FROM BARCODING TO METAGENOMICS: MOLECULAR IDENTIFICATION TECHNIQUES FOR ECOLOGICAL STUDIES OF ENDANGERED PRIMATES

AMRITA SRIVATHSAN

B. Sc. (Hons.)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

AND

DEPARTMENT OF LIFE SCIENCES

IMPERIAL COLLEGE LONDON

2014

DECLARATION

I hereby declare that this thesis is my original work. I have duly acknowledged all the sources of information which have been used in the thesis.

Amrita Srivathsan

31 July 2014

"I asked the question for the best reason possible, for the only reason, indeed that excuses anyone for asking any question - simple curiosity."

– Oscar Wilde

ACKNOWLEDGEMENTS

My most sincere gratitude goes to a number of people whose contributions were invaluable during my PhD studies:

Prof. Rudolf Meier, it seems only fitting to begin this with Wilde, who, through you, had a heavy hand in the first part of the thesis. Thank you. For nurturing scientific thinking in me, for being the most supportive supervisor one can imagine, and for a lot of reasons that would be difficult to list. Your immense knowledge and ideas in many different aspects of this field lead to discussions that give me perspective on the questions we are asking and always leave me instigated about research. I am fortunate to have an opportunity to be supervised by you.

Prof. Alfried Vogler, for hosting me in the Natural History Museum, which was my first time in such an environment, and it was a great experience. Thanks a lot for your extremely valuable suggestions, for giving me access to facilities that were critical for this thesis, for teaching me how scientific writing is done and most importantly showing me how to think big and work towards it. It is truly inspiring, and I learnt a lot about genomics and molecular systematics during my stay in NHM.

To both my supervisors I am really grateful for your support during the sudden change in timeline, so that that the thesis could be put together.

Andie Ang, without whom this project would not be possible. For collecting the samples, for plant collections, for molecular work, for data validation and for monkey-talk.

John Sha for providing the data for retention time for douc langurs. Members of Singapore Zoological gardens, for all their help with the feeding trials for douc langurs.

Members of Nee Soon Swamp Forest survey team and Mirza Rifqi Ismail for collecting plant samples, and vouchering them. Teo Li Young and Tay Ywee Chieh for their help with sequencing plant barcodes. Chong Kwek Yan for his insights into plant community in Nee Soon.

Simon Burbidge and Peter Foster, the people behind the scenes managing the servers where this work was done. I am fairly sure that I have spent more time with these servers than people in last two years and they have never left me frustrated.

AITBiotech members who generated the Next Generation Sequencing data for this project. NUS for funding my studies through President's Graduate Fellowship and for funding travel to and from London, Illumina for funding the MiSeq runs, Ministry of Education and National Parks Board for funding the project.

Members of two labs: From NUS, Jayanthi, for tea breaks and giving amazing company here, you will be dearly missed. Sujatha, who taught me how to run my first PCR and who was *there and back again*. Lei, my batchmate who is writing this along with me, your support during writing helped me a lot. Kathy, for all the coffee breaks and bearing with random statements from my corner and brainstorming with me. Yuchen, for never failing to entertain, and helping with formatting this. Darren and Youguang for helping me with crosschecking and forms. Denise, Wing Hing, Shiyang, Mindy, Youguang, Ywee Chieh, Diego, Gwynne, Gowri, Jinfa, Li Young, Bilge, Amy, you all made lab a really great place. From NHM, Chris, for sharing your pipeline that made the databases happen, Alex and Martijn, for sharing your ideas and valuable discussions, Ben, Kirsten, Nicole, Debora, Martin, Samia, Carmelo, Conrad, Paula, for your great company.

A number of friends, in particular, Shweta (whose gift of white tea was a constant company during writing), Akshat, Manali, Shefali, Aishwarya, Shy, Anupama, Janani, Eli, Seetha and Souvik, thanks for putting up with me especially through this last year of strange level of communication.

Anjali Ma'am and Hindustani greats for keeping me sane, even if momentarily.

Amma, Appa, and Atreya, for being great role models in life and academia and for being my backbone throughout.

TABLE OF CONTENTS

SUMMARY

DNA markers are increasingly used for studying environmental samples that contain DNA from multiple species. After sequencing a subsample of the extracted DNA, sequences are identified by matching them to so-called "DNA barcodes", i.e. short reference sequences from well-identified specimens. In my thesis, I address a range of methodological challenges that are associated with such matching of unidentified and identified sequences. I first demonstrate that some of the currently used identification techniques based on K2P distances are flawed and argue that simpler metrics such as uncorrected distances should be used (Chapter 2). Next, I reveal that DNA barcode species coverage for Metazoa in open-access databases remains poor (Chapter 3). I also generate large barcode databases for animals and plants needed for my studies on colobine monkeys and their diet (Chapter 4). The last two chapters (5 and 6) use these databases for studying fecal DNA from two species of colobine monkeys (*Pygathrix nemaeus* and *Presbytis femoralis*). Based on a set of plant barcode sequences, I identify the diet and obtain information on the genetics and parasite infestation of the host. While this was primarily based on direct shotgun sequencing ("metagenomics"), I also test an alternative PCR-based "metabarcoding" approach using deep sequencing of amplicons. I develop and optimize new methods for read-based identification and compare the results of either approach. I conclude that metagenomics is preferable because it simultaneously

provides information on diet, host genetics, and parasite infestation. In addition, metagenomic data provide more taxonomic precision for identifying plant species than the short barcodes used in metabarcoding. However, I find a correlation between read counts as obtained by either method so that the simpler metabarcoding may still be useful for diet quantification. When applied to fecal samples of endangered banded leaf monkeys (*Presbytis femoralis*), shot-gun sequencing reveals a diverse dietary profile, which recovers most of the diet identified by direct field observation. Upon characterizing the monkeys' mitochondrial genomes also present in the feces, I find very low genetic variability but I was able to detect heteroplasmy in the mitochondrial DNA. Lastly I find parasites such as *Strongyloides, Oesophagostomum, Entamoeba* and *Blastocystis* in the gut of these primates. Overall, these studies expose the enormous power of recent sequencing technologies in ecological research, to study species interactions and ecosystem function based on well constructed barcode reference data.

List of Figures

List of Tables

List of publications

(Listed by relevance to the thesis)

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

2. Kwong S., **Srivathsan A.**, and Meier, R. (2012) An update on DNA Barcoding: Low species coverage and an increasing number of unidentified barcodes. *Cladistics* 28(6): 639-644. Chapter 3

3. **Srivathsan, A.**, Sha, J.C.M., Vogler, A.P., and Meier R. (2014). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources.* doi:10.1111/1755-0998.12302. Chapter 5

4. Ang, A., **Srivathsan, A.**, Md-Zain, B., Ismail, M. and Meier, R. (2012). Low genetic variability in the recovering urban banded leaf monkey population of Singapore. *The Raffles Bulletin of Zoology* 60(2): 589-594.

5. Wong, H.F., Tan, S.Y., Koh, C.Y., Siow, H.J.M., Li, T., Heyzer, A., Ang, A., Ismail, M., **Srivathsan, A.,** Tan, H.T.W. (2013). *Checklist of the plant species of Nee Soon Swamp forest, Singapore: Bryophytes to Angiosperms.* National Parks Board and Raffles Museum of Biodiversity Research, National University of Singapore, Singapore. 521 pp.

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

7. Lei, Z., Ang, A.S.H., **Srivathsan, A.**, Su, K.F.Y., and Meier, R. (2013). Does better taxon sampling help? A new phylogenetic hypothesis for Sepsidae (Diptera: Cyclorrhapha) based on 50 new taxa and the same old mitochondrial and nuclear markers. *Molecular Phylogenetics and Evolution* 69(1): 153-164.

CHAPTER 1

General Introduction

Reconstructing history: why it is still relevant and what are the new opportunities

Over two thousand years ago, Archimedes of Syracuse wrote several treatises. Unfortunately, many were lost and considered irretrievable. When a copy of his *Method* was rediscovered as recently as 1906, it led to a great deal of excitement among, for example, mathematicians because there had been gaping hole in our understanding of how the Greeks came to discover their great theorems. Indeed, humans have a fascination with history and have developed many techniques for reconstructing events that took place in the past. Historians reconstruct history by deciphering records and reading and translating documents, but scientists have similar interest in history. Sometimes it is an interest in the history of science, but other times scientists use a variety of different tools for reconstructing biological change that happened in the past. Today's biologists often use DNA sequences to infer events that humans were not able to observe directly. For several decades, it has been routine to use these signatures for reconstructing the tree of life that reflects the relationships between organisms and species. Based on the trees, scientists have also been able to reconstruct evolutionary change of, for example, morphological and behavioural traits. These reconstructions have yielded much information about the origin of our planet's diversity. On the other hand, DNA sequences can also be used to reconstruct more recent, specific events. This is a familiar territory in forensics, where genetic information is regularly used to

reconstruct crime scenes using trace DNA left by victims, perpetrators, and innocent bystanders.

An ecosystem is not all that different from a crime scene. In principle, biologists can use DNA remnants to identify the DNA signatures left by the protagonists. However, until recently, getting a reasonable subset of all DNA contained in an environmental sample was far from trivial. Fortunately, recent advances in genomic technologies allow for generating large amounts of sequence data from biological samples, which has opened the door for sequencing complex environmental samples. DNA from such samples include signatures from numerous organisms belonging to many species. For instance, Venter *et al.* (2004) used samples from Sargasso sea and discovered nearly 1800 genomic species, with at least 148 of them being previously unknown. Over the years, sequencing DNA from such samples has become an exciting venture for scientists; especially for those who study the largely unknown diversity of microbial communities. For example, this approach has been applied to the microbiota of soil in order to understand and discover the diversity of microbes (East 2013). Similar studies have been carried out for samples of air and feces, in part because the microbial faunas can affect human health (Qin *et al.* 2010; Tringe *et al.* 2008).

These studies generally utilize two different approaches. Most still use PCR-based pre-amplification of a particular genetic marker. A good example is the use of 16S in microbial biology. After amplification, the thousands of sequences for different organisms that have been generated are sequenced using Next Generation Sequencing (NGS) technologies (e.g. Arboleya *et al.* 2012; Yu *et al.* 2012). The alternative approach uses shotgun sequencing where a subset of the extracted DNA is directly sequenced thus generating a very large number of random reads representing the entire genomes of the

community of organisms present in the samples (Wang *et al.* 2013; Xu *et al.* 2013). The second method is untargeted and thus allows for the identification of all genes instead of only genes and taxa that were targeted during pre-amplification (Eisen 2007). The downside is the higher cost and potentially more challenging bioinformatics given that the selection of genes is based on computation instead of relying on biochemical pre-selection techniques as in pre-amplification.

The term "metagenomics" has been applied to both approaches (Junemann *et al.* 2012; Rasheed *et al.* 2013). The definition of the term is further obscured because several authors have used it for the study of microbial communities only. For example, in the very first usage by Handelsman *et al.* (1998) they consider metagenomics to be the "analyses of collected genomes of (soil) microflora". Yet intuitively, "metagenome" carries a more inclusive meaning given that an environmental sample need not contain only a set of genomes that are microbial in origin. A broader definition of metagenomics would thus be similar to the one described by Thomas *et al.* (2012): "the direct genetic analyses of genomes contained with an environmental sample". A direct genetic analysis would be one where there is no enrichment for any taxon or gene. Essentially, this implies an untargeted approach where genomic data generated is categorised into various taxonomic and functional categories *after* data generation. In my thesis I adhere to this definition and distinguish the "metagenomic approach" from a "metabarcoding approach" that utilizes deep sequencing of PCR-based amplicons to generate taxonomic profiles of complex environmental samples.

Metagenomics has largely been made possible by the development and reduced cost of NGS technologies. In their earlier days, NGS technologies posed numerous limitations. Technologies that yielded long sequences (454 pyrosequencers) produced fewer reads so that

mostly the dominant taxa in a metagenomic community could be characterized (Liu *et al.* 2012). In contrast, newer, short read technologies (Solexa, SoLiD) produced reads of <50bp, generated a large number of reads, but struggled to provide enough information for deciphering the taxonomic composition of a sample (Liu *et al.* 2012). The situation has changed in recent years, with read lengths of the short-read technologies increasing (Liu *et al.* 2012). Today we are able to obtain large datasets where sequences are long enough for taxonomic assignments to family, genus, or even species (Thomas *et al.* 2012). Note, however, that metagenomics is often used to study DNA from samples such as feces, soil, fossils *etc.* which contain much degraded, short-length DNA. In such cases, read length is partially determined by sample origin and sequencing technology.

In its initial days, metagenomics was largely popular among microbiologists, because it filled a knowledge gap given that most microbial species and clades are unknown (Eisen 2007). Application of metagenomics to eukaryotes, particularly Metazoa, has been a recent phenomenon. For example, Zhou *et al.* (2013) characterized bulk arthropod samples using metagenomics in an attempt to reveal their diversity. Bon *et al.* (2012) studied the mitochondrial DNA from cave hyena coprolites and found a potential diet species for this extinct mammal. Presumably, the lack of studies using metagenomics is due to cost of sequencing (Andrew *et al.* 2013) because in many instances the proportion of DNA of desired taxa is very low. This means that the depth of sequencing has to be very high in order to capture sufficient DNA for the target taxa. However, with decreasing cost of NGS, I will argue that it is now becoming feasible to generate sufficient coverage to characterize rare DNA in metagenomes. I will demonstrate that the plant diet of phytophagous monkeys with long digestion times can be reconstructed although much fewer than <1% of all shotgun sequencing reads pertain to diet species.

Unfortunately, the nature of metagenomic data poses numerous challenges to taxonomic categorization (Eisen 2007). This is exacerbated when the data are generated using short-read technologies as short sequences may not have enough diagnostic information for classifying the sequences to an informative taxonomic level. In addition, the existing methods are largely designed and developed for microbial metagenomics (Thomas *et al.* 2012) and cannot be directly used for the purpose of identifying eukaryotes such as Metazoa and plants. This is due to several factors. Firstly, the genetic markers used for taxonomic identifications in these organisms are different and hence different reference databases have to be used (Hebert *et al.* 2003; Kress & Erickson 2007). Secondly, there is some consensus among microbiologists that sequences clustered at 3% can be used as species-equivalents. For Eukaryotes, fixed distance thresholds have also been used; for example in its first version Barcode of Life Datasystems used a 1% threshold (Ratnasingham and Hebert, 2007); and several studies have used a 2 or 3% threshold (Hebert *et al.,* 2003; Strutzenberger *et al.,* 2011; Ng'endo *et al.*, 2013; Song *et al.,* 2008). However, a universal threshold fails to delimit many groups of organisms (Meier *et al.,* 2006; 2008; Renaud *et al.*, 2012; Meyer and Paulay, 2005) and hence has met criticism (Collins *et al.,* 2012; Puillandre *et al.*, 2011). Thirdly, for many ecological questions involving animals and plants high-precision taxonomic information, i.e., identifications to species or genus are desired (Aylagas *et al.* 2014; Campos-Arceiz 2013). Lastly, the DNA for many eukaryotes is present in extremely low frequency in metagenomes which requires the development of identification methods that allow for the identification of low frequency reads with high reliability.

For identifying sequences from metagenomic data, one has to match sequences from the metagenome to reference databases containing identified sequences of known taxonomic

origin. For Metazoa and plants, this can be achieved using databases of DNA 'barcodes'. While the use of DNA for species identification has a long history (Will *et al.* 2005), the term "DNA barcoding" was proposed in 2003 when a 658 bp fragment of Cytochrome Oxidase Subunit I (COI) was first used for identifying sequences to species (Hebert *et al.* 2003). DNA barcoding has since evolved into a large scale initiative that intends to provide DNA barcodes for all described species on our planet. Since the initial proposal of COI as a barcode for Metazoa, different genes have been proposed as barcodes for other groups such as plants (Kress *et al.* 2005) and fungi (Schoch *et al.* 2012). For plants, it has been difficult to reach a consensus on which gene(s) should be used. Currently, two barcodes are recommended by CBOL (Consortium for the Barcode of Life). They are *rbcL* and *matK*. However, these barcodes overall lack taxonomic resolution and many closely related species pairs have identical sequences. Thus, a number of combinations of other genes have been proposed as alternatives (Hollingsworth 2011). Two proposals that have been widely adopted are the addition of *trnH-psbA* (Kress & Erickson 2007) and *nrITS* (Li *et al.* 2011) to the core barcodes. However, even with these additions, several lineages of plants require additional barcodes (Li *et al.* 2014). Nonetheless these efforts have led to the accumulation of barcode sequences for both plants and animals in public databases such as GenBank which now contain information on thousands of species/genera, many of which can now be identified based on DNA markers.

In this thesis, I propose to use a metagenomic approach to characterizing diet and other aspects of biology of endangered species. More specifically, I propose to do this for DNA generated from fecal samples which contain the DNA of endangered species (from shed cells from the gut lining), diet items (plants whose DNA was incompletely digested), microbes, and parasites residing in the intestine. Fecal samples have been used extensively to

study genetics (Munshi-South & Bernard 2011), diet composition (Mohammad *et al.* 1995), and microbial ecology (Ley *et al.* 2008). However, these studies analyzed only one particular aspect of an animal's biology. Even NGS based studies were using a targeted approach (Deagle *et al.* 2010; Deagle *et al.* 2009; Nossa *et al.* 2010; Taberlet *et al.* 2009). For example, if a researcher was interested in diet, he would pre-amplify genes for putative diet items; if a researcher was interested in the microbiome, he would use primers for a microbial marker. However, an untargeted metagenomic approach can address these questions simultaneously and this is what I pursue in two chapters of the thesis. Such an approach has the potential to reveal unexpected and genuinely novel information on biology. For example, if metagenomics is used to address diet, it could reveal carnivory in species that have been considered herbivorous. However, a metagenomic approach to studying diet is not without its problems. For example, before it can be used one has to develop very sensitive methods for finding rare reads in metagenomic data. This is necessary because many relevant DNA sequences will be present in only very small concentrations. In studies using amplification-based approaches, finding rare seqeunces is based on the high affinity primers to specific sequences; i.e., there is a biochemical filtering mechanism. In metagenomics, it is necessary to find an efficient bioinformatic filter.

The methodological problems are two-fold. Firstly, there has been considerable debate on methods of identifications of unknown sequences even when full-length barcode sequences are available (Little 2011; Little & Stevenson 2007; Meier *et al.* 2008). This has led to detailed discussions and refinements in the methods of identification of sequences. (Fan *et al.* 2014; Little 2011; Little & Stevenson 2007; Meier *et al.* 2006). Due to the complexity of metagenomes and thereby, the computational requirements, several of these methods cannot be directly utilized (e.g., those methods based on multiple sequence

alignments). Instead, it appears most promising to first optimize distance-based approaches for species identification. These approaches require the alignment of a query sequence to the reference sequence, and the subsequent calculation of distances between these two sequences. The simplest measure is uncorrected pairwise distances which measure the number of nucleotide differences between two sequences. On the other hand in the barcoding literature, K2P distance, i.e., distances measured after correction using the Kimura-2-parameter model, is used widely (Hubert *et al.* 2008; Zhang & Zhang 2014). In my thesis I demonstrate that this is an inappropriate use of K2P distances and discuss the problems with using the model. Throughout the rest of my thesis, I then use uncorrected distances for species identifications.

The second problem with identifying species in metagenomes is ensuring accuracy given that the reads of metagenomes tend to be very short. For microbes, short reads can generally be assembled prior to identifications because most microbe DNA sequences are present in large numbers (Mande *et al.* 2012; Qin *et al.* 2010). However, assembly-based approaches will generally not be suitable for detecting low frequency sequences because they will fail to assemble due to lack of overlap (Sharpton *et al.* 2011). In order to effectively scrutinize metagenomes for rare reads, low frequency sequences have to be identified directly. In this thesis, I develop strategies for the identification of such low abundance reads and demonstrate how they can be used to identify the diet items of two species of colobine monkeys that are phytophagous. I furthermore test whether different methods of alignment improve the accuracy of identification.

Of course, any new method has to justify its existence by first demonstrating that it is an improvement over an existing method. With regard to metagenomic approaches I can argue that they have the obvious advantage that they simultaneously characterize diet, host

genetics, parasites, and microbiome (Qin *et al.* 2010). Yet, in order to convince biologists that metagenomics should be used for diet characterization, it is important to show that this approach is preferable over the existing methods. Currently, the most widely used method for plant diet identification from fecal material is based on 'metabarcoding'. Plant sequences are first amplified for a particular barcoding gene by PCR. Afterwards, NGS is used to sequence the amplicons that may represent multiple species. For identifying plant diets, the choice of barcoding gene is the P6 loop of *trnL* (Taberlet *et al.* 2007; Valentini *et al.* 2009). Therefore, I test in my thesis whether metagenomics outperforms metabarcoding when applied to the same samples.

1.1 Aims and outline of the thesis

The main aim of this thesis is to optimize methods for a metagenomic approach to analysing the eukaryote DNA contained in fecal samples of an endangered population of leaf monkeys. The species of interest is the banded leaf monkey *Presbytis femoralis,* whose Singaporean population is critically endangered. Although I focus on diet and fecal samples, the methods used in my thesis are generally applicable to sequences in any environmental sample.

In chapter 2, I criticize the use of the Kimura-2-parameter model which is widely used in the DNA barcoding literature. I demonstrate that K2P is not an appropriate model for measuring distance in DNA barcodes and argue that uncorrected *p*-distance ought to be used. This article, published in *Cladistics,* is widely cited (Google Scholar: 63 citations as of 28.7.2014). Throughout the rest of the thesis, I identify sequences by the simple metric of "identity", i.e., % of identical nucleotides in an alignment.

In chapter 3, I discuss the issue of paucity of Metazoa barcodes in GenBank. I further discuss an important challenge for biologists trying to identify species based on barcode sequences: the lack of species coverage and the problem of 'dark' taxa, i.e., sequences without species names. I am a co-author of this publication in *Cladistics*, and I wrote the scripts required to extract the information from GenBank flatfiles, that led to the statistics described in the study. This study was conducted in early 2012, and thus reflects the status of the database two and half years ago.

In chapter 4, I describe the barcode databases used in this study. These databases are constructed using data from GenBank and new DNA barcodes that were generated at the National University of Singapore for the habitat of the Singaporean population of the banded leaf monkey. These sequences are part of an ongoing effort to barcode Nee Soon Swamp forest.

In chapter 5, I test whether a metagenomic approach can be used to address the diet of a species. The test involves feeding experiments with two individuals of douc langurs (*Pygathrix nemaeus*). Given that the diet was known, I was then able to test whether the diet items can be identified and how the identification techniques can be optimized to yield accurate genus-level identifications of diet taxa using sequences of 76 bp length. Furthermore I compare the identification success rates of the metagenomic approach with the success rates of metabarcoding. For the latter, the P6 loop of *trnL* was pre-amplified and then sequenced with NGS. I discuss the advantages and disadvantages of both methods. Among others, I demonstrate the multidimensionality of the metagenomic approach in that it characterizes parasites, the host mt-genome, as well as revealing an unexpected diet item. This manuscript was recently published in *Molecular Ecology Resources* and forms the methodological baseline for diet analyses of *Presbytis femoralis* in the next chapter.

In chapter 6, I present a case study based on six fecal samples of banded leaf monkey (*Presbytis femoralis femoralis*) from the critically endangered population of the species in Singapore. Based on previous research by Ang (2010), I had a preliminary list of diet species. I validate and compare the result of metagenomic analysis of the fecal samples with the observational data. I furthermore refine the metagenomic approach discussed in the previous chapter and characterize the diet, host mitochondrial genomes, and parasites present in the

gut of these primates. I discuss how these data will be important for the conservation of the small surviving population of banded leaf monkeys, which is estimated to consist of ~40 individuals (Ang 2010). The population seems to be recovering slowly from a low of <15 animals but the population viability remains unclear due to several constraints; availability of food resources (Ang 2010), carrying capacity (Yu *et al.* 2009), fertility, and genetic constraints (Ang *et al.* 2010; Ang *et al.* 2012).

CHAPTER 2[1]

On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA barcoding literature

2.1 Abstract

In this chapter, I present evidence based on ten data sets comprising 5,283 sequences for 200 genera that the use of the Kimura-2-parameter (K2P) model in DNA barcoding studies is poorly justified. I demonstrate that K2P is neither expected nor confirmed to be an appropriate model for closely related COI sequences. In addition, I show that the use of uncorrected distances yields higher or similar identification success rates for NJ trees and distance-based identification techniques. K2P also does not widen the barcoding gap for closely related sequences. I conclude that the spread of K2P through the barcoding literature is difficult to explain and urge the use of evidence-based approaches to DNA barcoding.

[1] A version of this chapter has been published as "**Srivathsan, A.,** and Meier, R. (2011). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. Cladistics 28(2): 1096-0031." I was the first author.

2.2 Introduction

DNA barcoding was proposed by Hebert *et al.* (2003) as a solution to the species identification problem caused by a mismatch between the number of employed taxonomists and species on our planet. High identification success rates were initially reported using a set of analytical techniques that remained largely untested given that the movement was in its infancy. What followed was a debate about, for example, the intellectual merits and analytical rigor of DNA barcoding (e.g. Sperling 2003; Moritz & Cicero 2004; Will & Rubinoff 2004; Will *et al.* 2005; Brower, 2006; Meier *et al.* 2006). Subsequently new analytical techniques were developed (e.g. Meier *et al.* 2006; Kuksa & Pavlovic 2007; Little & Stevenson 2007; Sarkar *et al.* 2008), but some of the poorly justified earlier methods continued to persist and flourish in the literature. I believe that it is important to address the methodological shortcomings of these techniques and I would argue that it is most effective to discuss them individually. For example, it was recently shown that the use of mean instead of closest interspecific distances leads to an overestimation of the so-called "barcoding gap" between the intra- and interspecific variability, thus giving investigators the erroneous and counterintuitive impression that species identification is getting easier as more species are sampled (Meier *et al.* 2008). In this chapter I will address the use of the Kimura-2-parameter model (Kimura 1980) in DNA barcoding studies. Here I describe some conceptual problems and test K2P using empirical data.

The original species identification method proposed by Hebert *et al.* (2003) involved the construction of neighbour-joining (NJ) trees based on K2P divergence which is measured in terms of nucleotide substitutions per site *d*, although the barcoding literature generally reports them as distances. *d* is given by:

$$d \equiv -\left(\frac{1}{2}\right)\ln(1-2P-Q)-\left(\frac{1}{4}\right)\ln(1-2Q) \qquad \text{… (Equation 3.1)}$$

$$\text{where } P = (\tfrac{1}{4})(1-2e^{-4(\alpha+\beta)t}+e^{-8\beta t})$$

$$Q = (\tfrac{1}{2})(1-e^{-8\beta t})$$

α and β represent the rate of transitional and transversional mutations per site per year and *t* is the time since divergence of the two sequences.

Sequences for the same species are generally considered to be correctly identified as long as they form a monophyletic cluster on an NJ tree and the intraspecific distances are below a threshold. It is beyond the scope of this chapter to summarize the arguments against the use of distances, the choice of *COI*, NJ trees, and monophyly as a criterion for determining identification success (see Will & Rubinoff 2004; DeSalle *et al.* 2005; Meyer & Paulay 2005; Will *et al.* 2005; Rubinoff 2006; Roe & Sperling 2007; Meier 2008; Meier *et al.* 2008; Ward *et al.* 2009), but the use of K2P requires more scrutiny given that distance-based techniques will continue to be popular in DNA barcoding. For example, a survey of the barcoding literature published in 2010 revealed that K2P was used in 106 publications (ca. 2/3 of all empirical barcoding studies published in 2010) which is probably partially due to the popularity of the Barcode of Life Datasystem (BOLD) (Ratnasingham and Hebert, 2007), which uses K2P for taxon ID trees.

This widespread use is surprising given that its justification in Hebert *et al.* (2003: 315) was brief: "For the species-level analysis, nucleotide-sequence divergences were calculated using the Kimura-2-parameter (K2P) model, the best metric when distances are

low (Nei & Kumar 2000) as in this study". The lack of justification has been pointed out before (e.g. Magnacca & Brown 2010; Moniz & Kaczmarska 2010), but in the absence of a comparative study based on data, the conceptual objections are unlikely to affect the barcoding literature. In addition, it is not uncommon that authors of DNA barcoding manuscripts are asked by reviewers and editors to use K2P. I hope that an explicit study of K2P's behavior will prove persuasive and encourage a more evidence-based approach to data analysis.

2.3 Materials and Methods

I tested the performance of K2P using the ten most recently published, suitable metazoan datasets from BOLD (accessed on March 4, 2011: Dettai *et al.* 2011; Ekrem, *et al.* 2010; Francis *et al*, 2010; James *et al*, 2010; Lakra *et al.* 2011; Mecklenburg *et al.* 2011; Pauls *et al.* 2010; Sweeney *et al.* 2011; Victor 2010). I deemed datasets unsuitable if they had <50 sequences, data for <10 species, very low identification success (Allcock *et al.* 2011), very short sequences (Baird *et al.* 2011), or major discrepancies between BOLD and the corresponding publication with regard to which sequences were identified to species (Baldwin *et al.* 2011). Unless mentioned otherwise, I removed sequences that were only identified to genus or family level (Ekrem *et al.* 2010; Dettai *et al.* 2011; Sweeney *et al.* 2011). However, species labeled with "sp." were retained as long as they formed monophyletic clusters on neighbor-joining trees.

I established the size of the barcoding gap for uncorrected and K2P distances using TaxonDNA (v 1.6.2) (Meier *et al.* 2006) by calculating the differences between the smallest inter- and the largest intraspecific distance for each sequence (see Meier *et al.* 2008). I tested whether K2P is an appropriate model for the data by submitting them to jModelTest (v 0.1.1) (Posada 2008) which uses an ML tree built by PhyML (v 2.4.4) (Guindon & Gascuel 2003). The appropriate models were chosen using the Akaike Information Criteria (AIC). Note that p-distances are not included in jModelTest and that the software only tests whether K2P is preferred over other models. I not only analyzed full datasets, but also each genus separately. The genus-level analyses were carried out because species identification relies on choosing appropriate models for closely related sequences. Note that for model-testing sequences identified to genus were included in the analyses.

In order to test whether analyses using K2P yield higher identification success than those based on p-distances, I constructed NJ trees for each of the ten datasets based on both kinds of distances using PAUP v 4.0 (Swofford 2003; with ties broken randomly). I then used a Python script to create a group membership character for each species and mapped it onto the tree in order to identify based on the consistency index of the characters which species were not monophyletic. I also conducted a "best close match" analysis as described in Meier *et al.* (2006) using TaxonDNA (v 1.6.2). The latter used the 1% threshold from BOLD and the 3% threshold that is often suggested in the barcoding literature. "Best Close Match" distinguishes between "correct", "incorrect", and "ambiguous" identifications; the latter is for sequences that have an equally good match to sequences from several species.

2.4 Results and Discussion

Any use of models requires justification. For K2P, Nei & Kumar (2005) argue that it ought to be used for sequences whose transition to transversion ratio is large, but the authors also stated that p-distances are preferable when sequences are short and derived from closely related species. Nei & Kumar (2005) point out that in these cases p-distances and model-based distances yield similar results and that a drawback of complex models is that they have more variance in estimating the model parameters (Nei & Kumar, 2005). Based on these comments, it already appears unlikely that K2P is appropriate for studies of short and closely related barcoding sequences. This prediction is confirmed by my model testing. jModeltest does not favor the use of K2P for any of the ten full or the 200 genus-level data sets (Fig. 2.1).

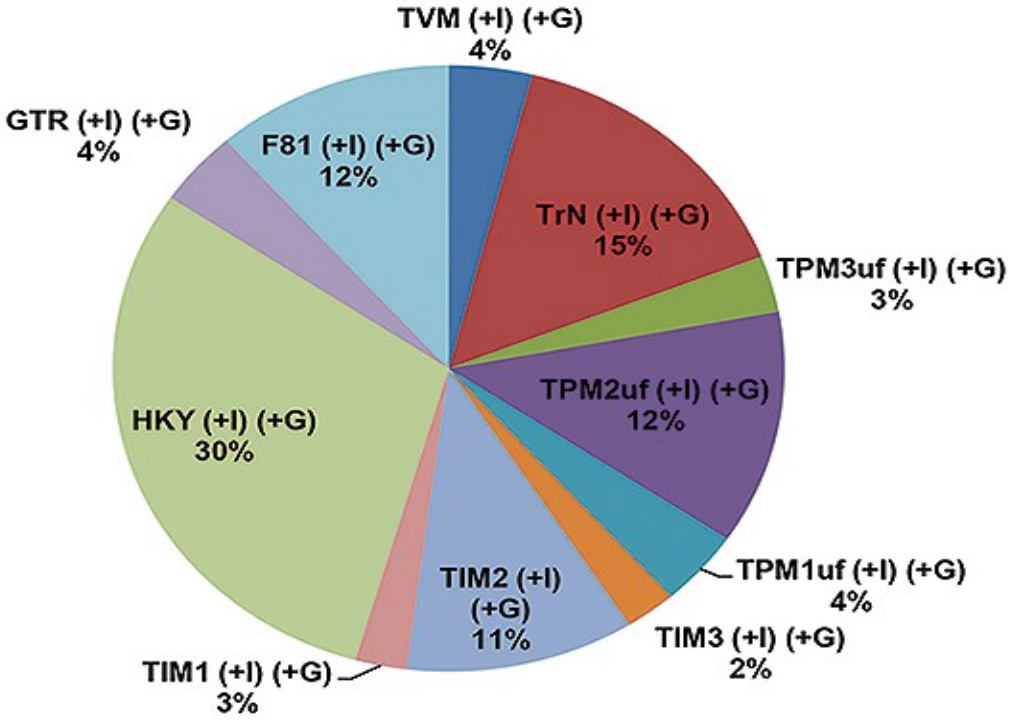


Figure 2.1: Models selected using Akaike Information Criterion (AIC) for the 200 genera in the ten datasets.

My analyses also reveal that the use of K2P does not increase species identification success rates. NJ trees remain the most popular method in distance-based barcoding studies although some studies suggest that it is the least accurate method (e.g. Meier *et al.* 2006; Little & Stevenson 2007). I found that for the ten datasets from BOLD, NJ trees based on p-distances performed better than those based on K2P. The former yielded more "monophyletic" species for the Dettai *et al.* (2011) dataset where two of the species were only paraphyletic on the K2P tree while being weakly supported as monophyletic on the NJ tree using p-distances (*Lycodichthys antarcticus*: bootstrap 36, *Paraliparis leobergi*: bootstrap 60). I can thus conclude that even if one were to adopt a utilitarian point of view of using whatever model increases identification success, K2P would not be a good choice.

Overall, K2P also does not increase the identification success when "best close match" is used (see Table 2.1 and Table 2.2). At a 1% threshold, two datasets have larger numbers of correct identifications using K2P and one using p-distances. At 3%, both methods yield the same result. A closer scrutiny of the analysis output reveals that the differences between K2P and p-distances are due to the probability of observing ambiguity. Given that transitions and transversions contribute differently to divergences based on K2P, K2P is less likely to yield ambiguous matches of a query sequence with sequences from multiple species. This may at first appear to be an advantage, but K2P more or less randomly breaks the ties identified by p-distances. For example, the Ekrem *et al.* (2010) and Lakra *et al.* (2011) datasets include two sequences for which the K2P analyses yield incorrect matches while the p-distances indicated ambiguity (using Best Match, i.e. without threshold).

Table 2.1: Summary of results of Best Close Match analysis with 1% threshold [numbers in brackets in column 1 are as follows: (Number of sequences/Number of sequences with at least one sequence overlapping by >300bp/Number of sequences with conspecifics/Number of species)].

Dataset	Correct		Incorrect		Ambiguities		Sequences without identities within 1%		Number of Singletons
	p	K2P	p	K2P	p	K2P	p	K2P	
Dettai *et al.* (555/555/540/73) (Actinopterygian fish)	96.1(519)	96.1(519)	0.7(4)	0.7(4)	2.2(12)	2.2(12)	0.9(5)	0.9(5)	15
Ekrem *et al.* (379/373/351/76) (Chironomidae)	95.4(337)*	95.4(335)	0(0)	0(0)	0(0)	0(0)	4(14)	4.6(16)*	22
Francis *et al.* (1896/1889/1862/160) (Mammals)	94.9(1768)	95(1770)*	0.43(8)	0.43(8)	0.05(1)*	0(0)	4.6(85)*	4.6(84)	27
James *et al.* (230/229/227/7) (Earthworms)	97.4(221)	97.4(221)	0(0)	0(0)	0(0)	0(0)	2.6(6)	2.6(6)	2
Lakra *et al.* (251/251/247/75) (Marine fish)	95.6(236)	95.6(236)	0.8(2)	0.8(2)	2.8(7)	2.8(7)	0.8(2)	0.8(2)	4
Mecklenburg *et al.* (684/684/649/111) (Marine fish)	96(623)	96(623)	0(0)	0(0)	3.1(20)	3.1(20)	0.9(6)	0.9(6)	35
Pauls *et al.* (463/463/452/42) (*Smicridea*)	94.5(427)	94.9(429)*	0.4(2)	0.4(2)	2(9)*	1.6(7)	3.1(14)	3.1(14)	11
Mitter *et al.* (72/70/64/11) (Butterflies)	100(64)	100(64)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	6
Sweeney *et al.* (686/686/680/45) (Aquatic macroinvertebrates)	87.2(593)	87.2(593)	1(7)	1(7)	7.1(47)	7.1(48)*	4.9(33)*	4.7(32)	6
Victor (67/67/60/20) (Actinopterygian fish)	98.3(59)	98.3(59)	1.7(1)	1.7(1)	0(0)	0(0)	0(0)	0(0)	7

* indicates the larger value

Table 2.2: Summary of results of Best Close Match analysis with 3% threshold [numbers in brackets in column 1 are as follows: (Number of sequences/Number of sequences with at least one sequence overlapping by >300bp/Number of sequences with conspecifics/Number of species)].

Dataset	Correct		Incorrect		Ambiguities		Sequences without identities within 3%		Number of Singletons
	P	K2P	P	K2P	p	K2P	p	K2P	
Dettai *et al.* (555/555/540/73) (Actinopterygian fish)	97.0 (524)	97.0 (524)	0.7 (4)	0.7(4)	2.2(12)	2.2(12)	0(0)	0(0)	15
Ekrem *et al.* (379/373/351/76) (Chironomidae)	98(344)	98(344)	0(0)	0(0)	0(0)	0(0)	2(7)	2(7)	22
Francis *et al.* (1896/1889/1862/160) (Mammals)	98.4(1832)	98.4(1832)	0.5(9)	0.5(9)	0.05(1)*	0(0)	1.1(20)	1.1(21)*	27
James *et al.* (230/229/227/7) (Earthworms)	99.1(225)	99.1(225)	0.4(1)	0.4(1)	0(0)	0(0)	0.4(1)	0.4(1)	2
Lakra *et al.* (251/251/247/75) (Marine fish)	95.5(236)	95.5(236)	0.8(2)	0.8(2)	2.8(7)	2.8(7)	0.8(2)	0.8(2)	4
Mecklenburg *et al.* (684/684/649/111) (Marine fish)	96.7(628)	96.7(628)	0(0)	0(0)	3.1(20)	3.1(20)	0.2(1)	0.2(1)	35
Pauls *et al.* (463/463/452/42) (*Smicridea*)	96.5(436)	96.9(438)*	0.4(2)	0.4(2)	2(9)*	1.5(7)	1.1(5)	1.1(5)	11
Mitter *et al.* (72/70/64/11) (Butterflies)	100(64)	100(64)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	6
Sweeney *et al.* (686/686/680/45) (Aquatic macroinvertebrates)	90.6(616)*	90.4(615)	1(7)	1(7)	6.9(47)	7(48)*	1.5(10)	1.5(10)	6
Victor (67/67/60/20) (Actinopterygian fish)	98.3(59)	98.3(59)	1.7(1)	1.7(1)	0(0)	0(0)	0(0)	0(0)	7

* indicates the larger value

My results raise the question why K2P may have been proposed in the first place. The application of nucleotide substitution models will generally yield distances that are larger than those based on uncorrected distances. This may have been an attractive property as model-based distances yielded higher values and thus implied a better interspecific separation. Indeed, I do find that the barcoding gaps tend to be larger under the K2P-model than with uncorrected p-distance (Fig. 2.2). However, upon closer inspection, K2P only makes a difference for sequences with large interspecific differences (Fig. 2.2).

