Mammal diversity across a human-modified tropical landscape in Borneo: a study based on the metabarcoding of terrestrial leeches

Rosalind Drinkwater

Supervisors: Prof Stephen J. Rossiter and Dr Elizabeth Clare

School of Biological and Chemical Sciences
Queen Mary University London
Mile End Road, London, E1 4NS

Thesis submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy, September 2018

Author's Declarations

I, Rosie Drinkwater, confirm that the research presented in this thesis is a product of my own work, and, that where work has been carried out in collaboration with others, this is appropriately acknowledged.

I attest that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law, contain confidential material, or infringe upon any third party's intellectual property. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author, and no material within it may be published without the prior written consent of the author.

This work is part of the Human Modified Tropical Forest (HMTF) research programme funded by the Natural Environment Research Council (NERC), with additional support from a Study Abroad Studentship awarded by The Leverhulme Trust, and a grant awarded by the Queen Mary Postgraduate Research Fund.

Details of collaborations and data acquisition:

- Twelve of the DNA sequences in my reference database, used in all data chapters, were generated by Prof Géraldine Veron of the Muséum National D'Histoire Naturelle.
- For Chapter 3 and Chapter 4, LiDAR data were provided by Dr Tom Swinfield and Prof David Coomes of the University of Cambridge.
- High throughput sequencing was undertaken by researchers at The Bart's and The London Genome Centre at Queen Mary University London or at the National High-throughput DNA Sequencing Centre at the University of Copenhagen.

Signature:

Date:

Abstract

Tropical forests are under intense anthropogenic pressure from activities such as logging, land conversion and unregulated fires. These practices are driving global forest loss and degradation, which have been particularly intense on the island of Borneo. Human-modified landscapes dominate Borneo and the need to assess and monitor the diversity of degraded forests is now of critical importance. The use of invertebrate-derived DNA (iDNA) sampling techniques is gaining popularity. In this approach, invertebrate blood meals are sequenced to identify the vertebrate hosts on which they have been feeding. In this study I applied an iDNA method to survey mammalian diversity in a degraded landscape in Sabah, Malaysian Borneo. I focused on two species of haematophagous terrestrial leech (*Haemadipsa picta* and *H. sumatrana*), which I sampled from primary forest and degraded forest sites of varying quality. First, I investigated the difference between the diet of the two focal leech species and show that *H. picta* detects a greater diversity of mammalian taxa. My findings emphasise the need to understand the ecology of the invertebrate sampler and the biases it introduces. Using *H. picta* I then conducted an in-depth analysis into the differences in mammalian diversity across a human-modified gradient over two years. Finally, I applied a hierarchical occupancy modelling framework to iDNA detection data to incorporate imperfect detections at two levels. This is a novel application of a classical statistical framework and to my knowledge the first study to do so. Although there are developments which need to be made, my results show the importance of accounting for imperfect detection to gain a nuanced understanding of species detections. Overall, I conclude that leech iDNA is a promising new sampling technique for mammals, but its use as a standard tool for conservation monitoring will rely on lowering sequencing costs and improved reference databases.

Acknowledgements

My biggest thanks go to my supervisors, Steve Rossiter and Beth Clare, for giving me so many amazing opportunities over the past few years and for all their endless support and guidance, even when I thought it was not possible. I would also really like to thank Steve Le Comber for all his academic and emotional support.

I am grateful for all the fieldwork in Sabah, and there are so many people to thank for making that possible. Firstly, my thanks go to Eleanor Slade, Matthew Struebig and Owen Lewis for their tireless effort in the LOMBOK project, and I want to thank Henry Bernard for his collaboration throughout. I also want to thank Tom Gilbert for hosting me for a year in Copenhagen. This was an amazing learning experience, and I also want to thank Kristine Bohmann, Ida Schnell, Martin Nielsen and all the technical staff for their advice and support while I was there.

My thanks go to the Sabah Biodiversity Council, Sabah Forestry Department, Benta Wawasan, Sime Darby, Yayasan Sabah, Danum Valley Management Committee and Maliau Basin Management Committee for given me permission to conduct my research in their forests. I am also grateful to Glen Reynolds and the South East Asian Rainforest Research Partnership for facilitating this research.

I have had the chance to work alongside so many amazing researchers and research assistants in the field. There are too many people to mention, but I have made so many great friends and the whole thing would have been much more difficult without them. In particular, I need to thank all the absolutely amazing LOMBOK research assistants for all their hard-work collecting leeches for mama pacat, in particular Noy, Lizzie, Loly, Kiky, Didy, Anis, Mudin, and especially to Unding for organising everyone and everything. My thanks also go to the amazing Ryan Gray SAFE camp would not have been the same without you.

My thanks go to all the PhD students and postdocs who have supportive throughout this PhD. I especially want to thank Tor Kemp for being the best jungle roommate, flatmate, yoga pal, ping pong partner and friend since day one! I have been so lucky to work in an amazing lab group, keeping me sane in the office. My

thanks go to: Dave Hemprich-Bennett, Tiago Teixeira, Josh Potter, Sandra Alverez, Ilya Levantis, Joe Williamson, Kim Warren, Tim Penny, Omar Khalilur Rahman, Hernani Oliveira, Georgia Tsagkogeorga, Michael McGowen, Nicolas Nesi and James Gilbert. Special thanks to Kalina Davies for all her amazing guidance and patience with me in the early days. My thanks also go to Monica Struebig, Phil Howard and Chloe Economou for all their support in the lab.

I want to thank all my wonderful friends for standing by me for all these years, even when the PhD took over and for the amazing adventures we have had: to Tessa Thomson, Emi Takahashi, Tilly Lenartowicz, Izy Neatrour, Storm Patterson, Steph Hogarth, Hari Holdsworth, Ed Mountjoy, Barney, Cam Bailey, Jamie Grundy, Nik Willi, Mark Smith, Alex Williamson and Eoghan MacManus. I am so proud of us all and I am so grateful to have such a supportive group of friends. I also want to thank Jamie Williamson, Allie Colaço and Karen Bosworth for looking after me in Copenhagen - I wouldn't have survived it without you.

Finally, my greatest thanks go to all the Drinkwater clan for their endless and unwavering support, to Fliss, Bob, Roxie and Anne-Marie for believing in me and to Tom and Amy for looking after me and letting me stay whenever I needed it. I could not have finished this without you all. I would not be where I am today without the help of my Dad, who has supported me mentally and financially for so many years while I ran around in jungles, and never once doubted me. I am also dedicating this thesis to my mum, who would have been so proud of everything I have achieved and would have thought it was hilarious I have ended up collecting leeches for four years.

