here address this shortcoming by quantifying the species

richness and by testing whether urban chironomid species invade and potentially displace the native fauna in the swamp forest habitat. My third aim is to understand species turnover within the swamp forest. I use the available environmental information to study the correlation of these parameters with the distribution of chironomids in a tropical swamp forest using multivariate statistical analyses. I specifically ask (i) what physicochemical variables determine the chironomid community in the different streams, (ii) whether sampling distance plays a role in chironomid community structure at this small landscape, (iii) whether there are any species intermixing between the habitats, and if so, (iv) whether the urban reservoir species invade the adjacent wild habitats?

4.3 Materials and Methods

4.3.1 Sampling

Swamp forest – Sampling larvae. Between October 2013 and September 2014, 40 freshwater streams in Nee Soon Swamp Forest were sampled (Table 4.1) by Tropical Marine Science Institute (TMSI). These sites are located within the Central Catchment Nature Reserve (CCNR) and thus protected. The CCNR covers 20 km^2 and is surrounded by highways and major roads as well as residential areas. For each sampling site, 12 physical and chemical parameters (cross-sectional area, stream width, stream order, stream velocity, stream discharge, temperature, conductivity, maximum depth, average depth, turbidity, dissolved oxygen, and pH) were collected using an Odeon turbidity meter, YSI 556 & YSI6600V2-4 Multimeter and a Hach FH950 velocity flow meter. Also, GPS coordinates for each site were recorded using TRIMBLE GeoXH 6000 series GPS. As the freshwater streams in Singapore are short, narrow and shallow (i.e., ranging from 1 to 2 m width and 10–80 cm depth) (Yeo & Lim, 2011), kick nets were used at each site, where chironomid larvae were collected along three replicates of 10 m stretches. All midge larvae (n= 6,620) were preserved in 70% EtOH

Table 4.1: Site name, code (numeric code) and location (geographical coordinates) for the study sites. Kick net sampling was used for Nee Soon Swamp Forest sites and sediment grab was used for the reservoir sites.

Site name	Site code	Latitude (°) decimal	Longitude (°) decimal	Year	Final analysis
Nee Soon	NS01	1.37606	103.80605	2013	✓
Nee Soon	NS02	1.37761	103.80515	2013	✓
Nee Soon	NS03	1.37994	103.80493	2013	✓
Nee Soon	NS04	1.37951	103.80166	2013	
Nee Soon	NS05	1.3803	103.80339	2013	✓
Nee Soon	NS06	1.38149	103.80500	2013	
Nee Soon	NS07	1.38255	103.80512	2013	✓
Nee Soon	NS08	1.38035	103.79716	2013	
Nee Soon	NS09	1.38151	103.80002	2013	
Nee Soon	NS10	1.38326	103.80240	2013	
Nee Soon	NS11	1.38462	103.80196	2013	
Nee Soon	NS12	1.38379	103.80410	2013	
Nee Soon	NS13	1.38516	103.80534	2014	✓
Nee Soon	NS14	1.38421	103.80502	2013	✓
Nee Soon	NS15	1.38584	103.80585	2013	✓
Nee Soon	NS16	1.38718	103.80716	2013	✓
Nee Soon	NS17	1.38854	103.80836	2013	
Nee Soon	NS18	1.39054	103.80915	2013	✓
Nee Soon	NS19	1.39136	103.80956	2014	✓
Nee Soon	NS20	1.3919	103.81075	2013	✓
Nee Soon	NS21	1.39454	103.81295	2014	✓
Nee Soon	NS22	1.39645	103.81330	2014	✓

Nee Soon	NS23	1.39677	103.81324	2014	
Nee Soon	NS24	1.39303	103.80400	2013	✓
Nee Soon	NS25	1.39487	103.80840	2013	✓
Nee Soon	NS26	1.3968	103.81040	2013	✓
Nee Soon	NS27	1.39875	103.81301	2014	✓
Nee Soon	NS28	1.39977	103.81283	2014	
Nee Soon	NS29	1.39925	103.80854	2014	✓
Nee Soon	NS30	1.39907	103.80996	2014	✓
Nee Soon	NS31	1.40002	103.81090	2014	✓
Nee Soon	NS32	1.40005	103.81171	2014	✓
Nee Soon	NS33	1.38188	103.81200	2013	✓
Nee Soon	NS34	1.38425	103.81377	2014	✓
Nee Soon	NS35	1.38346	103.81187	2014	
Nee Soon	NS36	1.38466	103.81123	2013	✓
Nee Soon	NS37	1.38675	103.81035	2013	
Nee Soon	NS38	1.38844	103.80964	2013	
Nee Soon	NS39	1.39029	103.81012	2013	
Nee Soon	NS40	1.39268	103.81170	2013	✓
Lower Peirce	LP	1.37236	103.81335		✓
Upper Peirce	UP	1.35787	103.79027		✓
Upper Seletar	USR	1.40516	103.80802	2013- 2014	✓

Swamp forest – Sampling adults. As part of a long-term insect biodiversity project, one site in the Nee Soon Swamp Forest was sampled for adults using two Malaise traps between 2012 and 2013. Alcohol-preserved chironomid adults (n= 1,551) were extracted from these samples.

Reservoirs. Reservoir chironomids have been continuously sampled as part of freshwater quality monitoring using a sediment grabber. I here include those samples that were collected during the same time periods that were covered by the swamp forest study. They are Upper Seletar (n = 3,647: October 2013 to June 2014), Upper Peirce (n=1,058: January to April 2014), and Lower Peirce (n =1,308; January to April 2014). Environmental variables were not collected in reservoirs. Therefore, I used the reservoir chironomids only for the species diversity and turnover analysis.

4.3.2 PCR amplification and NGS barcoding

I amplified NGS barcodes for each specimen using the direct polymerase chain reaction (directPCR) protocol described in Wong *et al.* (2014) that avoid the time-consuming and costly step of DNA extraction. PCR reactions were carried out in 20 µL volumes containing 2 µL of BioReady rTaq 10x Buffer, 1.5 µL of 2 mM dNTP mixture, 0.25 µL of BioReady rTaq DNA polymerase, 2 µL (1 mg/mL) of BSA and 2 µL of 10 uM forward and reverse primers. Sample-specific amplicon sequencing was carried out using unique combinations of tagged primers (Meier *et al.* 2015; Baloglu *et al.* unpublished). Degenerate metazoan primers [COI; mlCO1intF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' (Leray *et al.* 2013) and jgHCO2198: 5'-TAIACYTCIGGRTGICCRAARAAYCA-3' (Geller *et al.* 2013)] were used for the new PCR reaction conditions. The samples that failed at direct PCR

stage were processed with QuickExtract (Quick Extract DNA™). The specimens were immersed into 20 µl of the extraction solution and otherwise processed following the manufacturer's instructions. Final PCR products were pooled and sent for library preparation. NGS barcoding of specimens (n=12,633) was carried out on multiple MiSeq 2 X 300 cycle runs as part of multiple projects.

4.3.3 MOTU delimitation

I used Objective Clustering at 2-5% with uncorrected pairwise distances to delimit sequences into molecular operational taxonomic units (MOTUs) (Srivathsan & Meier, 2012). This range of thresholds has been shown to produce a stable number of clusters that are largely congruent with species boundaries as determined by morphology (Meier *et al.* 2015; Baloglu *et al.* unpublished). I was able to identify some of the resulting MOTUs to species using my available barcode database. That database included barcodes from specimens that were identified to species based on morphology.

4.3.4 Statistical analyses

Community analyses. To estimate chironomid species richness, I plotted species accumulation curves for each habitat with iNEXT package (Hsieh *et al.* 2016) and tested for significant differences between habitat types by assessing the overlap of the 95% confidence intervals (CIs). I treated individual habitats as samples and used sample-based rarefaction curves standardized to the sampling coverages to compare species richness between habitat types (Gotelli & Colwell, 2001). Distance matrices were generated from the site-species data matrices using the Bray-Curtis metric (Legendre & Gallagher, 2001). I used Mantel tests to assess correlations

among assemblage similarity matrices with the vegan package (Oksanen *et al.* 2017). The species overlap between the reservoirs and swamp forest were explored via the number of shared species and the number of specimens for each shared species. Furthermore, I investigated the directionality of the species intermixing (e.g. reservoirs to the swamp forest or swamp forest to the reservoirs) by comparing the abundances of each of the shared species in each habitat.

Multivariate analyses. I used a multivariate approach (redundancy analysis, RDA) using the vegan package to assess whether there are important local variables that explain the chironomid community structure at the swamp forest sites. I first standardized the samples at each site to 70% sampling coverage to minimize differences in abundance due to the different time/area sampled (see Final analysis; Table 4.1). As a result, there were only 26 of 40 sites for the following analysis, thus highlighting the incomplete sampling effort for the excluded sites (Fig. 4.1).

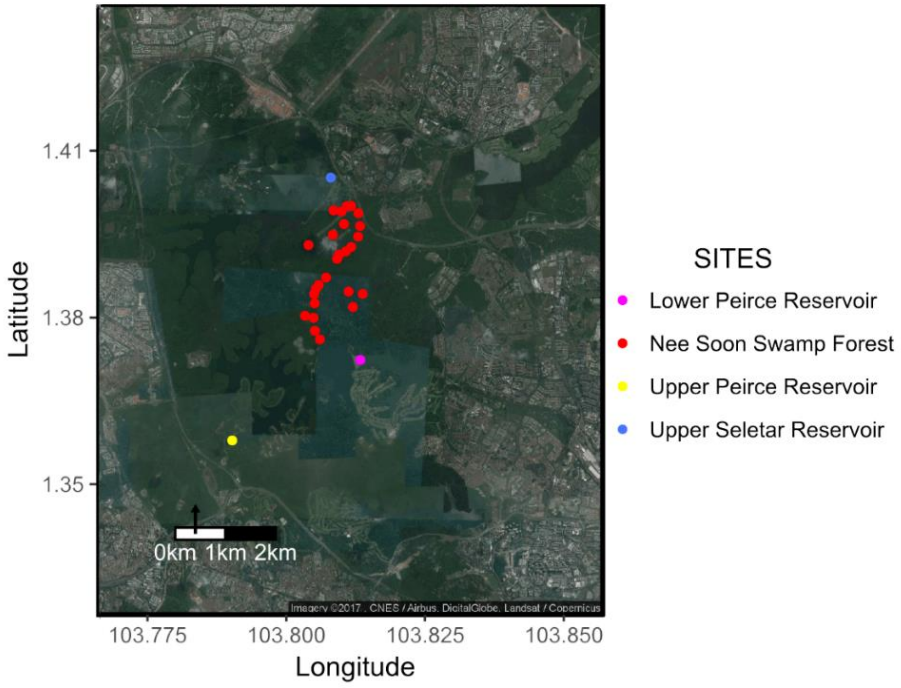

Figure 4.1: The distribution of the 29 sampling sites in the swamp forest and the three reservoirs in the Central Catchment Region of Singapore. Different colors are given for each habitat.

The species data matrix of 172 species in these areas was related to a total of 13 environmental (10 physicochemical, two spatial and one temporal) variables in RDA. Two variables (cross-sectional area and maximum depth) were highly correlated with the stream width and average stream depth, hence I removed them from the further analysis. I also used variance inflation factor (VIF) function in R for collinearity, but there were no VIF values larger than 10 (see Table 4.2). Thus I included all those variables. I used Monte Carlo permutation tests (n = 999) to assess the statistical power of all analyses.

Linear models. I used linear mixed effect (LME) analysis to extract common patterns of chironomid communities at the swamp forest. I selected this model because it can account for non-independence of errors (i.e., due to spatial autocorrelation) by differentiating between fixed versus random effects (Pinheiro & Bates, 2002). Spatial autocorrelation occurs when pairs of values, measured at given distances in space, are more or less similar than expected by chance alone (Legendre & Legendre, 2012). Models with spatial correlation structures were generated using the corrSpatial argument in nlme package (Pinheiro *et al.* 2017). Akaike information criterion (AIC) was used to compare the models. The model with the smallest value of AIC was preferred over the others. Hill numbers of order q: Species richness (q = 0), Shannon diversity (q = 1) and Simpson diversity (q = 2) were obtained with iNEXT package. These values were used as dependent variables for three separate linear mixed-effects models (Bates *et al.* 2014) using the lme function with maximum likelihood estimation. For each model, continuous physicochemical variables and one categorical variable (presence-absence of the non-native (reservoir) species) were used as fixed effects (without interaction term) nested within the sampling year as a

random effect. The variables that were log transformed before the analyses can be found in Table 4.2. Models were refined and validated following the guidance provided in Zuur *et al.* (2010): all parameters were included in the initial model with non-significant terms removed manually in a systematic, stepwise process to achieve the best goodness-of-fit with fewest factors, assessed by selecting the model with the lowest AIC value. If removal of a nonsignificant term increased the AIC value, the term was retained in the refined model. Once the final models were reached, a linear model was fitted after removing random effects, to assess the significance of each term in the model. The adjusted R2 value of the fitted model was calculated and compared with the adjusted R2 of models fitted with each parameter removed in turn. The relative contribution of each parameter in explaining the variance of the model was then calculated as a percentage of the total variance explained. p values for regression coefficients were obtained using the car package (Fox, 2002). Statistics and graphical outputs were computed with the *ade4* package (Dray & Dufour, 2007). I performed all statistical analyses in R Version 3.4.0 (R Core Team, 2017) unless otherwise stated.