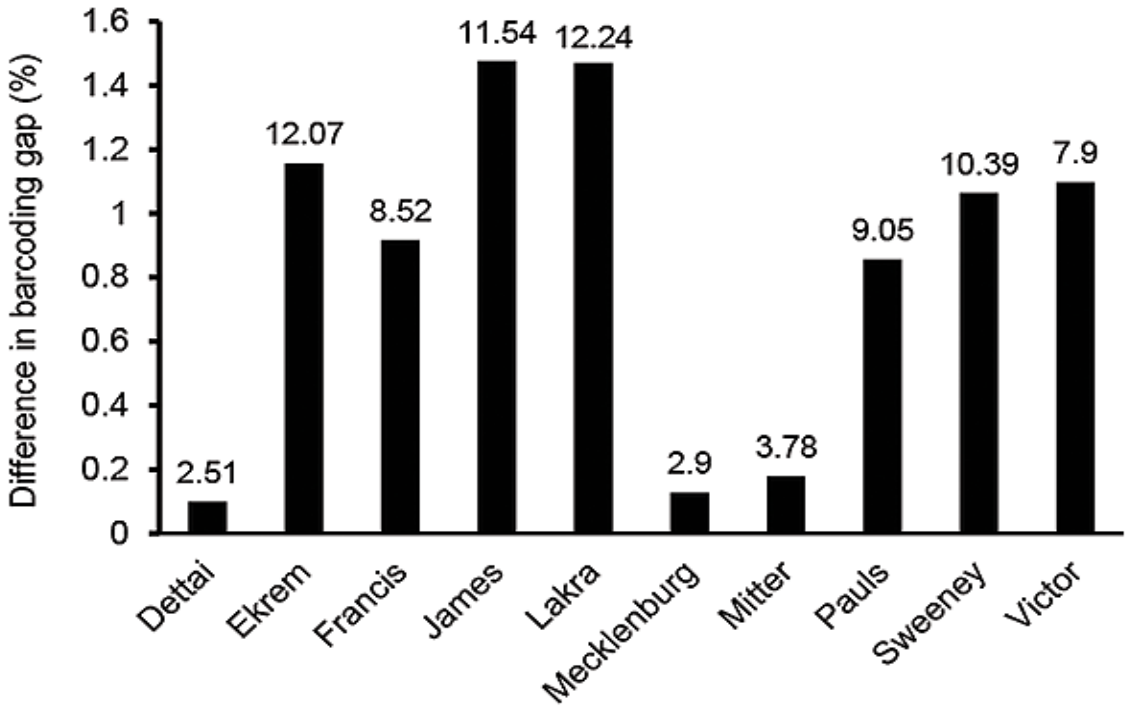


Figure 2.2: The difference between the K2P barcoding gap ($K2P_{inter}$-$K2P_{intra}$) and the uncorrected barcoding gap (p_{inter}-p_{intra}) is positively correlated with average interspecific distances (values above bars).

This becomes apparent when sequences with >5% smallest interspecific distances are excluded from analysis. Now, K2P and uncorrected distances yield similar barcoding

gaps (Table 2.3). This means that the use of K2P helps little with species identification given that the main challenge in DNA barcoding is to distinguish between closely related species. Species with large interspecific distances are readily identified using many techniques.

Table 2.3: Impact of K2P on barcoding gap: difference between K2P and p-distances for average intraspecific and average interspecific sequence divergences (* = all interspecific distances > 5%).

Dataset	All data		<5% interspecific distance	
	Average intraspecific (K2P-p)	**Average interspecific (K2P-p)**	**Average intraspecific (K2P-p)**	**Average interspecific (K2P-p)**
Dettai *et al.* (2011)	0.015	0.114	0.000	-0.143
Ekrem *et al.* (2010)	0.050	1.207	NA*	NA*
Francis *et al.* (2010)	0.006	0.923	0.003	0.013
James *et al.* (2010)	0.010	1.487	0.040	-0.022
Lakra *et al.* (2011)	0.000	1.470	NA*	NA*
Mecklenburg *et al.* (2011)	0.000	0.127	-0.001	0.038
Mitter *et al.* (2011)	0.000	0.180	0.000	0.034
Pauls *et al.* (2010)	0.005	0.861	0.000	0.005
Sweeney *et al.* (2011)	0.002	1.066	-0.029	-0.026
Victor (2010)	0.039	1.136	0.0765	0.0043

2.5 Conclusions

There are no obvious reasons why one should use K2P in DNA barcoding analyses. K2P is neither expected to perform better than uncorrected distances based on theoretical arguments, nor is its use supported by model testing or empirical evidence. K2P was introduced into the barcoding literature in 2003 and in contrast to most ideas in science it subsequently spread through copying with little further inquiry. It is particularly surprising that this copying has transcended the Metazoa barcoding literature. Although introduced for *COI* barcodes, K2P is now also used for plastid markers, including intron markers that are popular for barcoding plants (e.g. Lee *et al.* 2010; Ren *et al.* 2010; Pang *et al.* 2011). But surely these genes have very different evolutionary properties. I hope that the arguments and empirical data presented here can reverse the trend and inspire authors, reviewers, and editors to follow established criteria before using evolutionary models.

CHAPTER 3[2]

An update on DNA Barcoding: Low species coverage and an increasing number of unidentified barcodes

3.1 Abstract

DNA barcoding was proposed in 2003, the Consortium for the Barcode of Life was established in 2004, and the movement has since attracted more than $80 million funding. Here we investigated how many species of multicellular animals have been barcoded. We compared the numbers in a public database (GenBank as of January 2012) with those in the Barcode of Life Database (BOLD) and found that GenBank contained COI sequences for ca. 60,000 species while BOLD reported barcodes for ca. 150,000 species. The discrepancy was likely due to a large amount of unpublished data in BOLD. Overall, the species coverage was sparse, growth rates were low, and the barcode accumulation curve for Metazoa was linear with only 4,788 species added in 2011. In addition, the vast majority of species in the public database (73%) were barcoded by projects that were unlikely to be related to the DNA barcoding movement. Particularly surprising was the large number of DNA barcodes in GenBank that were not identified to species (Jan 2012: 74%), with insect barcodes often being identified only to order. Of these, several hundred thousand were then suppressed by NCBI because they did not satisfy the iBOL/GenBank

[2] A version of this chapter has been published as "Kwong, S., **Srivathsan, A.**, and Meier, R. (2012). An update on DNA Barcoding: Low species coverage and an increasing number of unidentified barcodes. *Cladistics* 28(6): 639-644." I wrote the scripts that generated the statistics from genbank flatfiles.

early release agreement. Species coverage was considerably better for target taxa of DNA barcoding campaigns (e.g. birds, fishes, Lepidoptera), although it also fell short of published campaign targets.

3.2 Introduction

This chapter was motivated by two observations in 2012:

(1) We used GenBank to blast a COI sequence from an unidentified species of chloropid flies and the top 100 BLAST matches were for sequences that had only been identified to order (labeled as "*Diptera sp.*"). These unidentified sequences had been overwhelmingly submitted by DNA barcoding projects. Once excluded, the informative sequences came predominantly from projects not associated with the DNA barcoding campaign. We thus decided to investigate how many identified and unidentified COI barcode sequences have been submitted to GenBank and what proportion came from DNA barcoding projects.

(2) Two papers in *Molecular Ecology Resources* highlighted discrepancies between the data in BOLD (DNA Barcode of Life Data System: http://www.boldsystems.org; Ratnasingham & Hebert 2007) and GenBank (Federhen 2011; Ratnasingham & Hebert 2011). For example, the numbers of species that had been barcoded according to BOLD differed considerably from the number of species for which there were sequences in GenBank. This was mostly due to unpublished data in BOLD that were available for query-matching but could not be downloaded. These data were included in the species counts on the BOLD websites that reported how many species have barcodes. We therefore investigated the species coverage in a public database such as GenBank where the data are available. Given that BOLD has only few identification tools, access to the original data is critical for most sophisticated analyses. Lastly, an update on the species coverage achieved by the DNA barcoding movement also appeared timely given that the technique was proposed almost ten years ago and Canadian agencies

alone had invested and/or pledged more than 80 million Canadian dollars to DNA barcoding (iBOL 2010).

The use of DNA sequences for species identification has a long history (e.g., Nanney 1982; Bartlett and Davidson 1991; see also Sperling 2003; Will & Rubinoff 2004; Will *et al.* 2005; Cameron *et al.* 2006; Meier 2008), but it only received much attention after it was formally proposed as "DNA Barcoding" in 2003 (Hebert *et al.* 2003). One year later an international Consortium for the Barcode of Life (CBOL) was established. This was followed by the creation of a specialized sequence database "BOLD" in 2007 (Ratnasingham & Hebert 2007). According to Marshall (2005), the goal of CBOL is "to tag every organism on Earth, starting with the 1.7 million species that have been named and moving on to the estimated 10 million to 20 million that have not" (see also Hajibabaei *et al.* 2005).

It is envisioned that this goal be accomplished in stages by initially concentrating the resources on particular branches of the Tree-of-Life. Two of the most prominent barcode campaigns are the "All Birds Barcoding Initiative (ABBI)" (Hebert *et al.* 2004) and "Fish Barcode of Life Initiative (FISH-BOL)" (Ward *et al.* 2009) whose explicit goals were outlined by Ratnasingham & Hebert (2007): "seek to deliver barcode coverage for all species of birds and fishes by 2012". Financially these goals were realistic given that as early as 2005, "funding [was] in place to ensure that the DNA barcode library for animals will grow by at least 500,000 records [the first] 5 years, providing coverage for some 50,000 species" (Hebert & Gregory 2005). These taxa were also fairly easy targets given that their taxonomy is comparatively well studied (Will *et al.* 2005).

Obtaining high identification success rates with DNA barcodes is critically dependent on having good species coverage in the DNA barcode databases against which unidentified sequences are queried (Little & Stevenson 2007; Meier 2008; Virgilio *et al.* 2008). In building these databases from scratch, one would expect that the number of barcoded species would initially increase rapidly given that tissues for common species are readily available. The growth would later plateau as rarer species have to be sampled (Lim *et al.* 2011); i.e., overall we would expect a "barcode accumulation curve" that resembles the kind of asymptotic collector's curve that is typically found in biodiversity studies (Colwell & Coddington 1994; Gotelli & Colwell 2001; Meier & Dikow 2004). Note however that this curve is not expected to be uniform given that sampling would be carried out in different geographical locations or taxonomic groups at different times.

In order to determine the progress of the DNA barcoding campaign, we here downloaded all GenBank COI sequences for Metazoa. We characterized the barcode (species) accumulation curves for all of Metazoa and three taxa that are targeted by specific DNA barcoding campaigns (birds, fishes and Lepidoptera). We furthermore determined the proportion of identified and unidentified sequences, investigated how many sequences were submitted by barcoding projects, and determined the number of identical sequences among those that are labeled as "*Diptera sp.*" in GenBank. Our study complemented a paper by Taylor & Harris (2012) who surveyed the DNA barcoding literature and found that most barcoding studies target invertebrates and continue to use NJ trees for sequence identification (Taylor & Harris, 2012). Information on the relative proportion of identified and unidentified sequences in GenBank can also be found in Page (2011).

3.2 Materials and Methods

In order to assess the growth of species coverage, a total of 855,442 metazoan COI sequences were downloaded from GenBank (January 2012). The COI sequences were identified using taxon-specific "taxonomy" searches combined with gene identifiers [COI(Gene Name) OR "cytochrome oxidase subunit 1"(Gene Name) OR "cytochrome *c* oxidase subunit 1"(Gene Name) OR "cytochrome *c* oxidase subunit I"(Gene Name) OR "cytochrome oxidase subunit I"(Gene Name) OR COX1(Gene Name)]. Note that this search strategy yields overestimates of species coverage because it does not exclude partial COI sequences that are very short and/or pertain to the non-barcoding portion of the gene. Sequences were considered to be the product of a DNA barcoding project if the words "barcode" or "barcoding" ("bar cod", "barcod", or "bar-cod") were present in the full GenBank entry. All other sequences were considered unrelated to DNA barcoding ("general systematics" henceforth). In order to determine barcode accumulation curves, we used the sequence submission dates to sort sequences by submission time and taxa. Barcode accumulation curves were then generated for all of Metazoa and the focal taxa of three DNA barcoding campaigns: (ABBI: birds, FISH-BOL: fishes, iBOL: Lepidoptera). The sequences for the latter were identified through taxonomy searches in GenBank (ABBI: "Dinosauria", FISH-BOL: "Chondrichthyes, Actinopterygii, and Hyperoartia", iBOL: "Lepidoptera"). The barcode coverage in GenBank was compared with the coverage reported in BOLD (accessed January 2012). In order to determine the proportion of unidentified sequences in GenBank, we identified those that were not identified to species. Given that we found many sequences that were identified only to order, we also determined the amount of redundant/repetitive

sequencing for 46,017 "*Diptera sp*." barcodes. We used objective clustering (Meier *et al.* 2006) to identify identical (0%) and near-identical (< 1%) sequences. Note that a large number of these unidentified sequences (341,978 according to http://iphylo.blogspot.sg/2012/04/dark-taxa-even-darker-ncbi-pulls-dna.html#disqus_thread) were then suppressed by NCBI because they did not meet the minimum data standard for an iBOL early release entry.

3.3 Results and Discussion

3.3.1 Species coverage: Metazoa

We found identified COI sequences for 60,930 species of Metazoa in GenBank; i.e., the species coverage was thus very sparse considering that the vast majority of described species are Metazoa, and Marshall (2005) estimated that there are 1.7 million described and 10-20 million undescribed species. Of course, some habitats and countries would be better represented than others, but our results suggested that a large number of DNA barcodes remained to be characterized even for the common species.

Arguably even more surprising than the sparse species coverage was that most of the growth came from projects unrelated to DNA barcoding (Fig. 3.1). In fact, even if all sequences from DNA barcoding projects were to be removed from GenBank, the number of species with COI sequences would have only dropped by ca. 16,000 because most identified sequences came from "general systematics" projects. Thus, by 2012, it appeared that the generously funded DNA barcoding projects had only generated barcodes for ca. 22,000 species of which 6,000 were shared with other GenBank projects (Fig. 3.1). This fell well short of the 50,000 species for which funding had been obtained by 2005 (Hebert & Gregory, 2005). Secondly, in terms of species accumulation in the database over years, we found a near-linear curve with an overall less than impressive slope (Fig. 3.1). Presumably, species accumulation would be influenced by both geographical locations of the funded projects, taxonomic group and the commonality of species. Nonetheless, we expected to see a rapid increase in the acquisition rate given that it should be straightforward to obtain identified specimens in the initial phase of the project.

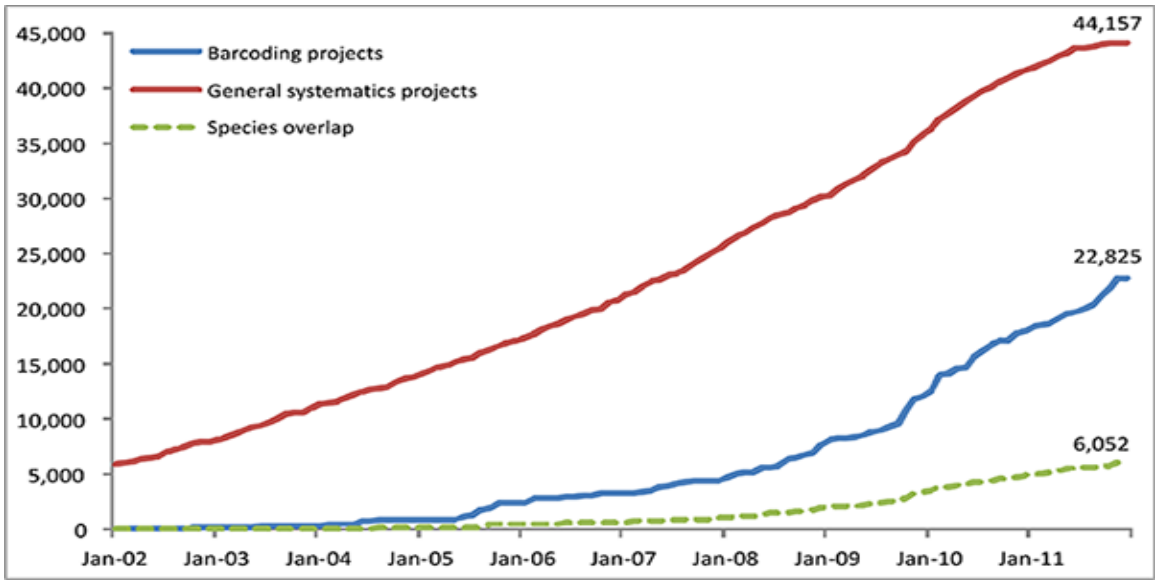


Figure 3.1: Species coverage and species overlap between sequences submitted by barcoding and other projects. y-axis represents number of species.

Table 3.1: Number of COI sequences submitted to Genbank since 2002. * includes sequences submitted before 2002.

Year	Barcode projects		General systematics projects	
	No. of submitted entries	Percentage of unidentified entries	No. of submitted entries	Percentage of unidentified entries
2002*	331	0.12%	22,574	6.25%
2003	196	1.53%	10,395	11.98%
2004	1,569	41.17%	15,934	6.06%
2005	7,978	30.10%	17,529	12.70%
2006	7,171	24.93%	22,635	9.92%
2007	6,330	7.74%	28,190	10.42%
2008	23,980	24.47%	40,292	9.95%
2009	49,391	28.02%	48,442	17.46%
2010	262,136	82.24%	54,685	16.61%
2011	212,915	85.73%	22,769	14.87%
Total	571,997	73.98%	283,445	12.69%

For example, between 12/2010 and 12/2011 only 4,788 new species were added. Note that this small number is unlikely due to manpower or equipment shortage given that during the same period more than 200,000 barcodes were submitted to GenBank (Table 3.1).

Our species coverage numbers deviated considerably from what was reported in BOLD which provided information on the "formally described species with barcodes". According to an update from early 2012, BOLD reported that the COI sequences for 146,067 species were known. Given that most sequences in BOLD were likely to belong to Metazoa, at least half of the Metazoa data were not publicly available. The user could query a sequence against the full database, but the underlying data remained hidden. This raised a number of issues – especially if identifications that were based on BOLD searches were to be used in publications (Federhen 2011). In a reply to Federhen (2011), Ratnasingham & Hebert (2011) thus "emphasize(d) the need for caution in the interpretation of identifications based on a reference library with entries that have seen limited validation." In the literature such identifications should be clearly attributed to BOLD with a specification of which version of the database was used.

3.3.2 Species coverage: BOLD campaign taxa

The contributions by DNA barcoding projects were more impressive for those taxa that were campaign targets. We evaluated three BOLD campaigns. The largest number of species with COI sequences in GenBank was found for Lepidoptera (7,742 species) with its ca. 160,000 described species (Kristensen *et al.*, 2007) (Fig. 3.2). We saw a significant increase in the number of barcoded species over 2010 and 2011. The fish campaign could draw on sequences from >4,244 species for the ca. 31,000 described species while the bird campaign was surprisingly far from being in its final stages given that only 2,838 of the >10,000 described species had COI sequences (Fig. 3.2).

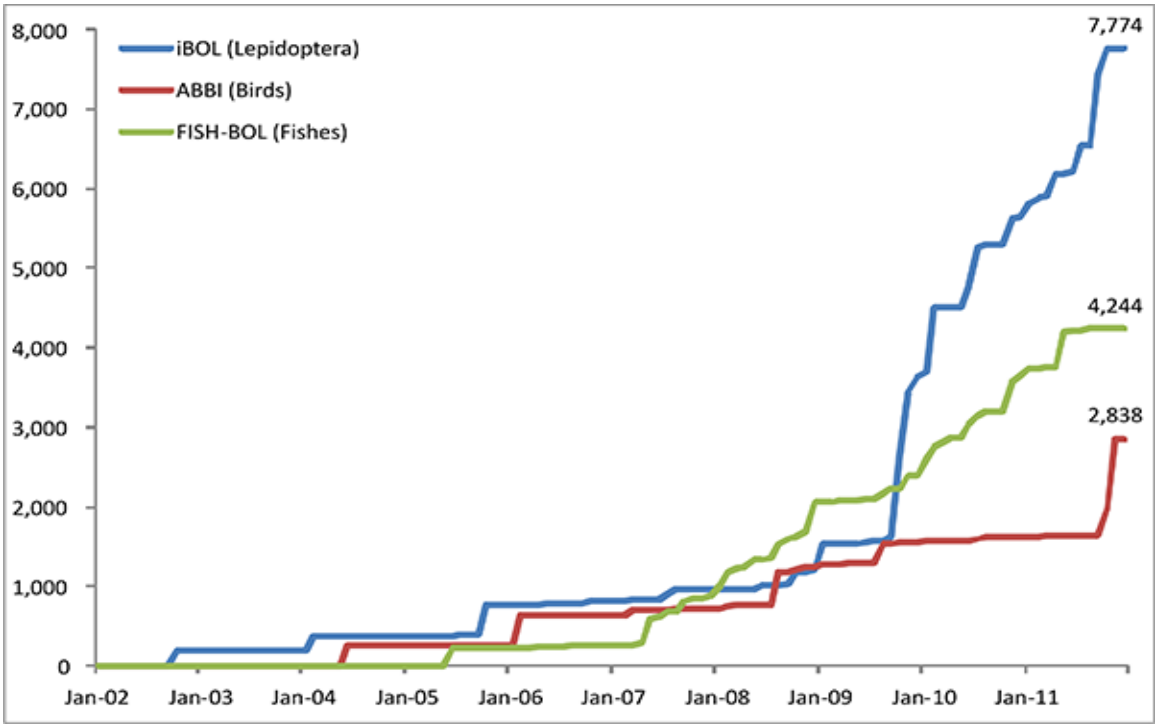


Figure 3.2: Species coverage for barcoding campaign taxa Lepidoptera, birds, and fishes.y-axis represents number of species

Note that Ratnasingham & Hebert (2007) had predicted complete species coverage for birds and fishes by 2012. This however was not completed by the end of 2012. Again, the species coverage reported in BOLD was higher than in GenBank. According to the

database, 66,430 species of Lepidoptera, 8,293 species of fishes, and 3,892 species of birds had been covered. There was additional information for FISH-BOL in the form of a progress report which indicated that as of July 2010 25% of the 31,000 species had barcodes (ca. 7,800 species) (Becker *et al.*, 2011); i.e., only ca. 500 species had been added in 1.5 years. This would imply that the growth in the number of barcoded species was slowing considerably before even half of the species diversity had been covered. Becker *et al.* (2011) indicated that the main problems were freshwater fishes (but see Collins *et al.* 2012), covering the species of certain geographic regions (Asia, South America, Africa), and obtaining properly identified tissues. Not surprisingly, the technical problems with obtaining sequences were comparatively minor.

Unfortunately, there were two problems with interpreting the species numbers in BOLD. First, it was unclear whether they pertained to described species or also "predicted" species although this made a big difference (Cameron *et al.* 2006) given that only barcodes for properly identified specimens can be used for the identification of future query sequences. Indeed, it would be surprising if BOLD had almost ten times as many identified barcodes of Lepidoptera than GenBank. Second, if the species counts in BOLD included predicted species, the reported coverage of, for example, 40% for the Lepidoptera was misleading because it was based on comparing predicted species with the number of described species.

Yet, the total number of described and undescribed Lepidoptera species was estimated to be 400,000-500,000 (Kristensen *et al.* 2007) and the existing barcodes for 66,430 species corresponded to a species coverage of approximately 15% as opposed to BOLD's reported figure of 40%. Similarly, the number of described fish species was 31,000, but

the true diversity was much higher given that about 4,000 of these species were described within the 10 years prior to this study (Becker *et al.* 2011). It would thus be desirable if BOLD were to distinguish between identified and predicted species and use appropriate denominators to quantify species coverage (number of described species for identified barcodes and estimated number of species for predicted species). Clearly distinguishing between described and undescribed diversity would also highlight the importance of DNA barcoding for the discovery of cryptic species (Bickford *et al.*, 2007). This will arguably be one of the more lasting contributions of the barcoding movement as long as potentially cryptic species are later confirmed based on additional data (Gomez *et al.* 2007; Tan *et al.* 2010).

3.3.3 Unidentified vs. Identified sequences

One of the most surprising features of the DNA barcodes in GenBank was the huge number of unidentified sequences (see also Page 2011). There were 571,997 COI barcodes in GenBank but a staggering 423,188 sequences (74%) were not identified to species (Table 3.1). The vast majority of these barcodes had only very approximate identifications. For example, the 49,629 barcodes for Diptera included 46,017 barcodes that were only identified to "*Diptera sp.*". Similarly, 195,348 of the 270,301 Lepidoptera barcodes were only identified to order ("*Lepidoptera sp.*"). Presumably many of these unidentified sequences came from environmental samples because we found a large number of identical or near-identical sequences. For example, clustering at 1% revealed that at least one of the "*Diptera sp.*" species had been sequenced 1,000 times while another had 305 identical sequences. This repetitive sequencing highlighted the problem of using DNA barcodes for evaluating environmental samples. Without presorting, processing such samples with molecular tools will be very costly and time-consuming.

Table 3.2: Number of COI sequences submitted to GenBank for barcoding campaign taxa after 2002. *includes sequences submitted before 2002.

Year	iBOL (Lepidoptera)		ABBI (Birds)		FISH-BOL (Fishes)	
	No. of submitted entries	Percentage of unidentified entries	No. of submitted entries	Percentage of unidentified entries	No. of submitted entries	Percentage of unidentified entries
2002*	216	1.39%	0	0.00%	0	0.00%
2003	0	0.00%	0	0.00%	0	0.00%
2004	852	60.21%	424	0.00%	0	0.00%
2005	4,289	35.09%	0	0.00%	760	0.53%
2006	213	7.04%	2,134	0.05%	196	0.00%
2007	872	3.44%	125	0.00%	1,770	2.43%
2008	3,031	22.47%	2,285	0.00%	8,397	1.13%
2009	31,466	29.31%	1,875	0.00%	3,905	9.35%
2010	139,591	85.31%	2,570	23.23%	16,579	57.43%
2011	89,771	90.41%	3,877	0.52%	9,895	39.84%
Total	270,301	78.51%	13,290	4.65%	41,502	9.50%

We previously mentioned that most species with COI sequences were sequenced by projects that were unlikely to be related to the DNA barcoding movement. The reverse was true for the unidentified sequences where the DNA barcoding projects contributed approximately three quarters of all unidentified barcodes (Table 3.1). Surprisingly the proportion of unidentified sequences was also very high for taxa that were subject to BOLD barcoding campaigns. The three BOLD campaign taxa evaluated here contributed ca. 50% of all barcodes in GenBank, but most Lepidoptera barcodes (78%) and many fish barcodes (34%) were not identified to species (Table 3.2). The only exception was the bird project whose sequences were overwhelmingly identified (95%).

Overall, the number of barcodes in GenBank was again much lower (571,997) than the numbers reported in BOLD, which reported 1,502,590 barcodes of which the vast majority were generated by the Canadian Centre (1,124,561 sequences). Note that the proportion of unidentified sequences submitted by DNA barcoding projects was not

only high but also rapidly increasing (Table 3.1). This may have reflected a change of emphasis in the movement from providing an identification tool to using sequences for biodiversity assessment. It also coincided with Schindel & Scott's (2010: 112) proposal of taxon labels ("a unique, stable, text-phrase applied to an unpublished taxon concept…") for taxa that have only provisionally been identified. Schindel & Scott (2010: 113) elaborated: "Ecologists and other non-taxonomists could publish results using taxon labels, thereby avoiding the delay often associated with waiting for taxonomists to put formal names on specimens." It appears that there was less emphasis on barcoding identified specimens and the importance of species descriptions.

3.4 Conclusions

The DNA barcoding campaign is one of the best-funded and most visible movements in biodiversity research. It promises easier ways to identify species based on data that do not require taxon-specific knowledge. However, we found that in terms of species coverage and accessibility of data, DNA barcoding leaves much to be desired and the number and quality of DNA barcodes will have to improve considerably in order to achieve the ambitious goals of the movement. In the past much scrutiny and criticism of DNA barcoding were devoted to the philosophical and methodological shortcomings (Will & Rubinoff, 2004; Will *et al.* 2005; Brower 2006; Cameron *et al.* 2006; Meier *et al.* 2008; Chapter 2), but our study highlighted the need to carefully evaluate whether the movement is capable of delivering sufficient species coverage to make the technique useful to a wide variety of users (Cameron *et al.* 2006).

3.5 An Update

Recently, some of the access problems to DNA barcodes existing in 2012 have been solved. Barcode of Life Datasystems (BOLD) made the database much easier to download data from, leading to access to additional sequences. Yet the several of these issues persists, e.g. as per BOLD public database, there are >2 million arthropod barcodes available for download of which only 667,092 are identified from ~81,000 species; currently from GenBank, there are 506,265 arthropod sequences of which ~250,000 are identified to species representing 48,074 species. These are creating a number of concerns developing tools for identifications using barcodes; and my hope is that proportion and number of identified sequences from existing barcodes increases. Furthermore, integration of public data from BOLD to GenBank is still desirable, given that BOLD is limited in its genes of choice and requires a separate set of tools for identification due to differences in format of taxonomy information as well as sequence data downloaded.

CHAPTER 4

The databases for diet and parasite analyses: barcoding the Nee Soon Swamp forest and the bioinformatic retrieval of barcode sequences from GenBank

4.1 Abstract

In order to identify DNA sequences obtained through shotgun sequencing or metabarcoding from environmental samples, the reads must be matched to DNA barcodes with known identity. In this chapter, I describe the databases that were used in the two chapters of my thesis that analyze DNA sequences from primate fecal samples (chapters 5 and 6). Both chapters intially analyze plant diets and for this purpose I utilize global and "local" databases. The latter comprise DNA barcodes for putative diet species and these barcodes were generated in the lab over the course of this thesis for the purpose of diet analyses. The global databases include all publicly available sequences in addition to all local sequences. The most recent version of the plant barcode databases, as of May 2014, comprised 28,680 species (7,539 genera) for *rbcL,* 37,068 species (7,894 genera) for *matK* and 22,820 species (5,053 genera) for *trnL-F.* However, the next two chapters of the thesis go beyond diet analyses and also match reads from the fecal samples to non-plant eukaryotes. For this purpose, I used publicly available rDNA databases. I also built a COI database, and a more targeted 18S rDNA database for non-human primate parasite

sequences. Here, I characterize the databases and describe the methods that were used for ensuring sequences homology of the downloaded data from GenBank.

4.2 Introduction

Databases of DNA barcodes are fundamental to identifying DNA sequence reads. In the following two chapters, the main focus is the diet of two colobine primate species. Being phytophagous primates, the diets can be identified using plant barcode databases that include DNA barcodes for putative food items. Unfortunately, the choice of plant barcode genes is not straightforward. Several proposed genes lack the desired taxonomic resolution, i.e., they fail to distinguish closely related species (Little & Stevenson 2007). The Consortium for Barcode of Life (CBOL) originally supported and approved *rbcL* and *matK* for the identifications of plants. This choice of markers was the result of testing the discriminatory power of a number of chloroplast markers in plants (Hollingsworth *et al.* 2009). However, it has become clear that these genes do not have sufficient taxonomic resolution. Different barcode genes have been suggested (Kress & Erickson 2007; Li *et al.* 2011), but apart from the CBOL recommendations there is no clear consensus which DNA barcode genes should be used for plants. This is in contrast to DNA barcodes for Metazoa where COI is used for most clades.

In my study, I use three plant barcodes: *rbcL, matK* and *trnL-F.* This choice was made based on several considerations: the most comprehensive barcode databases are available for *rbcL* and *matK*. Both have sequences for >7000 genera of plants. I also included *trnL-F* in my studies because in the next two chapters of my thesis I compare the performance of metagenomics and metabarcoding. The gene of choice for metabarcoding is the P6 loop of *trnL-F* (Taberlet *et al.,* 2007), so that this gene was also sequenced and analyzed for the metagenomic study. The P6 loop is nested within the longer *trnL-F* barcode sequences that can be generated using primers designed by Taberlet *et al.* (1991). Lastly for Chapter 5, note that I also included *trnH-psbA* given that it is has been widely

recognised to improve taxonomic resolution in a number of studies (Kress & Erickson 2007; Parmentier *et al,* 2013). This barcode was excluded in Chapter 6 due to the the large size of the datasets used.

In the next two chapters, I analyze the diet of two species of colobine primates. The first comprises diet analysis of captive primates in the Singapore Zoological Gardens (Chapter 5), and the second is a diet analysis of a wild population (Chapter 6). In both studies, I analysed diet using a combination of a "local" database of DNA barcodes for putative food plants and a "global" database that also includes all data from GenBank. For the captive primates, i.e. red shanked douc langurs (*Pygathrix nemaeus*), the diet was known and hence the main aim was to test diet analysis techniques after building a barcode database of sequences from the known diet plant species. Identified leaf samples were provided by the Singapore Zoo. In the case of the diet of a wild population of banded leaf monkeys, *Presbytis femoralis* in Singapore, a database of barcodes from the habitat of the primate was needed. However, obtaining a comprehensive barcode database for the rich flora of the native habitat is exceedingly difficult (Elliot & Davies 2014), given that it is a tropical rain forest with high species diversity (Brook *et al.* 2003). Based on current survey results, the native habitat contains ca. 730 tree and liana species (Wong *et al.* 2013). Despite extensive plant sampling that is still ongoing, we were only able to generate DNA barcodes for 248 species. For these, I contributed 287 sequences across three different barcodes.

Given the difficulty of obtaining complete DNA barcode databases, I complement my data with all angiosperm barcode sequences from GenBank. There has been a steady accumulation of sequences for *rbcL, matK* and *trnL-F* (Fig. 4.1) over the years. Many of

the sequences were gathered for phylogenetic purposes, but they are now also used as DNA barcodes (Gielly and Taberlet, 1994; Nepal and Fergusen, 2012). Ever since the setup of the Consortium for Barcode of Life (2004), there seems to be a slight increase in species accumulation for *rbcL* and *matK*. With increased sampling intensity, >20,000 species have been barcoded for these genes. However, even though data from GenBank is easy to access and download, the acquired sequences may not comprise only homologous regions. This is because over the years, different primer pairs have been used for sequencing the same genes. Another problem with downloads is that they often comprise more than the desirable gene fragment (e.g. full chloroplast genomes). One way to avoid such sequences would be to exclude any sequence that was not submitted by a barcoding study (limiting keyword to BARCODE). However, this will lead to the loss of a large number of data for many species for which barcodes were generated in phylogenetic studies. Therefore, it is preferable to bionformatically extract the regions homologous to the barcode segment from the set of downloaded sequences from GenBank.

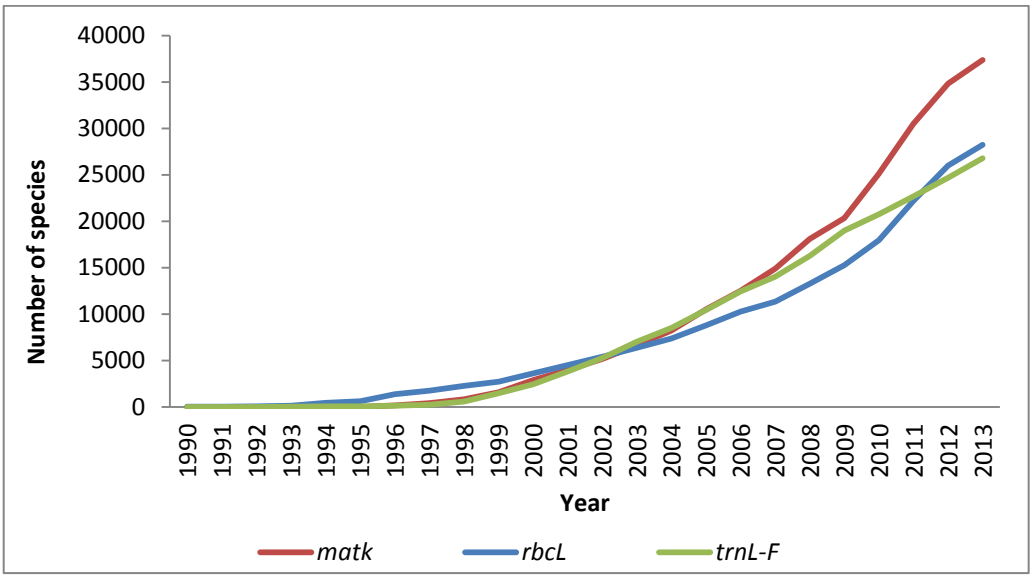


Figure 4.1: Accumulation of identified species in GenBank over years.

In this chapter, I describe how DNA barcodes were obtained for the tree and liana species occuring in Nee Soon Swamp forest. I furthermore describe procedures to obtain

a set of homologous sequences from GenBank. For this, I bioinformatically mined the data downloaded to obtain homologous regions by using a modified version of BLAST based on a pipeline built in Alfried Vogler's laboratory that was used in Hunt *et al.* (2007)'s study for Coleoptera phylogeny. The databases were designed to be suitable for a taxonomy assignment pipeline (as described in chapter 5) that was then used. This pipeline uses the NCBI taxonomy and thus links GI number information with NCBI taxid to generate taxonomic profiles for the sequences. I will demonstrate that the species recovery through this pipeline yielded similar taxonomic profiles as obtained by downloading plant barcode data from BOLD. This is promising given that any pipeline based on GenBank data will have greater utility because it is not limited to the recognized barcode regions only. I also built COI and rDNA databases, in order to allow for identification of additional eukaryotes in the metagenomes (e.g. intestinal parasites).

4.3 Methods

4.3.1 Local databases

Tissue Samples

For Chapter 5, tissues corresponding to seven foliage species provided as diet to primates (*Pygathrix nemaeus)* were obtained from the Singapore Zoo. For Chapter 6, samples obtained from the habitat of banded leaf monkeys (*Presbytis femoralis*) were used. Here, leaf sampling was first carried out for five plots of 25 x 50 metres in the forest; these plots were selected after estimating the range of *Presbytis femoralis.* Given that these primates mostly feed in the canopy (Bennett, 1983), the criteria for collecting vegetation was plants with a girth of ≥40cm at approximately 1.3m from the ground for trees and a minimum height of 5m for lianas. Furthermore, I collected 87 plant samples opportunistically when helping Andie Ang with her field work for *Presbytis femoralis*. Overall, 369 samples from 137 species were collected. I vouchered the specimens and aided in sample collection (see "AS" vouchers, Table 4.1).

More recently, another extensive survey of Nee Soon Swamp forest is being carried out by the Plant Systematics Laboratory at the National University of Singapore (Singapore). Plant tissue specimens (1,802 leaf specimens) have been collected from the swamp forest. The barcoding of the forest is ongoing, but I was able to already use some of the data ("Q" vouchers, Table 4.1).