Table of contents

List of abbreviations

AIC - Akaike Information Criteria

DVCA - Danum Valley Conservation Area

eDNA - Environmental DNA

GLM - Generalised linear model

GLMM - Generalised linear mixed effects model

HTS - High Through-put Sequencing

iDNA - Invertebrate derived DNA

LiDAR - Light Detection and Ranging

MCMC - Markov Chain Monte Carlo

NMDS - Non-Metric Multidimensional Scaling

OTU - Operational taxonomic unit

PCR - Polymerase Chain Reaction

PERMANOVA - Permutational Analysis of Variance

PPL - Posterior predictive loss

SAFE - Stability of Altered Forest Ecosystems project

WAIC - Watanabe-Akaike Information Criteria

List of figures

List of tables

Chapter 1: General Introduction

1.1. Biodiversity in human modified landscapes

1.1.1. Global trends

Tropical ecosystems cover 40% of the Earth's surface (Barlow *et al.* 2018) and make up vast regions of the Americas, Central Africa and Southeast Asia. Overall, tropical forest accounts for approximately 50% of total forest cover (Pan *et al.* 2011). These forests are crucial to the global carbon cycle (Pan *et al.* 2011), climate regulation, and primary production (Malhi 2012), and provide many other important ecosystem functions and services including the provision of natural resources such as timber and food (Toledo *et al.* 2003). High rates of primary productivity in the humid tropics allow these habitats to support higher levels of biodiversity than in other terrestrial biomes, and this is seen across broad groups such as trees (Slik *et al.* 2015), arthropods (Basset *et al.* 2012) and vertebrates (Schipper *et al.* 2008). Species endemism is also particularly high in tropical forests; for example, there are six times the number of endemic birds in the tropics compared to temperate forests (Barlow *et al.* 2018). Tropical forests are important not only for the hyper-biodiversity found within them, but also their roles as global carbon stores (Malhi 2012), which is of great importance in the context of current climate change. Deforestation causes the release of large amounts of carbon dioxide (CO_2) into the atmosphere and so the protection of tropical forests will be critical in the mitigation of global warming (Sullivan *et al.* 2017).

Very few habitats on Earth have escaped from human interference, and as human populations rise so does the demand for space and resources (Newbold *et al.* 2015). The driving forces behind tropical forest deforestation and degradation are complex and interlinked, but the most pervasive influences arguably are timber extraction, and land-use change for plantations and pasture (Malhi *et al.* 2013), practices which are widely sanctioned by governments but are also conducted through illegal means. Tropical forest loss is increasing by 2101 km^2 per year (Hansen *et al.* 2013) and 70% of all forests are now within 1 km of a forest edge (Haddad *et al.* 2015). Alongside this conversion there is increasing fragmentation

of the once intact forest, and the associated generation of forest edges also leads to changes in ecosystem functions (Fletcher *et al.* 2018) and opens up the forest for greater exploitation (Laurance *et al.* 2009). It has recently been shown that 85% of all forest vertebrates are in some way affected by fragmentation and edge effects, with often the most detrimental effects being felt by species of greater conservation concern (Pfeifer *et al.* 2017). Forest edges also effect species composition and turnover, with dramatically different communities found at the forest edges compared to forest interiors (Pfeifer *et al.* 2017). Understanding the effects of forest degradation on biodiversity is confounded by the historic legacy of land-conversion and shifting baselines (Lewis *et al.* 2015) which result in extinction debts and time-lagged CO_2 emissions being felt into the future (Rosa *et al.* 2016). Predictions for the future of tropical forests look particularly bleak, with the various detrimental effects of logging, climate change and overfishing for example, being compounded by socio-economic factors such as lack of governance and rapid market growth (Barlow *et al.* 2018).

1.1.2. Human-modified tropical landscapes

Human-modified landscapes now dominate the tropics and are characterised by a mosaic of different land use classes which often include protected areas, degraded forests, agricultural plantations and pastures. Within these landscapes, the conservation of intact primary rainforest is critically important in the maintenance of biodiversity and ecosystem functions (Gibson *et al.* 2011). This is often achieved through assigning forests as protected areas, and while there may be issues with illegal hunting and weak governance, in general protected areas world-wide have higher biodiversity than unprotected land (Gray *et al.* 2016). However, as large amounts of tropical forests have been degraded in some way (Hansen *et al.* 2010), it is crucial that these landscapes now play a role in the conservation of biodiversity (Putz *et al.* 2012). Degraded landscapes are heterogenous and difficult to define but can be characterised in some way by a gradient of deforestation, fragmentation due to roads and cleared areas and increased exploitation of natural resources (see Figure 1.1.) (Putz & Redford 2010). These degraded forests are also typically characterised by over-harvesting of timber, hunting and altered natural