4.4 Results

I found considerable variation in some of the environmental variables in Nee Soon Swamp Forest (Table 4.2). For instance, among physicochemical variables, water depth and turbidity ranged from 2.9 to 62.1 cm and from 0 to 1142.4 NTU, respectively.

Table 4.2: Selected environmental characteristics and variance inflation factor associated with each of the variables of 26 Nee Soon forest streams for redundancy analysis.

Variable	Abbreviation	Units	Min.	Max.	Median	VIF	Transformation
Latitude	Lat	DD°	1.37606	1.40005		4.30	None
Longitude	Long	DD°	103.80339	103.81377		3.19	None
Year	Year		2013	2014		2.82	None
pH	pH	log[H+]	3.83	5.88	4.46	4.14	None
Specific conductance	Cond	mS cm^{-1}	0.01	0.05	0.02	4.44	None
Water depth	Av dep	cm	2.9	62.1	14.55	3.99	Log
Temperature	Temp	°C	23.49	26.33	25.43	3.11	Log
Dissolved oxygen	DO	mg L^{-1}	0.33	6.62	5.6	2.84	Log
Width	Width	cm	50	500	200	2.22	Log
Average velocity	Av vel	cm s^{-1}	1.1	16.5	5.45	2.56	Log
Turbidity	Turb	NTU	2.79	1142.4	16.18	1.42	Log
Stream discharge	Tot dis	m3 s^{-1}	0.00	0.24	0.02	4.68	None
Stream type	Strahler	None	1	3		4.59	None
Non-native species presence	non_nat	None	0	1			None

4.4.1 Chironomid species richness and community structure

Chironomid species richness at the reservoirs.

For Lower Peirce Reservoir, 547 of 1,308 specimens were successfully barcoded (42%). A total of 19 species was observed, and the predicted species richness (Chao2) was 28 ± 10 species, respectively. For Upper Peirce Reservoir, 604 of 1,058 specimens (57%) were successfully barcoded. A total of 19 species was observed, and 24 ± 5 species was predicted. For Upper Seletar Reservoir, 2,318 of 3,647 specimens (63%) were successfully barcoded, and the estimated richness was 34 ± 14 species. The success rates are relatively low because the reservoir samples are of variable age and they were not preserved for the purpose of barcoding.

In total, 37 species were observed in the reservoirs. Across three reservoirs, the most common chironomid species was *Polypedilum quasinubifer*, accounting for 48% of 3,469 total chironomid specimens followed by *Polypedilum sp.* (near leei) (17%). *Polypedilum sp.* (near leei) is likely to be a cryptic species of *P. leei*, i.e., morphologically similar, however genetically more than 6% apart from each other.

Chironomid species richness at the swamp forest (Larva).

From a total of 6,620 specimens, 4,027 specimens were successfully barcoded. For 54 of these, the top 100 BLAST matches were for sequences that did not belong to Chironomidae, and only the remaining 3,973 specimens were retained for further analysis (60.5%: 3,973/6,566). A total of 259 species were observed based on a 4% clustering threshold, and 333 ± 24 species were estimated to be present based on Chao2 estimates. Singletons made up 26.6% of the total species occurrences for the larval community at Nee Soon.

Larval communities at Nee Soon and the reservoirs.

For a range of 3-5%, the number of MOTUs were as follows: 290, 276, and 271. Most of the MOTUs were congruent (253) between different thresholds and the overall stability at the MOTU level was at +/- 6%. The discrepancies were due to a total of 379 specimens lumping or splitting into a different MOTU; i.e., the assignment of only 5% of the total number specimens was uncertain. The remaining discussion is utilizing 3% MOTUs. Most species are only found in Nee Soon Swamp Forest (259 of 290 species) while the observed chironomid richness in the three reservoirs was low and did not differ much (Fig. 4.2).

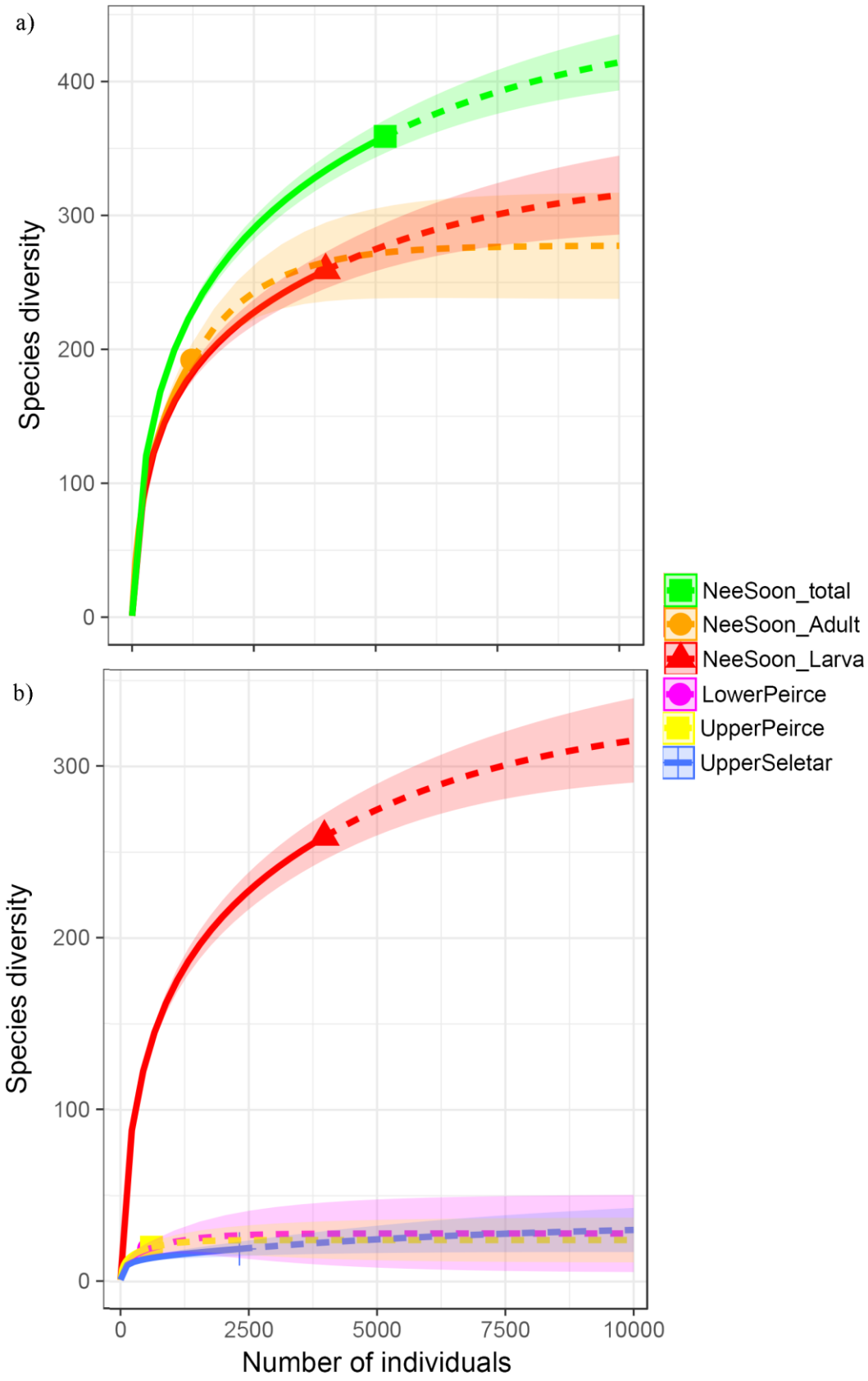

Figure 4.2: Rarefaction curves (solid line) and extrapolation (dashed line) for chironomid communities of a) only Nee Soon and b) Nee Soon and reservoirs in Singapore. The 95% confidence intervals (shaded areas) were obtained by a bootstrap method based on 200 replications.

Chironomid species richness at the swamp forest (Adult).

I attempted to barcode 1,551 specimens and 1,278 yielded sequences. The top 100 BLAST matches revealed that 55 of these specimens did not belong to Chironomidae. Hence, only 1,223 specimens were retained for further analysis (82%: 1,223/1,496). 192 chironomid species were observed, and 277 ± 27 species were estimated. The most common species contributed 13% of the total adult chironomid abundance. Singletons represented a large proportion of the adult dataset (74 species, 38.5%), indicating the need for more sampling.

Larval and adult communities at Nee Soon Swamp Forest.

For a range of 3-5%, I found the number of MOTUs as follows: 359, 347, and 341. Most of the MOTUs were congruent (322), and the overall stability at the MOTU level was at +/- 5%. The discrepancies were due to 235 specimens lumping or splitting, contributing only 4.5% of the total abundance. A total of 359 species were observed for the combined dataset of adult and larvae at Nee Soon Swamp Forest (n = 5,196). 92 adult and larval stages could be matched. These shared species were found in 39 of the 40 sampling locations. The estimated species richness for the combined adult and larval communities was 447 ± 24, respectively.

4.4.2 High species turnover between the reservoirs and the Swamp Forest

There was a large difference in the composition of the chironomid fauna between the swamp forest and the reservoirs. Only seven species were shared between the swamp forest and the reservoirs (Table 4.3). Overall species composition was not

significantly correlated (NSSF - USR: Mantel R= -0.03, NSSF - UP: R = -0.02, NSSF - LP: R = -0.02, p>0.05 for all). In contrast, the reservoirs had proportionally more shared species. However, for species composition, only Lower Peirce and Upper Peirce reservoirs showed significant but weak correlation with each other (R= 0.19, p<0.05) while they were dissimilar to Upper Seletar reservoir (LP - USR: R= -0.07, UP - USR: R= -0.11, p>0.05 for both).

4.4.3 Influence of reservoir species on overall species diversity in the swamp forest

Of the final 26 sampling sites, only eight sites had shared species with the reservoirs: seven sites each shared one species while one site sharing six species (Table 4.3). When I investigated the directionality of the species intermixing (e.g., reservoirs to the swamp forest or swamp forest to the reservoirs) by comparing the abundances of the shared species in each habitat, I found that only 1 of the 7 species, *Tanytarsus formosanus*, had higher abundance in the swamp forest (82 specimens) than in the reservoirs (only 4 specimens). All these shared species previously have been detected in Singapore reservoirs (Cranston *et al.* 2013; Wong *et al.* 2014; Lim *et al.* 2016; Baloglu *et al.* unpublished). I hypothesized that the Nee Soon sites sharing species with the reservoirs had overall lower species diversity than those without reservoir species. Using LME, I tested this hypothesis and found that there was no significant effect of the presence of non-native species on the overall species richness, Simpson, and Shannon diversity indices (see Table 4.7).

Table 4.3: Species shared between the reservoirs and Nee Soon communities. For clarity only the species that occurred in both the reservoirs and Nee Soon shown, and only the partial list of 92 shared species in Nee Soon communities provided.

Species	Sites/Communities				
	Nee Soon Larvae	Nee Soon Adults	Upper Peirce	Upper Seletar	Lower Peirce
Ablabesmyia typeTMSI	4	0	15	0	12
Cladotanytarsus sp4.	1	0	0	140	4
Polypedilum leei Freeman, 1961	4	0	1	1	2
Polypedilum quasinubifer Cranston sp. n.	2	7	24	1628	13
Tanytarsus formosanus Kieffer, 1912	82	0	0	4	0
Tanytarsus oscillans Johannsen, 1932	2	1	2	42	0
Tanytarsus ovatus Johannsen, 1932	23	1	0	50	0

Table 4.4: Weighted intraset correlation between the axes and the environmental variables following RDA of chironomid abundance data from Nee Soon Swamp Forest. Only the significant variables are shown. Significance of the axes by Monte Carlo test is given; p values for Monte Carlo test. All canonical axes: F = 1.94, p = 0.001.

	RDA1	RDA2	RDA3
Accumulated % of variance of species data explained	32.1	57.4	76.7
Correlation with axes			
Dissolved oxygen	-0.21	-0.61	-0.49
Depth	0.21	-0.83	0.22
pH	0.81	-0.12	0.03
Latitude	0.76	-0.02	0.34
Year	0.74	0.02	-0.26

4.4.4 Habitat characteristics and chironomid species composition in the swamp forest

Results of the RDA analysis (first three axes) are summarized in Table 4.4 and shown in Fig. 4.3, respectively. The environmental variables selected in the analysis are represented in the triplot with arrows, which point in the direction of maximum change in the value of the associated variable (Fig. 4.3).