DNA extraction and sequencing

For DNA extraction, tissues were ground using liquid nitrogen and extraction was carried out either with QIAGEN Blood and Tissue Kit (GmBH), or with the CTAB method

(Kutty *et al.,* 2007). Fragments of three chloroplast 'barcoding' regions were amplified representing *matK* (557-781 bp), *rbcL* (429-586 bp) and *trnL-F* (615-955 bp). The primer pairs used were as follows (annealing temperatures are in brackets): rbcLa_f and rbcLa_rev (54-55°C) (Kress and Erickson, 2007), 3F_KIM f and 1R_KIM r (52°C) (Kim Ki-Joong, unpublished) and trnL c and f (52-55°C) (Taberlet *et al.,* 1991). The PCR reactions were done using the following conditions: Initial denaturation at 95°C for 5 min, followed by 35 cycles of 94°C for 1 minute, annealing for 1 min and 72°C for 1 min 30 sec. Final extension was at 72°C for 5 min. Gel extractions were performed if there were multiple bands present after optimization of conditions. The amplified PCR products were purified with SureClean (Bioline, Randolph, MA). Cycle sequencing was performed using BigDye Terminator v3.1 and products were analysed in both directions on an ABI 3100 Genetic Analyser (Perkin Elmer, Waltham, MA). Sequences were edited with Sequencher v 4.6 (Gene Codes Crop, Ann Arbor, MI, USA).

4.3.2 Data mining from GenBank

In order to obtain sets of barcode sequences from GenBank, I first downloaded sequences using the following keywords for angiosperms: *rbcL*: (Magnoliophyta[Organism]) AND (rbcL[Gene Name] OR ribulose 1,5-bisphosphate carboxylase/oxygenase[Gene Name] OR Ribulose bisphosphate carboxylase[Gene Name] OR RuBisCO large subunit[Gene Name]); *matk*: (Magnoliophyta[Organism]) AND (matk[Gene Name] or maturase-k[Gene Name]) and *trnL-F*: (Magnoliophyta[Organism]) AND (trnL[Gene Name] OR trnL-F). For COI the keywords are described in Chapter 3.

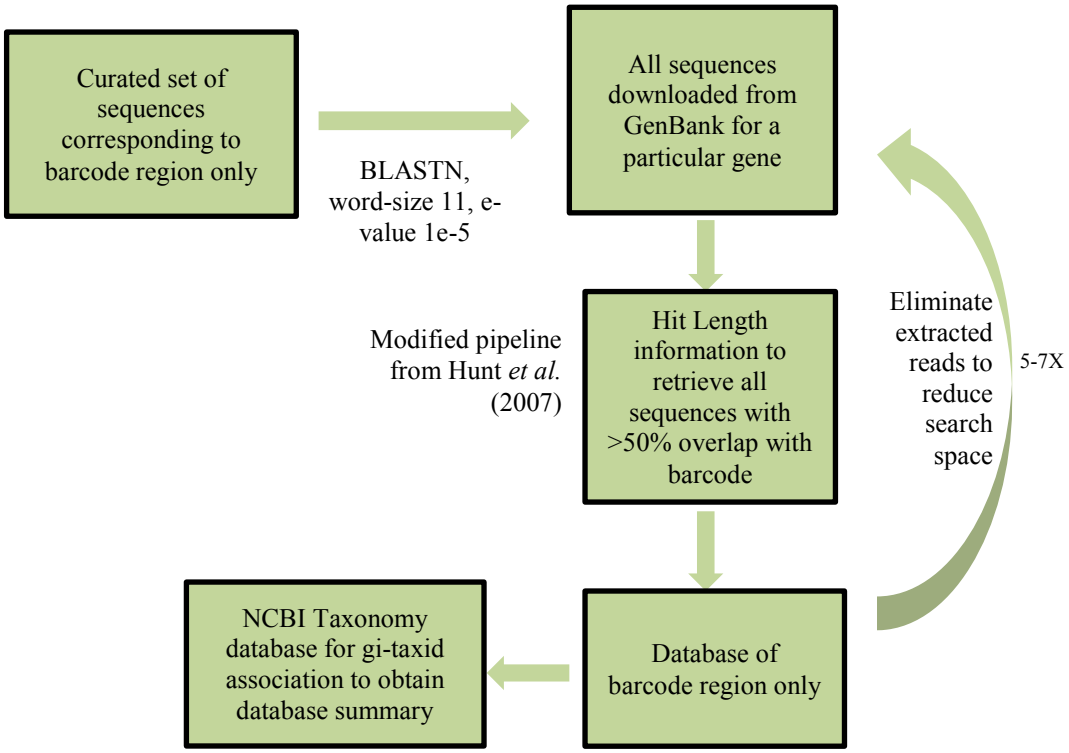


Figure 4.2 : Overview of database generation using data downloaded from GenBank.

I modified the BLAST based pipeline of Hunt *et al.* (2007) to obtain homologous regions (Fig. 4.2) of barcodes from the downloaded sequences. The principle behind this pipeline is to use a curated set of sequences corresponding to the desired gene region (i.e., regions we desire to extract). These sequences are then used to fish out homologous sequences from the downloaded material. For barcodes approved by CBOL, obtaining a curated set was easy because DNA barcodes could be downloaded from GenBank using the keyword ("BARCODE"). For *trnL-F,* I manually obtained a subset representing several different families that contained the target region. Using BLASTN, these sequences were matched to a database generated using sequences downloaded as mentioned above. The search was conducted using word-size 11, e-value 1e-5. Using the pipeline described above, I could then extract the region of interest based on BLAST start and end position. I excluded any

matches that were too short (<50% of the mean sequence length of the curated set of reads).

4.3.3 rDNA databases

In order to characterize rDNA sequences from the metagenome, in Chapter 5 I used MG-RAST's pipeline of rDNA prediction (Glass *et al.,* 2010). However, the procedure for upload of such large datasets to the MG-RAST online server is slow, and hence in the larger scale study described in Chapter 6, I analysed the data locally. I first examined the sequences using SILVA SSU and LSU rDNA databases (Pruesse *et al.,* 2007). I also built a local parasite database after doing a literature search on common parasitic infections in primates. The list of taxa included in the local parasite database is given in Appendix 4.

4.4 Results

For the diet barcode dataset used in Chapter 5, I generated 19 sequences for seven diet plant species (Appendix 1, Supplementary Table T1). The local dataset for the Nee Soon swamp forest comprises 248 taxa, of which 211 had ≥2 barcodes sequenced. This corresponded to 180 species (191 sequences) for *matK*, 223 species (248 sequences) for *rbcL*, and 207 species (211 sequences) for *trnL-F*. Of these 127 species and 7 genera did not have data in GenBank. Of the 650 barcodes included in chapter 6, I sequenced ~280 barcodes.

Table 4.1: List of species barcoded from Nee Soon. * represents multiple sequences where only a representative is listed.

SL No.	Species	Number of barcodes	*matK*	*rbcL*	*trnL-F*
1	*Adinandra dumosa*	3	AS003	AS003	AS003
2	*Aeschynanthus wallichii*	2	Q4U122	Q4U122	
3	*Agelaea borneensis*	3	AS166	AS166	AS166
4	*Agelaea macrophylla*	3	Q10U122*	AS191	Q10U122
5	*Aglaia elliptica*	3	Q4T59	Q4T59	Q4T59
6	*Aglaia leptantha*	2		AS056	AS056
7	*Aglaia odoratissima*	3	Q4T60	Q4T60	Q4T60
8	*Aglaonema simplex*	3	Q10120	Q10U120	Q10U120
9	*Agrostistachys borneensis*	3	Q2U38	Q2U38	Q2U38
10	*Alangium nobile*	3	Q3T63	Q3T63	Q3T63
11	*Albizia pedicillata*	3	AS283	AS283	AS283
12	*Ancistrocladus tectorius*	1			Q2U45
13	*Anisophyllea disticha*	3	AS286	AS286	AS286
14	*Anodendron candolleanum*	3	Q3U163	Q3U163	Q3U163
15	*Antidesma coriaceum*	2		Q10U133	Q10U133
16	*Antidesma cuspidatum*	2		Q10U114	Q10U114
17	*Aphanamixis polystachya*	2	AS156		AS156
18	*Aporosa falcifera*	3	Q4U159	Q4U159*	Q4U159
19	*Aporosa frutescens*	3	AS043	AS027	AS027
20	*Aporosa lucida*	2		AS058	AS058
21	*Aporosa symplocoides*	2		Q8T46	Q8T46
22	*Archidendron clypearia*	2	AS273		AS273
23	*Artabotrys suaveolens*	2	AS188		AS188
24	*Artocarpus integer*	2	AS127		AS127
25	*Artocarpus lacuca*	2		AS037	AS037
26	*Asystasia gangetica*	3	AS119	AS119	AS119
27	*Asystasia nemorum*	3	AS095	AS095	AS095
28	*Baccaurea bracteata*	1			Q3U164
29	*Baccaurea parviflora*	2		AS053	AS053
30	*Bauhinia semibifida*	2		AS026	AS026
31	*Breynia racemosa*	3	AS006	AS006	AS006
32	*Byttneria maingayi*	3	AS208	AS208	AS208

33	*Calophyllum dispar*	2		Q4T61	Q4T61
34	*Calophyllum ferrugineum*	3	Q10U129	AS285*	Q10U129
35	*Calophyllum pulcherrimum*	3	AS017	AS017*	Q10U146
36	*Calophyllum rubiginosum*	1			Q10U143
37	*Calophyllum wallichianum*	3	Q4U143	Q4U143	Q4U143
38	*Campnosperma squamatum*	2		Q3U179	Q3U179
39	*Canthium confertum*	3	Q10U98	Q10U98	Q10U98
40	*Carallia brachiata*	2		Q8U74	Q8U74
41	*Cayratia mollissima*	3	AS229	AS229	AS229
42	*Cinnamomum iners*	3	AS103	AS103	AS103
43	*Cissus nodosa*	3	Q8U92	Q8U92	Q8U92
44	*Clerodendrum deflexum*	3	AS098	AS075*	AS098
45	*Clerodendrum disparifolium*	3	Q1U06	Q1U06	Q1U06
46	*Cnestis palala*	2	AS347	AS347	
47	*Commersonia bartramia*	3	AS120	AS120	PAS120
48	*Connarus semidecandrus*	3	Q4U155	Q4U155	Q4U155
49	*Coptosapelta flavescens*	3	AS327	AS327	AS327
50	*Coptosapelta griffithii*	3	AS247	AS247	AS247
51	*Cratoxylum arborescens*	1			Q2T13
52	*Cratoxylum formosum*	2		AS129	Q10U147
53	*Cryptocarya ferrea*	2		Q3U167	Q3U167
54	*Cyathocalyx ramuliflorus*	3	Q8A	Q8A	Q8A
55	*Cyathostemma excelsum*	3	AS085	AS085	AS085
56	*Cyathostemma viridiflorum*	3	AS074	AS074	AS074
57	*Cyclea laxiflora*	3	Q8U109	Q8U109	Q8U109
58	*Dalbergia parviflora*	3	Q1U09	Q1U09	Q1U09
59	*Dalbergia rostrata*	3	AS135	Q3T44	Q3T44
60	*Dapania racemosa*	3	Q3U170	Q3U170	Q3U170
61	*Dasymaschalon wallichii*	3	Q3U127	Q3U127	Q3U127
62	*Dendrotrope varians*	2		Q1U32	Q1U32
63	*Derris maingayana*	3	Q3U193	Q3U193	Q3U193
64	*Dillenia excelsa*	3	Q3U165	Q3U165	Q3U165
65	*Dioscorea orbiculata*	1		Q2U04	
66	*Dioscorea pyrifolia*	2	AS022	AS022	
67	*Diospyros lanceifolia*	3	Q2U11	Q2U11	Q2U11
68	*Diospyros oblonga*	3	Q4T63	Q4T63	Q4T63
69	*Diospyros subrhomboidea*	3	Q8U199	Q8U99	Q8U99
70	*Dissochaeta echinulata*	3	AS203	AS203	PAS203
71	*Dissochaeta gracilis*	3	AS152	AS152	PAS152
72	*Dracaena porteri*	3	Q2U22	Q2U22	Q2U22
73	*Durio singaporensis*	3	Q10U121	Q10U121	Q10U121
74	*Dysoxylum cauliflorum*	1	AS161		
75	*Elaeocarpus salicifolius*	1		AS031	
76	*Elaeocarpus stipularis*	3	AS235	Q3U146	Q3U146
77	*Erycibe leucoxyloides*	3	Q8U76	Q8U76	Q8U76
78	*Erycibe tomentosa*	3	Q10U137	AS054*	AS141*
79	*Erythropalum scandens*	3	AS201	AS201	AS201
80	*Eurya acuminata*	3	Q1T06	Q1T06	Q1T06
81	*Fibraurea tinctoria*	3	AS004*	AS004*	Q2U09
82	*Ficus apiocarpa*	3	Q4U153	Q4U153	Q4U153
83	*Ficus aurata*	3	AS002	AS002	AS002
84	*Ficus fistulosa*	3	AS010	AS010*	AS121
85	*Ficus sagittata*	3	AS220	AS220	Q8U88
86	*Ficus variegata*	1			Q1U17
87	*Fissistigma latifolium*	3	AS007	AS007*	AS007
88	*Fissistigma manubathricum*	3	Q10U099	Q10U099	Q10U099
89	*Flacourtia rukam*	3	Q8U75	AS123	AS123
90	*Freycinetia angustifolia*	3	AS130*	Q4U105	Q4U105
91	*Freycinetia javanica*	2		Q4U129	Q4U130
92	*Friesodielsia borneensis*	3	AS133	AS133	Q3U140

93	*Friesodielsia glauca*	2		Q10U126	Q10U126
94	*Friesodielsia latifolia*	3	AS170	AS170	AS170
95	*Garcinia celebica*	1		Q8T36	
96	*Garcinia forbesii*	1		Q8U80	
97	*Garcinia nervosa*	1		Q4U145	
98	*Garcinia parvifolia*	1		AS021*	
99	*Gironniera nervosa*	3	AS018	AS018	AS018
100	*Glochidion borneense*	1	Q4U107		
101	*Glochidion zeylanicum*	3	Q10U131	Q10U131	Q10U131
102	*Gluta wallichii*	2	Q4U125	Q4U125	
103	*Gonystylus confusus*	3	Q3U190	AS046	AS046
104	*Grenacheria amantacea*	3	AS171	AS171*	AS171
105	*Grewia laevigata*	3	AS117*	AS023*	AS087
106	*Gynochthodes coriacea*	3	AS299	AS299	Q8U89
107	*Gynochthodes sublanceolata*	3	AS168	AS168	AS168
108	*Gynotroches axillaris*	1		AS034	
109	*Hornstedia leonurus*	3	Q1U14	Q1U14	Q1U14
110	*Horsfieldia polyspherula*	2	Q3U138	Q3U138	
111	*Horsfieldia sucosa*	1		Q3T7	
112	*Hypserpa nitida*	3	Q4U144	Q4U144	Q4U144
113	*Iodes ovalis*	3	AS101*	AS101	AS144
114	*Iodes velutina*	3	Q1U18	Q1U18	Q1U18
115	*Ixonanthes icosandra*	2		AS212	AS212
116	*Ixora congesta*	3	Q8U73	AS009	Q8U73
117	*Jasminum elongatum*	3	Q10U139	Q10U139	Q10U139
118	*Justicia vasculosa*	2	AS062	AS062	
119	*Kibatalia maingayi*	3	Q3U124	Q3U124	Q3U124
120	*Knema communis*	3	AS128	AS128	AS128
121	*Knema latericia*	3	Q4U134*	Q4U134	Q4U134
122	*Knema laurina*	3	AS164	AS164	AS164*
123	*Knema malayana*	3	AS066	AS066	Q3U147
124	*Koompassia malaccensis*	1			Q10U141
125	*Kopsia singapurensis*	3	Q4U109	Q4U109	Q4U109
126	*Kunstleria ridleyi*	3	AS196	AS196	AS196
127	*Lasianthus attenuatus*	1			Q10U102
128	*Leea indica*	3	AS028	AS028	Q3U186
129	*Leuconotis griffithii*	3	Q2U24	AS242	Q2U24
130	*Limacia scandens*	3	AS177	AS177	AS177
131	*Lindsaea cultrata*	1		Q1U02	
132	*Lithocarpus conocarpus*	3	AS083	AS083	AS083
133	*Lithocarpus lucidus*	2	Q3U126	Q3U126	
134	*Litsea erectinervia*	3	Q3U144	Q3U144	Q3U144
135	*Litsea grandis*	3	Q3T60	Q3T60	Q3T60
136	*Litsea machilifolia*	3	Q10U127	Q10U127	Q10U127
137	*Lophopetallum wightianum*	3	Q2U10	Q2U10	Q2U10
138	*Luvunga crassifolia*	2	Q4U149		AS217
139	*Lygodium logifolium*	1		AS060	
140	*Maasia glauca*	2	Q3U157	Q3U157	
141	*Macaranga bancana*	2		Q2U27	Q2U27
142	*Macaranga gigantea*	1			AS322
143	*Macaranga heynei*	2		AS036	AS036
144	*Maclurodendron porteri*	3	Q10U136	Q10U136	Q10U136
145	*Maesa ramentacea*	3	AS035	AS035	AS035
146	*Mallotus paniculatus*	2		AS049	AS049
147	*Matthaea sancta*	3	AS149	AS149	PAS149
148	*Melanochyla angustifolia*	3	Q3U155	Q3U155	Q3U155
149	*Melanochyla caesia*	3	AS207	AS207	AS207
150	*Melastoma malabathricum*	2		AS015*	AS099*
151	*Meliosma simplicifolia*	2		Q1U10	Q1U10
152	*Memecylon dichotomum*	3	Q8U84	Q8U84	Q8U84

153	*Memecylon garcinoides*	3	Q8U114	Q8U114	Q8U114
154	*Microcos latifolia*	2	Q1T47	Q1T47	
155	*Mikania micrantha*	3	AS116	AS116	AS116
156	*Mitrella kentia*	3	Q1T27	AS096	Q1T27
157	*Morinda citrifolia*	2		AS115	AS115
158	*Mussaendopsis beccariana*	3	Q3U168	Q3U168	Q3U168
159	*Myristica elliptica*	3	Q4T13	Q4T13	Q4T13
160	*Myristica iners*	2		Q8U94	Q8U94
161	*Myristica maxima*	2	Q8B	Q8B	
162	*Neoscortechinia sumatrensis*	3	AS126	AS126	AS126
163	*Nepenthes gracilis*	2	AS014	AS014	
164	*Nephelium laurinum*	3	Q4U120	Q4U120	Q4U120
165	*Osmelia phillippina*	2		AS082	AS082
166	*Paraderris montana*	1		Q3U145	
167	*Parameria polyneura*	3	AS352	AS352	AS352
168	*Paramignya scandens*	3	AS173	AS173	AS173
169	*Passiflora laurifolia*	3	AS143	AS143	AS143
170	*Pentace triptera*	1			Q10U128
171	*Piper caninum*	2	AS178	AS178	
172	*Piper flavimarginatum*	1		Q3U174	
173	*Piper macropiper*	1			Q10U107
174	*Piper pedicellosum*	3	Q1U31	Q1U31	Q1U31
175	*Piper porphyrophyllum*	2		AS045	AS045
176	*Polyalthia angustissima*	3	AS092	AS092	AS092
177	*Polyalthia cauliflora*	3	Q4U140	Q4U140	Q4U140
178	*Polyalthia glauca*	3	AS057	AS057	AS057
179	*Polyalthia lateriflora*	3	Q3U158	Q3U158	Q3U158
180	*Polyalthia rumphii*	3	Q2U32	Q2U32	Q2U32
181	*Pometia pinnata*	3	Q4T23	Q4T23	Q4T23
182	*Popowia fusca*	3	AS073	AS073	AS073
183	*Porterandia anisophylla*	3	Q3U133	Q3U133	Q3U133
184	*Pouteria malaccensis*	3	Q2U06	Q2U06	Q2U06
185	*Prunus arborea*	3	Q8U108	Q8U108	Q8U108
186	*Prunus grisea*	3	Q8U91	Q8U91	Q8U91
187	*Prunus polystachya*	3	AS293	AS293	AS293
188	*Psychotria ovoidea*	2		Q10U153	Q10U153
189	*Psychotria rhinocerotis*	2		Q10U130	Q10U130
190	*Psychotria sarmentosa*	2		AS169*	Q3U194
191	*Psydrax sp*	3	Q2U44	Q2U44	Q2U44
192	*Pterisanthes polita*	3	Q3U132	Q3U132	Q3U132
193	*Pternandra coerulescens*	1		AS011	
194	*Pternandra echinata*	3	AS005	AS005	AS005
195	*Pyramidanthe prismatica*	3	AS174	AS174*	AS175
196	*Radermachera pinnata*	3	Q3T2	Q2U28	Q2U28
197	*Rhaphidophora maingayi*	2	AS205	AS205	
198	*Rhaphidophora montana*	1		Q3U181	
199	*Rhodamnia cinerea*	2	AS274		AS274
200	*Rourea acutipetala*	3	Q2U35	Q2U35*	Q2U35
201	*Rourea asplenifolia*	1		Q10U118	
202	*Rourea fulgens*	2		AS239	AS239
203	*Rourea mimisoides*	1			Q10U115
204	*Rourea minor*	2	AS337		AS337
205	*Salacia korthalsiana*	1		Q3U128	
206	*Scaphium macropodum*	1		Q3U129	
207	*Securidaca phillippinensis*	3	AS100	AS100	AS100
208	*Smilax setosa*	2	AS269	AS269	
209	*Spatholobus ferrugineus*	3	AS067	AS067	AS067*
210	*Spatholobus ridleyi*	3	AS214	AS214	Q4U150
211	*Stenochlaena palustris*	1		AS167	
212	*Sterculia cordata*	3	AS059	AS059	AS059

213	*Sterculia lanceolata*	3	Q2U10	Q2U10	Q2U10
214	*Sterculia rubiginosa*	1		Q4T16	
215	*Streblus elongatus*	1		Q4U161	
216	*Strombosia ceylanica*	3	AS044	AS044	AS044
217	*Strophanthus caudatus*	3	AS048	AS048	AS048
218	*Symplocos fasciculata*	3	Q4U115	Q4U115	AS222
219	*Syzygium lineatum*	1			Q10U125
220	*Syzygium nemestrinum*	3	Q8U110	Q8U110	Q8U110
221	*Syzygium oblatum*	1		Q3T42	
222	*Syzygium pachyphyllum*	3	Q8T39	Q8T39	Q8T39
223	*Syzygium papillosum*	1		Q3U135	
224	*Syzygium pseudoformosum*	3	Q8U83	Q8U83	Q8U83
225	*Syzygium ridleyi*	3	AS042*	AS042	Q4U137
226	*Tetracera indica*	3	AS020*	AS020	AS020
227	*Tetracera macrophylla*	3	AS136	AS136	AS136
228	*Tetrastigma leucostaphylum*	2	AS038*	AS030*	
229	*Tinospora microcarpa*	3	AS213	AS213	AS213
230	*Uncaria attenuata*	3	Q10U110	Q10U110	Q10U110
231	*Uncaria cordata*	1			Q10U151
232	*Uncaria lanosa*	2		AS040	AS040
233	*Uncaria longiflora*	1			Q4U138
234	*Urophyllum sp*	3	Q3U150	AS013	AS013
235	*Uvaria cordata*	2	AS163	AS163	
236	*Uvaria griffithii*	3	Q10U119	Q10U119	Q10U119
237	*Uvaria lobbiana*	3	Q1U07	Q1U07	Q1U07
238	*Uvaria pauciovulata*	3	AS155	AS155	AS155
239	*Vanilla griffithii*	3	Q8U170	Q8U70	Q8U70
240	*Vatica pauciflora*	3	AS112	Q3U131	AS112
241	*Ventilago maingayi*	1	AS346		
242	*Vitex pinnata*	2	AS097		AS097
243	*Willughbeia coriacea*	3	AS243	AS138	AS138
244	*Xanthophyllum ellipticum*	3	Q4U154	Q4U154	Q4U154
245	*Xylopia magna*	3	Q2U18	Q2U18	Q2U18
246	*Xylopia malayana*	3	Q10U132	Q10U132	Q10U132
247	*Ziziphus calophylla*	3	AS114	AS176*	Q3U134
248	*Ziziphus elegans*	1			Q3U156

The current estimate of the diversity of species of trees and lianas in Nee Soon is ~720 species. Therefore in order to build a comprehensive database, much more sampling, vouchering and DNA sequencing needs to be carried out. In my thesis the problem of having only an incomplete barcode database was largely solved by including data from GenBank. Many of the Nee Soon genera that lacked DNA barcodes were fortunately represented in GenBank. After filtering out sequences that did not correspond to the barcode region and trimming the remaining sequences to the barcode region, I obtained 63,107 sequences for *rbcL*, 73,705 sequences for *matK* and 37,538 sequences for *trnL-F*. Overall this corresponded to data for 28,457 "species" for rbcL, 36,888 "species"

for *matK* and 22,613 "species" for *trnL-F*. However, among these 1,853, 1,915 and 758 contained "sp." as epithet. The genus level diversity was 7,532 (410 families), 7,888 (421 families) and 5,048 (280 families) for *rbcL, matK* and *trnL-F*, respectively. To this I added the locally sequenced data. Across all four databases the dominant families in terms of species diversity were Poaceae, Fabaceae, Asteraceae and Orchidaceae. Note that my method for sequence extraction was effective because the species numbers in my databases are similar to what is available in BOLD (which contains data largely mined from GenBank by the team responsible for BOLD, Ratnasingham and Hebert, 2007). BOLD public dataset currently comprises data for 23,346 species for *rbcL* and 31,362 species for *matK*. Lastly, the version of the database used in Chapter 5 was slightly older, and hence had fewer sequences, while in Chapter 6, I present results based on the latest database generated as of May 2014. Using the same procedures, I also included a database of *trnH-psbA* comprising 25,497 sequences for 2,638 genera in 251 families in Chapter 5.

Next, I built a COI database corresponding to the Metazoa barcode region. Overall, 900,499 sequences were downloaded from GenBank and 765,218 sequences were left after trimming to the barcode region and removing non-COI sequences. The dominant phylum in the database was Arthropoda, followed by Chordata, Mollusca and Annelida (Fig. 4.3). Lastly, I used different databases for characterizing rDNA sequences in the sample. While in Chapter 5, I used MG-RAST's annotation tools, in Chapter 6 I first used SILVA to assess the sequences. I also created a target non-human primate parasite database for SSU rDNA (18S) comprising 5,148 sequences from 25 genera. These databases were then used for characterizing the biology of species in Chapter 5 and Chapter 6.

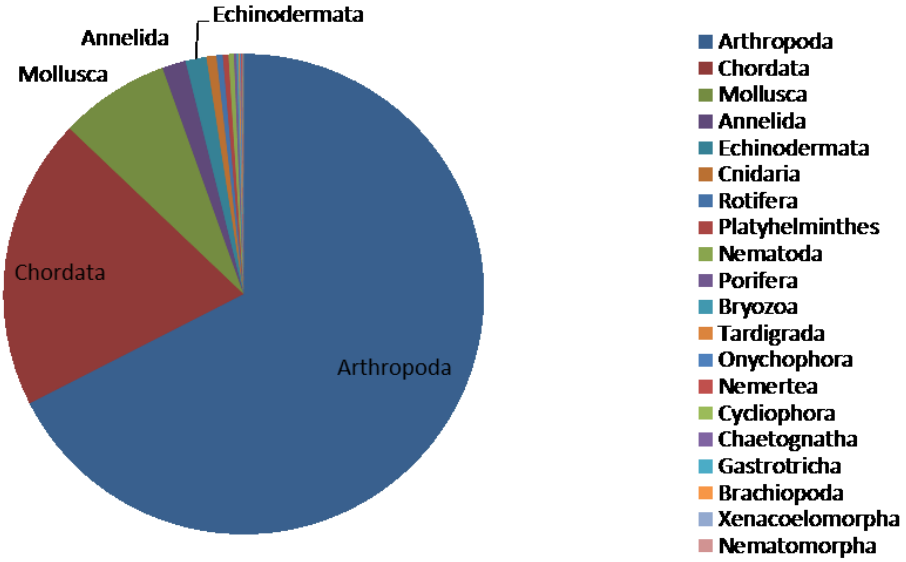


Figure 4.3: Distribution of sequences across various phyla in the COI database.

CHAPTER 5[3]

Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*)

5.1 Abstract

Fecal samples are of great value as a non-invasive means to gather information on the genetics, distribution, demography, diet, and parasite infestation of endangered species. Direct shotgun sequencing of fecal DNA could give information on these simultaneously, but this approach is largely untested. I used two fecal samples to characterize the diet of two Red-Shanked Doucs Langurs (*Pygathrix nemaeus*) that were fed known foliage, fruits, vegetables and cereals. Illumina HiSeq produced ~74 and 67 million paired reads for these samples, of which ~10000 (0.014%) and ~44000 (0.066%), respectively, corresponded to chloroplast genomes. Sequences were matched against a database of available chloroplast 'barcodes' for angiosperms. The results were compared with 'metabarcoding' using PCR amplification of the P6 loop of *trnL*. Metagenomics identified 7 and 9 of the likely 16 diet plants, against 6 and 5 identified by metabarcoding. Metabarcoding produced thousands of reads consistent with the known diet, but the

[3] A version of this Chapter has been published as "**Srivathsan A.**, Sha, J.C.M., Vogler, A.P., Meier R. (2014). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources.* doi:10.1111/1755-0998.12302" where I designed the study and conducted data analyses.

barcodes were too short to identify several plants to genus. Metagenomics utilized multiple, longer barcodes that combined had greater power of identification, but rare diet items were not recovered. Read numbers for diet species in metagenomic and metabarcoding data were correlated, indicating that both are useful for determining relative sequence abundance. Metagenomic reads were uniformly distributed across the chloroplast genomes; thus if chloroplast genomes are used as reference, the precision of identifications and species recovery would improve further. Metagenomics also recovered the host mitochondrial genome and numerous intestinal parasite sequences in addition to generating data useful for characterizing the microbiome.

5.2 Introduction

Rare, endangered and elusive animals are difficult to study in the field (Ang *et al.* 2010). Not only is it time-consuming to locate individuals but they may also stop behaving naturally once they discover the observer. In this situation fecal samples become important because they can provide ecological information (Kohn & Wayne 1997) even without direct observation. Such samples can be collected opportunistically or by using detection dogs (Reed *et al.* 2011) and hold a wealth of biological information (da Silva *et al.* 2012). DNA based methods have become important for characterizing these samples to obtain information on the genetics, diet, distribution, demography, gut parasites and intestinal flora of a species (Ang *et al.,* 2012, da Silva *et al.* 2012, Lamendella *et al.* 2011, Shehzad *et al.* 2012). In terms of diet, most DNA based studies currently adopt a metabarcoding approach, i.e., PCR-amplified short DNA 'barcodes' for diet items are sequenced using next generation sequencing (NGS) (Valentini *et al.* 2009; Shehzad *et al.* 2012). With the decreasing cost of NGS, the obvious alternative is PCR-free shotgun sequencing of genomic DNA (Taberlet *et al.* 2012). This yields large numbers of random sequence reads from which the relevant information can be extracted *in silico* (i.e., a metagenomic approach). Here I compare the power of metagenomic and DNA metabarcoding approaches to identify the food plants from feces of captive colobine Red Shanked Douc Langurs (*Pygathrix nemaeus*) that were fed a known diet. Additionally I used the fecal samples to recover sequences of the host, as well as other eukaryotic sequences that might indicate the presence of intestinal parasites.

Diet studies on fecal samples have traditionally been carried out using visual analyses of physical remains. Newer methods have included chemical analyses of plant cuticular wax (Dove & Mayes 1996), immunoassays (Pierce *et al.* 1990; Symondson

2002) or residual DNA of food items (Murray *et al.* 2011). The latter can be implemented by matching sequences from fecal samples against sequence databases of potential food sources. Initially, these sequences could be obtained via PCR amplification using lineage specific (Jarman *et al.* 2004; Deagle *et al.* 2007) or generic primers (Bradley *et al.* 2007), followed by cloning and sequencing. Recently, such amplicons are being sequenced using NGS. This approach has been applied to carnivorous (e.g. Shehzad *et al.* 2012), herbivorous (e.g. Valentini *et al.* 2009) and omnivorous animals (De Barba *et al.* 2014). One advantage of metabarcoding is that amplicons for multiple samples can be multiplexed. On the other hand, all PCR-based approaches have potential limitations due to amplification biases towards certain taxa (Pompanon *et al.* 2012), difficulty to obtain amplicons (Zarzoso-Lacoste *et al.* 2013), and the generation of PCR errors (Coissac *et al.* 2012) and chimerical sequences due to jumping PCR (Paabo *et al.* 1990).

Certain experimental procedures can mitigate these limitations (De Barba *et al.* 2014; Zarzoso-Lacoste *et al.* 2013), but the PCR step would be avoided altogether by metagenomics. In addition, a metagenomic approach would allow for characterization of reads bioinformatically to address not only the diet, but also the population genetics of the focal species (Ang *et al.* 2012), its intestinal parasites (Stensvold *et al.* 2011) and the microbiome of the gastrointestinal tract (Lamendella *et al.* 2011). While metabarcoding with multiple PCR primers could also be used for such multi-dimensional characterization of samples, it remains constrained by pre-determined choices of barcoding genes, which could preclude, for example, the detection of carnivory in species that are assumed to be phytophagous. In addition, the cost advantage of metabarcoding erodes as more genes are amplified.

The use of shotgun sequencing for diet characterization was championed by Bon *et al.* (2012) for coprolites. Here I apply metagenomics to fresh fecal samples of a phytophagous monkey and develop methods for identifying plant species from such data. By using captive animals I was able to test the methods against a known set of food plants with the greatest challenge being the low diagnostic power of plant barcodes (Hollingsworth *et al.* 2011) and the long digestion times of douc langurs (Lambert 1998) that are likely to favor the dominance of microbial DNA in the extraction (Lamendella *et al.* 2011). I address these issues by developing bioinformatic strategies for extracting plant sequences and comparing the results to metabarcoding data for the same samples.

5.3 Materials and Methods

5.3.1 Diet composition

Two individuals of *Pygathrix nemaeus* (PN1: male, 6 years old; PN2: male, 5 years old) were fed leaves of 7 species: *Acalypha siamensis* (acalypha), *Cinnamomum iners* (wild cinnamon), *Hibiscus rosa-sinensis* (hibiscus), *Hemigraphis sp., Leucaena leucocephala* (miracle plant), *Morus alba* (mulberry), and *Terminalia catappa* (ketapang). At the beginning of the third day of the trial, cinnamon was replaced by *Baphia nitida* (baphia). These plants were provided as a mixed bunch of leaves and eight non-foliage species were added to the diet, including *Malus domestica* (apple), *Daucus carota* (carrot), *Ipomoea batatas* (sweet potato), *Vigna unguiculata* (long bean), *Pyrus sp.* (pear), *Zea mays* (corn), *Cucumis sativus* (cucumber) and *Oryza sativa* (rice, provided as cooked rice balls). To optimize the time of sample collection, I used feeding trials to determine the Transit Time (TT) and Mean Retention Time (MRT) of the diet in the gut of the primate using bead markers (Appendix 2, Methods).

5.3.2 Sample preparation and Next Generation Sequencing

Fecal samples were collected 72 hours after the beginning of the feeding trial (see Results) and stored in -80°C prior to DNA extraction. Ten DNA extractions were conducted for each sample by randomly sampling the surface and interior of a single fecal pellet (QIAGEN DNeasy Blood and Tissue Kit with an additional wash step using Buffer AW2). The fecal samples were extracted on different days in a laboratory where no experimental work on plants was being conducted. DNA from the different extractions was quantified using Nanodrop and only those with A260/280 between 1.8 and 2.0 were

combined in equal amounts, after which the samples were split in separate sets to be used for metagenomics and metabarcoding respectively. For metagenomics, one library was constructed for each fecal sample (clone insert size 280-300bp). These were multiplexed in one lane of Illumina HiSeq 2000 and paired 76bp reads were obtained using TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3.

For metabarcoding, P6 loop of the chloroplast *trnL* intron from fecal DNA was amplified using primers *trnL*-g and *trnL*-h (Taberlet *et al.* 2007). Each sample was tagged using eight variable nucleotides at the 5' end of each primer that were designed using oligoTag (Coissac 2012; >= 5 variable sites; <3 bp homopolymers; additional dinucleotide CC was added to 5' end). PCR amplifications were carried out for 45 cycles as in Quéméré *et al.* (2013) using BioReady rTaq DNA polymerase (Bulldog Bio, Inc., Portsmouth, NH) with a reaction mixture of 2.5 µl Buffer, 1 µl dNTPs, 0.36 µM forward and reverse primers, 0.25 µl of rTaq polymerase. Three independent PCR replicates were obtained for each sample; PCR products were purified using the MinElute PCR Purification Kit (QIAGEN). Products were quantified with a Fragment Analyzer™ Automated CE System (Advanced Analytical) and combined in equimolar ratios before sequencing with lllumina MiSeq (Illumina Inc) using the TruSeq Nano DNA sample preparation kit (150 PE).

5.3.3 Diet database

A barcode database for known diet species comprising *rbcL, matK,* and *trnL-F*, (the latter containing the metabarcoding fragment) was prepared for the 16 plant species that were fed during the trial as well as 35 other "potential" diet plants (list of fodder plant species that are regularly fed by the Singapore Zoological Gardens; Appendix 1 Table T1). Seven of the 16 known diet species were sequenced with the Sanger method (Chapter 4). The *trnL* fragments were used to create the diet database for metabarcoding. I used ecoPCR (Ficetola *et al.* 2010) to only retain the fragments of *trnL* that corresponded to amplification productions generated with the *g-h* primer pair (Taberlet *et al.,* 2007). All sequences < 10 bp and > 200 bp were excluded.

5.3.4 Plant database

In order to assess the ability of the two approaches to identify plants even if the diet was not known *a priori*, I generated a database comprising all barcode sequences available at GenBank for *rbcL*, *matK*, *trnL-F* and *trnH-psbA* using the BLAST based pipeline of Hunt *et al.* (2007) (Chapter 4). This dataset was complemented with our *rbcL, matK* and *trnL-F* sequences for the seven foliage species used in the feeding trial to yield databases of the following sizes: *matK*: 55,996 sequences for 7165 genera in 401 families; *rbcL*: 48,831 sequences for 7,058 genera in 421 families; *trnL-F*: 37,241 sequences for 5,052 genera in 281 families; *trnH-psbA:* 25,497 sequences for 2,638 genera in 251 families). For metabarcoding, I obtained *trnL* fragments for 4,602 angiosperm genera corresponding to the region amplified by the g-h primer pair.

5.3.5 Data analysis

Metagenomic approach

An initial assessment of quality scores across the Illumina data was done using FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and sequences were analyzed either with or without assembly. For the assembly-free analyses, FASTQ sequences were converted to FASTA, and raw 'single-end' reads (SE analyses') or paired-end reads (PE analyses) were matched against the diet and plant barcode databases. For single-end analyses, each read was matched to the database using BLASTN (Altschul *et al.* 1990) as implemented in the BLAST 2.2.27+ suite under the default MEGABLAST settings (word-size 28). To check for false positives, all reads showing matches to the diet database were also tested for BLAST hits against the generic nucleotide (NT) database of GenBank to establish if these reads represent non-plant sequences. This analysis revealed that plant sequences were only reliably distinguished from bacterial sequences if the hit length exceeds 50 bp; removing shorter matches eliminated all matches to non-plant sequences. I then recorded identification success rates at 100% and 98% (=1bp mismatch) identity. Lower identity thresholds were rejected because they yielded identifications to plants that were not known to be fed to the animals in the Singapore Zoo.