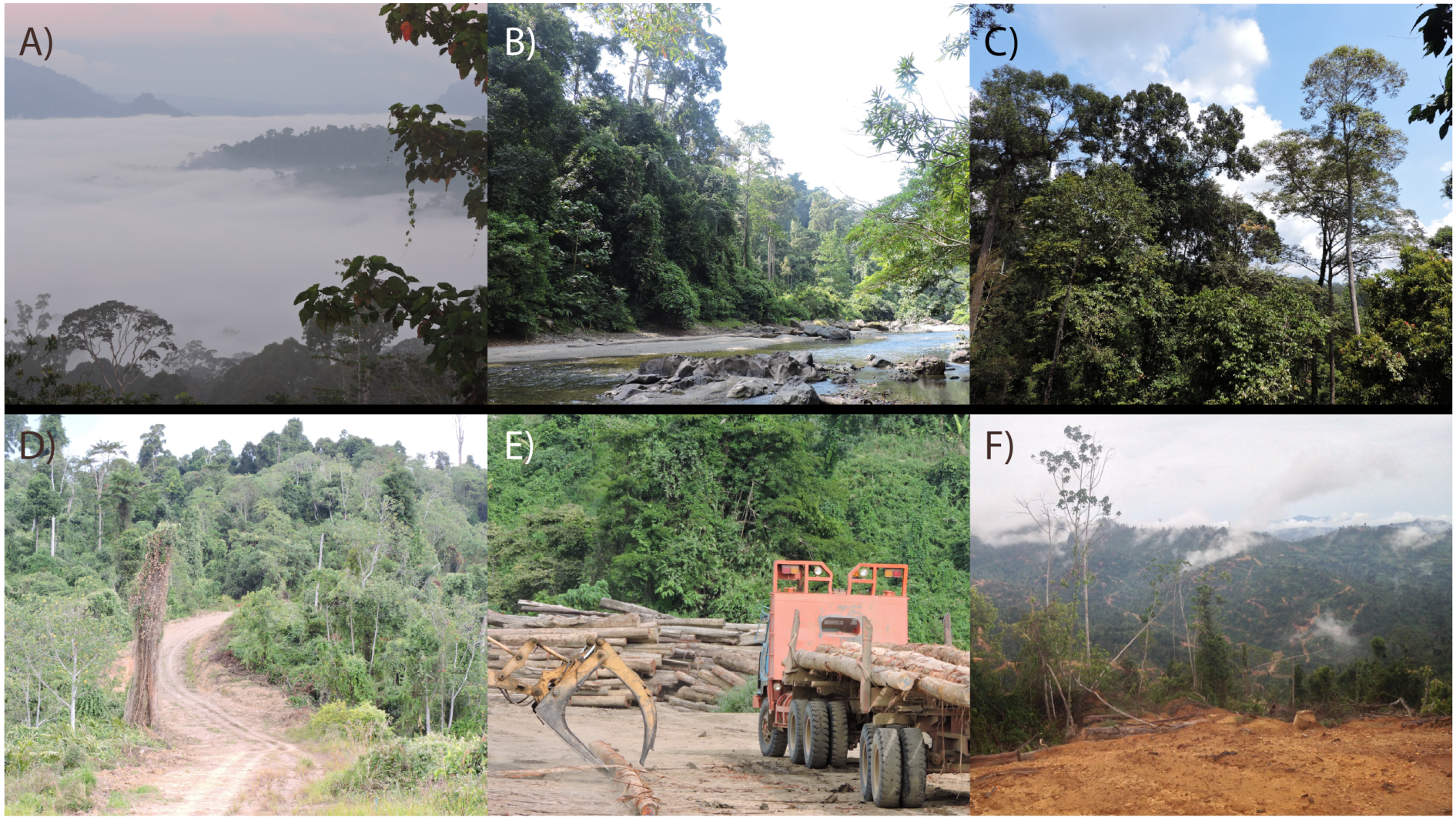


Figure 1.1. Top - Photographs showing primary rainforest at Danum Valley, Sabah, A) sunrise over the canopy, B) the Segama river, C) canopy showing emergent trees. Bottom - Degraded forest at the SAFE project, D) road cut through logged forests, E) cut timber and trucks at a sawmill, F) freshly cleared timber landing site (photo credit: R. Drinkwater).

fire regimes and different ecological communities of species compared to old growth habitats (Gardner *et al.* 2009).

1.1.3. Logging and land-conversion

Logging and land-use conversion for agriculture are perhaps the biggest drivers of forest degradation (Malhi *et al.* 2013). One logging rotation can cause collateral damage to the surround forests through the felling of trees, opening of roads and storing of timber (Putz *et al.* 2012). This can be through selective logging where only trees above a minimum trunk diameter are extracted compared to clear-felling, which involves the removal of all trees (Edwards *et al.* 2014). In general, the intensity of logging can have varying effects on biodiversity (Burivalova *et al.* 2014), and multiple rounds of logging continually degrade the forest, until it is deemed of little value. At this point, salvage logging takes place which includes the extraction of any valuable timber and then land-conversion often begins. There are methods of timber extraction which aim to reduce collateral damage during logging, termed reduced impact logging (RIL). RIL practices include: pre-felling inventories, vine-cutting to reduce the number of associated trees which are pulled down, directional felling and limits on skid trails and landing sites, among other activities (Edwards *et al.* 2012b). Adopting RIL can have lower impacts on diversity compared to a reduction in logging intensity alone (Bicknell *et al.* 2014). In the absence of high levels of hunting, forests that have only undergone one or two rounds of logging are still able to support biodiversity levels that are similar to those of primary forests (Putz *et al.* 2012). In comparison, other land-use categories such as plantations support much lower biodiversity (Edwards *et al.* 2014). Converting degraded forests to agricultural land is likely to cause a bigger loss in biodiversity compared to the initial round of logging (Gaveau *et al.* 2016). For example, mammalian species richness is lower in oil-palm dominated landscapes compared to riparian forest (Pardo *et al.* 2018) and selectively logged forest (Wearn *et al.* 2017).

1.1.4. Built infrastructure and roads

Forest degradation also occurs with built infrastructure in human-modified landscapes, for example the majority of deforestation in the Amazon occurs within

5.5km of a road (Barber *et al.* 2014). The negative effects of roads on biodiversity can be direct, such as through individual mortalities caused by collisions with vehicles, and indirect, such as through the creation of barriers to natural dispersal as well by facilitating the movement of hunters and invasive species (Laurance *et al.* 2009). Additionally, large roads are often the catalyst for opening up the forest, and lead to a proliferation of smaller roads that penetrate further into intact forests (Barber *et al.* 2014).

The impact of hunting associated with logging varies across taxa, with more detrimental consequences seen in groups such as ungulates and primates (Brodie *et al.* 2014b). Poaching for the illegal wildlife trade is an extremely lucrative business, decimating populations of Southeast Asian birds and mammals, for example (Wilcove *et al.* 2013). Hunting and the removal of vertebrates can alter ecosystem functions such as by reducing seed dispersal (Brodie *et al.* 2009), although the full impact of losing these species on seed dispersal and regrowth may take decades to come to light (Brodie *et al.* 2009).

1.1.5. Value of degraded forests

The sensitivity and resilience of forest taxa to habitat degradation has been shown to depend on aspects of their ecological traits. Generally, more specialist species with narrow niches are at greater risk from the impacts of forest degradation and fragmentation than generalists (Pfeifer *et al.* 2017). For example, birds with more specialised dietary niches are less abundant in logged forests (Edwards *et al.* 2013) and the same trends have also been reported for mammals (Wearn *et al.* 2017). However, there is now an appreciation that logged and degraded forests hold considerable conservation value and as such need protecting alongside primary forests for the safeguarding of unique and important biodiversity (Berry *et al.* 2010; Lewis *et al.* 2015; Costantini *et al.* 2016).

1.1.6. Conservation management and policy regarding human-modified forests

Due to its important roles in supporting biodiversity and storing carbon, tropical forests are the focus of numerous conservation policies at local, regional, national and international levels. Indeed, the United Nations recognises that global

deforestation and degradation are responsible for considerable carbon emissions, and, to mitigate the detrimental effects of these emissions, developed the "Reducing emissions from deforestation and forest degradation and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries" (REDD+) scheme. Specifically, REDD+ aims to financially incentivise developing countries to protect the carbon stored in their forests through various guidelines and policies (United Nations 2018). Importantly, this programme also references the conservation of biodiversity and, as such, opens up opportunities for conserving areas of degraded tropical forest that are both high in carbon and in biodiversity (Paoli *et al.* 2010). However, the benefits of REDD+ for biodiversity might not be so clear where structural definitions of forest are used, which encompass agroforestry and/or plantations (Harvey *et al.* 2010). This programme may also overlook, and thus afford less protection to, habitats that are characterised by low-carbon and high diversity such as the florally-diverse Cerrado region of Brazil (Paoli *et al.* 2010). Despite this, co-benefits between high carbon stock forest and vertebrates diversity have been found, particularly for threatened species (Deere *et al.* 2017).