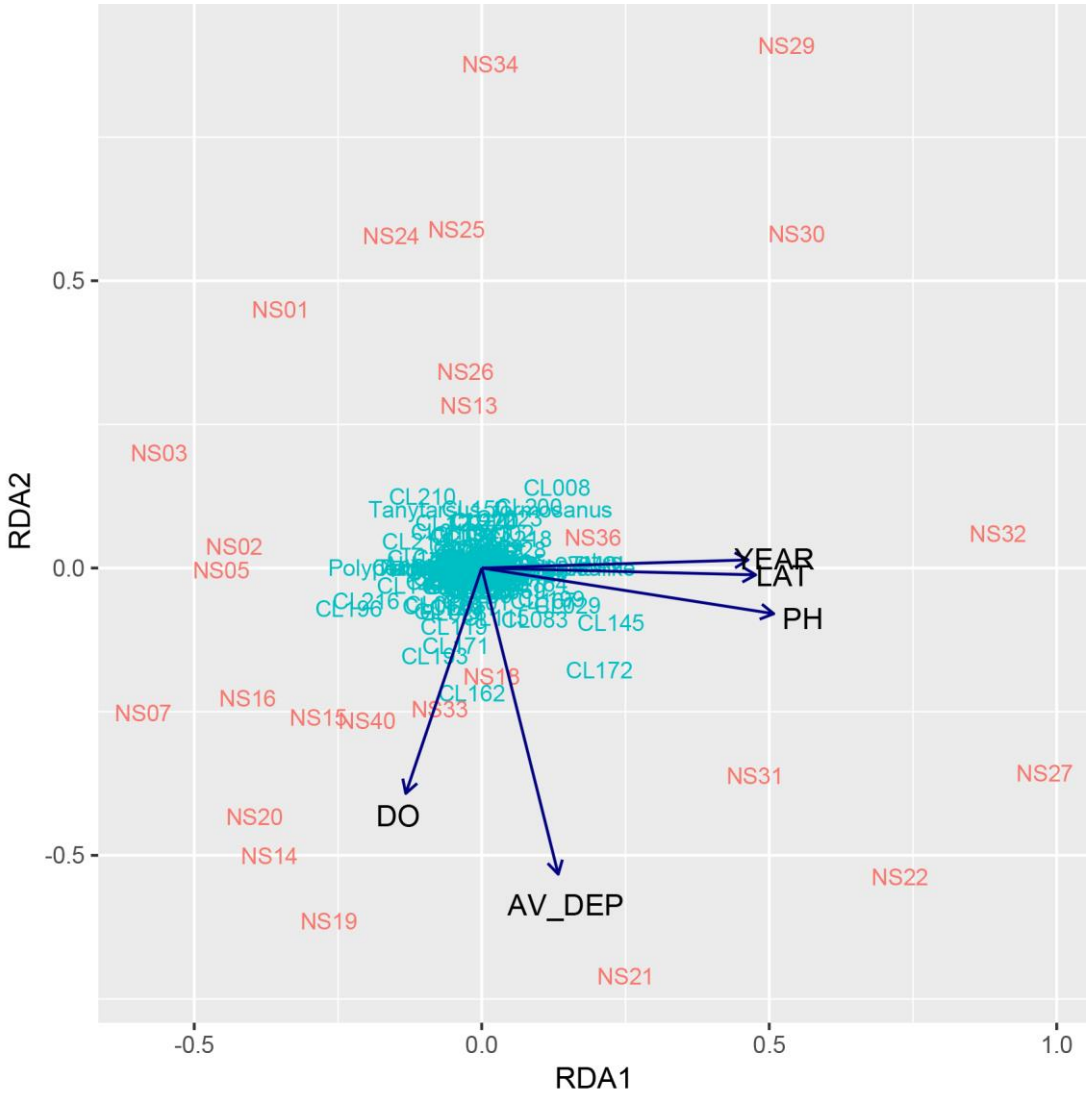

Figure 4.3: Ordination diagram from redundancy analysis (RDA) illustrating the relations between chironomid community composition and the four environmental variables that explained the most variance. Solid arrows indicate direction of sharpest increase in abundance of chironomid species. Sites are shown in red, and species are shown in blue. DO: Dissolved oxygen, AV_DEP: Stream depth, LAT: Latitude of the sampling site.

The first two axes of the ordination accounted for about 57% of the total variance in the chironomid community composition, with the first axis explaining 32.1% of the variation and the Monte Carlo tests were significant for all axes, respectively (Table 4.4). Axis 1 was positively correlated with pH, latitude, and year, while axis 2 was negatively correlated with stream depth and dissolved oxygen. I found that the samples taken at sites 21, 22, 27, 29-32, and 36 had an affinity for increasing pH. On the other hand, samples at sites 7, 14-16, 18-22, 27, 31, 33 and 40 showed seemed to correlate with higher depth and dissolved oxygen levels (Fig. 4.3).

The significant environmental (physicochemical, temporal, and spatial) variables explained 13.8% of the variance in the composition of chironomids at Nee Soon (F = 1.90, P = 0.001). Among the physicochemical variables, average depth, the amount of dissolved oxygen, and pH emerged as the most significant explanatory variables (Table 4.5). Other significant variables were latitude and the year of the sampling. Variation partitioning analyses revealed that 11% of the total variance was explained by physicochemical variables, and 16% of the total variance was explained by all the variables (Table 4.6).

Table 4.5: Results of RDA analyses with forward selection of environmental (physicochemical, spatial, and temporal) variables explaining the assemblage of chironomids in Nee Soon Swamp Forest.

	df	*F*-ratio	*P*-value
Water depth	1	2.18	0.001***
Dissolved oxygen	1	1.50	0.025*
Ph	1	2.67	0.001***
Latitude	1	1.73	0.007***
Year	1	1.53	0.027*

Table 4.6: Variation partitioning results: Percentage of variation explained (pure and shared effect) for each group of variables classified by scale

Effect	Adj. R^2 (%)
Physico-chemical only	0.11
Water depth	0.03
Dissolved oxygen	0.02
pH	0.05
Water depth*Dissolved oxygen	0.01
Physico-chemical, spatial, and temporal	0.16
Physico-chemical	0.09
Spatial	0.01
Temporal	0.02
Physico-chemical*spatial	0.01
Temporal*spatial	0.01
Physico-chemical*temporal*spatial	0.02

4.4.5 What explains chironomid species richness in the swamp forest?

There was no evidence of spatial autocorrelation of observations within groups, as the AIC values of the models with spatial error structures were higher than the null models (data not shown). Therefore, the models without the spatial autocorrelation structure were selected. I used a statistical modeling (LME) to identify the dominant physicochemical variables influencing the species richness, Shannon diversity, and Simpson diversity. I found that all three response variables were best predicted negatively by conductivity (i.e., salinity) and positively by turbidity of the water with these terms explaining most of the attributed variance in the model (Table 4.7). I retained the temperature and stream discharge in the final models as non-significant terms, but they only explained a small proportion of the variance (8.6% for Species richness, 7.2% for Shannon diversity, and <3% for Simpson diversity). pH, stream depth, and stream type also explained much of the model variance for all three

response variables, albeit non-significantly. These models indicate that species richness of chironomids is driven largely by the levels of salinity and turbidity in tropical freshwater streams, but is also influenced by a mixture of other physicochemical parameters.

Table 4.7: Linear mixed effects model to determine the relationships between three response variables (species richness, Shannon index and Simpson index) in separate models and the continuous physicochemical variables and one categorical variable in 26 Nee Soon Swamp Forest sites.

	Species richness		Shannon diversity		Simpson diversity	
Term	% Adj. R^2	*P*	% Adj. R^2	*P*	% Adj. R^2	*P*
Conductivity	37.1	*	37.3	*	36.8	*
Turbidity	15.1	*	24.5	*	42.4	*
Temperature	0	ns	0	ns	0	ns
Stream discharge	8.6	ns	7.2	ns	2.4	ns
Width	0	-	0	ns	0	ns
pH	14.4	-	12.3	-	8	-
Average velocity	0	-	1.3	-	3.1	-
Dissolved oxygen	4.7	-	2.9	-	0	-
Stream depth	9.6	-	7.3	-	2.3	-
Stream type	9	-	7.3	-	5.1	-
Non-native species presence	1.4	-	0	-	0	-
Total variance explained (Adj. R^2)	0.66		0.53		0.34	

The relative contribution (%) of each term in explaining model variance was calculated as % difference in adjusted R^2 comparing the full refined model and the model with each term removed. pH, dissolved oxygen, width, average velocity, turbidity, stream discharge, and stream type were removed during model refinement. Symbols indicate the presence or the significance of the term within the refined model: 0, negative adjusted R^2 values; –, not present in the refined model; ns, not significant, *=$P < 0.05$.

4.5 Discussion

4.5.1 Estimating chironomid species richness

My study reveals the tremendous diversity of chironomid species (>400 estimated species) in the slow flowing creeks and streams of a relatively small (90 ha) habitat that is the remnant of a previously much larger forest. To date, only 5,000 species of Chironomidae described (Cranston & Martin, 1989) but Armitage (1995) estimate the global diversity to be 15,000 species. Regardless of which point of comparison is used, my result of a diversity of >400 species is remarkable because it makes up 3-10% of the total estimated/described species richness. This either indicates that the global species richness estimate needs a major update, the swamp forest is a surprisingly rich hotspot for chironomid diversity or a combination of both. Regardless of which scenario is considered, it is important to remember that Nee Soon Swamp Forest is a tiny remnant of the kind of wet, lowland forest (Whitmore, 1985) that was part of a more extensive freshwater swamp forest covering 5% of Singapore (Corlett, 1991; Turner *et al.* 1996). Much of the forest has been lost due to forest clearance since the early 1800s. It made way to reservoir construction (early 20th century: O'Dempsey & Chew, 2011) and later to industrial and housing development (Ng & Lim, 1992). Nee Soon has a land area that is <0.28% of Singapore (Yee *et al.* 2011), but it is nevertheless the largest remnant of the pristine primary swamp forests and has significant ecological and conservation importance for the country. Given that the >400 species of chironomids were estimated to inhabit a habitat that is a fraction of its original size, the original species richness for chironomids must have been much higher. Brook *et al.* (2003) state that in the last two centuries, 30-80% of all species in Singapore have gone extinct. Based on these estimates, the original chironomid

species diversity in Singapore's swamp forest could have been as high as 1,300 species. The majority of the world's tropical swamp forests are found in Southeast Asia's Indo-Malayan region (Yule, 2010). These swamp forests collectively occupy a large area (13 million ha; Hooijer *et al.* 2006) and are found on many geographically separated peninsulas and islands; the chironomid midge diversity of these forests must be vast. Much of this diversity is threatened with destruction, however, because Southeast Asian peat swamps are disappearing fast (Yule, 2010) to make room for oil palm plantations. For instance, more than half of the original peat swamp forest in Sumatra and also Indonesian Borneo has been destroyed (Indonesia WWF, 2008).

According to Coffman *et al.* (1992) and Coffman & de la Rosa (1998), low-latitude and low-order tropical streams have similar numbers of chironomid species as temperate streams. However, my observed and estimated chironomid species richness values in the slow-flowing streams of the swamp forest exceed all values previously reported for lotic chironomids elsewhere in the world. For instance, maximum reported species richness values for temperate streams are 182 species for streams in the USA and 246 species in a river in Germany, respectively (Coffman *et al.* 1992). Moreover, maximum reported species richness values for tropical streams are 299 species from thirty-one 4th- to 6th-order West African streams, 250 species from thirteen 3rd- to 6th-order northwestern Costa Rican streams (Coffman, 1989, Coffman & de la Rosa 1998), and 195 species from fifteen 1st- to 2nd- order streams in Brazil (Roque *et al.* 2007). It has been suggested that the high richness values for tropical streams are mainly due to high numbers of rare species with very low abundances. This means that a large number of specimens has to be collected in order to estimate the true species richness (Melo & Froehlich, 2001). Indeed in my study, I detected a

high number of species that were present at low abundance and singletons made up nearly half of the total species richness.

Given that my observed species richness (359 species) is much higher than the previously reported values, could it be due to the use of NGS barcodes? I doubt so because several studies have shown the congruence between molecular and morpho-species for chironomids (Carew *et al.* 2011; Brodin *et al.* 2013; Silva & Wiedenburg, 2014; Montagna *et al.* 2016) and my results are largely insensitive to which clustering threshold is used for obtaining species estimates based on sequences (2-4%). It is more likely that by using NGS barcodes, my study revealed many species which are unknown to science. My estimates also appear reasonable because Southeast Asian swamp forests are known to host many new and undescribed species (Yule, 2010).