For single end analyses I used every read that gave a hit to single or multiple species (comprising taxon set $S1$ for a read that matched from end 1 and $S2$ for a read that matched from end 2). Species/genus level identifications were made only when the number of species/genera in $S1$ (or $S2$) were not >1, i.e. they were unequivocal with respect to other sequences in the database. Identifications were categorized as "ambiguous" when matches to multiple species/genera were obtained. For the paired-end

analysis up to 152-bp of sequence information could be used for identification; i.e., the set of genera identified was $S = S1 \cap S2$ (conflicting identifications from both ends were excluded). Both single- and paired-end analyses were repeated for each gene in the plant database and I recorded whether identifications were based on matches to one, two, or three barcode genes. The pipeline is available to download at https://github.com/asrivathsan/readsidentifier-1.0 and the details are provided in Appendix 3

Additionally I attempted a diet analysis based on assembled reads. Sequences were assembled using SOAPdenovo2 (Luo *et al.* 2012; Zhou *et al.* 2013) using three *k*-mer settings (31, 41, 51) before choosing the k-mer length that maximized species identification success (k=31, Appendix 1 Table T3-4). Assembled contigs were matched against the plant database using MEGABLAST with an overlap of >100bp and a 98% identity threshold for identifications.

Metabarcoding approach

For metabarcoding, I followed the methods in Quéméré *et al.* (2013) and De Barba *et al.* (2014). PE reads were first aligned and merged using *illuminapairedend* (http://www.grenoble.prabi.fr/trac/OBITools). Reads were assigned to the samples using *ngsfilter* under criteria of perfect match of tag sequence and a maximum of 2bp mismatch with the primer sequence, after which *obisplit* was used to divide the files. Identical sequences were clustered while retaining information on sequence counts using *obiuniq*. Sequences having length of <10bp were removed. I used two stringent filtering criteria after more relaxed criteria led to several erroneous identifications: (1) FC1: removing sequences with counts <0.1% of the most common sequence [~100 reads, similar to

Hilbert *et al.* (2013)] and (2) FC2: removing sequences with counts <1% of the most common sequence [~1000 reads, similar to Quéméré *et al.* (2013)]. Sequence variants were identified using *obiclean* and sequences were tagged as "head", "internal" and "singleton". These assignments of *obiclean* can be explained as follows: *obiclean* identifies all sequences that are 1 bp (or specified threshold) away from another sequence. Once identified, the sequences with the maximum counts are called "head" sequences while the variants are called "internal". Sequences that do not have any variants are then tagged as "singleton" (see Shehzad *et al.* 2012). Identifications were made using *ecotag*, and only "head" and "singleton" sequences were used for genus level identifications if identity was >98% (Quéméré *et al.* 2013) or >95% (De Barba *et al.* 2014).

Comparison of metagenomics and metabarcoding

I used three criteria to compare the performance of metagenomics and metabarcoding. Firstly, I tested whether diet sequences were recovered by matching reads to the diet database; this database is species-poor and most reads are sufficiently diagnostic for a particular species. Secondly, I determined if read abundances were correlated for the same diet species using diet database [Spearman's rho, R Development Core Team (2011)]. Thirdly, I tested whether the diet reads could be identified to species/genus using the plant database containing all angiosperm barcodes in GenBank.

Proportion of chloroplast reads in metagenomic data

I used BLAT searches (Kent 2002) with word-size 11 against all 366 full chloroplast genomes in NCBI Genomes (as of 6 Aug 2013) to extract all potential chloroplast reads. These reads were then filtered through BLASTN searches (word-size 11) against all non-human genomes in NCBI (*other_genomic* database) to retain only

those with matches exclusive to angiosperm chloroplast genomes. Distribution of these reads was studied by BLASTN (word-size 11) searches against a reference cp-genome using a 50bp overlap threshold. For each read, the position in the chloroplast genome was recorded to generate a map of hits. All best Score (S) matches were mapped; i.e., some reads were mapped multiply if they had tied S values for multiple sites.

Characterization of host mtDNA

To test whether host information can be retrieved, mitochondrial genomes were characterized by using MEGABLAST to match assembled contigs (see above) against the mitochondrial genome of *Pygathrix nemaeus* (JF293096.1). The matches were validated as non-human primate sequences and coding regions were translated and mapped using BRIG 0.95 (Alikhan *et al.* 2011). For quality checking, raw reads of each individual were also mapped using BWA (Li & Durbin 2009) and mismatches between contigs and reads were recorded as ambiguous bases.

Characterization of other eukaryotic DNA

For identifying other eukaryote reads, FASTQ files were submitted to MG-RAST (Glass *et al.* 2010) using the default pipeline with quality filtering, RNA and protein prediction, clustering and taxon assignment. In addition, the reads were matched against a sample database of 698,981 COI sequences downloaded from GenBank. SE and PE reads with matches to rDNA or COI were then identified against NT requiring a 70bp overlap. rDNA identifications were made at 98% identity, while COI identifications were summarized at both 98% and 95% identity. If multiple taxa had the same top similarity level, ambiguity was noted. The taxonomic classification of identified species was plotted at several hierarchical levels.

5.4 Results

5.4.1 Illumina sequencing

The *TT* of *P. nemaeus* was determined to be 27.8 hours and the *MRT* was 48.8 hours. After 72 hours, 80% of the bead markers had passed. Fecal samples were therefore collected at 72 hours after the beginning of the feeding trial. DNA extracted from these samples was sequenced to obtain 74,325,939 (11.3 Gb) and 67,127,731 (10.2 Gb) reads of 76 bp from PN1 and PN2, respectively (Table 5.1). The mean sequence quality was high (Phred score ~38), but decreased beyond 60 bp and showed very low scores beyond 70 bp. Across both samples, the mean, upper and lower quartiles of Phred scores were >20 for the first 60 bp. For metabarcoding 268,779 (PN1) and 289,834 (PN2) reads were available for variant calling and filtering

Table 5.1: Sequences used in metagenomic and metabarcoding analyses of samples. For metagenomics, data are summarized using the plant database.

	PN1	PN2
Metagenomics		
Total Number of reads	74,325,939	67,127,731
Single-end reads matching to barcode sequences	494	2001
Reads used for Single-End analyses (100% identity)	281	1107
DNA fragments overlapping barcode sequences	359	1257
DNA fragments with both ends overlapping barcode sequences	135	744
DNA fragments used for Paired-End analyses (98 %identity)	105	545
Metabarcoding		
Total Number of reads	268,779	289,834
Unique sequences	10,740	8,592
Unique sequences passing FC1	110	99
Unique sequences passing FC2	13	14

5.4.2 Comparison of metagenomics and metabarcoding

Best estimate of diet based on metagenomic and metabarcoding data

While the diet that was offered to the douc langurs is known, it remains unclear whether all species were consumed over the 72 hours of the trial. Our best estimate of diet thus has to be based on molecular evidence and I used all reads (metabarcoding and metagenomic) and our best identification criteria (see under "identification") for this purpose. For this, I matched the reads against the diet database. This showed that the metagenomic data included reads for ten (PN1) and fifteen (PN2) plants (Table 5.2, Green/Yellow; PE analysis). The corresponding numbers for metabarcoding (using FC1) were sixteen for PN1 and fourteen for PN2. Given the overlap between the data, our best estimate of the diet is 16 diet plant genera for each of the two samples.

Abundance

Next I correlated read numbers in the metagenomic data with read numbers of the corresponding metabarcoding data. Spearman's rank-correlation coefficient at >0.7 was highly significant when comparing metagenomic (paired-end) and metabarcoding (FC1) reads (Table 5.3). Across all analyses, most hits were for *Cinnamomum* followed by *Leucaena* and *Terminalia* (Table 5.2). The only major deviations were *Calophyllum* and *Mangifera* in PN1 that were only found in large numbers using metabarcoding.

Table 5.2: Genus level identifications using the various approaches tested in this study. Recovery of a genus and read quantifications were determined using the diet database comprising "known" (highlighted in bold) and "potential" diet genera. PE: Paired End; SE: Single End; Green: Recovered and unambiguously identified; Yellow: Recovered but ambiguous identification; Red: absent. *Ligustrum* was not included due to lack of data for potential diet species in GenBank. *Baphia* and *Daucus* identified using metabarcoding only at 95%.

	PE		SE			Metabarcoding (FC1)	
	PN1	PN2	PN1	PN2		PN1	PN2
Consistent Identifications							
Leucaena	10*	102*	19*	164*		34105*	59734*
Terminalia	11*	9*	19*	21*		11017	18049
Inconsistently identified							
Acalypha	2	1	4	4		2483*	2060*
Baphia		6		6		766*	1993*
Unambiguously identified by metagenomics							
Vigna	3*	19*	11*	29*		4177	4316
Cinnamomum	30	156	79	303		105975	107878
Ficus	4	1	7	3		9747	1630
Averrhoa	4	19*	8*	35*		5171	21191
Ipomoea	2	2	5	3		5223	1541
Daucus	1	7*		15*		12775	8945
Morus		1		3		383	2234
Unambiguously identified by metabarcoding							
Hibiscus		2		1			148*
Calophyllum			1			5698	
Present, but not identified							
Zea	1	4		3		577	1614
Pterocarpus				1		156	112
Malus		2		8			
Pyrus		1		1			
Mangifera						958	
Cucumis						190	

*identified to species as *Vigna unguiculata, Leucaena leucocephala, Averrhoa carambola, Daucus carota, Terminalia catappa, Acalypha siamensis, Hibiscus rosa-sinensis,* and *Baphia puguensis. Glycine max* (likely misidentification) was also identified to species for PN2 in metabarcoding.

Table 5.3: Correlation between the abundance of each genus using metabarcoding (FC1) and metagenomic (paired-end) approaches. Only identifications made under same identity threshold (98%) for the two approaches were considered.

Genera compared	Sample	Spearman's *ρ*	*n*	*p-value*
All genera identified*	PN1	0.822	14	0.0003
	PN2	0.717	14	0.004
Identified using both approaches	PN1	0.899	9	0.00097
	PN2	0.723	11	0.012

* If a genus is not identified by one approach it was represented by zero reads

Identification

Our best estimate of the diet was based on the diet database consisting of fairly distantly related species. However, in order to compare the performance of metagenomics and metabarcoding, it is more important to assess whether diet elements can be identified against a database of all (available) angiosperm barcodes. With regard to metagenomics, preliminary analyses based on SE and PE reads revealed that only identifications based on at least two different barcode loci were reliable, because matches based on single barcode genes yielded too many plant genera that were not part of the known diet (Appendix 1 Table T5). Once the 2 gene criterion for genus-level identification was applied, the PE analysis identified 7/16 (PN1) and 9/16 (PN2) genera, whereas the SE analyses identified 8/16 (PN1) and 7/16 (PN2) of the diet genera (Table 5.2, Fig. 5.1, Green). On the other hand assembly based analyses yielded fewest identifications (Fig. 5.1). For metabarcoding, I found that the FC1 criterion performed best, but the number of identifications was much lower than for metagenomics (PN1: 6/16 and PN2: 5/16) (Fig. 5.1). In contrast, using FC2 I found fewer diet genera but this criterion also avoided some misidentifications (Fig.5.1). The latter criterion for metabarcoding and assembly based analyses for metagenomics trades-off identification certainty against number of identifications; i.e., the more stringent criteria yielded only known diet genera, but there

were very few of them while the less stringent criteria identified a larger number of diet genera albeit at the expense of the occasional misidentification. For example, PE analyses (metagenomics) and FC1 (metabarcoding) yielded only one misidentification, while SE analyses yielded one misidentification per sample (Fig. 5.1). Thus, overall the results of SE and PE analyses are very compatible, but the PE matches are arguably more reliable. Hence I consider the PE analysis and FC1 as best identification criteria for metagenomics and metabarcoding. Note that these conclusions were not sensitive to choosing 95% or 98% identity threshold for metabarcoding.

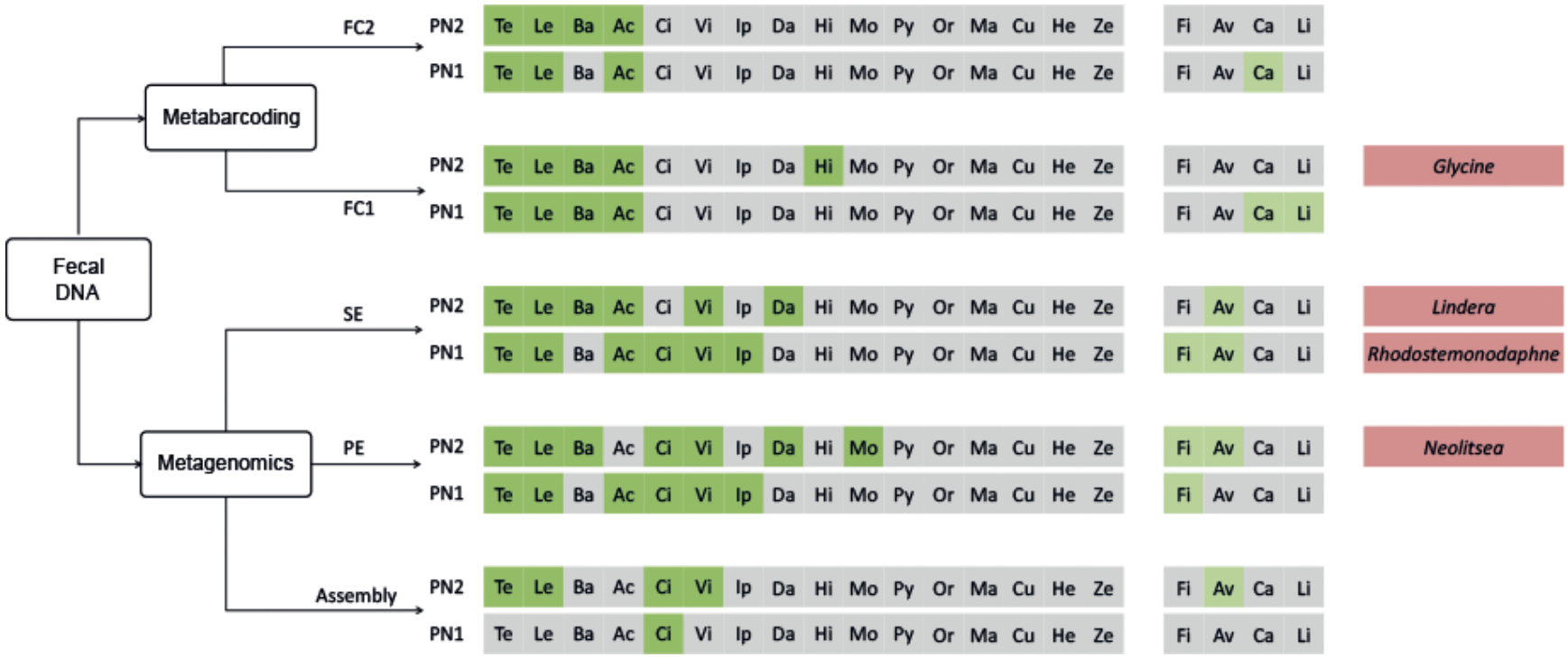


Figure 5.1: An overview of the genus level identification success for five approaches tested. *Te: Terminalia, Le: Leucaena, Ba: Baphia, Ci: Cinnamomum, Ac: Acalypha, Vi: Vigna, Ip: Ipomoea, Da: Daucus, Hi: Hibiscus, Mo: Morus, Py: Pyrus, Or: Oryza, Ma: Malus, Cu: Cucumis, He: Hemigraphis, Ze: Zea, Fi: Ficus, Av: Averrhoa, Ca: Callophyllum, Li: Ligustrum.* Dark green: known diet; light green: potential diet; red: others (potential misidentifications). SE: single end, PE: paired end, FC1: filtering criterion 1, FC2: filtering criterion 2.

At the species level, PE analyses identified three (PN1) and five (PN2) species, while SE analyses identified four (PN1) and five (PN2) species (Table 5.2). Metabarcoding (FC1) yielded three (PN1) and five (PN2) species-level identifications; however, one species was misidentified (*Baphia nitida* as *Baphia puguensis*) in both samples and *Glycine max* ("other" species; likely misidentification) was identified for PN2. Overall both metagenomics and metabarcoding yielded fewer identifications at species-level due to the poor species-level resolution of cp-DNA barcodes. Next, I assessed the genus-level overlap between the identifications made by the two methods. There was overlap for *Leucaena, Terminalia, Acalypha* and *Baphia* (Fig. 5.1, Table 5.2), while *Vigna, Ipomoea, Daucus, Ficus, Averrhoa* and *Morus* were identified only using metagenomics, and *Hibiscus, Calophyllum* and *Ligustrum* were identified only using metabarcoding. Note that *Ligustrum* was not recovered in the diet database, as data for this diet species was not available in GenBank.

The main problems with identifications based on the two methods differed. For metagenomics, the low read counts caused some species to remain undetected because they only had matches to one barcode gene and thus failed the multi-gene criterion. For example, two (PN1) and four (PN2) additional "known" and "potential" diet plants with low read counts satisfied only the single gene criterion (Appendix 1 Table T5). The main challenge for metabarcoding was the poor diagnostic value of the amplified barcode region. Even dominant diet elements such as *Cinnamomum, Ipomoea, Ficus, Vigna* and *Averrhoa* could not be identified to genus because the barcode for these species is not genus-specific; i.e., the PCR step had generated enough reads for these genera to be above the detection threshold but they could not be diagnosed reliably against a broad taxonomic database. Overall, species detection based on metabarcoding is thus limited by

the diagnostic power of the barcodes and several diet species could only be identified to family.

5.4.3 Number of chloroplast sequences

In order to test whether the metagenomic reads cover the chloroplast genome uniformly, I obtained 218,652 sequences for PN1 and 236,600 for PN2 with BLAT (see methods). Most were identified as bacterial sequences in BLAT searches against the genomic reference database, but 10,561 (0.014% of PN1) and 44,167 (0.066% of PN2) reads were likely genuine cpDNA. In order to test the efficiency of BLAT, I compared the recovered read numbers with those found via BLAST searches against the three barcode regions in the diet database. BLAT proved effective because it recovered ~85% (PN1) and 92% (PN2) of these reads. Overall, the chloroplast genome of *Magnolia denudata* (NC_018357.1) had the highest number of read hits and it was used to study the topological distribution of the sequences via BLASTN matches (yielding results for ~88-90% of reads). The sequences were overall uniformly distributed except that there were larger numbers of reads in the inverted repeat regions (Fig. 5.2). This shows that an expansion of the reference database from four barcode regions to full genomes would provide additional sequences that can be used to assess the diet.

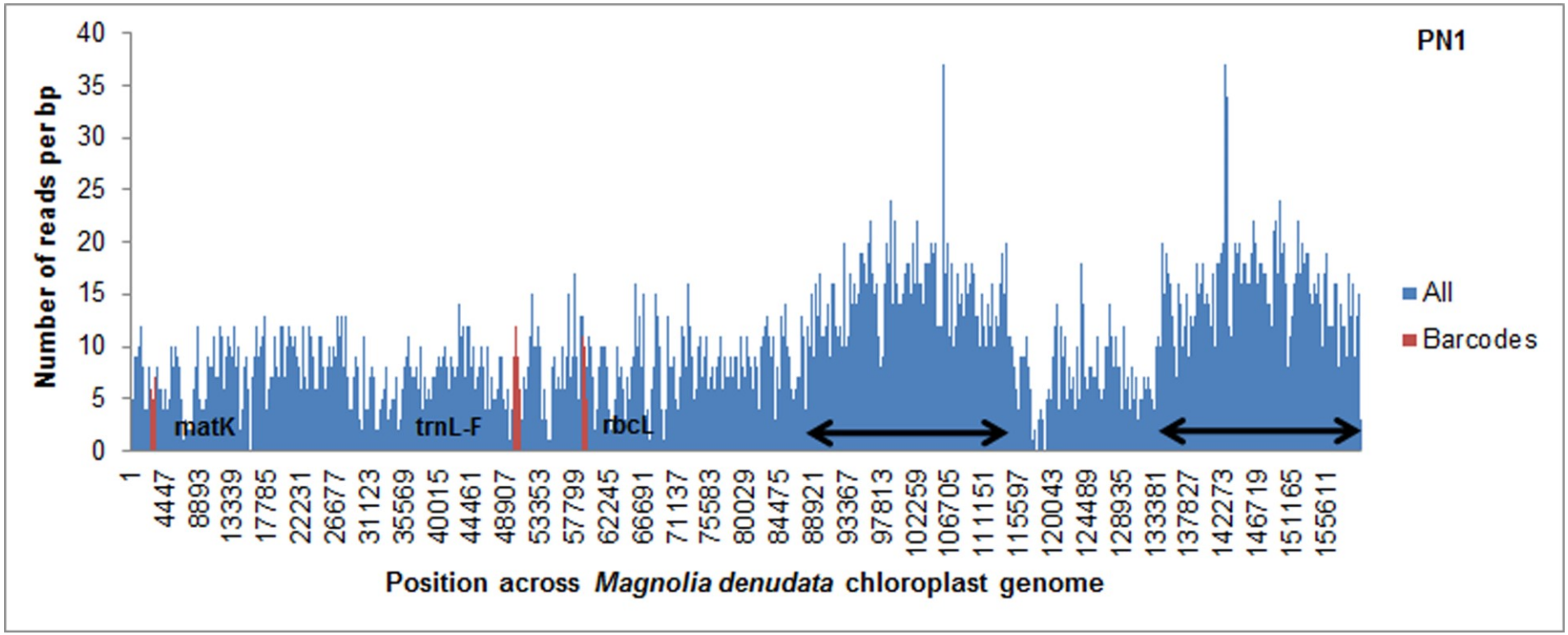


Figure 5.2 (a) Mapping of single-end reads of PN1 onto the *Magnolia denudata* chloroplast genome: Locations of inverted repeats are marked by arrows as estimated using the genome map from CpBase (http://rocaplab.ocean.washington.edu/tools/cpbase). Reads have approximately equal representation outside of the repeat region (see text).

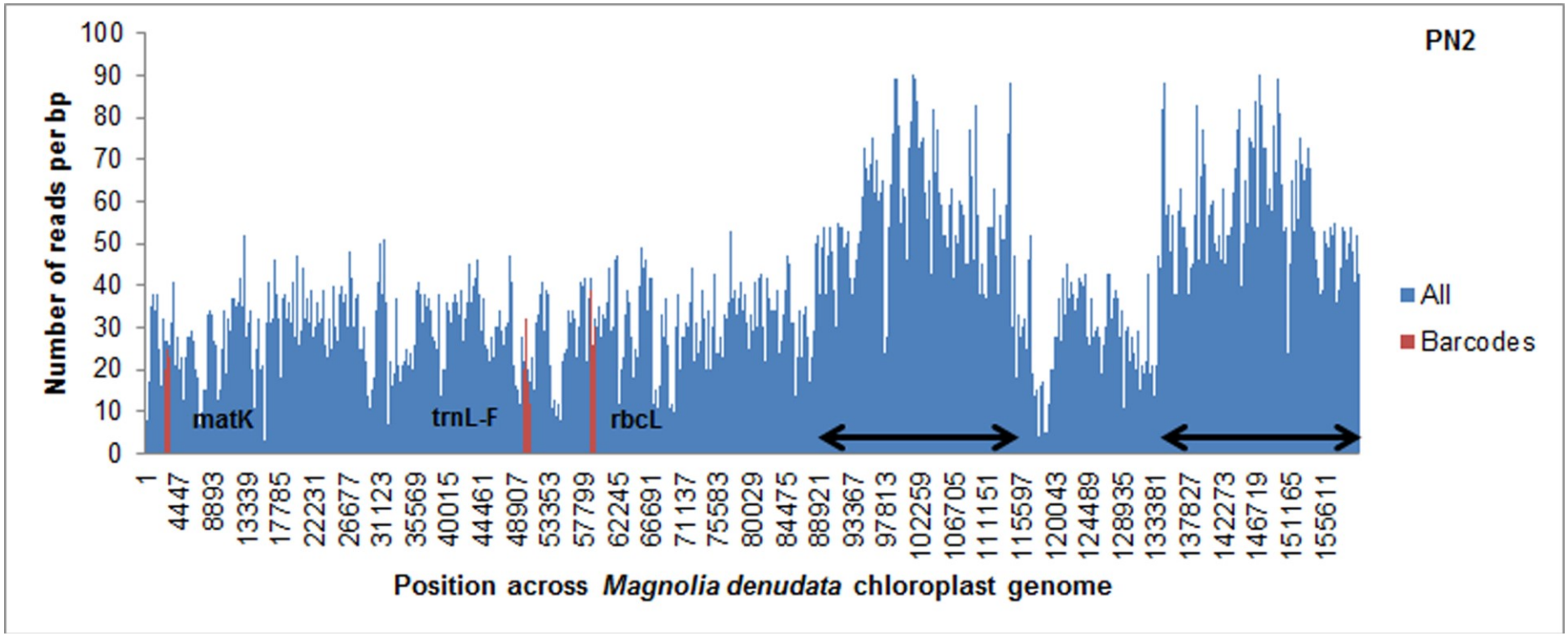


Figure 5.2 (b) Mapping of single-end reads of PN2 onto the *Magnolia denudata* chloroplast genome: Locations of inverted repeats are marked by arrows as estimated using the genome map from CpBase (http://rocaplab.ocean.washington.edu/tools/cpbase). Reads have approximately equal representation outside of the repeat region (see text).

5.4.4 Characterization of host mt-DNA and eukaryotic DNA

Assembly of PN1 and PN2 libraries from the maternally related monkeys produced 25 and 9 mitochondrial contigs of >100bp length which provided two identical, and nearly complete mitochondrial genomes with an average similarity of 99.1% to *Pygathrix nemaeus* (JF293096.1; Appendix 1 Fig. S1). Based on rDNA and COI sequences (Fig. 5.3, details in Appendix 1 Tables S6-9), numerous sequences from nematodes, protozoa, fungi and plants were obtained. Many corresponded to nematodes (4-12% of rDNA identified), specifically to *Strongyloides fuellerboni* (based on LSU, SSU rDNA and COI). Among the Protozoa, several hits were for the heterokont *Blastocystis sp.* (>3000 reads for PN2 based on SSU rDNA) and amoebozoan *Entamoeba sp.* (>2000 reads for PN1 and PN2 based on SSU rDNA). More precise species-level identification would require comparison to databases of homologous regions but there is evidence for the presence of "*Entamoeba sp.* RL3" which is known to be colobine-specific (Stensvold *et al.* 2011). Lastly, both rDNA (at 98% identity) and COI analyses (at 95% identity) revealed arthropod sequences, but they could not be identified beyond the order level based on COI. However, SE data suggested presence of *Ceratitis sp.* and *Drosophila sp.* in PN1, based on three sequences. COI analyses showed also the presence of *Gallus sp.* sequences suggesting that chicken had been ingested. Upon inquiry, the Singapore Zoo confirmed that the rice balls included cooked chicken. The LSU and SSU rDNA to plant sequences were largely congruent with the results of the barcode-based analyses particularly at higher taxonomic levels.

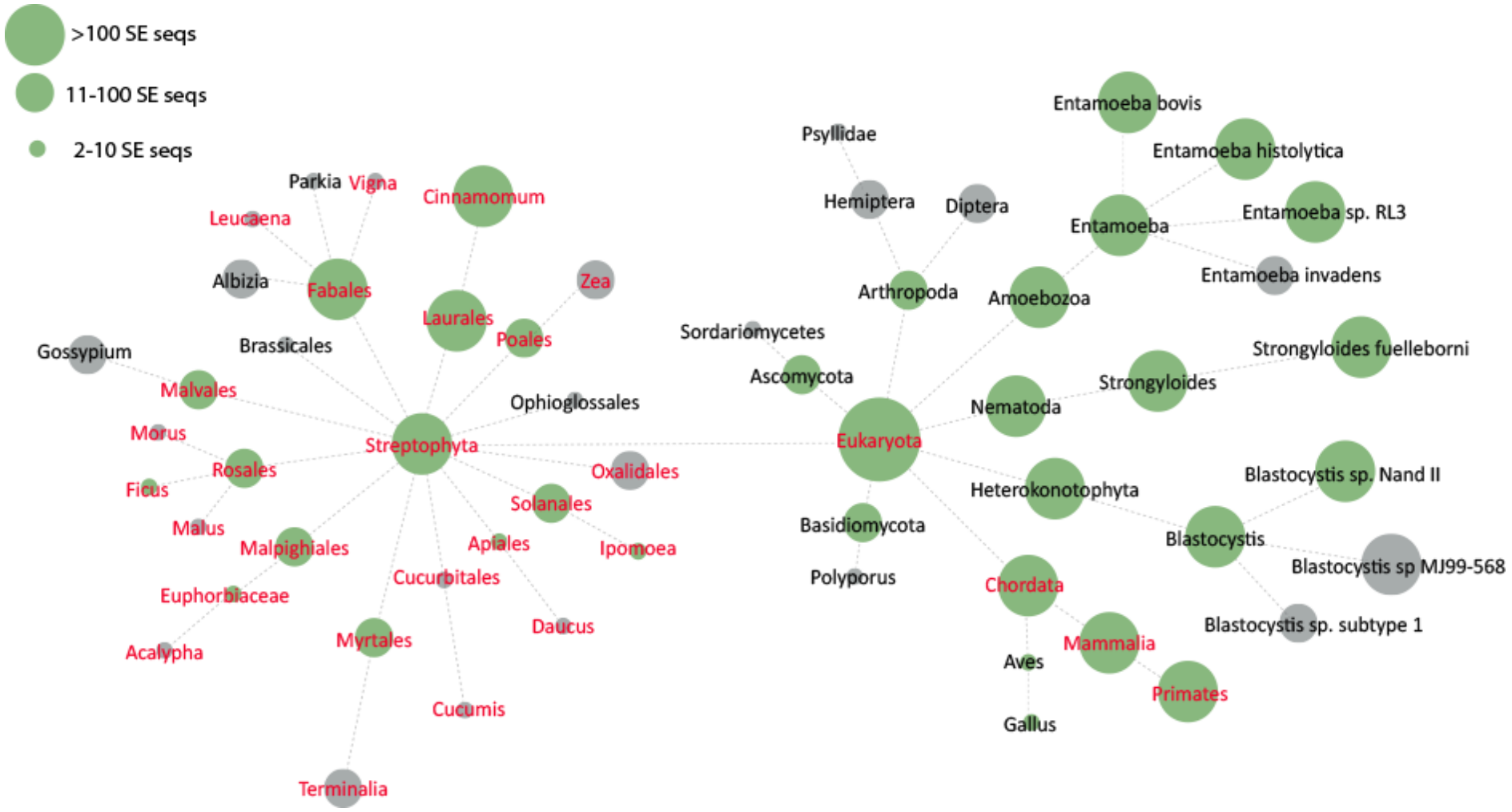


Figure 5.3: Eukaryote identifications based on COI and rDNA (pair-end, 98% identity, 70bp overlap): Green taxa present in both samples; Red=expected species (e.g., diet species, host). Species level identities shown only for *Strongyloides, Blastocystis* and *Entamoeba*. SE refers to Single End reads.

5.5 Discussion

Our study documents how a metagenomic approach can be used to identify food plants from fecal DNA of a mammal with long food retention times. I optimized bioinformatic procedures for plant identifications to genus-level based on paired end data and found chloroplast reads for many of the diet species. I also demonstrated that shotgun sequencing allows for a broad characterization of fecal samples. The same data that were used for diet analysis also document intestinal parasites and yield information on the genetics of the studied individuals. In addition, I revealed the unexpected presence of chicken in the monkey's diet, which could only be explained after the Singapore Zoo confirmed that rice balls that were fed to the monkeys contained chicken. This documents that unexpected diet items can be identified with metagenomics. Additional uses of the metagenomic data would include analyses of the gut microbiomes.

In this chapter I used samples from captive animals for which all potential diet items were known. Therefore I could test which bioinformatic strategy yields reliable results. I was also able to reject those strategies that either identified too few diet genera or mistakenly identified plant genera that are not part of the known diet. For my samples, I found that the following criteria yield the best results for the metagenomic data: (1) read identity with reference barcode ≥ 98%; (2) read overlap of 50 bp, and (3) use of two different reference DNA barcodes. Using these criteria, single-end (SE) and paired-end (PE) analyses yield largely compatible results with seven to nine of the diet taxa identified based on a genus-level identification against a broad database containing tens of thousands of plant barcodes (Fig. 5.1). Between SE and PE analyses I found that the latter marginally improved reliability. I also carried out an assembly-based analysis. It

again yields compatible results but only up to five diet genera are detected; i.e., much deeper sequence coverage would be needed for a complete characterization of the diet via assembled data. Thus, overall I identified seven and nine genera for the two samples using PE analyses. However, there is evidence that this list of diet genera is incomplete. Some reads are ambiguous when identified based on all data in Genbank but they can be assigned to diet species when matched against the species-poor 'diet database'. Moreover, a union of the set of plants recovered by metagenomics and metabarcoding revealed at least sixteen diet genera for each sample (Table 5.2). This dramatic increase and failure to detect low abundance diet items indicate that the coverage of my shotgun sequencing was insufficient for a complete characterization of the diet. Such a characterization would either require target enrichment, higher throughput, or identifying species based on more chloroplast genes. Once these additional data are available, my recommended bioinformatic techniques should be able to identify most diet plants.

Given the coverage problems with my metagenomic data, one may conclude that metabarcoding is a better and potentially cheaper technique because many samples can be multiplexed. However, the metabarcoding analyses using *trnL* are plagued by ambiguity problems which result in the identification of even fewer plant taxa to both genus and species (Table 5.2; Fig. 5.1). Particularly problematic are eight diet taxa that have fairly high read counts (in PN1 and/or PN2), but cannot be identified to genus because the reads have ambiguous matches. This includes the dominant diet item, *Cinnamomum*. The only reason why I can identify these reads in my study is because the potential diet species are known and distantly related. Alternatively, one could abandon taxonomic identifications and only determine the number of Operational Taxonomic Units (OTUs) as a measure of taxonomic diet diversity. As long as the species were distantly related, the number of

OTUs could be determined and they could be identified to family. However, depending on the number of species in the habitat, this level of information may not be useful for conservation purposes. There are several ways for obtaining better results with metabarcoding. Firstly, one could follow De Barba *et al.* (2014) who initially identified diet items to family before using more taxon-specific primers for variable genes such as *nrITS* for a second round of PCR and sequencing. However, this strategy erodes the potential cost advantage of metabarcoding. Alternatively, one could increase the number of amplified barcoding genes in the initial step itself. Such PCR-based amplification of multiple genes may also avoid misidentifications based on a single gene, as was observed in my study for *Baphia nitida.* Especially, barcoding variable markers such as ITS2 could be used for improving taxonomic resolution (Hollingsworth *et al.* 2011). However, the choice of barcodes will depend on degradation level of the samples and amplification efficiency of each primer. In my study, metabarcoding was based on an average of 51 and 53bp fragments of *trnL* for the PN1 and PN2 respectively (Hollingsworth *et al.* 2011), while metagenomic SE analyses uses only 76 bp reads. Such short reads can be used to characterize intact as well as highly degraded samples as insert sizes for library preparation of metagenomic samples can be adjusted to the nature of the sample.

One of the concerns with metabarcoding is amplification biases during the PCR stage. However, using *trnL,* I find very little evidence for such a bias. Most of the diet species in both samples are represented in the metabarcoding data. Indeed, there is overall a strong correlation between read numbers for the same genera in the metagenomic and metabarcoding datasets with only two major discrepancies (Table 5.3). The correlation of read counts is welcome news because one of the ultimate goals of NGS diet analysis is arguably to quantify biomass intake from counts of DNA reads. Metagenomics allows for

direct counts, but it currently comes at a higher cost than metabarcoding; i.e., it would be useful if metabarcoding reads could be used to estimate read quantity in DNA extractions. Of course, further research would be needed before read counts can reliably be correlated with feeding preferences and biomass intake. With regard to douc langurs in the Singapore Zoo, a recent study indicates that the species prefers *Leucaena*, *Terminalia* and *Morus* over *Acalypha*, *Hibiscus* and *Hemigraphis* (Xue & Sha 2010) and my results are consistent with two of the top choices (while *Cinnamomum* was not used in the preference test). This suggests a broad correlation between dietary preferences determined by direct observation and read recovery although the number of sequence reads is likely determined by many factors including differential rates of digestion (Deagle *et al.* 2010) which will be affected by structural differences, such as between leaves and fruits. As an example, *Baphia* was provided only on the third day of the feeding trial but the corresponding reads were already present in the feces at the end of the same day, despite the mean transit time of food being ~28 hrs. Surely, there will be complex interrelationships between biomass, read numbers, feeding preference, time of food intake, and retention time.

Given the advantages and disadvantages of metagenomics and metabarcoding, recommendations for future diet studies will be case specific. The advantage of metagenomics is that dominant diet taxa are identified with a greater resolution while the metabarcoding data has better coverage for rare diet taxa. Looking into the future, it is likely that NGS cost will decline and DNA barcode coverage will increase. Currently, *trnL* is only available for ~5000 angiosperm genera while *rbcL* is available for >7000 genera. This means that additional *trnL* diet reads that currently have definite matches at the genus-level will become ambiguous while those that are already ambiguous will

remain so because denser taxon sampling does not resolve ambiguity. Metagenomics, however, will benefit from lower cost and be less affected by ambiguity because it uses the signal of multiple barcodes (Li *et al.,* 2014). As more barcode regions are used for species identification, more metagenomic reads will become informative. For example, I could have used ~10,000 and ~44,000 sequences if whole chloroplast genomes had been available for identification. Given that my metagenomic reads are largely uniformly distributed across the cp-genome (Fig. 5.2), going from <3000bp of barcode sequence used here to full cp genomes would result in an ~50x increase of the data available for identification. Fortunately, more authors argue for longer and more barcodes as reference (Meier *et al.* 2006; Nock *et al.* 2011; Chapter 3), so that this development is already underway.

Overall, metabarcoding remains particularly attractive when a diet item has to be picked from a small number of distantly related, potential choices with discrete *trnL* barcodes while metagenomics is currently particularly valuable for the following cases: first, for species with little prior information on biology because selecting the correct primers for PCRs is difficult. Second, for endangered species where few samples are available that should be studied exhaustively. Metagenomics simultaneously provides data on the host, its intestinal parasites, and associated microbes. For example, I here characterized the eukaryotic reads and additional work could have been done on the gut microbiome (Lamendella *et al.* 2011). Particularly, interesting was the recovery of the mitochondrial genome of the host and a gut nematode belonging to *Strongyloides* (probably *S. fuellerboni*) (Fig. 5.3). This nematode has been found in Asian and African non-human primates (Labes *et al.* 2011). I also recorded the presence of *Entamoeba sp*., and in particular *Entamoeba sp.* RL3, a lineage found only in Colobinae (Stensvold *et al.*

2011). Additionally, I found sequences similar to the common fecal parasite *Blastocystis sp.* (Alfellani *et al.* 2013). Lastly a very small number of sequences matched insects revealing the presence of fruit (Tephritidae) and vinegar flies (Drosophilidae), which are likely to be plant-associated ingestions. Overall, this suggests that metagenomic data generated from wild samples allows for studies where a wider and more holistic picture of the ecology is desired. It can potentially give novel insights into multiple aspects of biology of endangered species and help with understanding pathogens that may be of conservation relevance. Generated from captive samples, it provides important veterinary information. In all, moving towards a metagenomic analysis of fecal DNA promises to provide numerous new insights into species-interactions that will go well beyond diet characterization.