Another popular scheme used in forest management is "green labelling" or the certification of sustainably harvested products (Edwards *et al.* 2012a). One example of this is the use of the 'High Conservation Value' (HCV) classification approach, which has been adopted by the Forestry Stewardship Council (FCS) and the Roundtable on Sustainable Palm Oil (RSPO) in assigning certification (Senior *et al.* 2015). For certification, land managers are required to perform HCV assessments and to retain high-conservation forests based on six core values, four of which are focused on biodiversity (Senior *et al.* 2015). The responses of groups such as birds and mammals can be used as indicators of compliance to certification guidelines. However, the benefits to biodiversity of the HCV approach have been questioned when considering the sustainable certification of agriculture as opposed to forestry (Edwards *et al.* 2012a).

1.2. Southeast Asian Island of Borneo, a biodiversity hotspot

1.2.1. The forests of Borneo

Southeast Asia is one of the tropical regions facing the highest relative rates of deforestation and degradation (Sodhi *et al.* 2004). The area contains four biodiversity hotspots, areas with high levels of species richness and endemism but that are also experiencing high levels of habitat loss (Myers *et al.* 2000). The biodiversity hotspot of Sundaland stretches from Peninsular Malaysia and Sumatra to the island of Borneo, and is particularly imperilled by habitat loss (Wilcove *et al.* 2013).

Borneo's lowland rainforests once covered vast areas of the island. These forests are dominated by tree species belonging to the highly speciose family Dipterocarpaceae, and also support unique animal diversity, distinguishing this island (and Southeast Asia) from other tropical regions (Corlett 2007). The fauna of Borneo includes a large number of endemic and specialist taxa which may be especially vulnerable to logging and poaching (Meijaard & Sheil 2008). Among mammals, these taxa include charismatic primate species such as the Bornean orangutan (*Pongo pygmaeus*) and red leaf monkey (*Prebytis rubicunda*), ungulates such as the yellow muntjac (*Muntiacus atherodes*), and carnivore species like the rare bay cat (*Catopuma badia*) and Hose's civet (*Diplogale hosei*). Other mammalian groups also show high levels of diversity; for example, Bornean forests support high alpha diversity of bats, as well as rodents, including squirrels from three major clades, i.e. giant, nocturnal and diurnal (tree and ground-dwelling) squirrels (Corlett 2007).

1.2.2. Threats to Bornean forests

Much of Borneo's unique fauna and flora is under threat from the effects of land-use change. Logging is an especially lucrative business on Borneo and, it has been estimated that over 30% of forest cover was lost between the 1970s and 2010 (Gaveau *et al.* 2014). Among the three nations that govern Borneo, there are differences in the percentage cover of forest that remains. The small, oil-rich nation of Brunei shows least percentage forest loss due its low economic reliance on logging (Bryan *et al.* 2013). In contrast, forest loss has been focused in

Malaysian Borneo (Sabah and Sarawak) and Indonesian Borneo (Kalimantan). Within Sabah and Sarawak, the majority of intact forest occurs within a few protected areas, with the remaining forest degraded or severely degraded (Bryan *et al.* 2013). Much of these degraded forests fall within designated production forests and are encroached by logging roads, and, in many cases, destined for future land conversion (e.g. Figure 1.1., Gaveau *et al.* 2014). Poaching for bushmeat and the wildlife trade is also high, notable Bornean examples include the trafficking of pangolins (Heinrich *et al.* 2016), helmeted hornbill 'ivory' casques (Beastall *et al.* 2016) and Sumatran rhinoceros horns (Havmøller *et al.* 2016) pushing these species towards extinction.

In recent decades, most forest clearance in Borneo has been for the expansion of agriculture, particularly for oil palm (*Elaeis guinensis*). Malaysia and Indonesia are the largest producers and exporters of palm oil globally (Food and Agriculture Organisation UN 2017). Oil palm plantations are especially problematic for forest-dependent wildlife because they are monocultures characterised by increased temperatures and reduced humidity (Hardwick *et al.* 2015). Compared with forest, these plantations tend to host much lower levels of diversity, (e.g. for birds, Edwards *et al.* 2010; and for small carnivores, Jennings *et al.* 2015) and the resilient taxa represent different assemblages (e.g. for dung beetles, Gray *et al.* 2014; and freshwater fish, Giam *et al.* 2015). Previously, the oil palm industry has asserted that new plantations were developed mainly on old cropland, yet between 1990 - 2005 over 50% of oil palm plantation expansion happened on forested land (Koh & Wilcove 2008). While sustainable practices surrounding oil palm agriculture have improved, for example, via certification schemes introduced by the Roundtable on Sustainable Palm Oil (RSPO), forests left degraded from logging are wrongly assumed to hold low conservation value and therefore are at a serious risk of conversion to agricultural plantations (Berry *et al.* 2010).

The impacts of habitat degradation on tropical forest species may be made worse by climate change. Tropical forests were previously thought to be buffered from the worst effects of increasing global temperatures, however, increasingly, the impact of tropical forest degradation needs to be viewed in the context of climate

change, where warming temperatures, increasing droughts and fires will affect species as they need to adapt or migrate to survive (Scriven *et al.* 2015). Human-modification puts Borneo's forests at greater risk of extreme dry seasons, fires and droughts (Brodie *et al.* 2012). Thus understanding how climate change will impact already vulnerable forests is critical for effective conservation planning (Struebig *et al.* 2015).

1.2.3. Biodiversity monitoring

To better predict the impacts of this human-driven forest degradation on biodiversity, species and populations need to be monitored and integrated with landscape-scale spatial data from satellites (Sodhi *et al.* 2010). Conservation and management strategies are most effective where there is a comprehensive understanding of species abundance and distribution, incorporating a cycle of evaluation and implementation (Nicholson *et al.* 2012). The resulting biodiversity data are used to inform and evaluate policies, such as those within the REDD+ framework and national government plans (e.g. for Malaysia; Ministry of Natural Resources and Environment 2016). Biodiversity monitoring also provides valuable data on species ranges and threats, which are then used to inform the conservation status classifications of species under the umbrella of the International Union for Conservation of Nature and Natural Resources (IUCN). Unfortunately, many vertebrate species remain classified as Data Deficient (DD) (IUCN 2001) and, without sufficient monitoring, such taxa are at risk of population declines and extinction before conservation actions can take place (Schipper *et al.* 2008)

1.2.4. Large-scale biodiversity experiments

To gain comprehensive insights into the effects of habitat degradation on tropical ecosystems - including biodiversity, biogeochemical processes, and ecosystem functioning - large-scale ecological experiments have been established. Two key examples include the Biological Dynamics of Forest Fragmentation Project (BDFFP) in Brazil, and the Stability of Altered Forest Ecosystems Project (SAFE) in Malaysian Borneo. BDFFP was established in the 1980s to investigate the relationship between minimum fragment size and ecosystem functioning (Bierregaard *et al.* 1992). Studies at this site have revealed important and varying

effects of fragmentation and the creation of edges on many taxonomic groups (Laurance *et al.* 2002).