4.5.2 Resilience of the swamp forest community

My results suggest that both reservoirs and the swamp forest were very resilient to each other, i.e., their chironomid species richness and community composition were very different. Of the 290 species collected during the study, only seven species were collected in both the forest stream and reservoir habitats, signaling nearly complete community turnover in just a few meters. The three reservoirs of Singapore were created in the early 20th century. Even in the nearly 100 years since the construction of the reservoirs, distinct differences in chironomid fauna have maintained between the two habitat types. This long-term resilience could be due to the pH differences between the swamp forest and the reservoirs, but this hypothesis would require further testing (see Table 4.2).

Species mixing was mostly one directional (reservoir to swamp forest), i.e., the abundances of the shared species were higher in the reservoirs than in Nee Soon, except for one species, *Tanytarsus formosanus* (see Table 4.3). One may have expected that the shared species would be found in swamp forest sites that were adjacent to the reservoirs, but this was not confirmed. Instead, shared species were found across several sampling sites in the swamp forest. This indicates that there was no edge effect, i.e., no influence of the reservoir on the adjacent swamp forest chironomid communities. It appears likely that a few chironomid adults are blown into the different habitats and are only able to establish temporary populations (we sampled larval midges). In other words, there is no evidence for a replacement of swamp forest species with widespread non-native species that are found in urbanized habitats; i.e., overall, the ecological integrity of Nee Soon appears secure with regard to chironomid midges that constitute an important component of the forests' macroinvertebrate community.

The question is whether these species could establish larger populations in the swamp forest? I assumed that the pH tolerance of the shared species should be high, given that Nee Soon is acidic when compared with the reservoirs. Indeed, one of the shared species is *T. formosanus* that is known from acidic rice fields in Malaysia although its abundance is positively correlated with water pH (range: pH 5.15-7.7; Al-Shami *et al.* 2010). To our knowledge, our study is the first to report the presence of *T. oscillans* and *T. ovatus* in an acidic aquatic environment. However, several other species of *Tanytarsus* are known to tolerate low pH such as *Tanytarsus* Pe15 (Langton, 1991; cited in Orendt, 1999: pH: 4.6-5.8), *Tanytarsus buchonius* Reiss & Fitt (pH: 3.3-5.7), and another *Tanytarsus sp.* in (Orendt, 1999: pH: 3-6.8).

Polypedilum leei, which is also shared, has previously been reported to be present in acidic aquatic environments (pH: 4-7, Outridge, 1987; pH: 4-7.1, Wright & Burgin, 2007). However, both *P. leei* and *P. quasinubifer* are also widely distributed in Singapore's alkaline reservoirs (Low, 2010; Baloglu *et al.* unpublished).

Overall, the reservoirs had lower species diversity than the natural habitat. This was not unexpected, because overall species richness, biotic interactions, and ecosystem complexity are known to decline between rural/native and more urbanized habitats (McKinney, 2006). Moreover, a large number of studies have shown that large reservoirs have an adverse impact on aquatic biodiversity (reviewed in Bunn & Arthington 2002). Take note, however, that the reservoirs in this study are arguably not species-poor. Instead, the swamp forest is exceptionally rich in species diversity.

4.5.3 Patterns of chironomid species richness in Nee Soon Swamp Forest

Only a relatively small amount of the variance could be explained by the environmental parameters that were measured (16%), but this may not be surprising given that no data were available for other variables such as competition and predation, food availability (Raposeiro *et al.* 2011), species interactions (Kohler, 1992), and the amount of vegetation cover (van der Berg *et al.* 1997). Similarly, sampling, random and stochastic events may also account for much of the variance given that we were studying a very diverse community (Ter Braak, 1987). However, it is also not atypical for studies of chironomid communities that local environmental factors explain a relatively small proportion (i.e., <30%) of the variation (Heino *et al.* 2009; Puntí *et al.* 2009). In my study, the most important environmental parameters

were water depth, pH, and the amount of dissolved oxygen. This is in agreement with the previous studies (Quinlan & Smol 2002; Molozzi *et al.* 2013).

Salinity was the most important individual variable which was negatively correlated with species richness. The direct physiological effect of high salinity of freshwater biota is osmoregulatory stress (Bayly, 1972) that can severely affect the growth, development or survival of the organisms (Cartier *et al.* 2011). Indeed, high salinities (conductivity >5000 μS/cm) were shown to lower the number of emergent adults, delay the time of emergence, and reduce larval growth rate (Hassell *et al.* 2006) in some chironomid species. Even though the Nee Soon stream was reported to be of low to medium salinity (see Table 4.2), small increases in the salinity may have contributed to the observed decrease in the species richness. Other studies have also shown a negative correlation between salinity and species numbers for chironomids (Williams *et al.* 1990; Brodersen & Anderson, 2002).

Turbidity was the second most important individual variable explaining chironomid species richness and this correlation was positive. Turbidity here refers to the measure of water clarity. Higher turbidity can reduce algal growth rates due to decreases in light availability, which can in turn negatively affect algal eating fish numbers. As certain fish species prey on chironomids, a positive correlation of chironomid species richness with increasing turbidity would be consistent with the hypothesis of reduced predation by predators that use vision to locate prey (e.g., fish, odonate larvae).

4.5.4 Effect of geography on chironomid distribution in the swamp forest

I found 92 shared species between the adult (sampled from one site) and larval chironomid communities (40 sites) of Nee Soon. There were no clear patterns between the localities where the adults and larvae were collected. The adults were all collected in one site and yielded specimens for 92 species. The corresponding larvae were collected in 39 of the 40 larval sampling sites suggesting that the species involved are distributed broadly in the 90 ha range of Nee Soon. Given that some of the reservoir species were also found in Nee Soon, albeit in small abundances, dispersal abilities of chironomids may not be the limiting factor. It is more likely that the heterogeneity of the microhabitats is responsible for the species-rich and yet very complementary adjacent chironomid communities in the swamp forest. Note, however, that only 26 sites could be included in the study because 14 sites had to be removed from analysis due to small sample size.

The only spatial influence was the latitude of the sampling sites in the swamp forest, which explains <3% of the total variance in chironomid community composition. Latitude here is likely to reflect the changes in the stream type (such as upstream or downstream) and the water flow, therefore influence the chironomid community. This is due to the fact that the streams flow from South to Northeast.

4.5.5 Effect of geography on chironomid distribution across the reservoirs

Overall, I would expect all three reservoirs to have similar community composition, because they cluster in one group based on their physicochemical variables for 13-years long measurements and there is water flow from Upper Seletar to Lower Peirce and Upper Peirce Reservoirs (Low, 2010). However, according to a Mantel test, chironomid communities of only the neighboring Lower and Upper Peirce reservoirs showed similarity to each other; i.e., only the neighboring reservoirs had similar communities which is consistent with an effect of geography on the reservoir chironomid communities. If the distance drives/limits the chironomid distribution, I would expect that chironomids would disperse freely between the reservoir sampling sites, and generate similar communities. This is because they can disperse widely in Nee Soon and the range of Nee Soon sampling sites is similar to the range of reservoir sites. However, it is likely that for chironomid distribution across the reservoirs, the wind/openness of habitat was more important than the actual distance between the sites. Hence, I believe that the swamp forest, with an environment hostile to reservoir midges, caused a barrier between the two similar reservoirs and their chironomid communities.

4.6 Implications for conservation of tropical swamp forests

My results showed that the tropical Nee Soon Swamp Forest has an outstanding chironomid species diversity that is dramatically different from the reservoirs. My study revealed that chironomid communities in the swamp forest were related to several physicochemical variables rather than geographic distance. These findings

have important conservation implications for other swamp forests in Southeast Asia, but it also suggests that even small or fragmented swamp forests can be suitable habitats for a rich and likely native chironomid community. As the chironomid communities in the swamp forest are complex, more research is needed to generate more in-depth insights into their ecology so that the conservation needs can be adequately addressed.

CHAPTER 5

Where Are We

And What Remains to be Done?

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time
T.S. ELIOT, "LITTLE GIDDING"

5.1 Traditional bioassessment

Traditionally, macroinvertebrates, mostly larval specimens, are collected for bioassessment programs worldwide, i.e., Australia, the United Kingdom, Canada, the European Union, the United States and the developing countries. These programs aim to collect as many taxa as possible, with an emphasis on obtaining a sufficiently large sample size for a robust analysis (Vlek *et al.* 2006). For most taxa, the material is only sorted to higher-level taxa (e.g., genus, family) because of constraints on time, budget, and availability of expertise. For instance, in Singapore, the Singscore (Blakely *et al.* 2014), which was mostly developed for Singapore's lotic ecosystems, uses only the presence-absence and abundance information from 74 macroinvertebrate families. On the other hand, my thesis shows that without species level information, some unique environmental responses by the taxa are lost, and species level information could

provide better precision and accuracy for measuring responses to environmental parameters than higher-level taxon information.

Most bioassessment programs rely on traditional morphology-based taxonomy. Just in the United States alone, 2-5 million specimens are being analyzed annually at the cost of $30 to $60 million in order to conduct traditional, morphology-based assessment (Stein *et al.* 2014). The high cost is related to inefficiencies caused by sampling techniques, specimen-sorting, and specimen identification steps (Nichols & Dyer, 2013). The cost is particularly high for difficult, yet commonly collected taxa, such as chironomids. In my thesis, I develop strategies to address the challenges mentioned above using DNA barcodes and show the promises and shortcomings of this method. The key conclusions are summarized here.

5.2 Optimizing bioassessment: From field to the lab

When this thesis started, next-generation sequencing (NGS) costs were comparable to or slightly more expensive than the traditional morphology-based bioassessment (Stein *et al.* 2014). However, during my Ph.D., I managed to bring down the cost sufficiently to include >30,000 chironomid specimens which were barcoded using NGS barcoding at a specimen cost of 20-50 cents (reagent cost). We now know NGS-based identification can be significantly cheaper than previously thought.

First, I show that sampling of the adult chironomids with emergence traps can provide sufficient information about the midge community of a reservoir. Larval sampling is the common practice but tedious and costly because the larvae have to be

removed from benthic mud sample and some samples require a week of manpower to complete the extraction (Chapter 2). Secondly, based on a dataset covering one-year, I demonstrate that sampling can be limited to two-months if the goal is an assessment of species diversity. Thirdly, I show that for characterizing the species profile of the midge fauna of a tropical reservoir, 600-1000 specimens are sufficient. The sample size can be reduced to 100-200 specimens if only the most abundant species, are to be identified. However, these recommendations still need to be empirically tested on temperate aquatic habitats.

My study optimizes specimen-based NGS barcoding for chironomids. However, at the moment, we do not know whether chironomids alone are sufficient for bioassessment. I suggest a comparative study, where one compares traditional bioassessment to emergence-trap based bioassessment based on chironomids using i) NGS barcoding as I have done in my thesis and ii) metabarcoding of bulk samples. The results of each method should be then correlated with water quality parameters. The comparative study can also include other macroinvertebrate groups as long as they are captured by emergence traps (e.g., mayflies). The results of such a pilot could be used to develop a biotic index that uses species-level information and would take advantage of the more cost-effective sampling that can be carried out with emergence traps.

5.3 Quick bioassessment in real-time and in the field

The existing methods for generating barcodes require a well-equipped molecular laboratory and can still be time-consuming and/or expensive. In chapter 3, I show that using a palm-sized Nanopore MinION^{TM} sequencer is sufficient for obtaining

sequence data for 50 chironomid midges within 2.5 hours and estimate the species composition within 1-1.5 hours. For this experiment, we estimate that a barcode can be generated for ~8.5 USD if 100 specimens are multiplexed. One of the drawbacks of MinION™ sequencing, as inferred from chapter 3, is the high error rates. The high coverage in our study substantially reduced the indel error rates. However, even at high coverage (30-100X), we observed that errors remained. Fortunately, they were concentrated in the homopolymeric regions of COI. With appropriate bioinformatic adjustment, one may be able to correct for many of these errors. Alternatively, one will have to wait for an improvement in technology.

Could it be that in the future, bioassessment can be automated from the field to the bench to the sequencing in 24 hours? Following the sampling of adults using emergence traps, an automated recognition approach can handle the pre-sorting of specimens into morpho-species, as suggested by Larios *et al.* (2008). Alternatively, without any pre-sorting process, all specimens can be spread onto a white background. A robotic handler can then collect a tissue sample from each specimen and drop it into a well on in a 96-well plate that contains the PCR master mix. Alternatively, the robot could drop each specimen into an extraction liquid (i.e., QuickExtract™ as I have used in my study), collect the DNA extract afterward, and set up a PCR plate. With the use of a unique combination of tagged forward and reverse primers for each specimen (i.e., tagged amplicon sequencing), a large number of amplicons could be produced with the minimal involvement of manpower. A liquid-handling robot could then collect the PCR products and pool them before purification. Even higher degrees of automation can be achieved with robots that provide more functions such as moving and changing plates, adding covers to the

plates, and transferring bulk liquid between workstations (Kong *et al.* 2012). Following an optimized lab protocol, I estimate that a single run of MinION run could generate 500-1000 barcodes, which would be sufficient for the bioassessment midge diversity in tropical reservoirs. Various government bioassessment programs could adopt this protocol that allows for large-scale, rapid, cost-effective, and species-level bioassessment.