CHAPTER 6

Metagenomics outperforms metabarcoding and field observations for diet characterization and yields additional information on host genetics and parasite infestation of the banded leaf monkeys (*Presbytis femoralis*)

6.1 Abstract

In this study I document how metagenomic data from fecal samples obtained in a Southeast Asian rainforest can be used to infer simultaneously the diet, mitochondrial genetics, and parasite community of the critically endangered Singapore population of the banded leaf monkey *Presbytis femoralis*. I compare the results of metagenomics with observational data collected in the field and find that metagenomics gives deeper dietary profiles than observational studies, which are likely to overlook rare feeding events for elusive animals. Furthermore, I compare the performance of metagenomics and metabarcoding and find that metagenomics outperforms metabarcoding because more species are represented in the data and they can be identified to a lower taxonomic level (species/genus). Based on our previous study on red-shanked douc langurs (*Pygathrix nemaeus*), recovering fewer species in the metabarcoding data is surprising while the

better taxonomic precision of metagenomic reads is not unexpected. I again find that the number of reads in the metagenomic and metabarcoding correlate. In the current study, I also refine the analysis of metagenomic data in order to provide more accurate dietary profiles using exact alignments. Overall, I obtain very diverse dietary profiles for banded leaf monkeys. I identify diet species from 60 genera from six samples and identify the dominant 21 plant genera that are present in ≥3 samples. I discuss the implications of the results for conservation and management of banded leaf monkeys. I furthermore obtain full mitochondrial genomes and optimize the assembly pipeline of such genomes from metagenomic data. Overall, I find very low genetic variation across the mt-genomes of the putatively highly inbred banded leaf monkeys in Singapore. Particularly interesting is the finding of heteroplasmy in five of six genomes, which prompts further investigation on the prevalence of heteroplasmy in wild populations. Lastly, in addition to *Entamoeba, Blastocystis* and *Strongyloides* that are prevelant (in 5-6 of the samples), I detect the presence of *Oesophagostomum* and *Trichostrongylus* in one, thus revealing the need to study these populations in greater detail for parasite prevalence.

6.2 Introduction

Obtaining information on the ecology of endangered species is critical for establishing effective conservation measures. In order to gather this information, numerous conservation biologists spend years of field work that yield information on population size, feeding ecology, social systems, and other behavioural traits (Ang *et al.* 2010; Smith & Smith 2013; Sommer & Mendoza 1995). Information obtained in such field studies can be supplemented by obtaining data from non-invasive samples such as feces, hair, etc (da Silva *et al.* 2012). Given that fecal samples have genetic material from diet, host, parasites as well as the microbiomes (Kohn & Wayne 1997), they are a very useful resource to characterize multiple aspects of ecology of a species. In chapter 5, I discussed how the small contributions of diet, host, and parasite species to the overall metagenome can be characterized reliably and provide useful biological information. I established the methodological procedures for analysing metagenomic data to characterize the diet from fecal samples and compared it with an existing metabarcoding approach to diet analyses using the P6 loop of *trnL* (Taberlet *et al.* 2007; Valentini *et al.* 2009). The comparison revealed the advantages as well as the limitations of both approaches. I found that metagenomics could identify plants with greater resolution by using longer reference barcode sequences for multiple genes. However, the data lacked information on rare diet species. On the other hand metabarcoding could recover reads for a larger number of plant species, but they could often not be identified to genus. Lastly the abundances of plant barcode sequences in the two approaches were correlated.

Based on the previous study, I predict that a comprehensive diet characterization from a fecal sample would require at least one of the following: (a) higher throughput in metagenomics, (b) a combination of both metagenomics and metabarcoding that would

combine the advantages of the two approaches, or (c) a two-step characterization of samples using metabarcoding as done by De Barba *et al.* (2014). The latter design requires an initial metabarcoding experiment to characterize diet species sequences to family and a follow-up metabarcoding experiment with family-specific primers for amplifying a short fragment of *nrITS* that would then be sequenced for identification to species/genus. Arguably, this approach would only be effective for characterizing low-diversity diets, because a diverse diet would require the laborious task of designing several family-specific primers, carrying out numerous PCR experiments, and sequencing many products using NGS. I have therefore argued that metagenomics may be a better approach because it offers the opportunity to not only characterise the diet without numerous amplification experiments, but also provides a wealth of other biological information based on complex environmental samples for endangered species. In this chapter, I use two of the three recommendations outlined above. I use a metagenomic approach to the study of fecal DNA, but I use greater coverage depth to identify a larger number of rare species. At the same time, I combine metagenomic and metabarcoding data for the same samples in order to study the biology of an endangered population of a colobine primate, the banded leaf monkey (*Presbytis femoralis femoralis*).

The Singapore population of the banded leaf monkey is critically endangered (Lim *et al.* 2008), but there are reasons why one has to be concerned about the species itself. Currently there are three recognised subspecies of *P. femoralis*, of which *P. femoralis femoralis* is found in the southern Malay peninsula and Singapore (Fig.5.1) (Ang *et al.* 2012). The second subspecies, *Presbytis f. robinsoni* ranges from the northwest Malay Peninsula extending north to Thailand and Myanmar, while the third, *Presbytis f. percura* is only found in eastern Sumatra. These recognised subspecies show variation in amounts

of black fur/white pigmentation. However, we have data demonstrating that there is high genetic divergence (~10% based on *Cytb*) between this subspecies and *P. femoralis robinsoni,* the subspecies found on the Northern Malay Peninsula (Meyer *et al.* 2011). This suggests that the Southern subspecies is likely to be a separate species. The main reason why these results have not been published is that there is no genetic information for the third subspecies *P. femoralis percura*. The type location for *Presbytis femoralis femoralis* is Singapore and as mentioned, the Singapore population is particularly endangered (Lim *et al.* 2008; Wilson & Reeder 2005). The population size is very small and the current population size estimate for these primates is only ~40 individuals (Ang *et al.* 2010; Lim *et al.* 2008). Over last 200 years Singapore has lost over 95% of its vegetation cover and as a consequence nearly 1/3 of its original plants and animal species (Ang *et al.* 2012; Brook *et al.* 2003). The banded leaf monkey is barely surviving. They were widespread across the island in the 19th century, but eventually became restricted to two forest fragments (Bukit Timah Nature Reserve, BTNR and Central Catchment Nature Reserve, CCNR). Upon the construction of the Bukit Timah Expressway in 1983, gene flow between the two populations ceased, and in 1987 the population in BTNR became extinct (Yang & Lua 1988). Thus these primates are currently limited to a small fragment of CCNR in Singapore.

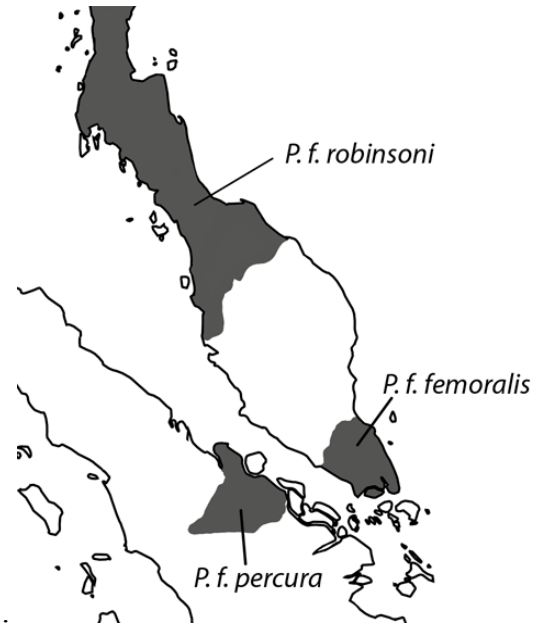


Fig. 6.1 : Distribution of the three currently recognized subspecies of *P. femoralis* (Ang *et al.* 2012)

Currently, little is known about the biology of *P. femoralis femoralis*. The main obstacle has been that these primates are difficult to study in field. They are very shy and elusive (Ang 2010) and the forests that they inhabit are dense, secondary and freshwater swamp forests that are difficult to traverse; thus observational study has been challenging (Ang 2010; Hüttche 1994). This is evident from the fact that a six-month study in the 1990s led to only 13 sightings (Hüttche 1994). More successful was a later, three-year study (2008-2011), during which 115 observations were made (Ang 2010). However, it has been particularly difficult to obtain meaningful ecological information. For example, the abovementioned study described only 31 feeding observations (Ang 2010) yielding an overall list of 27 plant species. Feeding behaviour was particularly difficult because direct observations are often obstructed by the canopy (Bennett 1983). In addition, it is difficult to obtain voucher material for the food trees given that much of the vegetation is out of reach, the observations are made from a distance, and tree and liana species diversity of the native habitat supports >700 species (Wong *et al.* 2013). This lack of information on the diet is unfortunate because a sound understanding of diet is important for an efficient

resource management of ecosystems that support endangered species (Cowlishaw & Dunbar 2000; Merlender *et al.* 1998). Moreover, these primates are found in a forest within an urban environment which creates serious concerns for viability of populations, as any alteration in habitat (e.g., loss of critical diet species) could lead to extinction of the species (Quéméré *et al.* 2013). Lastly, the urban setting also leaves potential for the transmission of parasites from and to humans.

One source for obtaining diet information of primates is fecal samples. The study of primate diets via feces has a nearly thirty-five year history. It started with morphological studies that were conducted on diet remnants in samples from baboons, vervets, Sykes' and colobus monkeys (Moreno-Black 1978). Later, DNA based approaches were utilized. Using PCR and cloning, Bradley *et al.* (2007) conducted diet analyses on chimpanzees. However, with the advent of Next Generation Sequencing (NGS), faster alternative methods are becoming popular. For example, Quéméré *et al.* (2013) used metabarcoding to investigate the dietary diversity and plasticity in the golden-crowned sifaka *Propithecus tattersalli*. Such NGS based studies consisted of two steps: (I) PCR based amplification of short fragments of DNA using generic primers, and (II) amplicon sequencing of these products using NGS technologies. However, such a PCR-based approach has limitations. Firstly, it depends on the availability of sufficiently general primers that can amplify the DNA of all potential dietary species. This requires *a priori* knowledge of the diet range and may interfere with genuinely new insights into the nutritional resources of a species (see Chapter 5). Secondly, there is a chance for an amplification bias that may skew the representation of the various taxa in the fecal samples. While for the red shanked douc langurs, I found the bias to be minimal, several authors expect these biases to be a significant problem (Hamad *et al.* 2014; Pompanon *et*

al. 2012). Lastly, the primer designed is required to amplify short fragments of the degraded DNA in fecal samples. However, these fragments often lack sufficient variability to classify several organisms to genus (Chapter 5). Typical examples for the short fragments used for the study of animal diets are ~100 bp of 12S (Shehzad *et al.* 2012) and for plant diets a 40-140 bp long piece of the P6 loop of *trnL* (Taberlet *et al.* 2007).

In the previous chapter I demonstrated that the issue of diet plant identification to genus can be largely resolved using a metagenomic approach. This approach has the additional advantage of being multi-dimensional and allows for the assembly of mitochondrial genomes and identification of intestinal parasites. An alternative method to obtain complete mitochondrial genome sequence from fecal DNA would involve PCR and sequencing; for example, Matsui *et al.* (2007) conducted PCR on 17 fragments of 300-2,000 bp lengths mt-DNA for *Propithecus verreauxi* (Verreax's sifakas) to characterize the mitochondrial genome (Matsui *et al.* 2007). Otherwise PCR-based approaches generally rely on single or a few gene fragments. For example, we previously sequenced the hypervariable region I of *d-loop* for our *P. femoralis* samples and found that the population was genetically impoverished (Ang *et al.* 2012). Using metagenomic data, population genetics studies can be based on entire mitochondrial genomes instead of being restricted to the short HV-I region of *d-loop*. In the future, the multi-dimensionality of metagenomic datasets can be used to look for other types of interactions. For example, parasitism can drive populations with low genetic variability to extinction (Whitehorn *et al.* 2011) and gut parasites have been shown to cause mortality (Chapman *et al.* 2005). Once more metagenomic data are available, the frequency of such correlation can be studied.

My first aim in this chapter is to address how the information obtained from NGS based analyses of environmental samples compares with traditional methods of studying feeding ecology using observational data. Secondly, I extend the comparison of metagenomics and metabarcoding from the study of captive Douc Langurs (see Chapter 5) to the much more complex diet analysis of samples collected in the wild. Thirdly, I here test whether a metagenomic approach is successful even if the available barcode database does not contain sequences for all potential food plants. Such cases are common because it is rare that barcode sequences are available for all potential diet species (Elliot & Davies 2014). Fourthly, I develop bioinformatic strategies that reduce misidentifications based on metagenomic data when only incomplete databases are available. Lastly, I characterize the biology of Singapore's banded leaf monkeys in terms of diet, host mitochondrial diversity, and parasites.

6.3 Materials and Methods

6.3.1 Fecal sample collection, DNA extraction and sample validation

Fecal samples used in the study were collected opportunistically by Andie Ang during her field studies on the Singapore population of the banded leaf monkeys. Groups of monkeys were observed followed and if defecation was observed, the sample was collected and brought back for storage at -70 °C. Note that samples were collected on different days and from places that were separated by man-made barriers (military infrastructure), thus increasing the likelihood that these were from different groups of monkeys (Ang *et al.* 2012). DNA was extracted as described in Chapter 5 using the QIAGEN DNeasy Blood and Tissue Kit according to the manufacturer's protocol with an extra wash step using Buffer AW2. Four independent extractions were carried out, and for each of these extractions, the interior of the feces was randomly sampled. The outside layer of the fecal sample was avoided in order to avoid contamination (Hamad *et al.* 2014).

In order to ensure that samples originated from *Presbytis femoralis*, a 12S fragment was amplified using primers L14724: CTGGGATTAGATACCCCACTAT and H15149: GAGGGTGACGGGCGGTGTGT (Ang 2010). PCR amplifications were carried out using Taqara ExTaq polymerase (Reaction mixture: 2.5 µL reaction buffer, 2 µL dNTPs, 1 µL$MgCl_2$, 1.2 µL of each primer and 0.15 µL of Takara ExTaq; reaction conditions: Initial denaturation of 95°C for 5 min followed by 35 cycles of 95°C for 1 min, annealing at 56°C, and extension 72°C for 1 min, final extension at 72°C for 5 min). The amplified products were purified using Bioline Sure-Clean solution (UK) using the manufacturer's protocol. Cycle sequencing reactions were carried out using BigDye

Terminator v3.1 and the sequences were analysed using the ABI3730xl DNA Analyzer. Sequences were edited using Sequencher 4.1 to obtain 12S fragments ranging between 300-400 bp. The sequences were validated as *Presbytis sp.* using BLAST against NCBI.

6.3.2 Next Generation Sequencing

DNA extractions from six validated banded leaf monkey samples (henceforth called BLM1-6) were sent for Next Generation Sequencing using Illumina HiSeq 2000 and MiSeq platforms. We used the same approach as in Chapter 5. For HiSeq sequencing, one library was constructed for each fecal sample (fragment size 280-300 bp). Two samples multiplexed in one lane of Illumina HiSeq 2000 and paired 76 bp reads were obtained using TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3. Additionally, we have datasets generated using the Illumina MiSeq platform for platform comparison purposes. I added this data to the analyses for diet and the mitochondrial genome. This dataset contained paired 300 bp reads. The libraries were prepared using the TruSeq Nano DNA sample preparation kit, with insert sizes of ~700 bp. Data were generated using one run of MiSeq per sample.

For the metabarcoding experiment, we used two sets of samples: the first set comprised four samples with the same extractions (BLM1, BLM3, BLM4, BLM6) that were used for metagenomics. The second set comprised different extractions from the same samples that were used for metagenomics (BLM2 and BLM5). For all six, the P6 loop of chloroplast *trnL* intron was amplified using primers *trnL*-g and *trnL*-h (Taberlet *et al.* 2007), that the latter were tagged using eight variable nucleotides at the 5' end of each primer that were designed using oligoTag (Coissac 2012) under the following criteria >= 5 variable sites; <3 bp homopolymers; the additional dinucleotide CC was added to 5'

end. The PCRs were carried out using BioReady rTaq DNA polymerase (Bulldog Bio, Inc., Portsmouth, NH) with reaction conditions and mixtures as described in Chapter 5. Three PCR replicates were obtained for each sample and the products were purified using a MinElute PCR Purification Kit (QIAGEN). Using Fragment Analyzer™ Automated CE System (Advanced Analytical), the products were quantified and pooled at equimolar ratios. Illumina MiSeq (Illumina Inc) was then used to obtain ~200,000-400,000 paired reads of 150bp; the libraries were prepared using the TruSeq Nano DNA sample preparation kit (150 PE).

6.3.3 Databases used in the study

Databases used in this study were previously described in Chapter 4. Briefly, the plant barcode databases consist of 73,892 sequences from 7,894 genera and 410 families for *matK*, 37,747 sequences from 5053 genera and 281 families for *trnL*, and 64,049 sequences from 7,539 genera, and 421 families for *rbcL*. This included 191, 248 and 211 barcodes for *matK, rbcL* and *trnL-F* respectively for species from Nee Soon Swamp (Chapter 4). For other eukaryotes and general characterization of metagenomes, COI (Chapter 4), SSU and LSU rDNA (SILVA, Pruesse *et al.* 2007) databases were used. A more targeted database of common non-human primate parasites was compiled based on a literature survey (Appendix 4). Sequences corresponding to SSU rDNA (18S) for these genera were downloaded from GenBank; this database comprised 5854 sequences from 26 genera (Appendix 4).

6.3.4 Data analysis

Prior to analyses, FASTQ files were trimmed using Trimmomatic v 0.32 (Bolger *et al.* 2014) under the following parameters: minimum average quality score=30, minimum length=50 after removal of all bases below average score of 20 at the start and end of sequences (LEADING=20, TRAILING=20). The obtained FASTQ files were converted to FASTA for further analyses.

Diet

For diet analyses using metagenomics, we followed the identification strategies developed in Chapter 5. For diet identification, MEGABLAST searches (word-size=28) for each end of the paired-end data were independently conducted against the three plant barcode databases. The results were filtered at 50 bp overlap and 98% identity threshold. The identification pipeline was then used to assign each read to different taxonomic levels (e.g., species, genus, family). Lastly, the results for the two ends were compared, conflicting matches were removed and the congruent genus-level identifications were reported (paired-end analyses). All genera identified using only single barcode were excluded (see Chapter 5: two gene criterion).

In the previous chapter I demonstrated that reliable plant identification requires matches to at least two barcoding genes. However, this criterion occasionally still yields misidentifications. Assembly-based approaches may be able to reduce misidentification although they would generally yield fewer identifications (see Chapter 5). Thus a further refinement of this methodology was tested. It was motivated based on the concern that BLAST creates local alignments that may lead to reads matching partially to reference sequence despite the read being in the interior of the reference sequence. A more precise

alignment could be obtained using global alignment algorithms, but they are computationally too expensive for mining entire metagenomes. Thus, I use a two-step approach, where the reads identified by BLAST were extracted and then aligned using the Needleman-Wunsch algorithm (Needleman & Wunsch 1970) to all reference sequences in the plant databases. This was done using *glsearch36* as implemented in the FASTA suite (Pearson 1990), which generates exact global alignments such that the alignment is global for the query and local for the reference sequences. Outputs were generated in the BLAST tabular format, thus I could apply the same identification pipeline described above for BLAST-based read identifications.

a. Comparison of metagenomic with field data

Observational data on the feeding ecology of the primates was obtained by Andie Ang during three years of field studies. 1,085 hours of field work were conducted in which 31 feeding observations were made. A feeding record was made whenever a monkey manually or orally handled a food item and brought it to the mouth (Ang 2010). The list of diet plant species based on direct observation was compiled and compared with identifications made based on metagenomic data (Table 6.1).

Table 6.1 List of known diet plants for *P. femoralis* in Singapore. Data obtained from A. Ang . * represent plants for which the corresponding genus does not have two or more barcodes, and thus cannot be identified using established criteria.

S.No.	Species	Material fed (leaves/fruits)	Number of observations
1	*Adinandra dumosa*	Flowers	1
2	*Agelaea macrophylla*	Fruits	1
3	*Artocarpus elasticus*	Fruits	1
4	*Bauhinia semibifida*	Leaves and flowers	1
5	*Erycibe tomentosa*	Leaves	1
6	*Fagraea fragrans*	Leaves	1
7	*Falcataria moluccana**	Leaves	1
8	*Fibraurea tinctoria*	Leaves and flowers	2
9	*Hevea brasiliensis*	Leaves	2
10	*Ixonanthes reticulata*	Fruits	1
11	*Knema malayana*	Fruits	1
12	*Litsea castanea*	Leaves	1
13	*Litsea elliptica*	Fruits	1
14	*Litsea firma*	Fruits	1
15	*Lophopetalum multinervium*	Fruits	1
16	*Madhuca sp.*	Fruits	1
17	*Nephelium lappaceum*	Fruits	1
18	*Nothaphoebe umbelliflora**	Leaves	1
19	*Palaquium xanthochymum*	Fruits	1
20	*Passiflora laurifolia*	Leaves	1
21	*Pellacalyx axillaris*	Fruits	1
22	*Prunus polystachya*	Fruits	2
23	*Pterocarpus indicus*	Leaves	1
24	*Syzygium grande*	Leaves	1
25	*Tetracera indica*	Fruits	1
26	*Xanthophyllum ellipticum*	Fruits	2
27	*Xanthophyllum eurhynchum*	Leaves	1

b. Comparison of metagenomic with metabarcoding data

In order to obtain a diet-estimate based on metabarcoding, paired-end reads were merged using *illuminapairedend*. Sequences were assigned to different samples using *ngsfilter* following which unique reads were obtained using *obiuniq*. All sequences <=10 bp were excluded using *obigrep*. Next, the sequences were tagged as “head”, “singleton” and “internal” using *obiclean* as in Chapter 5. Lastly I applied filtering criteria based on sequence counts (FC1, Chapter 5).

Next I compared the abundance of sequences corresponding to diet species between metagenomics and metabarcoding. In chapter 5, I used diet databases for a known diet. This was possible because more than 95% of the plant reads in the metagenomic data could be assigned to one diet plant species. Assigning reads to a species is much more challenging if the diet is unknown and not all diet species have been barcoded. In order to nevertheless compare the two approaches, I directly matched the sequences from the metabarcoding datasets to the metagenomic datasets. This was done by mapping the metabarcoding reads onto the metagenomic reads. In our case this is feasible because of the short length of the metabarcoding fragments (longest sequence retained after filtering was 64 bp long). In order to map the metabarcoding data I generated a fasta file containing the unique reads retained after the application of the FC1 criterion and variant calling. These sequences were mapped onto unassembled metagenomic data using BWA (Li & Durbin 2009) under criteria of perfect match criterion, allowing multiple mappings (up to 100,000) *(bwa aln –n 0 –k 0; bwa samse –r 100000)*.

This approach gives a direct read based correlation but has the following disadvantage: it can be used to correlate read counts only in cases where there is a metagenomic match for a metabarcoding read (i.e., abundance information is present for both metagenomics and metabarcoding). If a metabarcoding sequence does not map to the metagenomic dataset, i.e., if there are 0 reads in metagenomic datasets corresponding to the metabarcoding fragment, it could either be because a) it is not present in metagenomic dataset or b) a metabarcoding sequence is generated due to an artefact of PCR/sequencing

(Coissac *et al.*, 2012). In latter case, representing metagenomic data with the abundance information of "0" for the corresponding fragment would be misleading.

To account for the possibility that a valid metabarcoding read lacks a metagenomic match, I used a second strategy similar to Chapter 5, where I compared all the reads that have been identified to a plant family. At family-level most (94-96%) of metagenomic reads were identified. Moreover, this approach will overcome the above mentioned problem, as the variants are likely to be identified to the same family [under 95% identity threshold (Quéméré *et al.* 2013)] for metabarcoding. Since they are in much lower frequency than the original sequence, the variants will not modify the cumulative read count for the family.

Mitochondrial genomes

In order to assemble the mitochondrial genome of *Presbytis femoralis* I first compared 4 different assemblers (SOAPDENOVO2 (Luo *et al.* 2012), VELVET (Zerbino & Birney 2008), METAVELVET (Namiki *et al.* 2012) and IDBA-UD (Peng *et al.* 2012) to identify the best assembly algorithm for characterizing mitochondrial genomes using one sample (BLM6). Varying k-mer sizes (k=31, k=41 and k=51) were tested. This range of k-mer was selected based on our previous results for diet analyses for *Pygathrix nemaeus* and other studies focussing on fecal metagenomes (Rumen: Hess *et al.* 2011 and Giant Panda: Zhu *et al.* 2011). Additionally, SOAPDENOVO2 and IDBA-UD allow for a multi-k-mer approach; thus I also assembled datasets using multiple k-mers (ranging from k=31 to k=51 using these). Optimized assembly parameters were then used to construct the reference mitochondrial genome. Coding regions were annotated and validated.

Reads from HiSeq and MiSeq datasets for each sample were mapped back onto the reference genome using BWA (Li & Durbin 2009). I first identified variant sites by visual inspection of the mitochondrial genome. To validate the identified sites variant calling was done using GATK using UnifiedGenotyper with ploidy=1 (McKenna *et al.* 2010). Results obtained using HiSeq and MiSeq datasets from the same sample were cross checked for validation.

Parasites and other eukaryotes

In order to characterize which other Metazoa species were represented in the fecal samples, reads were matched against *COI* databases using settings identical to those in the diet analyses. Lastly, for other taxa, I matched sequences to SSU and LSU rDNA [SILVA, (Pruesse *et al.* 2007)] using BLASTN (word-size=11), and tested different percentage thresholds. While SILVA was useful in identifying microbial sequences, it lacked sufficient coverage to characterize eukaryotic gut parasites like helminths and nematodes. Therefore I generated a local database containing sequences from 18S for common, known non-human primate parasites (Appendix 5.4.1). For all identifications that were made using single barcodes (*COI*, SSU), I validated the match by extracting the reads and matching them to NT in GenBank that is the general nucleotide database (Chapter 5).

6.4 Results

6.4.1 Illumina sequencing

Illumina sequencing using HiSeq produced ~67 to ~108 million 76 bp reads per end for each sample (Table 6.2). Illumina MiSeq data comprised ~23 to ~29 million paired reads per sample; here the reads were of variable lengths with most reads having an average read length of 299 bp. Overall, this was equivalent to 10 to 16 Gb of data for each sample using HiSeq and 14to 17 Gb of data for each sample using MiSeq. After quality trimming at an average Phred score of 30, ~55-90 million reads per end for HiSeq and ~17-20 million paired reads for MiSeq were retained (Table 6.2). For metabarcoding 272,103 to 419,407 sequences per sample were generated that were subsequently filtered and subjected to variant calling and diet identification.

Table 6.2. Number of reads generated from each sample for Illumina HiSeq and Illumina MiSeq datasets and the metabarcoding experiment.

Sample	HiSeq (paired reads)		MiSeq (paired reads)		Metabarcoding
	Raw data	Post Q30	Raw data	Post Q30	
BLM1	107,675,433	90,201,101	28,715,570	16,869,060	338,131
BLM2	72,660,997	59,224,598	27,760,062	19,816,607	419,407
BLM3	85,963,340	72,349,546	26,595,637	19,029,890	371,919
BLM4	66,986,068	55,545,954	23,190,419	17,495,162	272,103
BLM5	68,188,666	55,310,058	27,840,572	17,516,063	294,907
BLM6	76,440,420	63,645,750	26,591,829	17,827,909	320,270

6.4.2 Diet analysis

The proportion of barcode reads used for paired-end analyses for HiSeq data was similar across the samples, and ranged between 0.004% - 0.008%; i.e., datasets contained a much larger proportion of diet reads compared to the study in Chapter 5 (Fig. 6.2, Chapter 5, Fig. 5.2). The lowest proportion was found for BLM3, which contained nearly five times as many barcode sequences as compared to the number of plant barcode sequences obtained for fecal samples of *Pygathrix nemaeus*. Thus with HiSeq data itself, these

samples had many more plant sequences for analyses. Adding MiSeq data gave additional reads; however the proportion of sequences in MiSeq data was variable for BLM5 and BLM6. Note that MiSeq data had been generated from different extractions for three samples so this appears to have affected the proportion of sequences. However given the preponderance of HiSeq sequences in the combined dataset, the proportion of sequences in each of the samples was similar and these datasets were used for further analyses.

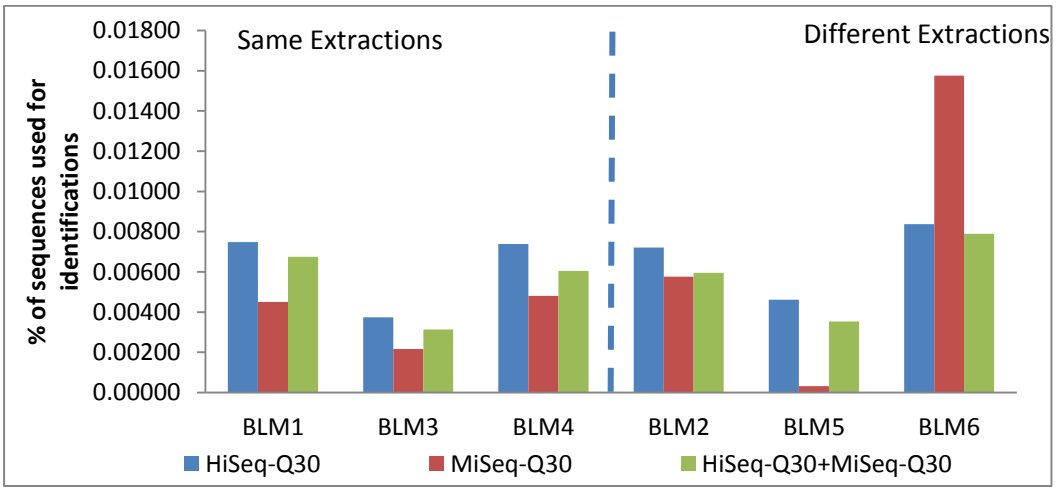
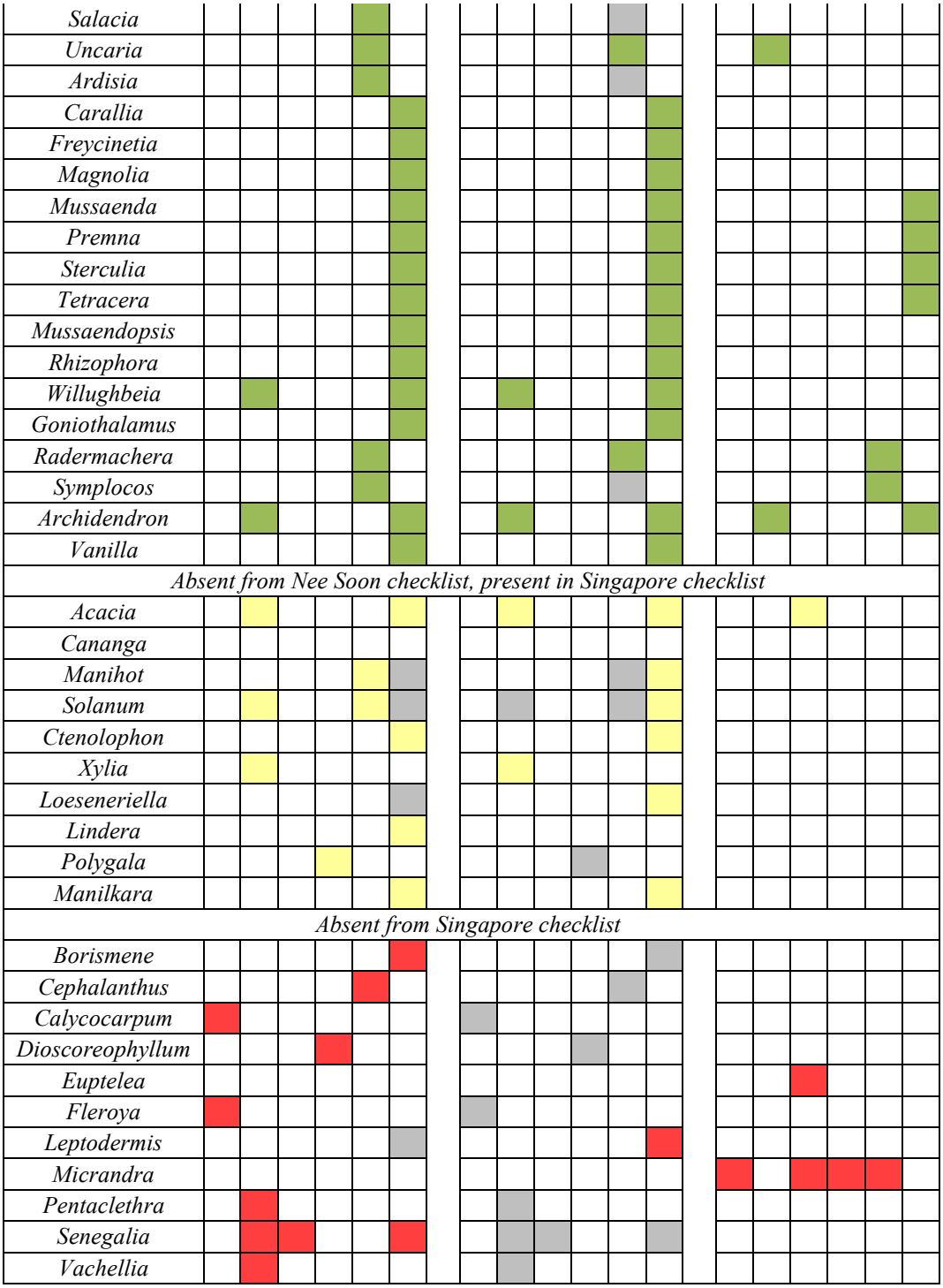


Fig. 6.2 Percentage of sequences used for paired-end analyses for plant identifications.

In terms of identifications, I found that number of genera identified by metagenomics was much larger than by metabarcoding (Table 6.3). Using metabarcoding 5-11 genera could be identified per sample, while using metagenomics 12–42 genera were identified. Another major factor influencing identification success rates was BLAST versus exact pairwise alignment. BLAST yielded a larger number of identified taxa (Table 6.3, green/yellow/red), but at the cost of eight, putative misidentifications given that the "identified" diet species are not known from Singapore's flora (Table 6.3, red). Only one such taxon is found when exact global alignments are used. However based on *glsearch* some plausible diet species are not identified so that a more conservative set of genera are identified [overall 146 "correct" identifications (Table 6.3, yellow and green) using *glsearch36* vs 157 using BLAST].

Table 6.3: Genus level identifications made using metagenomics and metabarcoding. MG: Metagenomics, MB: Metabarcoding. Green/ Yellow/ Red shaded cells represent identifications. Grey cells for metagenomics highlight the differences between BLAST-based and *glsearch*-based identifications (i.e. grey cells in MG: BLAST column represents identification made by *glsearch* only, and vice-versa). BLM1-6 represented as 1-6.

	MG: BLAST						MG: GLSEARCH						MB: FC1					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
	Present in Nee Soon checklist																	
Fibraurea	■	■	■	■	■	■	■	■	■	■	■	■						
Prunus	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Bauhinia		■	■	■		■		■	■	■		■		■		■		■
Ficus		■	■	■	■	■		■	■	■	■	■						■
Artocarpus	■		■		■	■	■		■		■	■						
Dalbergia	■	■			■	■	■	■			■	■	■	■				
Hevea	■		■	■	■		■		■	■	■		■					
Litsea	■			■	■	■	■			■	■	■						
Strychnos	■	■		■	■		■	■		■	■		■	■			■	
Xanthophyllum	■			■	■	■	■		■	■	■	■						
Knema		■		■	■	■		■		■	■	■				■	■	
Passiflora	■	■	■	■			■	■	■	■			■	■	■			
Cyathocalyx	■			■		■	■			■		■						
Securidaca	■	■		■		■	■	■		■		■						
Morinda	■				■	■	■				■	■						
Adenia		■	■			■		■	■			■						
Erythropalum		■		■		■		■		■		■		■		■		■
Psydrax				■	■	■				■	■	■						
Adinandra	■		■				■		■									
Pellacalyx	■		■			■	■		■			■						
Callerya		■			■			■			■							
Cassia		■	■					■	■									
Horsfieldia		■				■		■				■						■
Smilax		■				■		■				■						
Tinomiscium		■				■		■				■		■				
Tinospora	■			■	■	■					■	■						
Paederia					■	■						■						
Inga		■										■						
Erycibe		■			■	■		■			■	■			■		■	■
Aspidopterys	■						■											
Entada	■						■											
Pternandra	■						■						■					
Myristica		■						■										
Persea		■						■										
Hoya			■															
Pterocarpus			■						■									
Artabotys															■			
Macaranga										■								
Coscinium				■														
Agelaea				■	■													
Pertusadina					■													
Lophopetalum					■													
Ziziphus					■						■							
Dialium					■						■							


Salacia
Uncaria
Ardisia
Carallia
Freycinetia
Magnolia
Mussaenda
Premna
Sterculia
Tetracera
Mussaendopsis
Rhizophora
Willughbeia
Goniothalamus
Radermachera
Symplocos
Archidendron
Vanilla
Absent from Nee Soon checklist, present in Singapore checklist
Acacia
Cananga
Manihot
Solanum
Ctenolophon
Xylia
Loeseneriella
Lindera
Polygala
Manilkara
Absent from Singapore checklist
Borismene
Cephalanthus
Calycocarpum
Dioscoreophyllum
Euptelea
Fleroya
Leptodermis
Micrandra
Pentaclethra
Senegalia
Vachellia

Dietary profile for the banded leaf monkeys

Using a combination of metagenomic (*glsearch*) and metabarcoding techniques, I estimate the diversity of diet taxa to be 14 - 42 plant genera in the different samples (Fig. 6.3). Note this is conservative, firstly because I here use the exact alignments for metagenomics in combination with metabarcoding. Secondly, this approach is unlikely to include those diet species that do not have a reference barcode in the database. The smallest diversity of diet items was found for BLM3, and a surprisingly large diversity was observed for BLM6. Note that BLM3 also had the smallest proportion of reads recovered (Fig. 6.2).

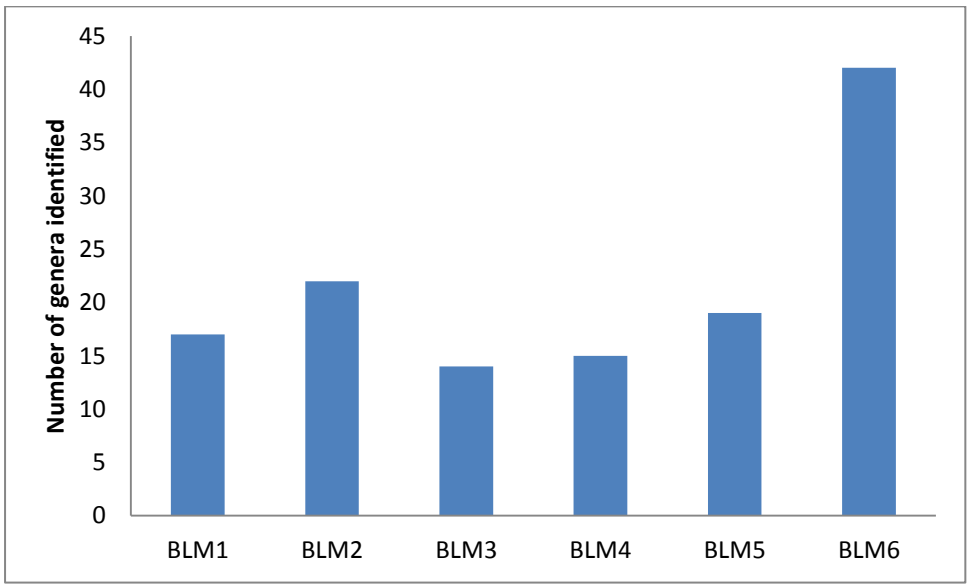


Figure 6.3 Number of genera identified per sample. Results are a combination of identifications made by *glsearch36* and metabarcoding.