In 2010, the Stability of Altered Forest Ecosystems (SAFE) project was established in Sabah, Malaysian Borneo, as a fragmentation experiment to investigate the effects of on-going land-use change (Figure 1.2.) (Ewers *et al.* 2011). The SAFE project is located within twice-logged forest which is now undergoing salvage logging before land-conversion, including the terracing of land and planting of new oil palms (Turner *et al.* 2012). In collaboration with multiple stakeholders, areas of degraded forest, in a replicated experimental design, are being set aside for research (Figure 1.2, Ewers *et al.* 2011). To date, there are many studies arising from the SAFE project documenting and recording the effects of forest degradation on diversity and ecosystem processes. For example, the conservation value of degraded forests has been demonstrated for taxonomic groups such as birds (Mitchell *et al.* 2017), mammals (Deere *et al.* 2017; Wearn *et al.* 2017), fish (Wilkinson *et al.* 2018), invertebrates (Gray *et al.* 2014; Ewers *et al.* 2015), and frogs (Konopik *et al.* 2015). Similarly, in this degraded landscape, negative changes post-logging have been reported in freshwater habitats (Luke *et al.* 2017), microclimate (Hardwick *et al.* 2015), carbon (Pfeifer *et al.* 2015) and primary productivity (Riutta *et al.* 2017). These large-scale studies allow for continuous and repeated monitoring to take place in heterogeneous degraded landscapes at multiple scales.

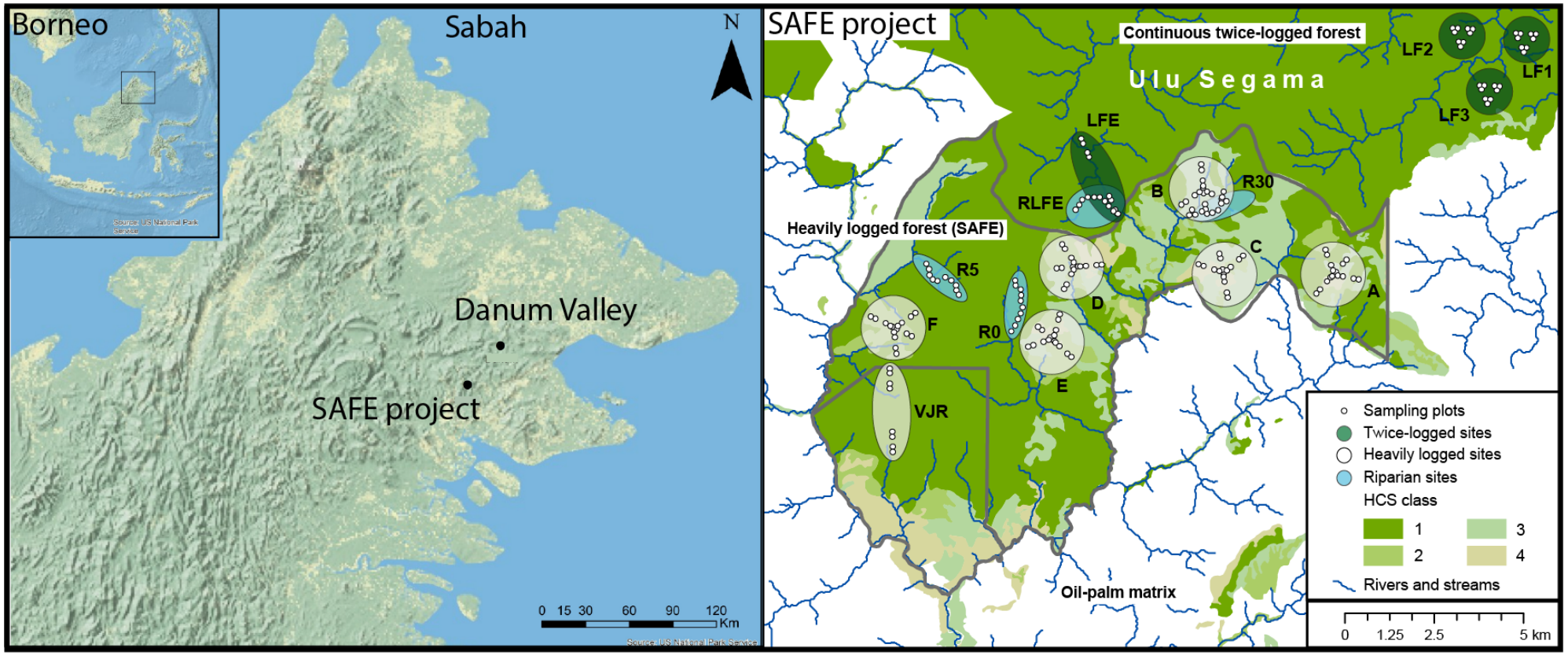


Figure 1.2. Maps of the field site. Left side showing the location of the Malaysian state of Sabah in North Borneo, with the locations of the SAFE project (Ewers *et al*. 2011) and Danum valley primary forest shown by black circles, inset map shows the location of Sabah on the island of Borneo. The map on the right side is a detailed schematic of the sampling design at the SAFE project used in this research. Twice-logged forest sites are shown by large dark green circles within the continuous forest (the Ulu Segama), heavily logged forest sites are shown as large white circles within the SAFE experimental area surrounded by oil palm matrix and the riparian sites are shown by blue ovals within the SAFE area. Locations of the vegetation plots, within each habitat type, are show by small white circles and rivers are shown with blue lines. The classification of the forest in terms of high carbon stock (HCS) is shown by the gradient of green shading. Adapted from LOMBOK consortium maps.

1.3. Non-invasive molecular sampling

1.3.1. Metabarcoding and environmental DNA

DNA barcoding was first proposed in the early 2000s as a method to speed up taxonomic inventories and identification of species. For animals, barcoding most commonly entails sequencing a part of the cytochrome c oxidase 1 (COI) gene, with the original idea that this would become a standard locus for species-level identification (Floyd *et al.* 2002). DNA barcoding is usually applied to high quality DNA samples for the identification of single species, and to date has typically relied on Sanger sequencing (Taberlet *et al.* 2012b). However, the development of high throughput sequencing technologies (HTS) has provided the means to simultaneously barcode multiple individuals from mixed samples (Taberlet *et al.* 2012b). This so-called 'metabarcoding' technique has been applied to many different ecological situations and for identifying different taxa (Yu *et al.* 2012; Pochon *et al.* 2017; Aizpurua *et al.* 2018).