5.4 DNA barcoding of invertebrates helps to understand habitat resilience

While invertebrates comprise over 80% of the world's biodiversity (species and biomass) (Clarke & Spier, 2003), most conservation studies focus on vertebrates. This is due to the lack of taxonomic, biological and distribution data for the vast majority of invertebrate species that are therefore excluded in many conservation assessments. I here demonstrate that much information can be obtained on species diversity and distribution of invertebrates through the analysis of DNA barcodes despite some problems such as the lack of a universal barcode gap. However, my thesis shows that species estimates for one habitat are relatively stable across multiple thresholds; hence the DNA barcode data can be used for ecological analysis without a need to correct for taxonomic problems caused by closely related species and distantly related allopatric populations of the same species. I demonstrate in my thesis that cost-effective barcodes can be used for studies that include as many as 14,000 specimens that can be sorted to species-level. This has important implications for work in invertebrate conservation biology as shown in my study of a tropical swamp forest remnant (Chapter 4). I reveal high diversity, study species turnover, and conclude that there is much resilience between the natural swamp forest and artificial reservoirs.

5.5 Towards a better understanding of the species diversity of our planet

In my thesis, I processed a substantial number of specimens. This allowed for finding rare species, which is essential for understanding the species diversity of our planet and for improving DNA barcode databases (Meier *et al.* 2016). The large-scale sample size also allowed for estimating the actual abundances of the species. Combined with the characterization of the community at the species level by considering environmental parameters, I am convinced that one will be able to improve conservation planning and move beyond approaches that rely on protecting flagship species. We are now closer to the 'future' that I envisioned in my introduction chapter where I outlined the vision that one would have ready tools for efficiently grouping organism into species and identifying them with relative ease so that one can concentrate on understanding biodiversity, studying the morphology and species interactions.

References

Al-Shami, S. A., Salmah, M. R. C., Hassan, A. A., & Azizah, M. N. S. (2010). Temporal distribution of larval Chironomidae (Diptera) in experimental rice fields in Penang, Malaysia. *Journal of Asia-Pacific Entomology*, *13*(1), 17-22.

Ambasht, R. S., & Ambasht, N. K. (Eds.). (2012). *Modern trends in applied aquatic ecology*. Springer Science & Business Media.

Ander, M., Troell, K., & Chirico, J. (2013). Barcoding of biting midges in the genus Culicoides: a tool for species determination. *Medical and veterinary entomology*, *27*(3), 323-331.

Armitage, P. D. (1995). Chironomidae as food. In *The Chironomidae* (pp. 423-435). Springer Netherlands.

Armitage, P. D., Pinder, L. C., & Cranston, P. (Eds.). (2012). The Chironomidae: Biology and ecology of non-biting midges. Springer Science & Business Media.

Arscott, D. B., Jackson, J. K., & Kratzer, E. B. (2006). Role of rarity and taxonomic resolution in a regional and spatial analysis of stream macroinvertebrates. *Journal of the North American Benthological Society*, *25*(4), 977-997.

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., ... & O'grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*, *33*(3), 296-300

Avise, J. C. (2012). *Molecular markers, natural history and evolution*. Springer Science & Business Media.

Balian, E. V., Segers, H., Lévêque, C., & Martens, K. (2008). The freshwater animal diversity assessment: an overview of the results. *Hydrobiologia*, *595*(1), 627-637.

Ball, S. L., Hebert, P. D., Burian, S. K., & Webb, J. M. (2005). Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society*, *24*(3), 508-524.

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

Bartram, J., & Ballance, R. (Eds.). (1996). *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes*. CRC Press.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42; 2011. *Reference Source*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bayly, I. A. E. (1972). Salinity tolerance and osmotic behavior of animals in athalassic saline and marine hypersaline waters. *Annual review of ecology and systematics*, *3*(1), 233-268.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology letters*, *15*(4), 365-377.

Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*, *5*(1), 4.

Bharucha, E. K., & Gogte, P. P. (1990). Avian profile of a man-modified aquatic ecosystem in the backwaters of the Ujjani Dam. *Journal of the Bombay Natural History Society*, *80087*(1), 73-90.

Blair, R. B. (2001). Birds and butterflies along urban gradients in two ecoregions of the United States: is urbanization creating a homogeneous fauna. *Biotic homogenization*, 33-56.

Blair, R. B., & Johnson, E. M. (2008). Suburban habitats and their role for birds in the urban–rural habitat network: points of local invasion and extinction? *Landscape Ecology*, *23*(10), 1157-1169.

Blakely, T. J., Eikaas, H. S., & Harding, J. S. (2014). The Singscore: a macroinvertebrate biotic index for assessing the health of Singapore's streams and canals. *Raffles Bulletin of Zoology*, *62*.

Blueweiss, L., Fox, H., Kudzma, V., Nakashima, D., Peters, R., & Sams, S. (1978). Relationships between body size and some life history parameters. *Oecologia*, *37*(2), 257-272.

Boulton, A. J., Boyero, L., Covich, A. P., Dobson, M., Lake, S., & Pearson, R. (2008). Are tropical streams ecologically different from temperate streams. *Tropical stream ecology*, 257-284.

Børsting, C., & Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics*, *18*, 78-89.

Brodersen, K. P., & Lindegaard, C. (1999). Classification, assessment and trophic reconstruction of Danish lakes using chironomids. *Freshwater Biology*, *42*(1), 143-157.

Brodersen, K. P., & Anderson, N. (2002). Distribution of chironomids (Diptera) in low arctic West Greenland lakes: trophic conditions, temperature and environmental reconstruction. *Freshwater Biology*, *47*(6), 1137-1157.

Brodin, Y., Ejdung, G., Strandberg, J., & Lyrholm, T. (2013). Improving environmental and biodiversity monitoring in the Baltic Sea using DNA barcoding of Chironomidae (Diptera). *Molecular Ecology Resources*, *13*(6), 996-1004.

Brook, B. W., Sodhi, N. S., & Ng, P. K. (2003). Catastrophic extinctions follow deforestation in Singapore. *Nature*, *424*(6947), 420.

Brundin, L. (1949). *Chironomiden und andere Bodentiere der südschwedischen Urgebirgsseen.* Institute of Freshwater Research Drottninholm Report 30, 1-914.

Bunn, S. E., & Arthington, A. H. (2002). Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. *Environmental management*, *30*(4), 492-507.

Burns, J. M., Janzen, D. H., Hajibabaei, M., Hallwachs, W., & Hebert, P. D. (2007). DNA barcodes of closely related (but morphologically and ecologically distinct) species of skipper butterflies (Hesperiidae) can differ by only one to three nucleotides. *Journal of the Lepidopterists Society*, *61*(3), 138-153.

Cairns, J., & Pratt, J. R. (1993). A history of biological monitoring using benthic macroinvertebrates. *Freshwater biomonitoring and benthic macroinvertebrates*, *10*, 27.

Carew, M. E., Pettigrove, V., & Hoffmann, A. A. (2005). The utility of DNA markers in classical taxonomy: using cytochrome oxidase I markers to differentiate Australian Cladopelma (Diptera: Chironomidae) midges. *Annals of the Entomological Society of America*, *98*(4), 587-594.

Carew, M. E., Pettigrove, V., Cox, R. L., & Hoffmann, A. A. (2007). DNA identification of urban Tanytarsini chironomids (Diptera: Chironomidae). *Journal of the North American Benthological Society*, *26*(4), 587-600.

Carew, M. E., Marshall, S. E., & Hoffmann, A. A. (2011). A combination of molecular and morphological approaches resolves species in the taxonomically difficult genus Procladius Skuse (Diptera: Chironomidae) despite high intra-specific morphological variation. *Bulletin of entomological research*, *101*(5), 505-519.

Carew, M. E., Pettigrove, V. J., Metzeling, L., & Hoffmann, A. A. (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in zoology*,*10*(1), 1.

Cartier, V., Claret, C., Garnier, R., & Franquet, E. (2011). How salinity affects life cycle of a brackish water species, Chironomus salinarius KIEFFER (Diptera: Chironomidae). *Journal of Experimental Marine Biology and Ecology*, *405*(1), 93-98.

CBD UNEP (2010) *CBD Strategic Plan for Biodiversity 2011–2020 and the Aichi Targets: 'Living in Harmony with Nature'*. www.cbd.int/doc/strategicplan/2011-2020/Aichi-Targets-EN.pdf

Chao, A., Ma, K. H., Hsieh, T. C. & Chiu, C. (2016). SpadeR: Species-Richness P rediction and Diversity Estimation with R. R package version 0.1.1. https:/ /CRAN.R-project.org/package=SpadeR

Clarke, G. M. (1993). Fluctuating asymmetry of invertebrate populations as a biological indicator of environmental quality. *Environmental Pollution*, *82*(2), 207-211.

Clarke, G., & Spier, F. (2003). A review of the conservation status of selected nonmarine invertebrates. *Environment Australia, Canberra*.

Clements, W. H., Carlisle, D. M., Courtney, L. A., & Harrahy, E. A. (2002). Integrating observational and experimental approaches to demonstrate causation in stream biomonitoring studies. *Environmental Toxicology and Chemistry*, *21*(6), 1138-1146.

Clews, E., Low, E. W., Belle, C. C., Todd, P. A., Eikaas, H. S., & Ng, P. K. (2014). A pilot macroinvertebrate index of the water quality of Singapore's reservoirs. *Ecological Indicators*, *38*, 90-103.

Coffman, W. P. (1989). Factors that determine the species richness of lotic communities of Chironomidae. *Acta Biologica Debrecina, Supplementum Oecologica Hungarica*, *3*, 95-100.

Coffman, W. P., De la Rosa, C., Cummins, K. W., & Wilzbach, M. A. (1992). Species richness in some neotropical (Costa Rica) and Afrotropical (West Africa) lotic communities of Chironomidae (Diptera). *Netherland Journal of Aquatic Ecology*, *26*(2-4), 229-237.

Coffman, W. P., & de la Rosa, C. L. (1998). Taxonomic composition and temporal organization of tropical and temperate species assemblages of lotic Chironomidae. *Journal of the Kansas Entomological Society*, 388-406.

Colwell, R. K. 2013. EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide and application published at: http://purl.oclc.org/estimates.

Corlett, R. T. (1991). Plant succession on degraded land in Singapore. *Journal of tropical forest science*, 151-161

Costello, M. J., May, R. M., & Stork, N. E. (2013). Can we name Earth's species before they go extinct?. *Science*, *339*(6118), 413-416.

Cranston, P. S., & Martin, J. (1989). 26. Family Chironomidae. *Catalog of the Diptera of the Australasian and Oceanian Regions. Bishop Museum Press and EJ Brill, Honolulu and Leiden*, 252-274.

Cranston, P. S., Cooper, P. D., Hardwick, R. A., Humphrey, C. L., & Dostine, P. L. (1997). Tropical acid streams–the chironomid (Diptera) response in northern Australia. *Freshwater Biology*, *37*(2), 473-483.

Cranston, P. S., Ang, Y. C., Heyzer, A., Lim, R. B. H., Wong, W. H., Woodford, J. M., & Meier, R. (2013). The nuisance midges (Diptera: Chironomidae) of Singapore's Pandan and Bedok reservoirs. *Raffles B Zool*,*61*(2), 779-793.

Darby, R. (1962). Midges associated with California rice fields, with special reference to their ecology (Diptera: Chironomidae). *California Agriculture*, *32*(1), 1-206.

Davis, G. E. (1989). Design of a long-term ecological monitoring program for Channel Islands National Park, California. *Natural Areas Journal*, *9*(2), 80-89.

Delettre, Y. R., & Morvan, N. (2000). Dispersal of adult aquatic Chironomidae (Diptera) in agricultural landscapes. *Freshwater Biology*, *44*(3), 399-411.

DeSalle, R., Freedman, T., Prager, E. M., & Wilson, A. C. (1987). Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian Drosophila. *Journal of Molecular evolution*, *26*(1-2), 157-164.

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, *22*(4), 1-20.

Drew, L. W. (2011). Are We Losing the Science of Taxonomy? As need grows, numbers and training are failing to keep up. *BioScience*, *61*(12), 942-946.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., ... & Sullivan, C. A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. Biological reviews, 81(2), 163-182.

Ekrem, T., Willassen, E., & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular phylogenetics and evolution*, *43*(2), 530-542.