Overall, sixty genus level identifications were made; 53 of these are from plants that have been recorded in the habitat of the primate, 7 of these are present in Singapore, even though they haven't been found in the habitat. These identifications were made using the exact alignments for metagenomics *(glsearch)* and metabarcoding, and I excluded identifications to plant genera not found in Singapore; these are likely to be erroneous. Of these, 19 plant genera were identified for three or more samples. Two

genera (*Fibraurea* and *Prunus*) were identified across all samples. *Xanthophyllum,* and *Ficus* were identified in five samples; while *Passiflora, Strychnos, Securidaca, Dalbergia, Hevea, Artocarpus, Litsea, Bauhinia* and *Knema* were identified in four samples (Fig. 6.4). Fourteen of the 53 diet plant genera identified based on fecal DNA were also observed to be diet genera by the field study of Ang (2010). Two others were identified only using the BLAST based approach. Eleven of these genera with molecular and observational evidence are amongst the 19 taxa identified for ≥3 fecal samples. Overall, I found that field observations led to a taxonomic profile that revealed mostly the dominant components of the primate's diets.

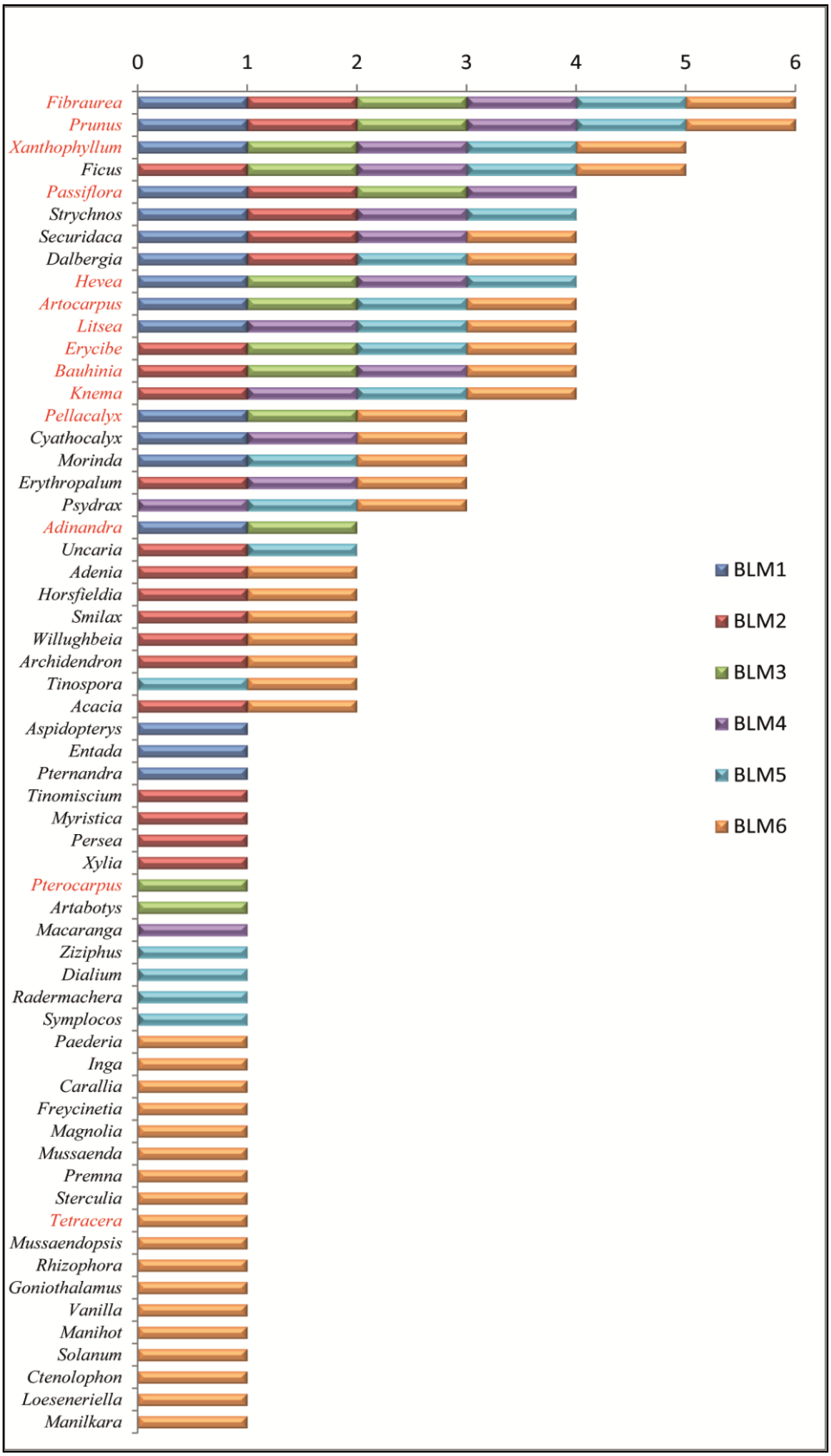


Fig. 6.4. Combined genus level identifications for the six samples using both HiSeq and MiSeq for metagenomics and metabarcoding data. Genera observed during field studies are highlighted in red. *Eupetalea, Leptodermis,* and *Micrandra* were excluded (likely misidentifications, Table 6.3)

Comparison with metabarcoding

a) Identifications

A number of metabarcoding sequences could only be identified to family, and hence this taxonomic category was chosen to compare the two approaches. This comparison is also fair because 94-96% of the metagenomic reads could be identified to family. Overall, the two approaches yielded congruent results with the majority of the identifications made by the two approaches being in agreement (Table 6.4). However, very few families were identified using only the metabarcoding approach (1-4 families per sample). Only for BLM3 did both approaches perform similarly while for all others metagenomics outperformed metabarcoding. Interestingly BLM3 also had the smallest proportion of plant reads, thus leading to a similar problem as that described in Chapter 5.

Table 6.4: Comparison of family level identifications of metagenomic and metabarcoding data. Green: Identified by both, Orange: identified using metagenomics only, Yellow: Identified using metabarcoding only. Values show number of barcodes identifying a particular family in metagenomics.

	Same extraction				Different extractions		Number of samples
	BLM1	**BLM3**	**BLM4**	**BLM6**	**BLM2**	**BLM5**	
Fabaceae	3	3	3	3	3	3	6
Menispermaceae	3	3	3	3	3	3	6
Moraceae	3	3	3	3	3	3	6
Rosaceae	3	3	3	3	3	3	6
Rubiaceae	3	2	3	3	3	3	6
Lauraceae	3		3	3	3	3	6
Apocynaceae	3	2	-	3	3	2	5
Euphorbiaceae	3	3	3	-		3	5
Annonaceae	3		3	3			5
Loganiaceae	3	-	2	2	3		5
Polygalaceae	2	-	2	2	2	2	5
Celastraceae	-	-		3	2	3	4
Convolvulaceae	-		-	3	2	3	4
Erythropalaceae	-	-	3	3	3	2	4
Myristicaceae	-	-	2	2	3	2	4
Passifloraceae	3	3	-	3	3	-	4
Connaraceae	-	-	3	3		3	3
Magnoliaceae	-	-	2	3	2	-	3
Rhizophoraceae	3	2	-	3	-	-	3
Sapotaceae	-	-	2	3	-	-	3
Smilacaceae	-	2	-	3	2	-	3
Asteraceae	-	-	-	2	-	2	2
Bignoniaceae	-	-	-	-	-	2	2
Malvaceae	-	-	-	3	-	2	2
Pentaphylacaceae	3	3	-	-	-	-	2
Primulaceae	-	-	-	-	2	3	2
Araceae	-	-	-	3	-	-	1
Berberidaceae	-	-	-	-	-	-	1
Cornaceae	-	-	-	-	-	2	1
Ctenolophonaceae	-	-	-	2	-	-	1
Dilleniaceae	-	-	-	3	-	-	1
Elaeagnaceae	-	-	-	-	-	2	1
Hamamelidaceae	-	-	-	-	-	2	1
Lamiaceae	-	-	-	3	-	-	1
Malpighiaceae	3	-	-	-	-	-	1
Melastomataceae	2	-	-	-	-	-	1
Pandanaceae	-	-	-	3	-	-	1
Phyllanthaceae	-	-	-	-			1
Rhamnaceae	-	-	-	-	-	3	1
Sapindaceae		-	-	-	-	-	1
Symplocaceae	-	-	-	-			1

b) Correlation between the number of metabarcoding and metagenomic reads for the same diet species

The correlation of read abundance in metagenomic and metabarcoding data was strong (rho >0.7, $p<0.05$), based on read-read mapping (Fig. 6.5). Read-read mapping also showed that the sequences corresponding to the metabarcoding fragments that were most abundant were almost always present in the metagenomic datsets. Of the ten most abundant sequences in the metabarcoding data, all were recovered in the metagenome for BLM1, BLM2, BLM6. 9/10 were recovered for BLM5 and 7/10 for BLM3 and BLM4; the latter two also had the fewest genera identified. Overall, I observed that, as expected, sequencing was deep enough in order to recover the P6 loop of *trnL* for dominant plants; i.e., the short ~50bp fragment was not present for rare taxa in the metagenomes. The strong correlation was further validated when the second approach was used, where I compared the number of reads identified to a given family using both metagenomics and metabarcoding. This approach is not based on mapping metabarcoding reads onto metagenomic sequences. Here for BLM4 and BLM6, ρ was within 0.5 to 0.6 (Table 6.5), while others showed strong correlation. For BLM4 and BLM6, the poorer correlation was likely a result of using a 95% threshold for metabarcoding, as a few dominant sequences could not be identified to family at this threshold. Lowering the identity threshold to 90% improved the correlation to $\rho = 0.728$ for BLM4 and $\rho = 0.775$ for BLM6. Note that there was no evidence that it mattered whether the DNA for metagenomics and metabarcoding was extracted once or at two different times.

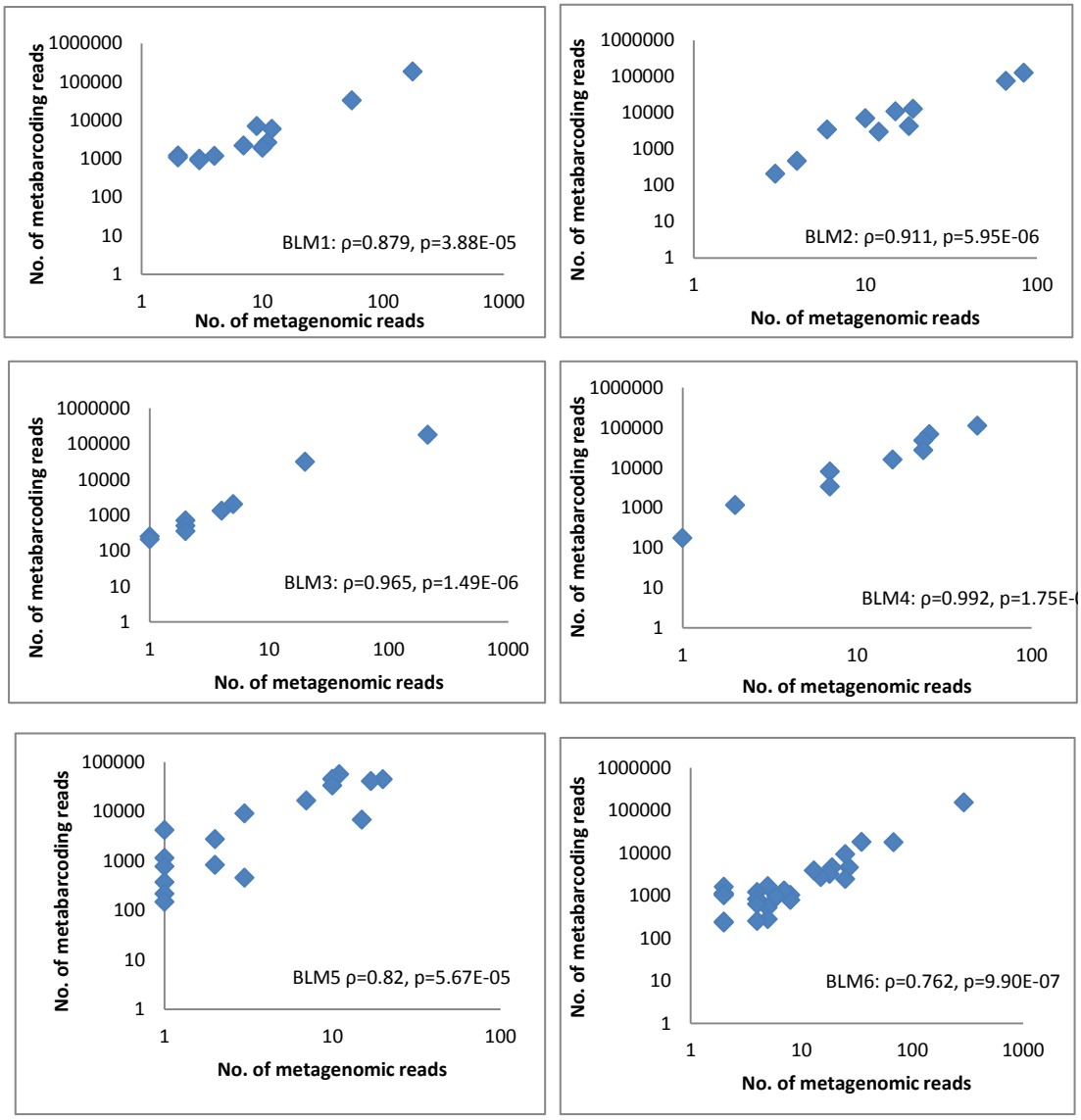


Figure 6.5. Scatterplot showing the number of metagenomic reads containing the P6 loop of *trnL*. The y-axis represents the read counts corresponding to the same sequence for metabarcoding.

Table 6.5 Spearman's ρ for correlation between number of reads corresponding to a family in the metagenomic and metabarcoding datasets.

	Identified using both approaches			**Identified using either approach**		
	ρ	p-value	Number of taxa	ρ	p-value	Number of taxa
BLM1	0.879121	1.90E-05	13	0.754911	8.33E-06	26
BLM2	0.916094	1.09E-05	13	0.746323	4.33E-05	23
BLM3	0.883333	0.003075	9	0.731707	0.000558	18
BLM4	0.811723	0.007889	9	0.52056	0.04665	15
BLM5	0.782372	0.001572	13	0.625952	0.000817	25
BLM6	0.871988	5.46E-07	20	0.765473	1.31E-06	29

6.4.3 Recovery of host mt-DNA

a) Assembly optimization

Four different algorithms were compared for one sample (BLM6) in order to determine the optimal approach for assembling the mitochodrial genome of the host based on the short-read in the HiSeq datasets (Table 6.6). Overall, I found that as the size of k-mer increases, N50 [defined as: "the length of the contig overlapping the midpoint of the length-order concatenation of contigs (Mäkinen *et al.,* 2012)] increases, but the number of scaffolds drops. Moreover, given that we were interested in mitochondrial genome assembly, the use of longer k-mers led to splitting of the mitochondrial contigs despite longer N50. This is likely due to splitting at low coverage regions (e.g. for SOAPdenovo2, number of mt-contigs was 6 (k=51), 6 (k=41) and 3(k=31)). A multi-kmer approach using SOAPDENOVO2 and IDBA-UD retained a maximal number of contigs while also maximizing the N50 and average scaffold length. Number of mitochondrial contigs obtained using SOAPDENOVO2 and IDBA-UD was 3 and 4, respectively. I chose IDBA-UD for assembling all datasets given that it yielded the complete mitochondrial genome without any gaps and also had the better N50 as compared to SOAPDENOVO2. The coding regions for the mitochondrial genome were validated by checking for stop codons in Artemis (Rutherford *et al.* 2000).

Table 6.6 Assembly statistics for BLM6 for the four software packages compared at k=31, k-41, k=51 and multi-k-mer approach with k varying between k=31 to 51.

	SOAP DENOVO 2	VELVET	META-VELVET	IDBA UD
		K31		
Number of scaffolds >100bp	1110149	1540719	1537360	NA
N50	649	519	520	
Longest scaffold	126390	14742	14742	
Mean length	351	347	348	
		K41		
Number of scaffolds >100bp	615381	506148	505363	NA
N50	1056	1185	1187	
Longest scaffolds	228740	70323	70323	
Mean length	487	552	555	
		K51		
Number of scaffolds >100bp	319068	198026	196001	NA
N50	1608	2035	2037	
Longest scaffolds	269962	101107	101107	
Mean length	620	763	765	
		K31-51		
Number of scaffolds >100bp	1193665	NA	NA	742840
N50	759			972
Longest scaffolds	172746			231523
Mean length	361			458

b) Low genetic variability and heteroplasmy in mitochondrial genomes

A complete reference mitochondrial genome of 16,548 bp was constructed for BLM6 using IDBA –UD as described. Reads for the other samples from MiSeq and HiSeq were mapped onto the reference mitochondrial genomes using BWA (Liu et al. 2012). The average coverage for the six samples was as follows (values are HiSeq/MiSeq), BLM1: 21.6/7.3, BLM2: 19.5/9.2, BLM3: 7.8/11.6, BLM4: 37.2/31.2, BLM5: 103.3/41.1, BLM6: 37.1/10.2. SNP calling using GATK led to identification of only three variable sites in the mitochondrial genomes. This was validated by manual inspection of the read mappings on to the reference genome constructed for BLM6. Two of the identified sites suggested the presence of biallelic heteroplasmy in the individuals. I tested whether this

heteroplasmy may be due to technical problems. This can happen (1) if errors were generated due to non-specific mapping in the *tRNA* sequences of the genome. However one of the sites was located in the hypervariable region or the *d-loop* while the other site was located in the CDS of ATP8. The heteroplasmy in *d-loop* had been previously observed for the same samples using PCR amplification and Sanger sequencing of the products thus further validating that these cases of heteroplasmy are not due to mapping errors. (2) The mapped sequences may not reflect heteroplasmy but are errors due to the mapping of NuMTs from the nuclear genomes. However, this is unlikely, given that the ratio of the polymorphisms is 50% in some samples (Table 6.7) and the nuclear genome would not be represented in similar abundance as the mitochondrial genome. (3) The polymorphisms could be observed due to contaminations during extraction and cross-lane contaminations in Illumina HiSeq. The latter was not the case given that MiSeq data yielded similar results. Furthermore, it is very unlikely that these results are due to genomic contaminations because three of the MiSeq runs (BLM2, BLM5 and BLM6) were generated from independent DNA extractions from the same fecal samples and they have the same polymorphisms.

Thus these heteroplasmic sites are likely to reflect the genetic make-up of the host monkeys. Based on the combination of polymorphisms 3 distinct "haplotypes" can be inferred: BLM2 and BLM4 are distinct from BLM1, 3, 5, 6. Furthermore the use of NGS allows us to quantify the level of heteroplasmy. Based on this, BLM6 showed a distinctly different profile from BLM1, 3, and 5 at both the positions (892 and 8673), and therefore I determine that there are at least four different genotypes represented in the six samples.

Table 6.7 SNPs identified and their position in the reference mitochondrial genome. (a) shows combined analyses of HiSeq and MiSeq data with potential heteroplasmic sites highlighted (b) provides the results by HiSeq and MiSeq separately.

(a)	By Depth			By Percentage		
Sample	Position 892 T/G	Position 8309 G/A	Position 8673 A/G	Position 892 T/G	Position 8309 G/A	Position 8673 A/G
BLM1	14/41	29/0	10/8	25/75	100/0	55/45
BLM2	0/29	0/35	0/29	0/100	0/100	0/100
BLM3	14/42	24/0	15/18	25/75	100/0	45/55
BLM4	8/65	51/0	2/79	11/89	100/0	2.5/97.5
BLM5	88/99	119/2	80/39	47/53	97.5/2.5	67.2/32.8
BLM6	30/7	17/0	24/3	81/19	100/0	89/11

(b)	HiSeq			MiSeq		
Sample	Position 892 T/G	Position 8309 G/A	Position 8673 A/G	Position 892 T/G	Position 8309 G/A	Position 8673 A/G
BLM1	6/32	21/0	8/6	8/9	8/0	2/2
BLM2	0/19	0/26	0/19	0/100	0/9	0/10
BLM3	2/19	3/0	4/9	4/14	13/0	9/7
BLM4	6/34	22/0	1/41	2/31	29/0	1/38
BLM5	57/69	78/2	57/24	31/30	41/0	23/15
BLM6	16/4	4/0	16/1	14/3	13/0	8/2

6.4.4 Parasites and others Metazoa in the fecal material

Reads and assembled contigs were first matched to SILVA SSU and LSU rDNA databases; however, no parasite sequences could be confidently detected, beside *Blastocystis* and *Entamoeba*. Therefore, I used a locally generated database of SSU rDNA for non-human primates and matched the reads at 98% identity, 50 bp overlap criteria. Once sequences were identified I matched these sequences to NT in GenBank and considered only those that were validated. The searches revealed presence of several protists and nematode sequences in these samples. Overall besides the common parasites, I found sequences for *Strongyloides sp., Oesophagostomum sp.* and *Trichostrongylus sp* in the database (Table 6.8). Most hits to *Strongyloides* were to *Strongyloides fuellerbonii,* or they were unidentified at species level. Using assembled data I found hits to at least four different species of *Entamoeba* in the various samples (corresponding to *E. bovis, E.moshkovskii, E. hartmanni,* and the colobine specific *Entamoeba sp. RL3*).

Table 6.8: Parasite sequences identified using paired end analyses and local non-human parasite database

Parasite	BLM1	BLM2	BLM3	BLM4	BLM5	BLM6
Blastocystis sp.	X	X	X	X	X	X
Entamoeba sp.	X	X	X	X	X	X
Strongyloides sp.*	X	X	X	X	X	X
Oesophagostomum sp.						X
Trichostrongylus sp.						X

Using a *COI* database I found that two samples (BLM3 and BLM6) had relatively high number of identifications to other eukaryotes. In both BLM3 and BLM6, I identified sequences from Drosophilidae, Muscidae, while identifications unique to each sample were BLM3: Sarcophagidae and BLM6: Sepsidae, Tortricidae. At genus level the closest hits were to*, Gatesclarkeana* (Tortricidae, BLM6), *Dicranosepsis* (Sepsidae, BLM6), *Leucophenga* and *Stegana* (Drosophilidae, BLM6). A few of sequences gave hits to

Ophyra/Coenosia/Sacrophaga (Muscidae and Sacrophagidae) however, these became ambiguous when the retrieved reads were BLASTed to all of nucleotide (NT) database.

6.5 Discussion

Studying diet is fundamental to understanding the ecology of a species. For endangered species, beyond ecology, characterization of diet is essential for designing effective conservation strategies. This is because the availability of food resources can play an important role in determining the geographical distribution of a species and its population density (Marshall 2009). Thus methods of characterizing diet have been of considerable interest for researchers studying these species (Marshall 2009, Ang 2010). Fecal samples have been a useful resource for obtaining this information as the DNA of the ingested food can be characterized. In chapter 5, I proposed a metagenomic approach to diet analysis using fecal DNA and established procedures for analysing these samples. In this chapter I consolidate this approach and assess the method for characterizing the diet of a wild population of an endangered primate in Singapore, *Presbytis femoralis.*

6.5.1 Evaluating NGS based diet analyses against "traditional" field studies

Traditionally, the diet of endangered mammals has been studied using field observations. The morphological and/or molecular characterization of fecal material was later added to the repertoire, but the evaluation of the DNA content is still rudimentary. In the red-shanked douc langur study (Chapter 5), I determined that plant diet taxa can be retrieved and identified using metagenomics, but this study was based on captive zoo animals with a known diet; i.e., the question whether similar analyses can be carried out in a more realistic setting remained un-answered. In the current study, I generated a diet profile consisting of 60 dietary plant genera for banded leaf monkeys (*Presbytis femoralis*) using NGS based techniques. Overall, I find good agreement between observational and DNA sequence evidence. Nearly half of the species obtained from

observational studies (11/24 genera) are also identified in ≥3 fecal samples. During any observational study of the diet of a species, researchers are more likely to observe feeding events involving important diet species. Therefore good overlap between observational and metagenomic evidence is expected and was here observed. In terms of frequency of feeding, Ang (2010) observed the monkeys feed on three genera (*Fibraurea, Hevea, Prunus*) twice and one other (*Xanthophyllum*) thrice. In metagenomic analyses of fecal samples, I found DNA from *Fibraurea* and *Prunus* in all six samples, *Xanthophyllum* in five and *Hevea* in four samples. Beyond the dominant taxa in three or more samples, two (*Adinandra* and *Pterocarpus*) were identified using the conservative exact alignment approach, while two others (*Lophopetalum* and *Agelaea)* were identified using BLAST. This suggests that overall we see an overlap of 16 genera between observational data and NGS based diet analyses. Amongst the remaining 8 that were observed, two cannot be identified due to lack of ≥2 barcode references in the database. Thus only six species/genera in the observational data remain unaccounted for. Clearly, the results from DNA based inference are reflecting what is observed in the field.

Upon combining results for metagenomic and metabarcoding data, I found a diet of 14-42 taxa per sample; most of these were identified using metagenomics. This richness is similar in diversity or greater than the diversityfound for another primate with a rich diet [golden sifakas, Quemere *et al.* (2013)]. In the latter study the average OTU richness was found to be 13.0±3.8. The high diversity in these fecal samples is partly due to the physiology of colobine guts which retains food for a long time (Lambert 1998); thus a fecal sample provides information of multiple days of diet. My analysis of only 6 samples adds 46 plant taxa to the observational data that required ~36 months of field work. When combined with the ten genera of plants identified using observational

techniques only, I obtained an overall diet list of 70 plant genera. At family level, the combination of NGS based studies and field techniques yield 44 families, 41 of which are found in NGS datasets. These results suggest that the banded leaf monkeys in Singapore utilize a very broad array of food plants. Obtaining such an extensive list through observation would be very expensive because it would require years of fieldwork given that the species is rare, shy and elusive. Given that metagenomics can extract information for a wide variety of diet items and provide taxonomic resolution, I would argue that it is the preferred method for characterizing the banded leaf monkey diet. Metagenomics has the additional advantage that it can identify lianas that are diet elements, while feeding observations on the latter are hard to obtain in the field (Ang 2010). Thus we identified climbers such as *Erythropalum scandens* (in 3 samples) which are relatively rare in Nee Soon Swamp forest (see section 6.5.5). Using these analyses, I can also determine that the likelihood of these primates feeding on animals is low; most of the *COI* based identifications made in the study are likely to be insects that are associated with either plants (Tortricidae) or feces (Drosophilidae, Sepsidae, Muscidae, Sarcophagidae). However, observational data still has the advantage that the observer can identify the plant parts that are being eaten and that species-level identifications can be obtained if voucher material can be retrieved from the food plants.

6.5.2 Comparison between metagenomics and metabarcoding for samples from wild

The high taxon diversity in the fecal samples for banded leaf monkeys provides an opportunity to compare the metagenomic and metabarcoding approaches at a broad taxonomic range. In terms of identifications, our results parallel the results for the diet analyses of douc langurs where metagenomics outperformed metabarcoding in providing greater taxonomic resolution. Only 5 to 11 genus level identifications were made using metabarcoding data in contrast to 12 to 42 using metagenomics. Note that most of these identifications are likely to be correct given that the identified genus is known to occur in Nee Soon Swamp Forest, the natural habitat of banded leaf monkeys in Singapore despite the search being conducted against all angiosperm barcode sequences

One major difference between my banded leaf monkey and the douc langur studies is that in most cases (5 of 6 samples) the metagenomic data yielded larger numbers of family level identifications. This is likely to be due to the fact that number and proportion of reads corresponding to the barcode regions for the six samples was ~5-fold greater than in the douc study in Chapter 5. In the latter study, I had concluded that the metagenome coverage was not sufficient for a complete diet profile. It appears that this problem is largely addressed by the much higher coverage in the banded leaf monkey samples. The reason for this larger coverage is unclear. There are multiple possible factors; firstly the biology of the two organisms may differ despite both being phytophagous colobine species. Colobines have evolved a complex digestive system presumably to retain food long enough to digest enzyme resistant polysaccharides (Kirkpatrick *et al.,* 2001), but there are differences even among colobine species. For examples the Transit Time for food for *Rhinopithecus bieti* (Kirkpatrick *et al.* 2001) and *Pygathrix nemaeus* (Chapter 5) are estimated to be 27-29 hours, while for *Trachypithecus*

cristatus and *Nasalis larvatus* it was found to be between 14-18 hours (Dierenfeld *et al.* 1992; Sakeguchi *et al.* 1991). These differences do not follow phylogenetic patterns (Wang *et al.* 2012), so that it is difficult to predict retention times and gut physiology of banded leaf monkeys. Secondly there may be differences in the types of materials ingested; for example animals in Singapore Zoo may have better access to younger and tender plant parts. Lastly the genomic smears for the extractions differed in the two studies, with the red-shanked douc langur samples having a larger proportion of long DNA strands while the field samples had more degraded profiles (Fig. 6.6). Recently, Cordona *et al.* (2012) showed that the degradation of fecal samples and storage conditions play an important role in affecting the taxonomic distributions of microbes in the feces. This is likely to also hold for the proportion of cp-DNA. Note that the distribution of size fragments was similar across the samples for *P. femoralis* (Fig. 5.2)

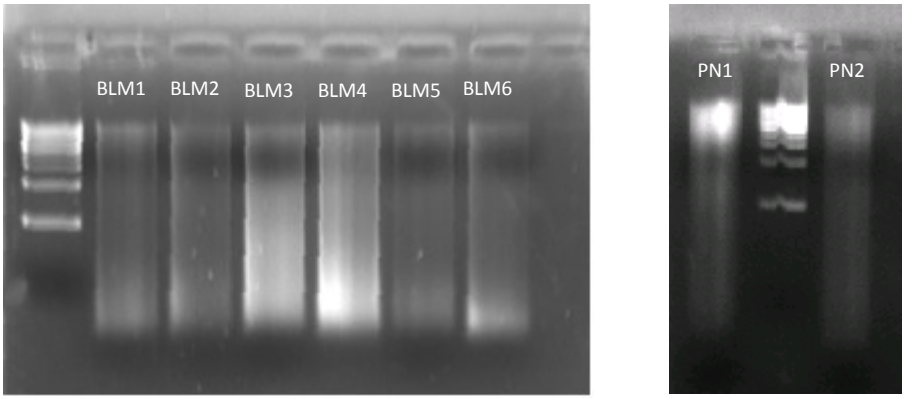


Fig. 6.6 Genomic smears of samples used in this study and Chapter 5.

6.5.3 Read counts are correlated between metagenomic and metabarcoding data.

One of the surprising results is that the read counts in the metagenomic and metabarcoding datasets are significantly correlated (Fig. 5.4). Such a correlation was already observed in my analysis of the red-shanked douc langur data. It suggested that the PCR step in metabarcoding did not bias the sequence profile greatly. This is heartening given that there has been considerable discussion of whether PCR biases abundance information (Pompanon *et al.* 2012). We find little evidence for this. Note that it has been recorded that PCR success decreases as the length of the amplification product increases (Little 2014) and therefore the lack of bias should not be generalised to other primer pairs.

Overall my results imply that if a large number of plant reads are recovered, metagenomics will outperform metabarcoding both in terms of taxon recovery and precision of identification. This increased recovery need not necessarily require greater throughput and is likely to depend either on species, the nature of the diet, and/or the nature of the samples. Indeed, the additional ~20 million reads obtained using the MiSeq platforms added at most 1-2 genus level identifications to the HiSeq based identifications which implies that higher coverage is unlikely to radically change the diet profile for the samples. Given that our results are based on colobine primates with long digestion times, an initial throughput of ~10 Gb appears to be a good starting point for the diet analysis of a phytophagous monkey with similar dietary complexity and food retention time as these monkeys.

6.5.4 Refining the metagenomic approach

In order to advance the utility of metagenomics for fecal samples, it is important to find analysis methods that are capable to identify many diet taxa while being robust against misidentifications. This is particularly difficult when the identifications have to utilize a DNA barcode database that lacks many of the putative diet taxa. Unfortunately, it is difficult to obtain a comprehensive barcode database for tropical rainforests (Elliot & Davies 2014). I was able to use and contribute to a database consisting of three barcode genes sequenced for ~250 species of tree and liana species, but the species estimate for the habitat of the monkeys is ~700 species (Wong *et al.* 2013).

In order to avoid misidentification, I tested the use of exact global alignments for read-based species identification (Ray *et al.* 2012). While the issues with BLAST-based taxonomic classification have been discussed in the past (Little 2011; Little & Stevenson 2007), it is still often preferred because it is faster and requires less computational power (Loh *et al.* 2012). The concerns behind using BLAST are two-fold: BLAST is heuristic such that it does not exhaustively search the sequence-space and is likely to miss sequences that are distant from the query sequences (Sharma and Mantri 2014). This is because it requires a perfect match to the seed sequence which is determined by the word-size in the BLAST search. Two of the settings commonly used are BLASTN and MEGABLAST. In the case of the first, the word-size is by default 11 and in the case of latter 28. These settings influence how distant sequences are retrieved. For diet analyses, I used the MEGABLAST settings given that I was identifying at 98% identity, i.e., a mismatch of 1-bp between the query sequence and the database sequence. This high identity threshold is necessary given the low variation amongst plant barcode sequences.

Given a 50-bp overlap and 98% identity threshold, the choice of the larger word size of 28 is thus not a problem. Indeed, when I tested varying word-sizes on the data from red-shanked douc langurs, the results varied by only 2 sequences in overall recovery when BLASTN or MEGABLAST were tried. MEGABLAST search on the other hand was much faster.

A second concern is that BLAST is a local alignment tool, such that it may terminate the alignment of sequences prematurely (Altschul et al. 1990, Heyn et al., 2010). This generates uncertainty in the results especially if they pertain to reads matched to plant barcodes with their low diagnostic values. Higher precision can be obtained with global alignments (pairwise or multiple sequence alignments). Optimizing multiple sequence alignment on raw data of 76 bp is computationally difficult. Thus, I utilized the Needleman Wunsch algorithm for pairwise exact alignments on sequences retrieved from blast based searches of the data (Pearson 1990). When applied to my data, the global-local algorithm in *glsearch36* reduced identification inaccuracies; however, it also reduced the number of identifications that are likely to be correct given that the plant genera occur in the habitat of banded leaf monkeys (for example *Lophopetallum* and *Agelaea* that were also in observational data). Some of these "lost" identifications are probably due to the requirement of full length overlap of sequences in global alignments instead of 50 bp in BLAST searches. Global alignments can therefore refine results, but these gains come at a cost. Nonetheless the approach should be used; especially when moving from genus-level to species-level identification.

Assembling the metagenome:

The assembly of metagenomes is difficult given that the data are derived from a complex mixture of DNA belonging to a number of organisms at varying amounts. There has been considerable discussion on the optimal parameters that should be used for characterizing the microbiome using metagenomic data. Using both single k-mer strategy and multi-k-mer strategies I compared the assembly parameters across four different softwares and optimized it for my HiSeq data. I chose the k-mer range of 31-51 for assessing the assemblies given a sequence length of 76 bp. This also corresponded to the range of k-mers in earlier studies of fecal metagenomes from rumen gut (Hess *et al.* 2011) or giant panda feces (Zhu *et al.* 2011). Overall, as the k-mer size increased the N50 of assembly improved; however it came at the cost of fewer contigs. As observed in the case of the red-shanked douc langurs, this is not necessarily an improvement given that it can lead to loss of the rarer fraction of metagenomes such as plant chloroplast fragments. Even for mitochondrial genomes, longer k-mers led to fragmentation of contigs (see Results). Therefore, to optimize the metagenomic assembly I used a combination of N50 and number of contigs. Overall, the multi-kmer approach yielded best results, with both SOAPDENOVO2 and IDBA-UD outperforming the other platforms. I used IDBA-UD to characterize the complete mitochondrial genome of *P. femoralis* because it yielded the larger N50 and yielded no gaps in the mitochondrial genome.

6.5.5 Implications on the biology and conservation of the banded-leaf monkeys

Diet

In this study I built a dietary profile for the *Presbytis femoralis* population in Singapore. Here I find that the diet is highly diverse, which corroborates earlier results by Ang (2010) based on few observations. The most common diet plants which are found in at least four of the samples are: *Fibraurea, Prunus, Ficus, Artocarpus, Dalbergia, Hevea, Litsea, Strychnos, Xanthophyllum, Bauhinia* and *Knema.* Of these 11, eight have been also been observed to be consumed (Ang 2010). Currently, we do not have precise information available for the abundance of plants in the forest; however, upon discussions with an expert botanist regularly working in the native habitat of the monkeys and referring to the checklist of common plants found in these forests (Tan *et al.* 2013), we find that a number of these plants are also common in Nee Soon suggesting that these primates are feeding on plants that are abundant in the forest (*Fibraurea, Prunus, Ficus, Artocarpus, Litsea, Strychnos, Xanthophyllum, Bauhinia, Knema).* This pattern is similar to what has been described for *Presbytis melalophos* (Davies et al. 1988), which was found to feed on the abundant taxa found in the forest. However, despite the consumption of common plants there is also some preference for certain dietary taxa such as *Erythropalum* and *Securidaca* that are not commonly present in the forest, suggesting that there are feeding preferences that should be considered in conservation programs. Overall, it is, however, difficult to infer whether this is because these primates are generalists when it comes to feeding ecology or whether they have adapted to the local flora for their diet. Nonetheless, this is encouraging for the primates as it implies that food resources are unlikely to be limiting factor for their survival. Yet the presence of rarer

plants like *Erythropalum* and *Securidaca* in three and four of the samples, respectively, suggests certain preferences for food plants that may be limiting for population growth.

Knowledge of such diet information is particularly relevant in the light of an ongoing project in Singapore that aims to reconnect forest fragments. Until 1987, Singapore's banded leaf monkey population inhabited two fragments of forest (BTNR and CCNR), which became separated after the construction of an expressway. The population in BTNR went extinct in 1987. Recently there has been an endeavour by Singapore's National Parks Board to reconnect the two forest fragments using an Eco-Link, "an ecological bridge that connects the two nature reserves". Currently trees are being planted on this link. In order to rehabilitate the banded leaf monkeys to BTNR, the preferred food plants should be planted to facilitate and encouragement the movement of banded leaf monkeys into BTNR. Particular focus should be given to the rarer plants as mentioned above. With ongoing efforts of vegetation sampling in Nee Soon Swamp forest, we will soon be obtaining abundance information of plants in the forest, such that these decisions can be made and implemented.

Mitochondrial DNA: low variability and heteroplasmy

I find alarmingly low variability across the entire mitochondrial genomes of Singapore's banded leaf monkeys. Previously, we had sequenced the *d-loop* sequences based on the fecal samples and found one variable site across the six samples. With six complete mitochondrial genomes now being assembled, I find only two additional variable sites. It is difficult to make any genetic inference on this variability given a limited number of samples. However, given the population size of ~40 individuals, this result is a reason for concern given that the primates sampled in my study were likely to be from different groups (Ang *et al.* 2012). Based on the genomic composition of the

mitochondrial genomes, we can determine that there are at least four different genotypes represented in these six samples. As documented by Ang *et al.* (2010), the banded leaf monkey population is expanding again, but the lack of genetic variability will make the population vulnerable against disease and parasites. Earlier, we pointed out that the low genetic variability is probably due to human disturbance over the last two hundred years where the monkey population went through a bottleneck, and is now only slowly recovering (Ang *et al.* 2010). Yet, the small population size and low genetic variability mean that the translocation of individuals from the southern Malaysian population should be discussed.

In addition to low variability, I found a number of gut parasite sequences. Most of them were from common parasites (*Blastocystis sp., Entamoeba sp*). *Strongyloides sp.* (likely to be *Strongyloides fuellerbonii*) sequences were found in all six samples. This is a common parasite in non-human primates and can cause strongyloidiasis, fatal cases of which have been reported in chimpanzees, gibbons, woolly monkeys etc. (Bennett *et al.,* 1998). It is also known to infect humans (King & Mascie-Taylor, 2004). BLM6 deviated from the other samples as it contained sequences from *Oesophagostomum* and *Trichostrongylus.* The sequences from *Oesophagostomum* matched to multiple species, which were *O. aculeatum, O. stephanostomum,* and *O. venulosum. O. aculeatum* has been reported in southeast Asia (Malaysia) (Arizono *et al.* 2012). To my knowledge this is the first indication of these two groups of nematodes in a non-human primate population in Singapore. The presence of these parasite sequences in the population calls for a closer monitoring of these primates and reveals the threat of potential infections. This is of critical importance for both conservation of the primates as well as assessing risks to human health (Chapman *et al.,* 2006). In future, more targeted characterization of

parasites can be carried out to determine the pathogenicity of these parasites in the population.