One of the most powerful applications of metabarcoding is for identifying species from environmental DNA (eDNA), defined as DNA sequences derived from the environmental samples such as water or soil, without the isolation of individuals (Taberlet *et al.* 2012a; Deiner *et al.* 2017). Although this method was originally used in microbiology, sequencing eDNA was also applied to the identification of plant or animal DNA from different diverse substrates (e.g. permafrost sediment, Willerslev *et al.* 2003; water, Ficetola *et al.* 2008). Since then, metabarcoding has been used for sequencing eDNA to identify whole communities from many environmental samples, including seawater (Sigsgaard *et al.* 2017), freshwater (Thomsen *et al.* 2012), soil (Andersen *et al.* 2012) and even from pitcher plants (Littlefair *et al.* 2018). These techniques can also be used to identify invasive species (Jerde *et al.* 2013) and for both ancient and modern eDNA (Pedersen *et al.* 2015). Metabarcoding has also been instrumental in the sequencing of dietary samples (Pompanon *et al.* 2012) and bulk arthropod samples, such as those obtained from malaise traps (Yu *et al.* 2012). By metabarcoding DNA recovered from faecal samples or stomach contents, it has been possible to gain a greater understanding of species interactions (e.g. Bell *et al.* 2017; McInnes *et al.* 2017; Arrizabalaga-Escudero *et al.* 2018).

One of the many benefits of sampling with environmental DNA is that the target species does not need to be confirmed with audio or visual methods (e.g. through point counts or camera traps). This can be especially important for monitoring species of conservation concern which tend to be rare and challenging to monitor (Thomsen *et al.* 2012). It is also a technique increasingly being used to identify invasive species, pinpoint invasion fronts, and develop early warning systems (Comtet *et al.* 2015).

Recently, the metabarcoding of haematophagous invertebrate blood-meals has been proposed for use in wildlife monitoring in conjunction with other techniques such as camera traps. For example, the World-Wide Fund for Nature has begun collection of terrestrial leeches in Vietnam and Laos, alongside their camera trapping campaigns and scat surveys (WWF 2013). They hope that this molecular technique might aide in the detection of the saola (*Pseudoryx nghentinhensis*) a Critically Endangered antelope, thought to inhabit the forests in the Annamite Range but has evaded detection for over 20 years (WWF 2013). Even though the molecular analyses have not yet revealed the presence of Saola DNA, the initial study resulted in the detection of other extremely rare and newly discovered native species important for the conservation of this area (Schnell *et al.* 2012).

1.3.2. Known limitations with metabarcoding

Despite the popularity of molecular biodiversity surveys, there are limitations associated with metabarcoding that apply to many of the different types of sample (e.g. bulk arthropod, environmental, dietary). In particular, the ability to draw robust conclusions relies on an understanding of how mixed DNA samples behave (Barnes & Turner 2015). For example the amount of DNA and the rate at which it is shed in the environment, and its rate of subsequent degradation, will all vary among species (Deiner *et al.* 2017). Additionally, when using dietary samples, differential digestion must be considered (Clare 2014). To reduce the impacts of some known issues on the resulting taxonomic assignments and conclusions, various steps can be taken (described below).

Contamination

Environmental and dietary DNA samples typically consist of short DNA sequences at low copy numbers, and thus are at great risk of contamination both in the field and in the laboratory. This poses a problem because normally the contents of mixed samples are unknown and, as such, DNA contaminants can erroneously be recorded as "true" diversity (false positives) resulting in inflated estimates (Bohmann *et al.* 2014). To reduce this risk, many eDNA studies have adopted field and laboratory protocols similar to those used for ancient DNA sequencing. For example, steps taken can include strict cleaning of field equipment, field blanks (negative controls), physical separation of field sites and samples (e.g. US Fish and Wildlife 2013; Thomas *et al.* 2018), the use of clean (PCR-free) laboratories, sample blanks, and matched sample tagging (Pedersen *et al.* 2015; Schnell *et al.* 2015b)

Primer bias

The choice of primers can introduce additional biases in mixed samples, in which the target template DNA is often rare. This can lead to the preferential amplification of more abundant sequences (Elbrecht & Leese 2015) or the amplification of a particular taxonomic group (Alberdi *et al.* 2017). Where DNA is particularly degraded, primers may preferentially amplify higher quality DNA fragments (Deagle *et al.* 2006), such as human DNA from researchers. Finally, although primers that amplify short DNA fragments are usually needed for metabarcoding of fragmented eDNA, these unavoidably limit taxonomic resolution (Pompanon *et al.* 2012). These issues are especially relevant to studies of taxa such as arthropods, where sequence diversity can be very high, and where reference sequences are not always available. One way to improve taxonomic resolution and reduce primer bias is through the use of multiple loci.

Reference databases

Identifying species based on sequences from environmental or mixed samples relies heavily on access to comprehensive and accurate reference databases, however, this is not always possible (Blaxter 2016). Most often, reference sequences are retrieved from GenBank, National Centre for Biotechnology

Information (NCBI) (Benson *et al.* 2013), however, this is an un-curated repository, which if used unfiltered can lead to erroneous taxonomic assignments (Mioduchowska *et al.* 2018). For DNA barcoding (and metabarcoding) data, the Barcode of Life Database (BOLD, Ratnasingham & Hebert 2007) represents a second source of reference sequences, although this only includes the standardised barcoding region, COI. However, for studies of broad taxonomic scope, it is highly likely that some species will be missing from existing reference sequence databases. In such cases, new sequences must be generated (Mohd Salleh *et al.* 2017) and taxonomic resolution should be moderated to a higher classification, e.g. genus, where databases remain incomplete (Kocher *et al.* 2017b).

Bioinformatics and parameter choices

As a consequence of the huge amount of sequence data generated using HTS, bioinformatic pipelines are used for filtering and quality checks on the sequence data. There are many decisions applied at the quality control stage which will have an impact on the diversity outcomes (Alberdi *et al.* 2017). Conservative filtering risks the removal of rare sequences but if the filtering is not strict enough, sequencing error can be retained (De Barba *et al.* 2014). Filtered sequences can then be clustered into operational taxonomic units (OTUs), sometimes referred to as molecular OTUs (MOTUs) in genetics studies, to reduce error from sequencing artefacts (Floyd *et al.* 2002; Alberdi *et al.* 2017). Generating OTUs involves clustering sequences together based on similarity thresholds in order to generate comparable units of diversity instead of assigning traditional species identification (Clare *et al.* 2016). Using a threshold that is too low will result in OTUs being under-split, such that too few OTUs are generated, potentially containing sequences from different taxa Grouping multiple taxa in this way will mask the true sample diversity. On the other hand, if a threshold is set that is too high, the OTUs will be over-split, i.e. there will be too many OTUs designated, which will artificially inflate the apparent diversity of the sample (Clare *et al.* 2016).