Ekrem, T., Stur, E., & Hebert, P. D. (2010). Females do count: Documenting Chironomidae (Diptera) species diversity using DNA barcoding. Organisms Diversity & Evolution, 10(5), 397-408.

Epler, J. H. (2001). *Identification manual for the larval Chironomidae (Diptera) of North and South Carolina: a guide to the taxonomy of the midges of the southeastern United States, including Florida* (p. 526). Palatka, FL: St. Johns River Water Management District.

Esteves, F. D. A. (1988). Fundamentos de limnologia. In *Fundamentos de limnologia*. Interciência/Finep.

Failla, A. J., Vasquez, A. A., Fujimoto, M., & Ram, J. L. (2015). The ecological, economic and public health impacts of nuisance chironomids and their potential as aquatic invaders. *Aquatic invasions*, *10*(1).

Fattorini, S. (2011). Insect extinction by urbanization: a long term study in Rome. *Biological Conservation*, *144*(1), 370-375.

Fausch, K. D., Lyons, J. O. H. N., Karr, J. R., & Angermeier, P. L. (1990). Fish communities as indicators of environmental degradation. In *American fisheries society symposium* (Vol. 8, pp. 123-144).

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. *Mol. Mar. Biol. Biotechnol. 3*, 294–299.

Fox, J. (2002). An {R} Companion to Applied Regression, Second Edition. Thous and Oaks, CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/ Companion

Friberg, N., Bonada, N., Bradley, D. C., Dunbar, M. J., Edwards, F. K., Grey, J., ... & Woodward, G. U. Y. (2011). Biomonitoring of human impacts in freshwater ecosystems: the good, the bad and the ugly. *Advances in Ecological Research*, *44*, 1-68.

Fu, Z., Yoshizawa, K., Yoshida, N., Kazama, F., & Hirabayashi, K. (2012). Bathymetric distribution of chironomid larvae (Diptera: Chironomidae) in Lake Saiko, Japan. Lakes & Reservoirs: Research & Management, 17(1), 55–64.

Fuller, R. A., Irvine, K. N., Devine-Wright, P., Warren, P. H., & Gaston, K. J. (2007). Psychological benefits of greenspace increase with biodiversity. *Biology letters*, *3*(4), 390-394.

Garcia-Moreno, J., Harrison, I. J., Dudgeon, D., Clausnitzer, V., Darwall, W., Farrell, T., ... & Tubbs, N. (2014). Sustaining freshwater biodiversity in the Anthropocene. In *The Global Water System in the Anthropocene* (pp. 247-270). Springer International Publishing.

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular ecology resources*, *13*(5), 851-861.

Gleick, P. H. (1993). Water in crisis: a guide to the worlds fresh water resources.

Gonçalves, P. F., Oliveira-Marques, A. R., Matsumoto, T. E., & Miyaki, C. Y. (2015). DNA barcoding identifies illegal parrot trade. *Journal of Heredity*, *106*(S1), 560-564.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333-351.

Gotelli, N. J., & Colwell, R. K. (2011). Estimating species richness. *Biological diversity: frontiers in measurement and assessment*, *12*, 39-54.

Grant, I. F. (2002). Aquatic invertebrates. In: I.F. Grant and C.C.D. Tingle (eds.), Ecological Monitoring Methods for the Assessments of Pesticide impact in the Tropics. London: The University of Greenwich, p.183-193.

Graul, W. D., Torres, J., & Denney, R. (1976). A species-ecosystem approach for nongame programs. *Wildlife Society Bulletin (1973-2006)*, *4*(2), 79-80.

Greffard, M. H., Saulnier-Talbot, É., & Gregory-Eaves, I. (2011). A comparative analysis of fine versus coarse taxonomic resolution in benthic chironomid community analyses. Ecological Indicators, 11(6), 1541-1551.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., ... & Dodd, R. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome medicine*, *7*(1), 99.

Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Uriz, M. J., & Turon, X. (2016). Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ*, *4*, e2807.

Haenel, C., & Chown, S. L. (1998). The impact of a small, alien invertebrate on a sub-Antarctic terrestrial ecosystem: Limnophyes minimus (Diptera, Chironomidae) at Marion Island. *Polar Biology*, *20*(2), 99-106.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., & Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS one*, *6*(4), e17497.

Hargreaves, A. D., & Mulley, J. F. (2015). Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ*, *3*, e1441.

Hassell, K. L., Kefford, B. J., & Nugegoda, D. (2006). Sub-lethal and chronic salinity tolerances of three freshwater insects: Cloeon sp. and Centroptilum sp.(Ephemeroptera: Baetidae) and Chironomus sp.(Diptera: Chironomidae). *Journal of Experimental Biology*, *209*(20), 4024-4032.

Hebert, P. D., Cywinska, A., & Ball, S. L. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1512), 313-321.

Hecnar, S. J., & M'Closkey, R. T. (1996). Regional dynamics and the status of amphibians. *Ecology*, *77*(7), 2091-2097.

Heinis, F., & Davids, C. (1993). Factors governing the spatial and temporal distribution of chironomid larvae in the Maarsseveen lakes with special emphasis on the role of oxygen conditions. *Netherland Journal of Aquatic Ecology*, *27*(1), 21-34.

Heino, J., & Paasivirta, L. (2008). Unravelling the determinants of stream midge biodiversity in a boreal drainage basin. *Freshwater Biology*, *53*(5), 884-896.

Heino, J., Tolonen, K. T., Kotanen, J., & Paasivirta, L. (2009). Indicator groups and congruence of assemblage similarity, species richness and environmental relationships in littoral macroinvertebrates. *Biodiversity and Conservation*, *18*(12), 3085.

Helson, J. E., Williams, D. D., & Turner, D. (2006). Larval chironomid community organization in four tropical rivers: human impacts and longitudinal zonation. *Hydrobiologia*, *559*(1), 413-431.

Hilty, J., & Merenlender, A. (2000). Faunal indicator taxa selection for monitoring ecosystem health. *Biological conservation*, *92*(2), 185-197.

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., ... & Wollenberg, K. R. (2016). Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerging infectious diseases*, *22*(2), 331.

Holmlund, C. M., & Hammer, M. (1999). Ecosystem services generated by fish populations. *Ecological economics*, *29*(2), 253-268.

Holway, D. A., & Suarez, A. V. (2006). Homogenization of ant communities in mediterranean California: the effects of urbanization and invasion. *Biological conservation*, *127*(3), 319-326.

Hooijer, A., Silvius, M., Wösten, H. & Page, S. (2006). PEAT-CO2, Assessment of CO2 emissions from drained peatlands in SE Asia. Delft Hydraulics report Q3943.

Hourigan, T. F., Timothy, C. T., & Reese, E. S. (1988). Coral reef fishes as indicators of environmental stress in coral reefs. In *Marine organisms as indicators* (pp. 107-135). Springer New York.

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: iNterpolation and EXTrapola tion for species diversity. R package version 2.0.12. URL: http://chao.stat. nthu.edu.tw/blog/software-download/.

Hutto, R. L., Reel, S., & Landres, P. B. (1987). A critical evaluation of the species approach to biological conservation. *endangered species*, *4*(12).

Hynes, H. B. N. (1960). *The biology of polluted waters* (No. 574.52632). Liverpool University Press.

Indonesia WWF. (2008). Deforestation, forest degradation, biodiversity loss and CO2 emissions in Riau, Sumatra, Indonesia. *One Indonesian Province's Forest and Peat Soil Carbon loss over a Quarter Century and its Plans for the Future*.

Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., ... & Piazza, P. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, *4*.

Ivanova, N. V., Borisenko, A. V., & Hebert, P. D. (2009). Express barcodes: racing from specimen to identification. *Molecular Ecology Resources*, *9*(s1), 35-41.

Jacobsen, R. E., & Perry, S. A. (2007). Polypedilum nubifer, a chironomid midge (Diptera: Chironomidae) new to Florida that has nuisance potential. *Florida Entomologist*, *90*(1), 264-267.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature methods*, *12*(4), 351-356.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, *17*(1), 239.

Johnson, R. K., Wiederholm, T., & Rosenberg, D. M. (1993). Freshwater biomonitoring using individual organisms, populations, and species assemblages of benthic macroinvertebrates. *Freshwater biomonitoring and benthic macroinvertebrates*, 40-158.

Johnson, K. P., Cruickshank, R. H., Adams, R. J., Smith, V. S., Page, R. D., & Clayton, D. H. (2003). Dramatically elevated rate of mitochondrial

substitution in lice (Insecta: Phthiraptera). *Molecular phylogenetics and evolution*, *26*(2), 231-242.

Jones, R. I., & Grey, J. (2004). Stable isotope analysis of chironomid larvae from some Finnish forest lakes indicates dietary contribution from biogenic methane. *Boreal Environment Research*, *9*(1), 17-24.

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363-375.

Judge, K., Harris, S. R., Reuter, S., Parkhill, J., & Peacock, S. J. (2015). Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *70*(10), 2775-2778.

Kallis, G., & Butler, D. (2001). The EU water framework directive: measures and implications. *Water policy*, *3*(2), 125-142.

Kaplan, R. (2001). The nature of the view from home: Psychological benefits. *Environment and behavior*, *33*(4), 507-542.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, *15*(5), 266-274.

Kiester, A. R., & Eckhardt, C. (1994). *Review of wildlife management and conservation biology on the Tongass National Forest: a synthesis with recommendations*. Pacific Northwest Research Station, USDA Forest Service.

Kohler, S. L. (1992). Competition and the structure of a benthic stream community. *Ecological Monographs*, *62*(2), 165-188.

Kong, F., Yuan, L., Zheng, Y. F., & Chen, W. (2012). Automatic liquid handling for life science: a critical review of the current state of the art. *Journal of laboratory automation*, *17*(3), 169-185.

Kurtzman, C. P. (1994). Molecular taxonomy of the yeasts. *Yeast*, *10*(13), 1727-1740.

Kuehn, I., & Klotz, S. (2006). Urbanization and homogenization–comparing the floras of urban and rural areas in Germany. *Biological conservation*, *127*(3), 292-300.

Kwong, S., Srivathsan, A., Vaidya, G., & Meier, R. (2012a). Is the COI barcoding gene involved in speciation through intergenomic conflict?. *Molecular phylogenetics and evolution*, *62*(3), 1009-1012.

Kwong, S., Srivathsan, A., & Meier, R. (2012b). An update on DNA barcoding: low species coverage and numerous unidentified sequences. *Cladistics*, *28*(6), 639-644.

Lake, P. S., Schreiber, E. S. G., Milne, B. J., & Pearson, R. G. (1994). Species richness in streams: patterns over time, with stream size and with latitude. *Internationale Vereinigung fur Theoretische und Angewandte Limnologie Verhandlungen*, *25*(3), 1822-1826.

Langton, P. H. (1991). *A key to pupal exuaviae of West Palaearctic Chironomidae*. Langton. PE 17 1 YH, 386 pp.

Larios, N., Deng, H., Zhang, W., Sarpola, M., Yuen, J., Paasch, R., ... & Shapiro, L. G. (2008). Automated insect identification through concatenated histograms of local appearance features: feature vector generation and

region detection for deformable objects. *Machine Vision and Applications*, *19*(2), 105-123.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular detection and quantification*,*3*, 1-8.

Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, *129*(2), 271-280.

Legendre, P., & Legendre, L. F. (2012). *Numerical ecology* (Vol. 24). Elsevier.

Leggett, R. M., Heavens, D., Caccamo, M., Clark, M. D., & Davey, R. P. (2015). NanoOK: Multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics*, btv540.

Lenat, D. R., & Resh, V. H. (2001). Taxonomy and stream ecology—the benefits of genus-and species-level identifications. *Journal of the North American Benthological Society*, *20*(2), 287-298.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., ... & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in zoology*, *10*(1), 1.

Lewis Jr, W. M. (1996). Tropical lakes: how latitude makes a difference. *Perspectives in tropical limnology*, *4364*.

Li, C., Chng, K. R., Boey, E. J. H., Ng, A. H. Q., Wilm, A., & Nagarajan, N. (2016). INC-Seq: accurate single molecule reads using nanopore sequencing. *GigaScience*, *5*(1), 34.

Lim, N. K., Tay, Y. C., Srivathsan, A., Tan, J. W., Kwik, J. T., Baloğlu, B., ... & Yeo, D. C. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society open science*, *3*(11), 160635.

Lin, Y. J., & Quek, R. F. (2011). Observations on mass emergence of chironomids (Diptera: Chironomidae) in Bedok, Singapore with notes on human–chironomid interactions. *Nature in Singapore*, *4*, 339-347.

Lin, X., Stur, E., & Ekrem, T. (2015). Exploring genetic divergence in a species-rich insect genus using 2790 DNA Barcodes. *PloS one*, *10*(9), e0138993.