Particularly interesting are the heteroplasmic sites in the mitochondrial genomes in 5/6 samples from these primates. Currently there is very little information available about the extent of prevalence of heteroplasmy across non-human animals, although several sporadic reports exist (e.g. Volmer *et al.* 2011; Shigenbou *et al.* 2005). Heteroplasmy refers to the presence of at least two different mitochondrial genomes within an individual. A common occurence in aged individuals is somatic heteroplasmy where mutations accumulate over time in certain types of cells. However, the pattern of heteroplasmic sites in banded leaf monkeys suggests that these are inherited as the same polymorphisms are present in sequences that represent different individuals. Such heritable heteroplasmy could be due to either mutation in germline tissues or leakage of paternal mitochondria ("paternal leakage") during fertilization of the egg (Kvist *et al.*, 2003). In terms of its biological significance, while heteroplasmy has been discussed in relation with human disease, it is currently unknown whether it is associated with other factors such as biology of a species or population size and thereby, inbreeding. Presumably this lack of information is due to the fact that for years polymorphisms in mt-DNA sequences were masked by Sanger sequences where "double peaks" were often represented by ambiguity codes. With the advent of NGS based analyses, these aspects of genomes can now be studied in greater detail.

6.5.6 Future directions

In this chapter I show the promise of a metagenomic approach to obtain an understanding of the biology of an endangered species in terms of diet, host genetics and

parasites. I have tested this on a phytophagous colobine primate in Chapter 5 and 6 using 8 samples and 14 datasets (8 HiSeq, 6 MiSeq). The next step in optimizing this approach would be to test the metagenomic approach on other mammals and beyond mammals to birds, insects etc on larger sample sizes, which was a limitation in the case of the banded leaf monkeys. Furthermore it remains to be seen how effective the metagenomic approach would be for different types of feeding strategies; i.e., carnivory, insectivory and omnivory. In order to effectively test this approach, the recommendation would be to test it on organisms where there is observational data on feeding and where there is opportunity to compare with metabarcoding. This will help us understand whether this method is robust across different dietary types and different lifestyles. It will also help us understand the question of throughput required for different organisms so that recommendations can be made for studying them in larger numbers.

The second key question that comes out of the current study is the question of diversity. I evaluate metagenomic data by identifying reads against barcode databases. Generally, I then used lists of identified reads as measures of diversity. However, such lists are underestimates because some metagenomic reads remain unidentified but represent additional species. Thus obtaining measures of diversity without using reference sequences would be desirable (Quemere *et al.* 2013). Currently this is being done in two ways; first, several studies use the concept of Molecular Operational Taxonomic Units (or MOTUs) where DNA sequences are clustered and the number of sequence clusters is used as a measure for diversity (Ratnasingham & Hebert 2013). Others have employed a tree based approach to finding potential taxonomic groups (Pons *et al.* 2006; Zhang *et al.* 2013). However, applying these approaches to small fractions of metagenomes requires a number of considerations. For microbiomes, researchers have developed methods to use

rRNAs from shotgun metagenomic data, to either cluster or reconstruct trees based on rRNA sequences (Sangwan *et al.* 2012). However, read based clustering can only be achieved when sequences are long (Mande *et al.* 2012); which is unlikely to be the case when the focus is on degraded fractions of the samples (Valentini *et al.* 2009). While sequences can be assembled, these will not depict the diet diversity, given that low frequency reads will not assemble (Chapter 5, Thomas *et al.* 2012). Even if short reads are aligned, lack of homology across sites could lead to an overestimation of species because the same species may be represented by several clusters. Other signature based tools exist but it is currently difficult to determine how these would perform for assessing the species/genus diversity of plant sequences. Finding appropriate methods for estimating diversity without DNA barcodes is thus one of the frontiers although from a conservation point of view, identifications will remain important.

6.6 Conclusions

The proposal of sequencing ~10 Gb of data per sample to infer diet of a species cannot be defended without a discussion of the cost effectiveness of this approach. Recently, 1 Gb of sequence data costs 40 USD (Zhou *et al.* 2013) and thus ~10 Gb of data amounts to nearly 500 USD after inclusion of cost for library preparation. This seems large compared to metabarcoding where in a recent study 50 million paired sequences corresponding to P6 loop of *trnL* were generated for 91 samples; this amount of data corresponds to the cost of sequencing one sample. But this excludes manpower cost that comes prior to sequencing. In a recent study optimizing accuracy of metabarcoding, De Barba *et al.* (2013) recommend that multiple replicates (4, as per De Barba *et al.* (2013)) per sample with different barcodes ought to be sequenced for data quality purposes. Such procedures are very labour intensive. Moreover, the use of minibarcodes (Little 2014; Taberlet *et al.* 2007) limits taxonomic resolution such that additional steps (such as family specific primer design for nrITS, PCR optimization and second round of sequencing) are required before taxonomic resolution is achieved. Given that bioinformatic procedures, although intensive, can largely be automated, I would argue that metagenomics is cost-effective because it saves the manpower cost. Besides, even though the molecular cost per sample is higher for metagenomics, I consider it the preferred choice if taxonomic resolution is desired. This is particularly desirable for plants where barcodes are fraught with ambiguity problems (Hollingsworth 2011). It is also useful when diet is not the only focus and where any PCR based approach may require several rounds of optimization for every dimension, each of which would require deep sequencing.

CHAPTER 7

A foray into the future of environmental forensics

This thesis started at a time when species identification via molecular markers largely consisted of generating DNA barcoding databases for taxa or habitats (see datasets for Chapter 2). Diet analyses using NGS had just started via metabarcoding (2009: Valentini *et al.* and Deagle *et al.*). Four years later, run-of-the-mill DNA barcoding studies still dominate the literature, but more publications are starting to appear that make use of the better taxon coverage of DNA barcodes in metabarcoding studies (Baamrane *et al.* 2012; De Barba *et al.* 2014; Hilbert *et al.,* 2013; Quéméré *et al.,* 2013; Shehzad *et al.* 2012; Soininen *et al.* 2013) . Beyond diet, metabarcoding is now used for characterizing various types of environmental samples including arthropod "soups" (Ji *et al.* 2013), soil (Andersen *et al.* 2012), and leaf litter (Yang *et al.* 2014). During my PhD research, researchers started discussing the possibility of doing diet analyses using direct sequencing via a metagenomic approach. For example, looking into the future, in 2012 Taberlet *et al.* stated: "A simpler possibility to avoid PCR would be to directly sequence the eDNA extract with NGS platforms, which can produce several billion sequence reads per experiment (e.g. using the Illumina HiSeq 2000 platform)… However, at the moment, we do not know the proportion of potentially informative sequence reads (i.e. the proportion of mitochondrial, chloroplast and nuclear ribosomal DNA) that is possible to obtain in such a sequencing experiment." My thesis provides information on this point for phytophagous primates (<<1% of the reads are of chloroplast origin). More recently,

Andrew *et al.* (2013) suggested the elimination of the PCR amplification step as one of the challenges for NGS based studies for trophic interactions, and called it "theoretically possible" should cost of sequencing come down and should genome sequences become available for individual species. In this transition, where researchers have started considering possibilities for metagenomics, I investigate such an approach in depth, develop strategies to address the above mentioned questions using DNA barcodes and show the promises and current shortcomings of this method. The key conclusions are summarised here.

7.1 Optimizing metagenomics under challenging conditions

The bioinformatic strategies optimized in this thesis are designed for reconstructing the diet of species from fecal samples under challenging conditions: first, the species in question were colobines which have long digestion times (Lambert 1998, Chapter 5). Second, the monkeys are phytophagous and I had to use plant barcodes which are harder to generate and have lower species-specificity than COI (Hollingsworth *et al.*, 2011). Third, Chapter 6 applies these techniques to a wild population of primates living in a species-rich tropical forest. Despite these problems, I was able to characterize the diet at the genus level with all common reads being identified. Given that diet identification could be achieved with good reliability under these circumstances, metagenomics is a promising approach even under challenging conditions. Of course, this prediction needs to be empirically tested on different organisms from different habitats. In particular library coverage is likely to need adjustments in order to account for different types of diets. One of the drawbacks of metagenomics as inferred from Chapter 5 is that it may not be able to identify rare taxa even with ~10 Gb coverage. However, coverage appeared to

be no problem for the samples in Chapter 6 which documents the need to adjust sampling conditions to specific circumstances.

7.2 Metagenomics and metabarcoding correlate, at least for *trnL*

PCR amplification biases have been a concern in diet analyses using metabarcoding. While there has been considerable interest in quantifying read abundance to infer dominant dietary components (Deagle *et al.* 2013), researchers have continued to question whether this is because it is often assumed that the PCR step skews read counts (Andrew *et al.,* 2013). This is because no study has systematically studied the correlation between read counts in metabarcoding and metagenomics. Fortunately, my datasets could be used for this purpose and overall I observed strong correlations suggesting that sequences generated after amplifying *trnL* P6 loop reflect the original DNA sequence abundances in the extracted DNA. This suggests that, while metabarcoding may have other problems (low taxonomic resolution for plants), it may be possible to use it to quantify read numbers. Of course, read counts alone will not solve the problem of how to translate them to biomass. This will require a lot more research into DNA content and differential digestion rates (Deagle *et al.* 2010).

7.3 DNA barcoding: how to go forward in an NGS era?

While Andrew *et al.* (2013) specified that reference genomes would be needed for a characterization of metagenomes, I demonstrate that much can be gained through the analysis of DNA barcodes for plant identifications. When DNA barcoding was initially proposed, the idea was to identify sequences from unknown individuals using a database of barcodes (Hebert *et al.* 2003). While there was a lot of debate about taxonomic

implication of barcodes when it was proposed (Moritz & Cicero 2004), there was little doubt that a standardized sequencing of a single gene (or two genes for plants) would enable researchers to use molecular markers for sorting unidentified specimens. NGS-based studies have further pushed the frontiers in this field, but many problems remain. One is of the paucity of identified DNA barcodes for the 1.5 million described species of Metazoa, which renders many species unidentifiable via DNA barcodes. Second, methodological practices in DNA barcoding should be justified and the use of techniques such as K2P NJ trees should be abandoned. Unless the DNA barcoding movement adopts more rigorous analysis techniques, it will be difficult to implement large-scale bioinformatic pipelines that will be respected outside of the field. Given the amount of data generated by NGS, it is preferable if the analytical techniques are computationally tractable. Lastly, it is important that the barcode databases come with structured taxonomic databases, so that bioinformatic pipelines yield information on taxonomic hierarchies. This was the strategy pursued in my thesis where my pipeline used GenBank data and NCBI taxonomy. rDNA databases satisfying these criteria are already available (e.g., SILVA: Pruesse *et al.* 2007), and similar tools should be developed for DNA barcodes of all eukaryotes.

7.4 Towards a holistic characterization of eDNA

In this thesis I have characterized the diet, parasites and host mitochondrial genomes for captive red shanked douc langurs (*Pygathrix nemaeus*) and for individuals of a wild population of the banded leaf monkeys (*Presbytis femoralis*). Yet, I ignored most of my data; i.e., the microbiome that was represented by >90% of the reads. Characterizing the microbiome is the next logical step of my study, given that some assemblies have already been generated. The microbial flora living in the gut of these

colobines is particularly interesting. For years, it has been known that colobine guts contain diverse bacteria specialized for digesting plant material and degrading cell walls for the release of nutrients (Kay *et al.* 1976). Like ruminants, they have multi-chambered guts with symbiotic bacteria in the fore-stomach (Kay and Davies 1994). Yildrim *et al.* (2010) provided a preliminary 16S taxonomic profile; nonetheless little else is known about the bacterial community. Based on my shotgun metagenomes, I will now be able to look at the functional profile of these sequences and determine the genes that aid in digestion. The high throughput data generated in my study allows for the comparison of leaf monkey microbiomes with the rumen gut microbiome generated from cows using ~250 Gb of data (Hess *et al.* 2011). Combined with the characterization of diet, parasites and host, we will be able to address a multitude of questions about the biology of an organism about which we knew very little and important information could be obtained before seeing the species.

References

Adedokun OA, Adedokun RAM, Emikpe BO, Ohore OG, Oluwayelu DO, Ajayi OL. Concurent fatal helminthosis and balantidosis in red monkey (*Erythrocebus patas*) in Ibadan, Nigeria. *Nigerian Veterinary Journal*, **23**, 56-59.

Allcock AL, Barratt I, Eleaume M, Linse K, Norman MD, Smith PJ, Steinke D, Stevens DW, Strugnell JM (2011) Cryptic speciation and the circumpolarity debate: A case study on endemic Southern Ocean octopuses using the COI barcode of life. *Deep-Sea Research Part. II*, **58**, 242-249.

Alfellani MA, Jacob AS, Perea NO *et al.* (2013) Diversity and distribution of *Blastocystis sp.* subtypes in non-human primates. *Parasitology*, **140**, 966-971.

Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, **12**, 402.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

Ang A (2010) *Banded leaf monkeys in Singapore: Preliminary data on taxonomy, feeding ecology, reproduction and population size*, National University of Singapore.

Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjaer KH, Orlando L, Gilbert MT, WIllerslev E (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology,* **21**, 1966-1979.

Andrew RL, Bernatchez L, Bonin A*, et al.* (2013) A road map for molecular ecology. *Molecular Ecology*, **22**, 2605-2626.

Ang A (2010) Banded leaf monkeys in Singapore: preliminary data on Taxonomy, feeding ecology, reproduction and population size. *Masters Thesis,* National University of Singapore, Singapore.

Ang A, Ismail MRB, Meier R (2010) Reproduction and infant pelage coloration of the banded leaf monkey, *Presbytis femoralis,* (Mammalia: Primates: Cercopithecidae) in Singapore. *The Raffles Bulletin of Zoology*, **58**, 411-415.

Ang A, Srivathsan A, Md-Zain BM, Ismail MRB, Meier R (2012) Low genetic variability in the recovering urban banded leaf monkey population of Singapore. *The Raffles Bulletin of Zoology*, **60**, 589-594.

Arboleya S, Ang L, Margolles A*, et al.* (2012) Deep 16S rRNA metagenomics and quantitative PCR analyses of the premature infant fecal microbiota. *Anaerobe*, **18**, 378-380.

Arizono N, Yamada M, Tegoshi T, Onishi K (2012) Molecular identification of *Oesophagostomum* and *Trichuris* eggs isolated from wild Japanese macaques. *Korean Journal of Parasitology*, **50**, 253-257.

Aylagas E, Borja A, Rodriguez-Ezpelata (2014) Environmental status assessment using DNA metabarcoding: towards a genetics based marine biotic index (gAMBI). *PLoS One*, **9**, e90529.

Baamrane AAB, Shehzad W, Quhammou A, Abbad A, Naimi M, Coissac E, Taberlet P, Znari M (2012) Assessment of the food habits of the Moroccan Dorcas Gazelle in M' Sabih Talaa, West Central Morocco, using trnL approach. *PLoS One*, **10**, e35634.

Baird DJ, Pascoe TJ, Zhou X, Hajibabaei M (2011) Building freshwater macroinvertebrate DNA-barcode libraries from reference collection material:

formalin preservation vs specimen age. *Journal of the North American Benthological Society*, **30**, 125-130.

Baldwin CC, Castillo CI, Weigt LA, Victor BC (2011) Seven new species within western Atlantic *Starksia atlantica*, *S. lepicoelia*, and *S. sluiteri* (Teleostei, Labrisomidae), with comments on congruence of DNA barcodes and species. *ZooKeys*, **79**, 21-72.

Bartlett SE, Davidson WS (1991) Identification of *Thunnus* Tuna species by the polymerase chain reaction and direct sequence analysis of their mitochondrial cytochrome b genes. *Canadian Journal of Fisheries and Aquatic Sciences*, **48**, 309-317.

Becker S, Hanner R, Steinke D (2011) Five years of FISH-BOL: Brief status report. *Mitochondrial DNA*, **22**, 3-9.

Bennett BT, Abee CR, Henrickson R (1998) Nonhuman primates in biomedical research: Diseases, *p.* 133. Academic Press, San Diego.

Bennett C (1994) The ecology, taxonomy and conservation status of the banded leaf monkey *(Presbytis femoralis femoralis)* in Singapore, Free University Berlin.

Bennett EL (1983) *The banded langur: ecology of a colobine in West Malaysian rainforest.*, University of Cambridge.

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I (2007) Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution*, **22**, 148-155.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. doi: 10.1093/bioinformatics/btu170.

Bon C, Berthonaud V, Maksud F, Labadie K, Poulain J, Artiguenave F, Wincker P, Aury JM, Elalouf JM (2012) Coprolites as a source of information on the genome and

diet of the cave hyena. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 2825-2830.

Bradley BJ, Stiller M, Doran-Sheehy DM*, et al.* (2007) Plant DNA sequences from feces: potential means for assessing diets of wild primates. *American Journal of Primatology*, **69**, 699-705.

Brook BW, Sodhi NS, Ng PK (2003) Catastrophic extinctions follow deforestation in Singapore. *Nature*, **424**, 420-426.

Brower AVZ (2006) Problems with DNA barcodes for species delimitation: 'ten species' of *Astraptes fulgerator* reassessed (Lepidoptera : Hesperiidae). *Syst. Biodivers.*, **4**, 127-132.

Brook BW, Sodhi NS, Ng PK (2003) Catastrophic extinctions follow deforestation in Singapore. *Nature* **424**, 420-426.

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

Campos-Arceiz (2013) Next Generation Poo Studies. *Tapir Conservation*, **22**, 4-5.

Chapman CA, Bowman DD, Gha RR, Gogarten JF, Goldberg TL, Rothman JM, Twinomugisha D, Walsh C (2011). Protozoan Parasites in Group-Living Primates: Testing the Biological Island Hypothesis. *American Journal of Primatology*, **72**, 1-8.

Chapman CA, Gillespie TR, Goldberg TL (2005) Primates and the ecology of their infectious diseases: How will anthropogenic change affect host-parasite interactions? *Evolutionary Anthropology* **14**, 134-144.

Chapman CA, Speirs ML, Gillespie TR, Holland T, Austad KM (2006) Life on the edge: gastrointestinal parasites from the forest edge and interior primate groups. *American Journal of Primateology,* **68**, 397-409.

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

Coissac E (2012) OligoTag: a program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods Mol Biol*, **888**, 13-31.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals (2012). *Molecular Ecology*, **21**, 1834-1847.

Collins RA, Armstrong KF, Meier R, Yi Y, Brown SDJ, Cruickshank RH, Keeling S, Johnston C (2012) Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *Plos One*, **7**, e28381.

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B Biol Sci*, **345**, 101-118.

Cordona S, Eck A, Cassellas M *et al.* (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiology,* **12**, 158.

Cowlishaw G, Dunbar RIM (2000) *Primate Conservation Biology* University of Chicago Press, Chicago.

da Silva F, Minhos MJ, Sa RM, Bruford M (2012) Using genetics as a tool in primate conservation. *Nature Education Knowledge*, **3**, 89.

Davies AG, Bennett EL, Wateman PG (1988) Food selection by two South-east Asian colobine monkeys (*Presbytis rubicunda* and *Presbytis melalophos*) in relation to plant chemistry. *Biological Journal of Linnean Society*, **34**, 33-56.

De Barba M, Miquel C, Boyer F*, et al.* (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, **14**, 306-323.

Deagle BE, Chiaradia A, McInnes J, Jarman SN (2010) Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics* **11**, 2039-2048.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology* **28**, 2022-2038.

Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources*, **13**, 620-633.

Deagle BE, Gales NJ, Evans K, Jarman SN, Robinson S, Trebilco R, Hindell MA (2007) Studying seabird diet through genetic analysis of faeces: a case study on macaroni penguins (*Eudyptes chrysolophus*). *PLoS One*, **2**, e831.

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **360**, 1905-1916.

Dettai A, Lautredou AC, Bonillo C, Goimbault E, Busson F, Causse R, Couloux A, Cruaud C, Duhamel G, Denys G, Hautecoeur M, Iglesias S, Koubbi P, Lecointre G, Moteki M, Pruvost P, Tercerie S, Ozouf C (2011). The actinopterygian diversity of the CEAMARC cruises: Barcoding and molecular taxonomy as a multi-level tool for new findings. *Deep-Sea Res. Pt. II*, **58**, 250-263.

Dewit I, Dittus WPJ, Vercruysse J, Harris EA, Gibson DI (1991) Gastro-intestinal helminths in a natural population of *Macaca sinica* and *Presbytis* spp. at Polonnaruwa, Sri Lanka. *Primates*, **32**, 391-395.

Dierenfeld ES, Koontz FW, Goldstein RS (1992) Feed intake, digestion and passage of the proboscis monkey *(Nasalis larvatus)* in captivity. *Primates*, **33**, 399-405.

Dove H, Mayes RW (1996) Plant wax components: A new approach to estimating intake and diet composition in herbivores. *Journal of Nutrition*, **126**, 13-26.

East R (2013) Microbiome: Soil science comes to life. *Nature*, **501**, S18-19.

Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS* Biol, **5**, e82.

Elliot TL, Davies J (2014) Challenges to barcoding an entire flora. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12277.

Ekanayake DK, Arulkanthan A, Horadagoda NU, Saneevani GK, Kieft R, Gunatilake S, Dittus WP (2006) Prevalence of cryptosporidium and other enteric parasites among wild non-human primates in Polonnaruwa, Sri Lanka. *American Journal of Tropical Medicine and Hygien*,. **74**, 322-329.

Ekrem T, Stur E, Hebert PDN (2010) Females do count: Documenting Chironomidae (Diptera) species diversity using DNA barcoding. *Org. Divers. Evol.*, **10**, 397-408.

Fan L, Hui JHL, Yu ZG, Chu KH (2014) VIP Barcoding: composition vector-based software for rapid species identification based on DNA barcoding. *Molecular Ecology Resources*, **14**, 871-881.

Federhen S (2011) Comment on 'birdstrikes and barcoding: can DNA methods help make the airways safer?'. *Molecular Ecology Resources*, **11**, 937-938.

Ferreira da Silva MJ, Minhos T, Sa RM, Bruford MW (2012) Using genetics as a tool in primate conservation. *Nature Education Knowledge*, **3**, 89.

Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessière J, Taberlet P, Pompanon F. (2010). An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomic*, **11**, 434.

Francis CM, Borisenko AV, Ivanova NV, Eger JL, Lim BK, Guillen-Servent A, Kruskop SV, Mackie I, Hebert PD (2010) The role of DNA barcodes in understanding and conservation of mammal diversity in southeast Asia. *PLoS One*, **5**, e12575.

Gielly L, Taberlet P (2014) The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbc*L sequences. *Mol. Bio. Evo.*, **5**, 769-777.

Gillespie TR, Greiner E, Chapman CA (2005) Gastrointestinal parasites of the colobus monkeys of Uganda. *Journal of Parasitology,* **91**, 569-573.

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, prot5368.

Goldsmid JM (1974). The intestinal helminthzoonoses of primates in Rhodesia. *Annales de la Societe belge de medecine tropicale*, **54**, 87-101.

Gomez A, Wright PJ, Lunt DH, Cancino JM, Carvalho GR, Hughes RN (2007) Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon. *Proceedings of the Royal Society B-Biological Sciences*, **274**, 199-207.

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

Gotoh S (2000). Regional differences in the infection of wild Japanese macaques by gastrointestinal helminth parasites. *Primates*, **41**, 291-298.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate earge phylogenies by maximum likelihood. *Syst. Biol.*, **54**, 696-704.

Hajibabaei M, DeWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, Hebert PDN (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **360**, 1959-1967.

Hamad I, Delaporte E, Raoult D, Bittar F (2014) Detection of termites and other insects consumed by African great apes using molecular fecal analysis. *Scientific Reports* **4**.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, **5**, R245-R249.

Hasegawa H, Kano T, Mulavwa M (1983) A parasitological survey on the feces of pygmy chimpanzees, *Pan paniscus*, at Wamba, Zaire. *Primates*, **24**, 419–423.

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

Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B.*, **270**, 313-321.

Hess M, Sczyrba A, Egan R*, et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**, 463-467.

Hilbert F, Taberlet P, Chave J, Scotti-Saintagne C, Sabatier D and Richard-Hansen C (2013). Unveiling the diet of elusive rainforest herbivores in Next Generation Sequencing era? The Tapir as a case study. *PLoS One*, **8**, e60799.

Hollingsworth PM (2011) Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 19451-19452.

Hollingsworth ML, Andra Clark A, Forrest LL*, et al.* (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources*, **9**, 439-457.

Hollingsworth PM, Graham SW and Little DP (2011), Choosing and using a plant DNA barcode. *PLoS* One, **6**, e19254.

Hubert N, Hanner R, Holm E*, et al.* (2008) Identifying Canadian freshwater fishes through DNA Barcodes. *PLoS One*, **3**, e1371.

Hunt T, Bergsten J, *et al.* (2007) A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science, **318**, 1913-1916.

iBOL (*International Barcode of Life*) (2010) *International Barcode of Life*. (Available from: http://ibol.org/big-funding-boost-for-ibol*;* accessed 28.3.2012).

James SW, Porco D, Decaens T, Richard B, Rougerie R, Erseus C (2010) DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): resurrection of *L. herculeus* (Savigny, 1826). *PLoS One*, **5**, e15629.

Jarman SN, Deagle BE, Gales NJ (2004) Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Molecular Ecology*, **13**, 1313-1322.

Ji Y, Ashton L, Pedley M *et al.* (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245-1257.

Jones-Engel L, Engel GA, Schillact MA, Froehlich J, Paputungan U, Kyes RC (2004). Prevalence of enteric parasites in pet macaques in Sulawesi, Indonesia. *American Journal of Primatology*, **62**, 71-82

Junemann S, Prior K, Szczepanowski R*, et al.* (2012) Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. *PLoS One*, **7**, e41606.

Karere GM, Munene E (2002). Some gastro-intestinal tract parasites in wild De Brazza's monkeys (*Cercopithecus neglectus*) in Kenya. *Veterinary Parasitology*, **110**, 153-157.

Kay RNB, Hoppe P, Maloiy GMO (1976). Fermentative digestion of food in the colobus monkey. *Experientia,* **32**, 485-487.

Kay RNB, Davies G (1994) Digestive physiology. *In Colobine Monkeys: their ecology, behaviour and evolution.* (eds G Davies & J Oates), pp 229-250. Cambridge University Press, Cambridges, UK,

Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Research*, **12**, 656-664.

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

Kirkpatrick RC, Zou RJ, Dierenfeld ES, Zhou HW (2001) Digestion of selected foods by Yunnan snub-nosed monkey *Rhinopithecus bieti* (Colobinae). *American Journal of Physical Anthropology*, **114**, 156-162.

Kohn MH, Wayne RK (1997) Facts from feces revisited. *Trends in Ecology and Evolution*, **12**, 223-227.

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One*, **2**, e508.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 8369-8374.

Kristensen NP, Scoble MJ, Karsholt O (2007) Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. *Zootaxa*, **1668**, 699-747.

Kvist L, Martens J, Nazarenk AA, Orell M (2003) Paternal leakage of mitochondrial DNA in the great tit (*Parus major*)

Marshall E (2005) Taxonomy. Will DNA bar codes breathe life into classification? *Science*, **307**, 1037.

Kuksa P, Pavlovic V (2007) Fast kernel methods for SVM sequence classifiers. *Lectures in Bioinformatics*, **4645**, 228-239.

Kutty SN, Bernasconi MV, Sifner F, Meier R (2007) Sensitivity analysis, molecular systematics and natural history evolution of Scathophagidae (Diptera: Cyclorrhapha: Calyptratae). *Cladistics*, **23**, 64-83.

Labes EM, Nurcahyo W, Deplazes P, Mathis A (2011) Genetic characterization of *Strongyloides spp*. from captive, semi-captive and wild Bornean orangutans (*Pongo pygmaeus*) in Central and East Kalimantan, Borneo, Indonesia. *Parasitology*, **138**, 1417-1422.

Lakra WS, Verma MS, Goswami M, Lal KK, Mohindra V, Punia P, Gopalakrishnan A, Singh KV, Ward RD, Hebert P (2011) DNA barcoding Indian marine fishes. *Molecular Ecology Resources*, **11**, 60-71.

Lambert JE (1998) Primate digestion: interactions among anatomy, physiology, andfeeding ecology. *Evolutionary Anthropology*, **7**, 8-20.

Lamendella R, Domingo JW, Ghosh S, Martinson J, Oerther DB (2011) Comparative fecal metagenomics unveils unique functional capacity of the swine gut. *BMC Microbiology*, **11**, 103.

Lee SC, Chiou SJ, Yen JH, Lin TY, Hsieh KT, Yang JC (2010) DNA barcoding *Cinnamomum osmophloeum* Kaneh. based on the partial non-coding ITS2 region of ribosomal genes. *Journal of Food and Drug Analyses*, **18**, 128-135.

Legesse M, Erko B (2004). Zoonotic intestinal parasites in Papio anubis (baboon) and Cercopithecus aethiops (vervet) from four localities in Ethiopia. *Acta Tropica*, **90**, 231-236.

Levecke BP, Dorny T, Geurden F, Vercammen, Vercruysse J (2007). Gastrointestinal protozoa in nonhuman primates of four zoological gardens in Belgium. *Veterinary Parasitology*, **148**, 236-246.

Ley RE, Hamady M, Lozupone C*, et al.* (2008) Evolution of mammals and their gut microbes. *Science*, **320**, 1647-1651.

Li X, Yang YJ, Henry R*, et al.* (2014) Plant DNA barcoding: from gene to genome. *Biological Reviews*. doi: 10.1111/brv.12104

Li DZ, Gao LM, Li HT*, et al.* (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 19641-19646.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Lim GS, Balke M, Meier R (2011) Determining species boundaries in a world full of rarity: singletons, species delimitation methods. *Systematic Biology*, **61**, 165-169.

Lim KKP, Subaraj R, Yeo SH*, et al.* (2008) Mammals. In: *The Singapore Red Data Book: Threatened plants and animals in Singapore* (eds. Davison GWH, Ng PKL, Ho HC), p. 198. The Nature Society, Singapore.

Little DP (2011) DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS One*, **6**, e20552.

Little DP (2014) A DNA mini-barcode for land plants. *Molecular Ecology Resources*, **14**, 437-446.

Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, **23**, 1-21.

Liu L, Li Y, Li S*, et al.* (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, **2012**, 251364.

Luo R, Liu B, Y. X*, et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**, 18.

Magnacca KN, Brown MJ (2010) Mitochondrial heteroplasmy and DNA barcoding in Hawaiian *Hylaeus* (Nesoprosopis) bees (Hymenoptera: Colletidae). *BMC Evol. Biol.*, **10**, 174.

Mak JW, Inder-Singh, Yen PKF, Yap LF (1980) Dipetalonema digitatum (Chandler, 1929) infection in the leaf monkey, *Presbytis obscura* (Reid). *Southeast Asian Journal of Tropical Medicine and Public Health*, **11**, 141.

Mäkinen V, Salmela L, Ylinen J (2012) Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics,* **13**, 255.

Mande SS, Mohammed MH, Ghosh TS (2012) Classification of metagenomic sequences: methods and challenges. *Brief Bioinform*, **13**, 669-681.

Matsui A, Rakotondraparany F, Hasegawa M, Horai S (2007) Determination of a complete lemur mitochondrial genome from feces. *Mammal Study*, **32**, 7-16.

McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303.

Mecklenburg BW, Moller PR, Steinke D (2011) Biodiversity of arctic marine fishes: taxonomy and zoogeography. *Mar. Biodiv.*, **41**, 109-140.

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

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

Meier R, Dikow T (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology*, **18**, 478-488.

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**, 809-813.

Merlender A, Kremen C, Rakotondratisma M, Weiss A (1998) Monitoring impacts of natural resource extraction on lemurs of the Masoala Peninsula, Madagascar. *Ecology and Society*, **2**, 5.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.*, **3**, e422.

Meyer D, Rinaldi ID, Ramlee H*, et al.* (2011) Mitochondrial phylogeny of leaf monkeys (genus Presbytis, Eschscholtz, 1821) with implications for taxonomy and conservation. *Molecular Phylogenetics and Evolution*, **59**, 311-319.

Mitter KT, Larsen TB, de Prins W, de Prins J, Collins S, Weghe GV, Safian S, Zakharov EV, Hawthorne DJ, Kawahara AY, Regier JC (2011) The butterfly subfamily Pseudopontiinae is not monobasic: marked genetic diversity and morphology reveal three new species of *Pseudopontia* (Lepidoptera: Pieridae). *Systematic Entemology*, **36**, 139-163.

Mohammad AG, Pieper RD, Wallace JD, Holechek JL, Murray LW (1995) Comparison of fecal analysis and rumen evacuation techniques for sampling diet botanical composition of grazing cattle. *Journal of Range Management*, **48**, 202-205.

Moniz MBJ, Kaczmarska I (2010) Barcoding of diatoms: nuclear encoded ITS revisited. *Protist*, **161**, 7-34.

Moreno-Black G (1978) The use of scat samples in primate diet analysis. *Primates* **19**, 215-221.

Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. *PLoS Biol.*, **10**, e354.

Munene E, Otsyula M, Mbaabu DAN, Mutahi WT, Muriuki SMK, Muchemi GM (1998). Helminth and protozoan gastrointestinal tract parasites in captive and wild-trapped African non-human primates. *Veterinary Parasitology*, **78**, 195-201.

Munshi-South J, Bernard H (2011) Genetic diversity and distinctiveness of the proboscis monkeys (Nasalis larvatus) of the Klias Peninsula, Sabah, Malaysia. *Journal of Heredity*, **102**, 342-346.

Muriuki SMK, Murugu, RK, Munene E, Karere GM, Chai DC (1998). Some gastro-intestinal parasites of zoonotic (public health) importance commonly observed in old world non-human primates in Kenya. *Acta Tropica*, **71**, 73-82.

Murray DC, Bunce M, Cannell BL, Oliver R, Houston J, White NE, Barrero RA, Bellgard MI, Haile J (2011) DNA-Based faecal dietary analysis: A comparison of qPCR and High Throughput Sequencing approaches. *PLoS One*, **6**, e25776.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, **40**.

Nanney DL (1982) Genes and phenes in *Tetrahymena*. *Bioscience*, **32**, 783-788.

Nath BP, Islam S, Chakraborty A (2012). Prevalence of parasitic infection in captive non human primates of Assam state zoo, India. *Veterinary world*, **4**, 614-616.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

Nei M, Kumar S (2005) *Molecular evolution and phylogenetics*. Oxford University Press, New York.

Nepal MP, Ferguson CJ (2012). Phylogenetics of *Morus* (Moraceae) inferred from ITS and *trnL-trnF* sequence data. *Systematic Botany,* **37**, 442-450.

Ng'endo RN, Osiemo ZB, Brandl R (2013). DNA barcodes for species identification in the hyperdiverse ant genus *Pheidole* (Formicidae: Myrmicinae). *Journal of Insect Science,* **13**, 27.

Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ (2011) Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal*, **9**, 328-333.

Nossa CW, Oberdorf WE, Yang L*, et al.* (2010) Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World Journal of Gastroenterology*, **16**, 4135-4144.

Paabo S, Irwin DM, Wilson AC (1990) DNA damage promotes jumping between templates during enzymatic amplification. *Journal of Biological Chemistry*, **265**, 4718-4721.

Page RDM (2011) *iPhylo. "Dark taxa: GenBank in a post-taxonomic world".* (Available from: http://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html' accessed 28.3.2012).

Palmieri JR, Purnomo, Lee VH, Dennis DT, Marwoto DHA (1980). Parasites of the silvered leaf monkey, Presbytis cristatus Eschscholtz 1921, with a note on a Wuchereria-like nematode. *Journal of Parasitology*, **66**, 170-171.

Pang X, Song J, Zhu Y, Hongxi X, Huang L, Chen S (2011) Applying plant DNA barcodes for Rosaceae species identification. *Cladistics*, **27**, 165-170.

Parmar SM, Jani RG, Mathakiya RA (2012). Study of parasitic infections in non-human primates in Gujarat state, India. *Veterinary World*, **5**, 362-364.

Parmentier I, Duminil J, Kuzmina M, Philippe M, Thomas DW, Kenfack D, Chuyong GB, Cruaud C, Hardy OJ (2013). How Effective Are DNA Barcodes in the Identification of African Rainforest Trees? *PLoS One*, **8**, e54921.

Pauls SU, Blahnik RJ, Zhou X, Wardwell CT, Holzenthal RW (2010) DNA barcode data confirm new species and reveal cryptic diversity in Chilean *Smicridea* (Smicridea) (Trichoptera:Hydropsychidae). *Journal of the North American Benthological Society*, **29**, 1058-1074.

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, **183**, 63-98.

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420-1428.

Pierce GJ, Boyle PR, Diack JSW, Clark I (1990) Sandeels in the diets of seals - Application of novel and conventional methods of analysis to feces from seals in the Moray Firth area of Scotland. *Journal of the Marine Biological Association of the United Kingdom*, **70**, 829-840.

Pompanon F, Deagle BE, Symondson WOC*, et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931-1950.

Pons J, Barraclough TG, Gomez-Zurita J*, et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595-609.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253-1256.

Pruesse E, Quast C, Knittel K*, et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, **35**, 7188-7196.

Qin J, Li R, Raes J*, et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59-65.

Quemere E, Hibert F, Miquel C*, et al.* (2013) A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS One* , **8**, e58971.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 1-10.

Ratnasingham S, Hebert PDN (2011) BOLD's role in barcode data management and analysis: a response. *Molecular Ecology Resources*, **11**, 941-942.

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) System. *PLoS One*, **8**, e66213.

Raye G, Miquel C, Coissac E*, et al.* (2011) New insights on diet variability revealed by DNA barcoding and high-throughput pyrosequencing: chamois diet in autumn as a case study. *Ecological Research*, **26**, 265-276.

Rasheed Z, Rangwala H, Barbara D (2013) 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. *BMC Systems Biology*, **7** (S4).

Ray J, Dondrup M, Modha S*, et al.* (2012) Finding a needle in the virus metagenome haystack--micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. *PLoS One*, **7**, e34238.

Reed SE, Bidlack AL, Hurt A, Getz WM (2011) Detection distance and environmental factors in conservation detection dog surveys. *The Journal of Wildlife Management*, **75**, 243-251.

Remfry J (1978). The incidence, pathogenesis and treatment of helminth infections in rhesus monkeys (*Macaca mulatta)*, *Laboratory* animals, **12,** 213-218.

Remis MJ and Dierenfeld ES (2004). Digesta passage, digestibility and behavior in captive gorillas under two dietary regimes. *International Journal of Primatology*, **25,** 825-845.

Ren BQ, Xiang XG, Chen ZD (2010) Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers. *Molecular Ecology Resources*, **10**, 594-605.

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

Rubinoff D (2006) Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology*, **20**, 1026-1033.

Rutherford K, Parkhill J, Crook J*, et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944-945.

Sakeguchi E, Suzuki K, Dotera S, Ehara A (1991) Fibre digestion and digesta retention time in macaque and colobus monkeys. In: *Primatoogy today* (eds. Ehara A, Kimura T, Takenaka O, Iwamoto M), pp. 671-674. Elsevier Science, New York.

Sangwan N, Lata P, Dwivedi V*, et al.* (2012) Comparative metagenomic analysis of soil Microbial Communities across Three Hexachlorocyclohexane Contamination Levels. *PLoS One*, **7**, e46219.