1.3.3. Invertebrate-derived DNA and invertebrate samplers

One source of DNA that has been used for metabarcoding is the blood-meals of haematophagous insects or other vertebrate-feeding invertebrates (Schnell et al.

2015). The detection of such DNA has been termed invertebrate derived DNA (iDNA), and it has been used previously to investigate host specificity and disease vectors (Malmqvist *et al.* 2004; Konnai *et al.* 2008; Kent 2009) and, more recently, to quantify biodiversity (Schnell *et al.* 2012; Calvignac-Spencer *et al.* 2013b). Various invertebrates have been tested for their utility as samplers of biodiversity, for example terrestrial leeches (Weiskopf *et al.* 2017; Tessler *et al.* 2018), carrion flies (Schubert *et al.* 2015; Rodgers *et al.* 2017), dung beetles (Gómez & Kolokotronis 2016), mosquitoes and sand-flies (Kocher et al. 2017b; Kocher, et al. 2017c).

Sampling with iDNA could provide several key additional benefits for surveying difficult to spot/trap vertebrates. First, field expenses for invertebrate sampling tend to be low, and invertebrate traps can be hand-made and very little specialist equipment is needed, compared to, for example, camera traps (see method comparison in Lee *et al.* 2016; Weiskopf *et al.* 2017). Also, expert taxonomic expertise is also not needed in the field, which some sampling techniques rely on (e.g. point counts for birds, Mitchell *et al.* 2017). The reduction of field costs increases the potential scope of iDNA studies, for covering both a wide geographic area and for maximising temporal sampling efficiency. For example, Schnell *et al.* (2018) sampled across most of the haemadipsid leech range, across the Palaeotropics detecting diversity across multiple vertebrate classes. Additionally, Weiskopf *et al.* (2017) found comparable richness using leech iDNA compared to camera traps in only 12 days sampling with leech iDNA. While a generalist diet is an ideal characteristic for a non-specific vertebrate sampler, especially for surveying broad diversity of a particular area and also for the detection of rare species (Schnell *et al.* 2012), another approach is to target detections towards specific species. For example, Schubert *et al.* (2015) developed an assay to detect the sooty mangabey (*Cercocebus atys*) using carrion-fly iDNA. Aside from their use in detecting species, carrion-flies are often the first to monopolise carcasses, and thus carrion-fly iDNA could potentially be used to look at mortality rates and mass die-off events (Calvignac-Spencer *et al.* 2013a).

One of the drawbacks of sampling with iDNA is that the time of feeding is unknown. Unlike some other methods, where the time or day of detection is known, such as through a time-stamped photo, or a point count survey, there could be a large discrepancy between the times of invertebrate feeding and capture, which may influence conclusions regarding biodiversity. To try and understand this, preliminary controlled experiments on rates of DNA degradation have been conducted. For leeches (using the medicinal leech - *Hirudo medicinalis*), Schnell *et al.* (2012) found that goat DNA could still be detected after 3-4 months post-feeding, and for the DNA of mammalian viruses, Kampmann *et al.* (2017) found detectable levels 50 days post exposure. In contrast, however, DNA was found to persist in blow fly guts for less than four days (Lee *et al.* 2015). As well as these questions regarding DNA persistence, the utility of the sampler will also depend on aspects of the invertebrate's ecology, such as its dietary breadth, host preferences, and dispersal behaviour (Calvignac-Spencer *et al.* 2013a). For some invertebrates, particularly those with no economic importance for humans (e.g. terrestrial leeches or carrion flies), very little is known about their behaviour compared to disease vectors such as mosquitoes (Logue *et al.* 2016) or blackflies (Malmqvist *et al.* 2004). When monitoring for conservation, the location of detections is important, and this is difficult to establish for flying dipterans, which may have fed and then dispersed far from their prey. In comparison, leeches appear not to move far independently of their host. One important question is the extent to which leeches can be used to detect variation in mammal assemblages across local scales, where conservation monitoring is taking place, compared to regional scales where studies have pooled information from multiple leech species (Schnell *et al.* 2018; Tessler *et al.* 2018).

1.3.4. Haemadipsid leeches and biodiversity monitoring

Haemadipsid (family Haemadipsidae) leeches number around 70 species distributed across the Indo-Pacific tropics (Borda *et al.* 2008). All haemadipsids are blood-feeding and terrestrial (Borda & Siddall 2010), in contrast to most other leech species, which are semi- or fully-aquatic. Species of haemadipsid can be classified as either duognathous or trignathous (two or three-jawed), that latter of which includes members of the genus *Haemadipsa spp.*, which are found in the

Indian subcontinent, Southeast Asia and Japan (Borda & Siddall 2010). Haemadipsids are confined to areas of high humidity, probably linked to their evolution from an amphibious ancestor (Borda & Siddall 2010). These leeches are abundant in forests of north Borneo, where they are found within primary and logged forest habitats (Kendall 2012), however, they are absent from drier habitats such as agricultural plantations.

Studies focusing on their host preferences of *Haemadipsa spp.* have shown that they are generalists (Schnell *et al.* 2018; Tessler *et al.* 2018). Member of this genus in Borneo also have large bodies and can ingest and store large volumes of blood in their digestive structures, allowing the prey DNA to remain available for detection over long time-periods (Calvignac-Spencer *et al.* 2013a). These traits, combined with their ease of sampling in the forest (Schnell *et al.* 2012), makes haemadipsid terrestrial leeches an ideal choice for sampling mammalian DNA. Here, I focus on two leech species (*Haemadipsa picta* and *H. sumatrana*, Figure 1.3.) that are abundant in Sabah, North Borneo, and address the question of whether leech-iDNA can be used for biodiversity monitoring in degraded habitats.