Loke, L. H., Clews, E., Low, E. W., Belle, C. C., Todd, P. A., Eikaas, H. S., & Ng, P. K. (2010). Methods for sampling benthic macroinvertebrates in tropical lentic systems. *Aquatic Biology*, *10*(2), 119-130.

Loman, N. J., & Watson, M. (2015). Successful test launch for nanopore sequencing. *Nature methods*, *12*(4), 303.

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, *12*(8), 733-735.

Loose, M., Malla, S., & Stout, M. (2016). Real time selective sequencing using nanopore technology. *bioRxiv*, 038760.

Low, E. (2010). *Singapore reservoirs: quantifying water quality through physicochemical, algae, and invertebrate analyses* (Doctoral dissertation).

Lucca, J. V., Pamplin, P. A. Z., Gessner, A. F., Trivinho-Strixino, S., Spadano-Albuquerque, A. L., & Rocha, O. (2010). Benthic macroinvertebrates of a tropical lake: Lake Caçó, MA, Brazil. *Brazilian journal of biology*, *70*(3), 593-600.

MacArthur, R. H. (1972). *Geographical ecology: patterns in the distribution of species*. Princeton University Press.

Mace, G. M., Norris, K., & Fitter, A. H. (2012). Biodiversity and ecosystem services: a multilayered relationship. *Trends in ecology & evolution*, *27*(1), 19-26.

Maclean, I. M., & Wilson, R. J. (2011). Recent ecological responses to climate change support predictions of high extinction risk. *Proceedings of the National Academy of Sciences*, *108*(30), 12337-12342.

Maller, C. J., Henderson-Wilson, C., & Townsend, M. (2009). Rediscovering nature in everyday settings: or how to create healthy environments and healthy people. *EcoHealth*, *6*(4), 553-556.

Marchant, R. (2002). Do rare species have any place in multivariate analysis for bioassessment?.*Journal of the North American Benthological Society*, *21*(2), 311-313.

Margalef, R. (1983). Limnologia. Barcelona. *Omega*, *1*(010).

Marshall, T. R., Ryder, R. A., Edwards, C. J., & Spangler, G. R. (1987). Using the lake trout as an indicator of ecosystem health-application of the dichotomous key. *Techn. rep./Great Lakes fishery commiss.*

Marziali, L., Armanini, D. G., Cazzola, M., Erba, S., Toppi, E., Buffagni, A., & Rossaro, B. (2010). Responses of Chironomid larvae (Insecta, Diptera) to ecological quality in Mediterranean river mesohabitats (South Italy). *River research and applications*, *26*(8), 1036-1051.

Marx, V. (2015). PCR heads into the field. *Nature methods*, *12*(5), 393-397.

McKinney, M. L. (2006). Urbanization as a major cause of biotic homogenization. *Biological conservation*, *127*(3), 247-260.

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic biology, 55(5), 715-728.

Meier, R. (2008). DNA sequences in taxonomy - Opportunities and challenges. In: *The New Taxonomy. Systematics Association Special Volume.* (ed. Wheeler QD), pp. 95-128. CRC Press, New York.

Meier, R., Zhang, G., & Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic biology*, *57*(5), 809-813.

Meier, R., Wong, W., Srivathsan, A., & Foo, M. (2016). $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics.

Melo, A. S., & Froehlich, C. G. (2001). Evaluation of methods for estimating macroinvertebrate species richness using individual stones in tropical streams. *Freshwater Biology*, *46*(6), 711-721.

Metzeling, L., Robinson, D., Perriss, S., & Marchant, R. (2002). Temporal persistence of benthic invertebrate communities in south-eastern

Australian streams: taxonomic resolution and implications for the use of predictive models. *Marine and Freshwater Research*, *53*(8), 1223-1234.

Metzker, M. L. (2010). Sequencing technologies—the next generation. Nature reviews genetics, 11(1), 31-46.

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS biology*, *3*(12), e422.

Mikheyev, A. S., & Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources*, *14*(6), 1097-1102.

Millennium ecosystem assessment (2005). *Ecosystems and human wellbeing: a framework for assessment. Washington, DC: Island Press*.

Molozzi, J., Feio, M. J., Salas, F., Marques, J. C., & Callisto, M. (2013). Maximum ecological potential of tropical reservoirs and benthic invertebrate communities. *Environmental monitoring and assessment*, *185*(8), 6591-6606.

Montagna, M., Mereghetti, V., Lencioni, V., & Rossaro, B. (2016). Integrated taxonomy and DNA barcoding of Alpine midges (Diptera: Chironomidae). *PloS one*, *11*(3), e0149673.

Monteiro, A., & Pierce, N. E. (2001). Phylogeny of Bicyclus (Lepidoptera: Nymphalidae) inferred from COI, COII, and EF-1α gene sequences. *Molecular phylogenetics and evolution*, *18*(2), 264-281.

Morais, S. S., Molozzi, J., Viana, A. L., Viana, T. H., & Callisto, M. (2010). Diversity of larvae of littoral Chironomidae (Diptera: Insecta) and their role as bioindicators in urban reservoirs of different trophic levels. Brazilian Journal of Biology, 70(4), 995–1004.

Moriyama, E. N., & Powell, J. R. (1997). Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. *Journal of Molecular Evolution*, *45*(4), 378-391.

Morse, J. C., Bae, Y. J., Munkhjargal, G., Sangpradub, N., Tanida, K., Vshivkova, T. S., ... & Yule, C. M. (2007). Freshwater biomonitoring with macroinvertebrates in East Asia. Frontiers in Ecology and the Environment, 5(1), 33-42.

Murakami, M., & Nakano, S. (2002). Indirect effect of aquatic insect emergence on a terrestrial insect population through by birds predation. *Ecology Letters*, *5*(3), 333-337.

Naeem, S., Duffy, J. E., & Zavaleta, E. (2012). The functions of biological diversity in an age of extinction. *Science*, *336*(6087), 1401-1406.

Nazarova, L. B., Riss, H. W., Kahlheber, A., & Werding, B. (2004). Some observations of buccal deformities in chironomid larvae (Diptera: Chironomidae) from the Ciénaga Grande de Santa Marta, Colombia. Caldasia, 26(1), 275–290.

Ng, P. K., & Lim, K. K. (1992). The conservation status of the Nee Soon freshwater swamp forest of Singapore. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *2*(3), 255-266.

Ng'endo, R. N., Osiemo, Z. B., Brandl, R., & Kondo, T. (2013). DNA barcodes for species identification in the hyperdiverse ant genus Pheidole (Formicidae: Myrmicinae). *Journal of Insect Science*, *13*(1).

Nicacio, G., & Juen, L. (2015). Chironomids as indicators in freshwater ecosystems: an assessment of the literature. Insect Conservation and Diversity.

Nichols, S. J., & Dyer, F. J. (2013). Contribution of national bioassessment approaches for assessing ecological water security: an AUSRIVAS case study. *Frontiers of Environmental Science & Engineering*, *7*(5), 669-687.

Nisbet, E. K., Zelenski, J. M., & Murphy, S. A. (2011). Happiness is in our nature: Exploring nature relatedness as a contributor to subjective well-being. *Journal of Happiness Studies*, *12*(2), 303-322.

O'Dempsey, T., & Chew, P. T. (2011). The Freshwater Swamp Forests of Sungei Seletar Catchment: A Status Report. In *Proceedings of Nature Society, Singapore's Conference on 'Nature Conservation for a Sustainable Singapore'–16th October* (pp. 121-166).

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B ., ... & Oksanen, M. J. (2017). vegan: Community Ecology Package. R pac kage version 2.4-3. https://CRAN.R-project.org/package=vegan

Orendt, C. (1999). Chironomids as bioindicators in acidified streams: a contribution to the acidity tolerance of chironomid species with a classification in sensitivity classes. *International Review of Hydrobiology*, *84*(5), 439-449.

Outridge, P. M. (1987). Possible causes of high species diversity in tropical Australian freshwater macrobenthic communities. *Hydrobiologia*, *150*(2), 95-107.

Patton, D. R. (1987). *Is the Use of" management Indicator Species" Feasible?*. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station.

Paulsen, S. G., & Linthurst, R. A. (1994). Biological monitoring in the environmental monitoring and assessment program. *Biological monitoring of aquatic systems. CRC Press, Boca Raton, Florida, USA*, 297-322.

Pearson, W. R. (1990). [5] Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in enzymology*, *183*, 63-98.

Pentinsaari, M., Vos, R., & Mutanen, M. (2017). Algorithmic single-locus species delimitation: effects of sampling effort, variation and nonmonophyly in four methods and 1870 species of beetles. *Molecular ecology resources*, *17*(3), 393-404.

Pereira, H. M., Leadley, P. W., Proença, V., Alkemade, R., Scharlemann, J. P., Fernandez-Manjarrés, J. F., ... & Chini, L. (2010). Scenarios for global biodiversity in the 21st century. *Science*, *330*(6010), 1496-1501.

Pettigrove, V., & Hoffmann, A. (2005). A field-based microcosm method to assess the effects of polluted urban stream sediments on aquatic macroinvertebrates. *Environmental Toxicology and Chemistry*, *24*(1), 170-180.

Pfenninger, M., Nowak, C., Kley, C., Steinke, D., & Streit, B. (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. Molecular ecology, 16(9), 1957-1968.

Pinder, L. C. V. (1986). Biology of freshwater Chironomidae. *Annual review of entomology*, *31*(1), 1-23.

Pinder, L. C. V. (1995). The habitats of chironomid larvae. In *The Chironomidae* (pp. 107-135). Springer Netherlands.

Pinheiro, J., & Bates, D. (2002). Mixed-Effects modelling in S and S-PLUS. *Statistics and Computing Series*.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Team, R. C. (2017). nlme: linear and nonlinear mixed effects models. R package version 3.1-131. *R Foundation for Statistical Computing, Vienna*.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular ecology*, *21*(8), 1864-1877.

Puntí, T., Rieradevall, M., & Prat, N. (2009). Environmental factors, spatial variation, and specific requirements of Chironomidae in Mediterranean reference streams. *Journal of the North American Benthological Society*, *28*(1), 247-265.

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*(6783), 212.

Rainbow, P. S., Hildrew, A. G., Smith, B. D., Geatches, T., & Luoma, S. N. (2012). Caddisflies as biomonitors identifying thresholds of toxic metal bioavailability that affect the stream benthos. *Environmental pollution*, *166*, 196-207.

Raposeiro, P. M., Costa, A. C., & Hughes, S. J. (2011). Environmental factors–spatial and temporal variation of chironomid communities in oceanic island streams (Azores archipelago). In *Annales de Limnologie-International Journal of Limnology* (Vol. 47, No. 4, pp. 325-338). EDP Sciences.

Ratnasingham, S., & Hebert, P. D. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PloS one*, *8*(7), e66213.

Raunio, J., & Muotka, T. (2005). The use of chironomid pupal exuviae in river biomonitoring: the importance of sampling strategy. *Archiv für Hydrobiologie*, *164*(4), 529-545.

Raunio, J., Heino, J., & Paasivirta, L. (2011). Non-biting midges in biodiversity conservation and environmental assessment: findings from boreal freshwater ecosystems. Ecological Indicators, 11(5), 1057-1064.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Real, M., Rieradevall, M., & Prat, N. (2000). Chironomus species (Diptera: Chironomidae) in the profundal benthos of Spanish reservoirs and lakes: factors affecting distribution patterns. *Freshwater Biology*, *43*(1), 1-18.

Research Matters. (2017). Retrieved from https://researchmatters.in/article/brief-history-understanding-biodiversity

Resh, V. H. (2007). Multinational, freshwater biomonitoring programs in the developing world: lessons learned from African and Southeast Asian river surveys. *Environmental Management*, *39*(5), 737-748.

Resh, V. H. (2008). Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. *Environmental Monitoring and Assessment*, *138*(1-3), 131-138.

Reynoldson, T. B., & Metcalfe-Smith, J. L. (1992). An overview of the assessment of aquatic ecosystem health using benthic

invertebrates. *Journal of Aquatic Ecosystem Stress and Recovery (Formerly Journal of Aquatic Ecosystem Health)*, *1*(4), 295-308.

Ricciardi, A., & Rasmussen, J. B. (1999). Extinction rates of North American freshwater fauna. *Conservation Biology*, *13*(5), 1220-1222.

Ridder, B. (2008). Questioning the ecosystem services argument for biodiversity conservation. *Biodiversity and Conservation*, *17*(4), 781-790.

Risse, J., Thomson, M., Patrick, S., Blakely, G., Koutsovoulos, G., Blaxter, M., & Watson, M. (2015). A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience*, *4*(1), 60.

Riva-Murray, K., Bode, R. W., Phillips, P. J., & Wall, G. L. (2002). Impact source determination with biomonitoring data in New York State: concordance with environmental data. *Northeastern Naturalist*, *9*(2), 127-162.