Sarkar IN, Planet PJ, Desalle R (2008) CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*, **8**, 1256-1259.

Schoch CL, Seifert KA, Huhndorf S*, et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 6241-6246.

Sharpton TJ, Riesenfeld SJ, Kembel SW*, et al.* (2011) PhylOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *Plos Computational Biology*, **7**, e1001061.

Shehzad W, Riaz T, Nawaz MA*, et al.* (2012) Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (Prionailurus bengalensis) in Pakistan. *Molecular Ecology* **21**, 1951-1965.

Shigenbou Y, Saitoh K, Hayashizaki K, Ida H (2005) Nonsynonymous site heteroplasmy in fish mitochondrial DNA.*Genes and Genetic systems,* **80**, 297-301.

Smith DAE, Smith YCE (2013) Population density of red langurs in Sabangau tropical peat-swamp forest, Central Kalimantan, Indonesia. *American Journal of Primatology*, l**75**, 837-847.

Sperling F (2003) DNA Barcoding: Deus ex Machina. *Newsletter of the Biological Survey of Canada (Terrestrial Arthropods)*, **22**, 50-53.

Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, Brochmann C, Brysting AK, Sonstebo JH, Ims RA, Yoccoz GN, Taberlet P (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, **6**, e1186.

Sommer V, Mendoza GD (1995) Play as indicator of habitat quality: a field study of langur monkeys (*Presbytis entellus*). *Ethology*, **99**, 177-192.

Song H, Buhay JE, Whiting MF, Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *PNAS,* **105**,13486-13491.

Stensvold CR, Lebbad M, Victory EL, Verweij JJ, Tannich E, Alfellani M, Legarraga P, Clark CG (2011) Increased sampling reveals novel lineages of *Entamoeba:* Consequences of genetic diversity and host specificity for taxonomy and molecular detection. *Protist*, **162**, 525-541.

Strutzenberger P, Brehm G, Fiedler K (2011). DNA barcoding-based species delimitation increases species count of *Eois* (Geometridae) moths in a well-studied tropical mountain forest by up to 50%. *Insect Science,* **18**, 349-362.

Symondson WO (2002) Molecular identification of prey in predator diets. *Molecular Ecology*, **11**, 627-641.

Sweeney BW, Battle JM, Jackson JK (2011) Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, **30**, 195-216.

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taberlet P, Coissac E, Pompanon F*, et al.* (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* **35**, e14.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045-2050.

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of 3 noncoding regions of chloroplast DNA. *Plant Molecular Biology*, **17**, 1105-1109.

Tan DSH, Ang Y, Lim GS, Ismail MR, Meier R (2010) From 'cryptic species' to integrative taxonomy: an iterative process involving DNA sequences, morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae: Diptera). *Zoologica Scripta*, **39**, 51-61.

Tan SY, Koh CY, Siow J*, et al.* (2013) *100 common vascular plants of the Nee Soon Swamp forest, Singapore.* Raffles Museum of Biodiversity Research, National University of Singapore, Singapore.

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, **12**, 377-388.

Thatcher VE, Porter JA (1968). Some helminth parasites of Panamanian primates. *Transactions of American Microscopical Society*, **87**, 186-196.

Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, **2**, 3.

Tringe SG, Zhang T, Liu X*, et al.* (2008) The airborne metagenome in an indoor urban environment. *PLoS One* **3**, e1862.

Valentini A, Miquel C, Nawaz MA*, et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51-60.

Venter JC, Remington K, Heidelberg JF*, et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.

Victor BC (2010) *Emblemariopsis carib* and *Emblemariopsis arawak*, two new chaenopsid blennies from the Caribbean Sea: DNA barcoding identifies males, females, and uveniles and distinguishes sympatric cryptic species. *Journal of the Ocean Science Foundation*, **4**, 1-30.

Virgilio M, Backeijau T, Barr N, De Meyer M (2008) Molecular evaluation of nominal species in the *Ceratitis fasciventris*, *C-anonae*, *C-rosa* complex (Diptera : Tephritidae). *Molecular Phylogenetics and Evolution*, **48**, 270-280.

Volmer NL, Viricel A, Wilcox L, Katherine Moore M, Rosel PE (2011). The occurrence of mtDNA heteroplasmy in multiple cetacean species. *Current Genetics,* **57,** 115-131.

Wang XP, Yu L, Roos C, Ting N, Chen CP, Wang J, Zhang YP (2012). Phylogenetic Relationships among the Colobine Monkeys Revisited: New Insights from Analyses of Complete mt Genomes and 44 Nuclear Non-Coding Markers. *PLoS One,* **7**, e36274.

Wang JF, Qi J, Zhao H*, et al.* (2013) Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Scientific Reports*, **3,** 1843**.**

Ward RD, Hanner R, Hebert PD (2009) The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, **74**, 329-356.

Whitehorn PR, Tinsley MC, Brown MJF, Darvill B, Goulson D (2011) Genetic diversity, parasite prevalence and immunity in wild bumblebees. *Proceedings of the Royal Society B-Biological Sciences*, **278**, 1195-1202.

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, **54**, 844-851.

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

Wilson DE, Reeder DM (2005) *Mammal Species of the World. A Taxonomic and Geographic Reference* 3rd edn. Johns Hopkins University Press.

Wong HF, Tan SY, Koh MHJ*, et al.* (2013) *Checklist of the Plant Species of Nee Soon Swamp Forest, Singapore: Bryophytes to Angiosperms* National Parks Board and Raffles Museum of Biodiversity Research, National University of Singapore, Singapore.

Xu B, Xu WJ, Yang FY*, et al.* (2013) Metagenomic Analysis of the pygmy loris fecal microbiome reveals unique functional capacity related to metabolism of aromatic compounds. *PLoS One*, **8**, e56565.

Xue YT, Sha CM (2010) *Dietary and digestive differences in primates at the Singapore Zoo: Folivory and frugivory in relation to feed preference, intake, digest retention and nutrition*. Final Year Project. Nanyang Technological University, Singapore.

Yang C , Ji Y, Wang X, Yang C, Yu DW (2013) Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. *Science China Life Sciences,* **56**, 76-81.

Yang CM, Lua HK (1988) A report of a banded leaf-monkey found dying near the Bukit Timah Nature Reserve. *The Pangolin*, **1**, 23.

Yildrim S, Yeoman C, Sipos M, Torralba M, Wilson BA, Goldberg TL, Stumpf RM, Leigh SR, White BA, Nelson KE (2010). Characterization of the Fecal Microbiome from Non-Human Wild Primates Reveals Species Specific Microbial Communities. *PloS One,* **5***,* e13963.

Yu DW, Ji YQ, Emerson BC*, et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613-623.

Yu SX, Li L, Ren BP*, et al.* (2009) A study on the carrying capacity of the available habitat for the Rhinopithecus bieti population at Mt. Laojun in Yunnan, China. *Environmental Science and Pollution Research*, **16**, 474-478.

Zarzoso-Lacoste D, Corse E, Vidal E (2013) Improving PCR detection of prey in molecular diet studies: importance of group-specific primer set selection and extraction protocol performances. *Molecular Ecology Resources*, **13**, 117-127.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821-829.

Zhang JJ, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, **29**, 2869-2876.

Zhang RL, Zhang B (2014) Prospects of using DNA barcoding for species identification and evaluation of the accuracy of sequence databases for ticks (Acari: Ixodida). *Ticks and Tick-Borne Diseases*, **5**, 352-358.

Zhou X, Li Y, Liu S*, et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

Zhu LF, Wu Q, Dai JY, Zhang SN, Wei FW (2011) Evidence of cellulose metabolism by the giant panda gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 17714-17719.

Appendices

Appendix 1

Supplementary tables and figure for chapter 5

Tables T1-9

Supplementary Table T1: List of accession numbers of sequences used in the diet database. *sequenced locally (known foliage plants)

Species	matK	rbcL	trnL-F
Acacia auriculiformis	GU134998	JX856621	-
Acalypha siamensis	-	KM029997*	**KM030006***
Adenanthera pavonina	GU135053	GU135287	**AF278486**
Averrhoa bilimbi	-	-	**AJ582291**
Averrhoa carambola	FJ670048	FJ670180	**JN620114**
Azadirachta indica	AY128180	JX856639	**EF489263**
Bambusa multiplex	EF125166	M91626	**DQ137347**
Baphia nitida	EU361867	AM234261	**AY232777**
Bauhinia blakeana	JN881361	JX856641	**FJ801074**
Calophyllum inophyllum	HQ331553	HQ332016	**AB817676**
Carica papaya	JX092002	JQ025026	**JX091823**
Caryota rumphiana	JF344997	JF738928	-
Cenchrus purpureus	JQ588784	JQ593414	**AB817696**
Cinnamomum iners	KM030013*	KM029998*	**KM030005***
Cocos nucifera	JQ586726	JQ590456	**AM113647**
Cratoxylum formosum	HQ331588	AF518395	AY389798
Cucumis sativus	AJ970307	AJ970307	**AJ970307**
Daucus carota	HM850728	HM849948	FJ490764/ **HQ323879**
Dillenia suffruticosa	-	FJ860354	-
Dimocarpus longan	JN407209	JN407382	**EU721213**
Ficus auriculata	JQ773629	JQ773647	-
Ficus benjamina	JQ773509	JQ592814	**AF501605**
Garcinia mangostana	HQ331601	JX664049	GQ456077
Hemigraphis sp.	KM030014*	KM029996*	**KM030009***

Hibiscus rosa-sinensis	KM030012*	KM029999*	**KM030008***
Ipomoea batatas	JX629287	JX011625	**AY101071**
Leucaena leucocephala	KM030010*	KM030000*	**KM030003***
Malus domestica	AF309207	-	**JX122471**
Mangifera indica	JQ586472	JF739088	**KC479210**
Manihot esculenta	NC_010433	NC_010433	**NC_010433**
Manilkara zapota	GU135011	JX856724	**DQ924309**
Morus alba	KM030011*	KM030001*	**KM030004***
Moringa oleifera	JX092021	JX091931	**JX091843**
Munitingia calabura	JQ589354	JQ594271	**AY328166**
Murraya paniculata	GU135010	GU135173	**AY295280**
Musa acuminata	FJ871652	FJ871827	**FJ621283**
Myristica fragrans	EU669472	AY298839	**AY145351**
Nephelium lappaceum	EU720584	-	**EU721175**
Oryza sativa	NC_008155	NC_008155	**NC_008155**
Polygonum chinense	JN407191	JN407357	**HQ843150**
Psidium guajava	JQ024987	JQ025077	-
Pyrus communis	JQ391389	JQ391389	**AM157400**
Pyrus pyrifolia	AP012207	AP012207	**AP012207**
Pterocarpus indicus	JN083546	JF739158	**AF208953**
Samanea saman	JQ587830	JQ592000	**AF522965**
Swietnia macrophylla	JQ588350	JQ592736	**EF489262**
Syzygium zeylanicum	DQ088619	-	-
Tamarindus indica	JQ587876	JQ592062	**AF365206**
Terminalia catappa	-	KM030002*	**KM030007***
Vigna unguiculata	NC_018051	NC_018051	**NC_018051**
Zea mays	NC_001666	NC_001666	**NC_001666**

Accession numbers in bold represent sequences used for metabarcoding experiment.

Supplementary Table T2: Analyses of metagenomic sequences against diet database: summary of number of reads by barcode (Paired end analyses).

	PN1					PN2			
	rbcL	*trnL-F*	*matK*	Total		*rbcL*	*trnL-F*	*matK*	Total
Acalypha siamensis	-	2	-	2		1	-	-	1
Averrhoa sp.	2	2	0	4		.4	13	2	19
Baphia nitida	-	-	-	-		-	2	4	6
Cinnamomum iners	8	14	8	30		35	79	42	156
Cucumis sativus	-	-	-	-		-	-	-	-
Daucus carota	0	1	0	1		1	1	5	7
Ficus sp.	1	2	1	4		-	1	-	1
Hemigraphis sp.	-	-	-	-		-	-	-	-
Hibiscus rosa-sinensis	-	-	-	-		-	-	2	2
Ipomoea batatas	0	1	1	2		1	1	-	2
Leucaena leucocephala	4	5	1	10		25	44	33	102
Malus domestica	-	-	-	-		-	2	-	2
Morus alba	-	-	-	-		-	-	1	1
Oryza sativa	-	-	-	-		-	-	-	-
Pyrus sp.	-	-	-	-		-	-	1	1
Terminalia catappa	5	6	-	11		3	6	-	9
Vigna unguiculata	2	0	1	3		3	1	15	19
Zea mays	1	0	0	1		-	2	2	4
Total	23	33	12	68		67	143	88	332

Supplementary Table T3: Assembly of PN1 and PN2 metagenomic dataset using SOAPdenovo2 at various *k*-mers. Identifications for diet species were made using diet database under criteria of 98% identity and 100bp overlap.

	PN1			PN2		
	K31	K41	K51	K31	K41	K51
Number of contigs >100bp	1,359,037	836,554	509,432	1,786,760	1,023,421	459,282
Mean Size/ Median size	320/166	406/192	449/234	291/154	304/178	371/209
N50	469	655	614	423	331	423
Longest Contig	27,630	247,751	142,410	23,818	131,275	191,115
Number of known diet species identified	3	3	1	5	3	2

Supplementary Table T4: Number of contigs matching to the diet database. Assembly was done using K=31, K=41 and K=51. K=51 revealed poorer results than other two (Supplementary Table T3) and hence was not shown here. Values shown are: Number of contigs/ Average coverage/ length of longest contig.

Species	PN1		PN2	
	K31	K41	K31	K41
Acalypha siamensis	-	-	-	-
Baphia nitida	-	-	-	-
Cinnamomum iners	5/2.4/435	3/2/275	5/9.9/1656	5/6.6/1217
Cucumis sativus	-	-	-	-
Daucus carota	-	-	1/3.0/122	-
Hemigraphis sp.	-	-	-	-
Hibiscus rosa-sinensis	-	-	-	-
Ipomoea batatas	-	-	-	-
Leucaena leucocephala	3/1.3/177	1/2/126	7/6.4/617	8/5.2/2221
Malus domestica	-	-	-	-
Morus alba	-	-	-	-
Oryza sativa	-	-	-	-
Terminalia catappa	3/2.4/273	1/1/147	1/2.0/167	-
Vigna unguiculata	-	-	2/2.8/152	1/2/152
Zea mays	-	-	-	-
Averrhoa carambola.			2/2.4/168	-

Supplementary Table T5: Genera identified from metagenomic datasets of PN1 and PN2 when tested against plant database (Paired end). Known and Potential diet taxa are highlighted in bold.

Genus	Number of genes giving identification	
	PN1	PN2
Terminalia	4	4
Leucaena	3	4
Vigna	3	4
Ficus	3	2
Averrhoa	-	3
Cinnamomum	2	2
Ipomoea	2	1
Acalypha	2	1
Baphia	-	2
Morus	-	2
Daucus	1	3
Neolitsea	-	2
Pyrus	-	1
Hibiscus	-	1
Rhodostemononodaphne	1	1
Oxalis	1	-
Dorstenia	1	-
Ligustrum	1	-
Dapania	1	1
Anadenanthera	-	1
Atherosperma	-	1
Austrobuxus	-	1
Cotoneaster	-	1
Glycine	-	1
Lindera	-	1
Litsea	-	1
Mimozyganthus	-	1
Persea	-	1
Pterocyclus	-	1
Tripsacum	-	1

Supplementary Table T6: Eukaryote identification based on rRNA sequences (MEGABLAST, 98% identity, 70bp overlap); summary derived from top 500 hits of NT. Values in bracket represent (number of Paired end reads from PN1 / number of End 1 reads from PN1) and (number of Paired end reads from PN2 / number of End 1 reads from PN2).

Phylum	Order	Family	Genus	Species
Amoebozoa (1336/2776) (1492/2912)	Nil	nil	Entamoeba (1336/2776) (1492/2912)	Entamoeba bovis (4/56) (11/56)
Amoebozoa (1336/2776) (1492/2912)	Nil	nil	Entamoeba (1336/2776) (1492/2912)	Entamoeba histolytica (137/890) (150/886)
Amoebozoa (1336/2776) (1492/2912)	Nil	nil	Entamoeba (1336/2776) (1492/2912)	Entamoeba sp. RL3 (27/193) (30/199)
Amoebozoa (1336/2776) (1492/2912)	Nil	nil	Entamoeba (1336/2776) (1492/2912)	Entamoeba invadens (1/27) (0/36)
Arthropoda (5/20) (9/58)	Diptera (1/13) (0/4)	*multiple*	*multiple*	*multiple*
Arthropoda (5/20) (9/58)	Hemiptera (0/0) (8/17)	Psyllidae (0/0) (3/4)	*multiple*	*multiple*
Ascomycota (4/11) (2/6)	Sordariomycetes (0/0) (2/4)	*multiple*	*multiple*	*multiple*
Basidiomycota (1/4) (1/24)	Ustilaginales (1/1) (0/0)	Ustilaginaceae (1/1) (0/0)	*multiple*	*multiple*
Basidiomycota (1/4) (1/24)	Polyporales (0/0) (1/3)	Polyporaceae (0/0) (1/3)	Polyporus (0/0) (1/3)	Polyporus umbellata (0/0) (1/3)
Chordata (4/10) (11/42)	Primates (0/3) (8/20)	Cercopithecidae (0/0) (5/14)	Macaca (0/0) (2/8)	Macaca fascicularis (0/0) (2/8)
Heterokontophyta (42/57) (2652/3521)	Blastocystida (42/57) (2652/3521)	Blastocystidae (42/57) (2652/3521)	Blastocystis (42/57) (2652/3521)	Blastocystis sp. MJ99-568 (0/2) (23/132)
Heterokontophyta (42/57) (2652/3521)	Blastocystida (42/57) (2652/3521)	Blastocystidae (42/57) (2652/3521)	Blastocystis (42/57) (2652/3521)	Blastocystis sp. NandII (16/18) (1253/1649)
Heterokontophyta (42/57) (2652/3521)	Blastocystida (42/57) (2652/3521)	Blastocystidae (42/57) (2652/3521)	Blastocystis (42/57) (2652/3521)	Blastocystis sp. subtype 1 (0/0) (4/27)
Nematoda (338/458) (248/353)	Rhabditida (333/453) (244/350)	Strongyloididae (306/426) (224/310)	Strongyloides (293/419) (212/304)	Strongyloides fuelleborni (31/85) (37/87)
Streptophyta (249/400) (1398/2123)	Laurales (13/34) (106/227)	Lauraceae (12/31) (89/196)	Cinnamomum (11/27) (69/160)	Cinnamomum camphora (8/20) (48/119)
Streptophyta (249/400) (1398/2123)	Apiales (1/3) (1/7)	Apiaceae (1/3) (1/4)	Daucus (1/3) (0/3)	Daucus carota (1/3) (0/3)
Streptophyta (249/400) (1398/2123)	Malvales (6/11) (2/8)	Malvaceae (6/11) (1/6)	Gossypium (6/11) (0/1)	Gossypium hirsutum (6/11) (0/0)
Streptophyta (249/400) (1398/2123)	Piperales (1/1) (0/0)	Saururaceae (1/1) (0/0)	Saururus (1/1) (0/0)	Saururus cernuus (1/1) (0/0)
Streptophyta (249/400) (1398/2123)	Fabales (4/20) (30/96)	Fabaceae (4/20) (27/89)	Vigna (1/4) (0/6)	*multiple*
Streptophyta (249/400) (1398/2123)	Fabales (4/20) (30/96)	Fabaceae (4/20) (27/89)	Parkia(0/1) (4/8)	*multiple*
Streptophyta (249/400) (1398/2123)	Malpighiales (1/5) (1/13)	Euphorbiaceae (1/1) (1/4)	Acalypha (0/1)(1/3)	*multiple*
Phylum	**Order**	**Family**	**Genus**	**Species**
Streptophyta (249/400) (1398/2123)	Rosales (1/5) (10/21)	Moraceae (1/2) (7/9)	Ficus (1/1) (1/1)	*multiple*

Streptophyta (249/400) (1398/2123)	Solanales (1/5) (2/19)	Convolvulaceae (1/2) (2/11)	Ipomoea (1/2) (1/6)	Ipomoea purpurea (0/0) (1/3)
Streptophyta (249/400) (1398/2123)	Sapindales (1/1) (0/1)	Anacardiaceae (1/1) (0/0)	*multiple*	*multiple*
Streptophyta (249/400) (1398/2123)	Poales (1/9) (27/66)	Poaceae (1/8) (26/62)	Zea (0/0) (14/42)	Zea mays (0/0) (13/40)
Streptophyta (249/400) (1398/2123)	Myrtales (1/10) (5/31)	*multiple*	*multiple*	*multiple*
Streptophyta (249/400) (1398/2123)	Brassicales (0/1) (1/9)	Brassicaceae (0/1) (1/9)	*multiple*	*multiple*
Streptophyta (249/400) (1398/2123)	Myrtales (1/10) (5/31)	Combretaceae (0/8) (4/25)	Terminalia (0/8) (4/20)	Terminalia catappa (0/8) (4/16)
Streptophyta (249/400) (1398/2123)	Fabales (4/20) (30/96)	Fabaceae (4/20) (27/89)	Albizia (0/3) (6/18)	Albizia julibrissin (0/3) (6/17)
Streptophyta (249/400) (1398/2123)	Rosales (1/5) (10/21)	Moraceae (1/2) (7/9)	Morus (0/0) (6/8)	Morus nigra (0/0) (1/2)
Streptophyta (249/400) (1398/2123)	Ophioglossales (0/0) (2/3)	Ophioglossaceae (0/0) (2/3)	*multiple*	*multiple*
Streptophyta (249/400) (1398/2123)	Oxalidales (0/0) (7/11)	Oxalidaceae (0/0) (6/9)	*multiple*	*multiple*
Streptophyta (249/400) (1398/2123)	Cucurbitales (0/0) (1/5)	Cucurbitaceae (0/0) (1/5)	Cucumis (0/0) (1/2)	Cucumis sativus (0/0) (1/1)
Streptophyta (249/400) (1398/2123)	Fabales (4/20) (30/96)	Fabaceae (4/20) (27/89)	Malus (0/0) (2/3)	Malus domestica (0/0) (2/2)
Streptophyta (249/400) (1398/2123)	Fabales (4/20) (30/96)	Fabaceae (4/20) (27/89)	Leucaena (0/0) (1/3)	*multiple*
Streptophyta (249/400) (1398/2123)	Malpighiales (1/5) (1/13)	Euphorbiaceae (1/1) (1/4)	Acalypha (0/0) (1/3)	*multiple*
Streptophyta (249/400) (1398/2123)	Malvales (6/11) (2/8)	Thymelaeaceae (0/0) (1/1)	Gonystylus (0/0) (1/1)	Gonystylus bancanus (0/0) (1/1)
Streptophyta (249/400) (1398/2123)	Myrtales (1/10) (5/31)	Melastomataceae (0/0) (1/1)	Clidemia (0/0) (1/1)	Clidemia dentata (0/0) (1/1)

Supplementary Table T7: Eukaryote identification based on COI sequences (Megablast, 98% identity, 70bp overlap) summary derived from top 500 hits of NT.

Phylum	**Order**	**Family**	**Genus**	**Species**
Chordata (60/65) (81/95)	Primates (55/59) (76/86)	Cercopithecidae (54/58) (76/86)	*Pygathrix* (53/57) (73/83)	*Pygathrix nemaeus* (1/8) (3/7)
Chordata (60/65) (81/95)	Galliformes (2/3) (2/6)	Phasianidae (2/3) (2/6)	*Gallus* (2/2) (2/5)	multiple
Nematoda (0/5) (3/5)	Rhabditida (0/4) (3/5)	Strongyloididae (0/4) (3/5)	*Strongyloides* (0/4) (3/5)	*Strongyloides fuelleborni* (0/4) (3/5)
Streptophyta (9/12) (83/99)	Fabales (1/1) (2/6)	Fabaceae (1/1) (2/6)	multiple	multiple

Supplementary Table T8: Eukaryote identification based on COI sequences (Megablast, 95% identity, 70bp overlap) summary derived from top 500 hits of NT.

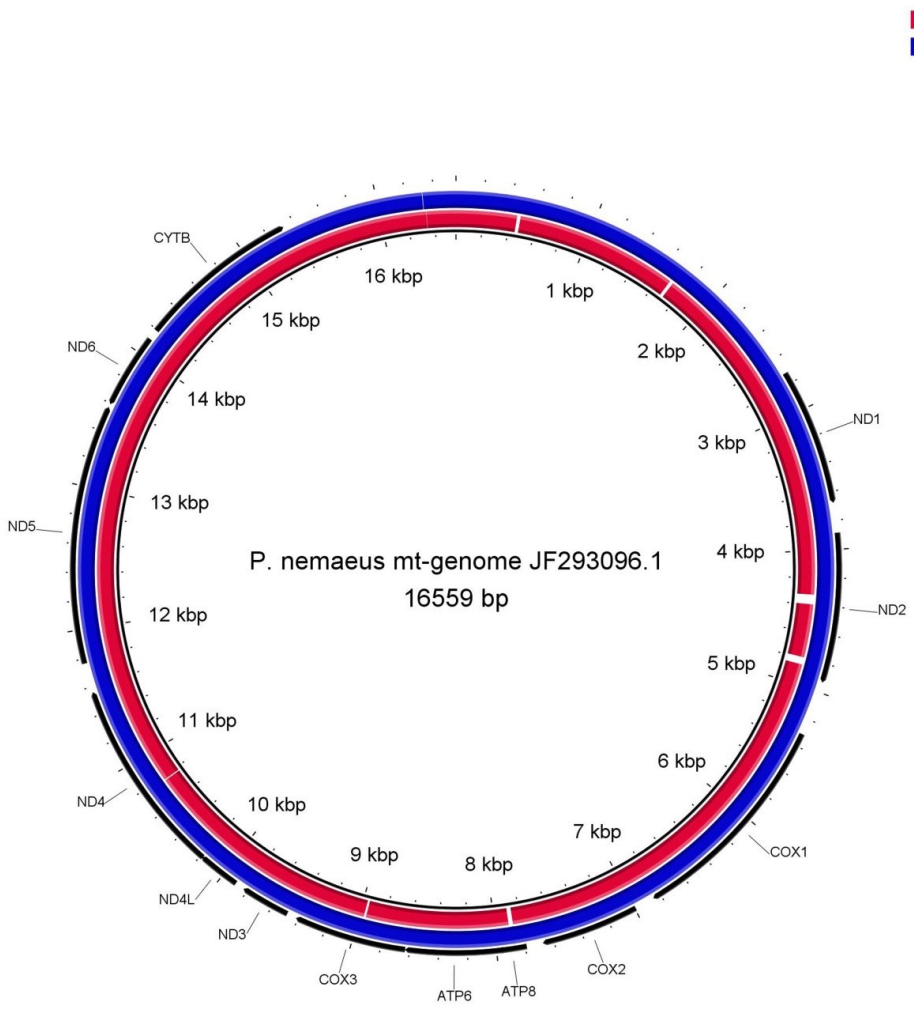
Phylum	**Order**	**Family**	**Genus**	**Species**
Arthropoda (3/6) (1/1)	Diptera (2/4) (0/0)	multiple	multiple	multiple
Chordata (66/69) (97/102)	Primates (61/63) (89/93)	Cercopithecidae (60/62) (89/93)	*Pygathrix* (59/61) (86/90)	*Pygathrix nemaeus* (1/8) (4/8)
Chordata (66/69) (97/102)	Galliformes (2/3) (3/6)	Phasianidae (2/3) (3/6)	*Gallus* (2/2) (3/5)	multiple
Nematoda (7/16) (5/9)	Rhabditida (7/14) (5/8)	Strongyloididae (5/12) (5/8)	*Strongyloides* (5/12) (5/8)	*Strongyloides fuelleborni* (5/12) (5/8)
Streptophyta (12/13) (97/109)	Fabales (1/1) (5/9)	Fabaceae (1/1) (5/9)	multiple	multiple
Streptophyta (12/13) (97/109)	Cucurbitales (0/0) (1/2)	Cucurbitaceae (0/0) (1/2)	multiple	multiple

Supplementary Table T9: Arthropod identifications based on single end analyses using COI sequences (Megablast, 95% identity, 70bp overlap), summary derived from top 500 hits of NT. Sequences in bold represent those that had paired end match to same insect order.

Sequence (Total=6)	Species	Genus	Family	Order	Class	Phylum
1	***Ceratitis capitata***	***Ceratitis***	**Tephritidae**	**Diptera**	**Insecta**	**Arthropoda**
2	***Drosophila fraburu***	***Drosophila***	**Drosophilidae**	**Diptera**	**Insecta**	**Arthropoda**
3	*Ceratitis curvata*	*Ceratitis*	Tephritidae	Diptera	Insecta	Arthropoda
4	**Unidentitfied**	**Unidentitfied**	**Unidentified**	**Diptera**	**Insecta**	**Arthropoda**
5	*Brachycaudus* sp. C1760	*Brachycaudus*	Aphididae	Hemiptera	Insecta	Arthropoda
6	Unidentified	Unidentified	Unidentified	Unidentified	Insecta	Arthropoda

Supplementary Figure

Figure S1: Contigs of PN1 and PN2 identified as mitochondrial sequences were mapped onto *P. nemaeus* genome.



Appendix 2

Supplementary Methods for Chapter 5

Feeding trial

Captive *P. nemaeus* at the Singapore Zoo were studied to obtain information on transit and retention time. Rate of food processing was determined by feeding bead markers whereby transit time (TT) was assessed as the first appearance of the bead markers in the feces while the mean retention time (MRT) referred to the average time taken for the passage of 10-90% of recovered plastic beads marker (Remis & Dierenfeld 2004):

$$MRT(h) = \sum_{i=1}^{n} M_i T_i \,/\, \sum_{i=1}^{n} M_i$$

Where M_i is the amount of markers excreted in the *i*th defecation at time T_i and n is the total number of defecations.

Appendix 3: Description of taxonomic categorization pipeline

This pipeline has been scripted to summarize taxonomy information from blast outputs. It can be downloaded from https://github.com/asrivathsan/readsidentifier-1.0 where the source code is available.

The steps followed are as follows, and the functions written in the script are in the brackets, and can be accessed from the abovementioned link.

(I) BLAST is conducted against a database, and summary of top hits is obtained per sequence (by default BLAST generated 500 best hits). The output format 6 is chosen.

(II) Every subject sequence that the query has a matched to is linked with the TAXID information as obtained by gi_taxid.dmp (matchdb)

(III) Any hit below a user specified minimum overlap threshold are removed. Once removed the best identity hit is retained per sequence. (parse)

(IV) The user is asked to define an identity threshold. All sequence below this threshold are removed (best_by_id)

(V) Taxonomy assignment is conducted.

a. First taxid information is matched to the various hierarchical levels available for a particular taxid. (tax_to_cat)

b. For every blast hit a consistency profile is created. Here if a sequence matches to multiple taxa at a particular taxonomical hierarchy number of taxa are recorded (consist)

c. Taxid are converted to Taxonomic names to generate output files (cat_to_name)

d. Similar procedure is followed for paired-end analyses. However, prior to generating a profile per sequence, I identify the taxon set S1 and S2 and an intersection of these two is then used to generate the taxonomic information for the sequence (consistpe).

Appendix 4: Parasite database

Parasite	Host species	Location	Reference
Ancyclostoma	*Erythrocebus patas*	Africa	Adedokun *et al.* (2002)
Ascaris	*Macaca nigra, Macaca mulatta, Presbytis entellus*	Asia	Jones-Engel *et al.* (2004), Remfry (1978), Parmar *et al.* (2012)
Balantidium	*Cercopithecus aethiops, Cercopithecus mitis, Cercocebus torquatus, Cercocebus albigena, Papio cyanocephalus, Cercopithecus neglectus, Pan troglodytes, Hylobates leucogenys, Erythrocebus patas*	Africa	Muriuki *et al.*, 1998, Munene *et al.* (1998), Karere and Munene (2002), Adedokun *et al.* (2002)
Bertiella	*Colobus guereza, Cercopithecus ascanius, Macaca fuscata, Papio ursinus, Trachypithecus cristatus*	Africa,Asia	Chapman *et al.*(2005), Gotoh (2000), Goldsmid (1974), Palmieri *et al.* (1980)
Blastocystis	*Macaca nigra, Macaca nigrescens, Macaca hecki, Macaca tonkeana, Macaca Maura, Macaca ochreata, Macaca fascicularis, Macaca nemestrina, Papio cyanocephalus, Cercopithecus aethiops, Lophocebus albigena, Procolobus rufomitratus*	Asia, Africa, Captive	Jones-Engel *et al.* (2004), Legesse *et al.* (2004), Chapman *et al.* (2011), Chapter 5
Cryptosporidium	*Cercopithecus aethiops,Papio cyanocephalus, Macaca sinica, Semnopithecus priam, Trachypithecus vetulus*	Asia, Africa	Legesse *et al.* (2004), Ekanayake *et al.* (2006)
Dicrocoeliidae	*Colobus guereza, Cercopithecus ascanius*	Africa,	Chapman *et al.*(2005),
Dipetalonema	*Saguinus geoffroyi, Aotus trivirgatus, Ateles fusciceps, Ateles geoffroyi, Cebus capucinus, Presbytis obscura*	Asia, Neotropics	Thatcher and Porter (1968), Mak *et al.* (1980)
Endolimax	*Macaca nigra, Lophocebus albigena, Procolobus rufomitratus, Cercopithecus ascanius, Colobus guereza, Cercopithecus mitis*	Asia, Africa	Jones-Engel *et al.* (2004), Chapman *et al.* (2011)
Entamoeba	*Piliocolobus tephrosceles, Colobus guereza,Cercopithecus ascanius, Papio cyanocephalus, Cercopithecus aethiops, Cercopithecus mitis, Cercocebus torquatus, Cercocebus albigena, Macaca nigra, Micaca nigrescens, Macaca hecki, Macaca tonkeana, Macaca Maura,*	Africa,Asia, Captive	Chapman *et al.*(2005), Muriuki *et al.*, 1998, Jones-Engel *et al.* (2004), Munene *et al.* (1998), Legesse *et al.* (2004), Parmar *et al.* (2012), Chapman *et al.* (2011), Karere and Munene (2002), Gillespie *et al.*

	Macaca ochreata, Macaca fascicularis, Presbytis entellus, Lophocebus albigena, Procolobus rufomitratus, Cercopithecus neglectus, Colobus angolensis, Hylobates syndactylus, Gorilla gorilla, Pan troglodytes, Hylobates lar		(2005), Levecke *et al.* (2007), Chapter 5
Enterobius	*Cercopithecus ascanius, Macaca sinica, Macaca mulatta, Papio cyanocephalus, Cercopithecus mitis, Papio ursinus, Trachypithecus cristatus, Erythrocebus patas*	Africa,Asia	Chapman *et al.*(2005),Dewit *et al.* (1991), Remfry (1978) Munene *et al.* (1998), Goldsmid (1974), Palmieri *et al.* (1980), Adedokun *et al.* (2002)
Giardia	*Cercopithecus ascanius, Procolobus rufomitratus, Lophocebus albigenus, Hylobates syndactylus, Gorilla gorilla, Hylobates lar, Hylobates leucogenys*	Africa, Captive	Chapman *et al.*(2005), Chapman *et al.* (2011), Levecke *et al.* (2007)
Hymenolepis	*Presbytis entellus, Macaca sinica, Macaca mullatta*	Asia	Dewit *et al.* (1991), Remfry (1978)
Oesophagostomum	*Piliocolobus tephrosceles, Colobus guereza,Cercopithecus ascanius, Macaca arctoides Macaca sinica, Macaca mulatta, Cercopithecus mitis, Macaca fuscata, Papio ursinus, Trachypithecus cristatus, Pan paniscus*	Africa, Asia	Chapman *et al.*(2005), Nath *et al.*, 2012, Remfry (1978), Munene *et al.* (1998), Gotoh (2000), Goldsmid (1974), Palmieri *et al.* (1980), Hasegawa *et al.* (1983)
Physaloptera	*Macaca sinica, Macaca mulatta, Saguinus geoffroyi*	Asia, Neotropics	Dewit *et al.* (1991), Thatcher and Porter (1968)
Plasmodium	*Pan troglodytes, Gorilla gorilla*	Africa	Prugnolle *et al.* (2009)
Schistosoma	*Papio cyanocephalus, Cercopithecus mitis, Papio ursinus*	Africa	Munene *et al.* (1998), Goldsmid (1974)
Spirometra	*Papio cyanocephalus, Presbytis entellus, Macaca mullatta, Saguinus geoffroyi*	Africa, Asia, Neoptropics	Nobrega-Lee *et al.* (2007), Parmar *et al.* (2012), Thatcher and Porter (1968)
Streptopharagus	*Cercopithecus ascanius, Macaca sinica, Macaca mulatta, Papio cyanocephalus, Macaca fuscata, Papio ursinus, Cercopithecus neglectus*	Africa,Asia	Chapman *et al.*(2005), Dewit *et al.* (1991), Munene *et al.* (1998), Gotoh (2000), Goldsmid (1974), Karere and Munene (2002)
Strongyloides	*Piliocolobus tephrosceles, Colobus guereza,Cercopithecus ascanius, Papio cyanocephalus, Cercopithecus aethiops, Cercopithecus mitis, Cercocebus torquatus, Cercocebus albigena, Macaca*	Africa,Asia, Captive	Chapman *et al.*(2005), Muriuki *et al.*, 1998, Dewit *et al.* (1991), Remfry (1978), Legesse *et al.* (2004), Gotoh (2000), Goldsmid (1974), Paramar *et*

	sinica, Macaca mulatta, Macaca fuscata, Papio ursinus, Presbytis entellus, Cercopithecus neglectus, Colobus angolensis, Hylobates syndactylus. Hylobates leucogenys, Pygathrix nemaeus, Erythrocebus patas, Pan paniscus		*al.* (2012), Karere and Munene (2002), Gillespie *et al.* (2005), Levecke *et al.* (2007), Chapter 5, Adedokun *et al.* (2002), Hasegawa *et al.* (1983)
Taenia	*Cercopithecus aethiops, Erythrocebus patas*	Africa	Sulaiman *et al.* (1986)
Trichostrongylus	*Macaca sinica, Macaca mulatta, Papio cyanocephalus, Cercopithecus mitis, Papio ursinus*	Asia, Africa	Dewit *et al.* (1991), Munene *et al.* (1998), Goldsmid (1974)
Trichuris	*Piliocolobus tephrosceles, Colobus guereza,Cercopithecus ascanius, Papio cyanocephalus, Cercopithecus aethiops, Cercopithecus mitis, Cercocebus torquatus, Cercocebus albigena, Trachypithecus geei, Macaca sinica, Macaca hecki, Macaca tonkeana, Macaca fuscata, Papio ursinus, Presbytis entellus, Colobus angolensis, Hylobates concolor, Trachypithecus francoisi, Hylobates hoolock, Erythrocebus patas, Pan paniscus*	Africa,Asia, Captive	Chapman *et al.*(2005), Muriuki *et al.*, 1998, Dewit *et al.* (1991), Jones-Engel *et al.* (2004), Gotoh (2000), Goldsmid (1974), Parmar *et al.* (2012), Gillespie *et al.* (2005), Levecke *et al.* (2007), Liu *et al.* (2013), Nath *et al.* (2012), Adedokun *et al.* (2002), Hasegawa *et al.* (1983)
Trypanosoma	*Macaca silenus, Lemur catta, Saimiri sciurius, Macaca mullatta*	Captive, Asia, Neotropics	Pung *et al.* (1998), Ziccardi and Lorenco-de-Oliveira (1997), Fulton and Harrison (1946)
Trypanoxyuris	*Alouatta villosa, Aotus trivirgatus, Ateles fusciceps, Ateles geoffroyi, Saguinus geoffroyi,*	Neotropics	Thatcher and Porter (1968)