Robinson, C. T. & Minshall, G. W. (1986). Effects of disturbance frequency on stream benthic community structure in relation to canopy cover and season. *Journal of the North American Benthological Society*, *5*(3), 237-248.

Roura-Pascual, N., Bas, J. M., & Hui, C. (2010). The spread of the Argentine ant: environmental determinants and impacts on native ant communities. *Biological Invasions*, *12*(8), 2399-2412.

Roque, F. O., Trivinho-Strixino, S., Milan, L., & Leite, J. G. (2007). Chironomid species richness in low-order streams in the Brazilian Atlantic Forest: a first approximation through a Bayesian approach. *Journal of the North American Benthological Society*, *26*(2), 221-231.

Roque, F. O., Siqueira, T., & Escarpinati, S. C. (2009). Do fallen fruit-dwelling chironomids in streams respond to riparian degradation? Pan-American Journal of Aquatic Sciences, 4(3), 357–362.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., & Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, *238*(1), 195-209.

Sandifer, P. A., Sutton-Grier, A. E., & Ward, B. P. (2015). Exploring connections among nature, biodiversity, ecosystem services, and human health and well-being: Opportunities to enhance health and biodiversity conservation. *Ecosystem Services*, *12*, 1-15.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, *74*(12), 5463-5467.

Shaffer, M. L. (1981). Minimum population sizes for species conservation. *BioScience*, *31*(2), 131-134.

Shin, J., Lee, S., Go, M. J., Lee, S. Y., Kim, S. C., Lee, C. H., & Cho, B. K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific reports*, *6*, 29681.

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular ecology resources*, *14*(5), 892-901.

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., & Hajibabaei, M. (2015a). A DNA mini-barcoding system for authentication of processed fish products. *Scientific reports*, *5*.

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... & Hajibabaei, M. (2015b). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, *5*.

Silva, F. L., Ekrem, T., & Fonseca-Gessner, A. A. (2013). DNA barcodes for species delimitation in Chironomidae (Diptera): a case study on the genus Labrundinia. *The Canadian Entomologist*, *145*(6), 589-602.

Silva, F. L., & Wiedenbrug, S. (2014). Integrating DNA barcodes and morphology for species delimitation in the Corynoneura group (Diptera: Chironomidae: Orthocladiinae). *Bulletin of entomological research*, *104*(1), 65-78.

Simberloff, D. (1998). Flagships, umbrellas, and keystones: is single-species management passé in the landscape era?. *Biological conservation*, *83*(3), 247-257.

Sinclair, C. S., & Gresens, S. E. (2008). Discrimination of Cricotopus species (Diptera: Chironomidae) by DNA barcoding. *Bulletin of entomological research*, *98*(6), 555-563.

Smith, M. A., & Fisher, B. L. (2009). Invasions, DNA barcodes, and rapid biodiversity assessment using ants of Mauritius. *Frontiers in Zoology*, *6*(1), 31.

Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the national academy of sciences*, *105*(36), 13486-13491.

Sović, I., Šikić, M., Wilm, A., Fenlon, S. N., Chen, S., & Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature communications*, *7*, 11307.

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

Stein, E. D., Martinez, M. C., Stiles, S., Miller, P. E., & Zakharov, E. V. (2014). Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States. PloS one, 9(4), e95525.

Strayer, D. L. (2006). Challenges for freshwater invertebrate conservation. *Journal of the North American Benthological Society*, *25*(2), 271-287.

Strayer, D. L., & Dudgeon, D. (2010). Freshwater biodiversity conservation: recent progress and future challenges. *Journal of the North American Benthological Society*, *29*(1), 344-358.

Stribling, J. B. (2006). Environmental protection using DNA barcodes or taxa?. *AIBS Bulletin*, *56*(11), 878-879.

Strutzenberger, P., Brehm, G., & Fiedler, K. (2011). DNA barcoding-based species delimitation increases species count of Eois (Geometridae) moths

in a well-studied tropical mountain forest by up to 50%. *Insect Science*, *18*(3), 349-362.

Stur, E., & Ekrem, T. (2011). Exploring unknown life stages of Arctic Tanytarsini (Diptera: Chironomidae) with DNA barcoding. *Zootaxa*, *2743*(18), 27-39.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular ecology*, *21*(8), 1789-1793.

Taenzler, R., Sagata, K., Surbakti, S., Balke, M., & Riedel, A. (2012). DNA barcoding for community ecology-how to tackle a hyperdiverse, mostly undescribed Melanesian fauna. *PLoS One*, *7*(1), e28832.

Takahashi, M. A., Higuti, J., Bagatini, Y. M., Zviejkovski, I. P., & Velho, L. F. M. (2008). Composition and biomass of larval chironomid (Insecta, Diptera) as potential indicator of trophic conditions in southern Brazil reservoirs. *Acta Limnologica Brasiliensia*, *20*(1), 5-13.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, *30*(12), 2725-2729.

Tarkowska-Kukuryk, M., & Mieczan, T. (2014). Distribution and Environmental Determinants of Chironomids (Diptera, Chironomidae) in Sphagnum Microhabitats. *Polish Journal of Environmental Studies*, *23*(2).

Ter Braak, C. J. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, *69*(1), 69-77.

Thienemann, A. (1921). *Biologische Seetypen und die Gründung einer hydrobiologischen Anstalt am Bodensee*. Archiv für Hydrobiologie 13, 347-70.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4-18.

Tilman, D. (2000). Causes, consequences and ethics of biodiversity. *Nature*, *405*(6783), 208.

Townsend, S. A. (1999). The seasonal pattern of dissolved oxygen, and hypolimnetic deoxygenation, in two tropical Australian reservoirs. *Lakes & Reservoirs: Research & Management*, *4*(1-2), 41-53.

Turner, I. M., Boo, C. M., Wong, Y. K., Chew, P. T., & Ibrahim, A. B. (1996). *Freshwater swamp forest in Singapore, with particular reference to that found around the Nee Soon Firing Ranges*. National Parks Board.

Tsui, C. K., Woodhall, J., Chen, W., Andrélévesque, C., Lau, A., Schoen, C. D., ... & de Hoog, S. G. (2011). Molecular techniques for pathogen identification and fungus detection in the environment. *IMA fungus*, *2*(2), 177-189.

Quick, J., Quinlan, A. R., & Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience*, *3*(1), 1.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., ... & De Pinna, E. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome biology*, *16*(1), 114.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., ... & Ouédraogo, N. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228.

Quinlan, R. and Smol, J. P. (2002). Regional assessment of long-term hypolimnetic oxygen changes in Ontario (Canada) shield lakes using subfossil chironomids. *Journal of Paleolimnology*, *27*(2), 249-260.

Vaughn, C. C. (2010). Biodiversity losses and ecosystem function in freshwaters: emerging conclusions and research directions. *BioScience*, *60*(1), 25-35.

Vié, J. C., Hilton-Taylor, C., & Stuart, S. N. (Eds.). (2009). *Wildlife in a changing world: an analysis of the 2008 IUCN Red List of threatened species*. IUCN.

Vlek, H. E., Šporka, F., & Krno, I. J. (2006). Influence of macroinvertebrate sample size on bioassessment of streams. *The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods*, 523-542.

Voigt, O., Eichmann, V., & Wörheide, G. (2012). First evaluation of mitochondrial DNA as a marker for phylogeographic studies of Calcarea: a case study from Leucetta chagosensis. *Hydrobiologia*, *687*(1), 101-106.

Yee, A. T. K., Corlett, R. T., Liew, S. C., & Tan, H. T. (2011). The vegetation of Singapore-an updated map. *Gardens' Bulletin Singapore*, *63*(1&2), 205-212.

Yeo, D. C. J., & Lim, K. K. P. (2011). Freshwater ecosystems. *Singapore biodiversity: an encyclopedia of the natural environment and sustainable development. Editions Didier Millet, Singapore*, *5263*.

Yule, C. M. (2010). Loss of biodiversity and ecosystem functioning in Indo-Malayan peat swamp forests. *Biodiversity and Conservation*, *19*(2), 393-409.

Walker, B. H. (1992). Biodiversity and ecological redundancy. *Conservation biology*, *6*(1), 18-23.

Walshe, B. M. (1947). Feeding mechfanisms of Chironomus larvae. *Nature*, *160* (4066), 474-474.

Ward, R. D. (2009). DNA barcode divergence among species and genera of birds and fishes. *Molecular ecology resources*, *9*(4), 1077-1085.

Wazbinski, K. E., & Quinlan, R. (2013). Midge (Chironomidae, Chaoboridae, Ceratopogonidae) assemblages and their relationship with biological and physicochemical variables in shallow, polymictic lakes. *Freshwater Biology*, *58*(12), 2464-2480.

Wheeler, Q. D. & Pennak. S. (2011). State of Observed Species. International Institute for Species Exploration, 11pp. Retrieved from http://species.asu.edu/SOS, 13th July 2017.

Whitmore, T. C. (1985). Rain Forests. (The State of Ecology: Tropical Rain Forests of the Far East). *Science*, *228*, 874-875.

Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist?–a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in zoology*, *4*(1), 8.

Will, K. W., & Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, *20*(1), 47-55.

Will, K. W., Mishler, B. D., & Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Systematic biology*, *54*(5), 844-851.

Williams, W. D., Boulton, A. J., & Taaffe, R. G. (1990). Salinity as a determinant of salt lake fauna: a question of scale. *Hydrobiologia*, *197*(1), 257-266.

Wilson, K. H. (1995). Molecular biology as a tool for taxonomy. *Clinical infectious diseases*, *20*(Supplement_2), S117-S121.

Wong, W. H., Tay, Y. C., Puniamoorthy, J., Balke, M., Cranston, P. S., & Meier, R. (2014). 'Direct PCR' optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction. Molecular ecology resources, 14(6), 1271-1280.

Wright, I. A., & Burgin, S. (2007). Species richness and distribution of eastern Australian lake chironomids and chaoborids. *Freshwater Biology*, *52*(12), 2354-2368.

Zaaijer, S., Gordon, A., Piccone, R., Speyer, D., & Erlich, Y. (2016). Democratizing DNA Fingerprinting. *bioRxiv*, 061556.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational biology*, *7*(1-2), 203-214.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2013). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614-620.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3-14.

Appendices

Appendix 1

Supplementary tables and figure for chapter 2

Tables T1-2

Supplementary Table T1: The correlation coefficients between communities with different sampling intervals, determined by random permutation testing (999 permutations). † Comparison among sites, only 3 sampling intervals are shown. ˜ p < .10, * p < .05, ** p < .01.

Sampling frequency	Sites †	Mantel's *r*	Significance level
every two weeks	FAD-FDA	0.20	(0.009**)
	FAD-WBA	0.46	(0.001**)
	FDA-WBA	0.38	(0.001**)
every four weeks	FAD-FDA	0.12	(0.09˜)
	FAD-WBA	0.56	(0.002*)
	FDA-WBA	0.29	(0.01**)
every six weeks	FAD-FDA	0.17	(0.04*)
	FAD-WBA	0.55	(0.001**)
	FDA-WBA	0.39	(0.001**)

Supplementary Table T2: Matrix with similarity values for midge communities sampled at two weeks interval across sites. Midge communities, similarity values between each bi-weekly sample: Center (1-4), and edge sites: FAD (5-8), FDA (9-12) and WBA (13-16). All similarity values were calculated using the multiple community similarity index in Spade program in R; q value = 2 and bootstrapping set to 200 replicates. Data shown for the first four sampling occasions only.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	1	1	1	1	0.08	0.26	0.72	0.29	0.00	0.19	0.00	0.06	0.03	0.15	0.15	0.00
2		1	0.99	1	0.08	0.26	0.73	0.29	0.00	0.19	0.00	0.07	0.03	0.15	0.15	0.00
3			1	0.99	0.08	0.24	0.71	0.28	0.00	0.18	0.00	0.06	0.03	0.15	0.15	0.00
4				1	0.08	0.26	0.72	0.29	0.00	0.19	0.00	0.06	0.03	0.15	0.15	0.00
5					1	0.61	0.53	0.81	0.95	1	0.87	0.65	0.99	0.99	0.96	0.88
6						1	0.76	0.89	0.59	0.60	0.74	0.44	0.57	0.62	0.66	0.56
7							1	0.84	0.42	0.59	0.52	0.36	0.44	0.56	0.56	0.39
8								1	0.74	0.80	0.95	0.55	0.71	0.78	0.78	0.62
9									1	0.94	0.93	0.70	0.96	0.95	0.97	0.93
10										1	0.86	0.71	0.97	0.99	0.97	0.85
11											1	0.72	0.82	0.87	0.92	0.80
12												1	0.66	0.70	0.76	0.60
13													1	0.98	0.95	0.94
14														1	0.99	0.88
15															1	0.92
16																1

Supplementary Figure

Figure S1: Automatic Barcode Gap Discovery (ABGD) results