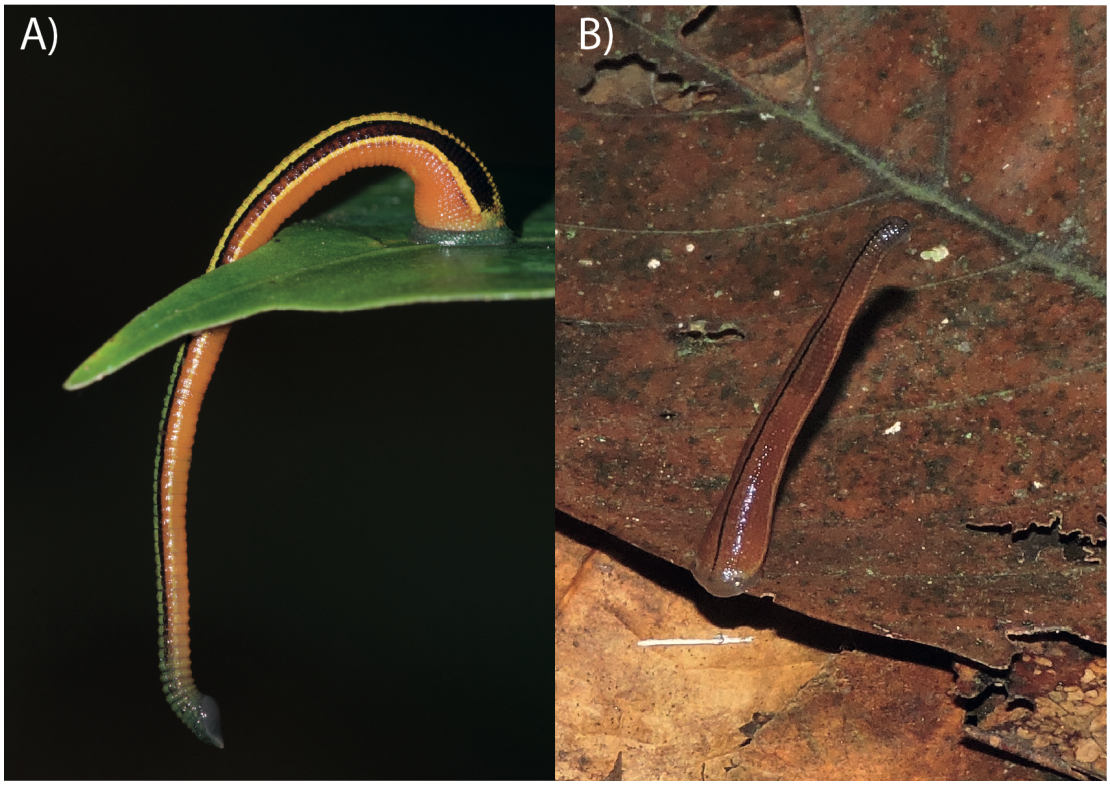


Figure 1.3. Photographs of the focal terrestrial leech species collected for iDNA sequencing to detect mammalian DNA in the blood-meals. A) *Haemadipsa picta*, tiger leech (photo credit: S. J. Rossiter). B) *Haemadipsa sumatrana*, brown leech (photo credit: R. Drinkwater)

1.4. Aims and objectives

In this thesis, I aim to combine the use of terrestrial leech iDNA and metabarcoding for biodiversity monitoring in a vulnerable human-modified tropical forest landscape. To do this, I have performed field collections in Sabah and performed analyses of leeches collected during two field seasons. By isolating and amplifying mammalian DNA from the bloodmeals, I show that it is possible to identify and quantify mammalian diversity across a degraded landscape. In the face of large-scale tropical land-use change, it is increasingly important to gain information on species and populations. Leech blood meal iDNA (along with iDNA from other invertebrates) presents a new opportunity for developing rapid, non-invasive molecular sampling for mammals. Combined with the ability to sample within an experimental and replicated framework, I use leech iDNA to test ecological questions regarding the effects of forest degradation.

In Chapter 2, I describe a comparative study of the utility of the two most abundant blood-feeding terrestrial leeches in Sabah, *Haemadipsa picta* and *Haemadipsa sumatrana* for mammal surveys (Figure 1.3). I show that overall *H. picta* detects a greater relative abundance of diversity and thus, is considered the better sampling tool. This chapter provides an essential foundation in the behaviour of these two understudied species.

In Chapter 3, I use one of the leech species: *H. picta*, to perform the first landscape-scale study of habitat quality on mammalian diversity and community composition based on iDNA data. To date most iDNA studies have focused on the testing of protocols and cataloguing of species. I show iDNA sequenced from *H. picta* can detect local scale differences in mammalian diversity.

In Chapter 4, I apply the statistical framework of hierarchical occupancy modelling to mammalian detections from leech iDNA. To my knowledge, this is the first time this technique has been used in this way. My results reveal that taxa show various responses to habitat quality, sampling effort and technical replication. I show the importance of including imperfect detections in iDNA studies, along with highlighting potential limitations.

Finally, I bring all my findings together in a General Discussion. Here, I aim to evaluate the benefits of leech-based iDNA studies, and the current gaps in our technical and ecological understanding of terrestrial leeches as samplers.

Chapter 2: Using metabarcoding to compare the suitability of two blood-feeding leech species for sampling mammalian diversity in North Borneo

The following chapter has published in:

Drinkwater, R., Schnell, I. B., Bohmann, K., Bernard, H., Veron, G., Clare, E., Gilbert, M. P. T. & Rossiter, S. J. (2018). Using metabarcoding to compare the suitability of two blood-feeding leech species for sampling mammalian diversity in North Borneo. *Molecular Ecology Resources*. DOI: 10.1111/1755-0998.12943

2.1. Abstract

The application of high throughout sequencing (HTS) for metabarcoding of mixed samples offers new opportunities in conservation biology. Recently the successful detection of prey DNA from the guts of leeches has raised the possibility that these, and other blood-feeding invertebrates, might serve as useful samplers of mammals. Yet little is known about whether sympatric leech species differ in their feeding preferences, and whether this has a bearing on their relative suitability for monitoring local mammalian diversity. To address these questions, I collected spatially-matched samples of two congeneric leech species *Haemadipsa picta* and *H. sumatrana* from lowland rainforest in Borneo. For each species, I pooled ~500 leeches into batches of ten individuals, performed PCR to target a section of the mammalian 16S rRNA locus, and undertook sequencing of amplicon libraries using an Illumina MiSeq. In total sequences from 14 mammalian genera, spanning nine families and five orders were identified. Greater numbers of detections, and higher diversity of OTUs, were found in *H. picta* compared with *H. sumatrana*, with rodents only present in the former leech species. However, comparison of samples from across the landscape revealed no significant difference in mammal community composition between the leech species. My findings therefore suggest that *H. picta* is the more suitable iDNA sampler in this degraded Bornean forest. I conclude that the choice of invertebrate sampler can influence the detectability of different mammal groups, and that this should be accounted for when designing iDNA studies.

2.2. Introduction

The rapid assessment of biodiversity through metabarcoding offers new opportunities in ecology. In particular, the ability to amplify and deep sequence the DNA from mixed sources contained within environmental samples has led to renewed interest in applying non-invasive molecular techniques to address questions in conservation. DNA metabarcoding is now a common technique to catalogue diversity (Deiner *et al.* 2016) and infer species interactions, including trophic connections (Salinas-Ramos *et al.* 2015).

One application of DNA metabarcoding that has shown particular promise for biodiversity monitoring is the screening of invertebrate-derived DNA (iDNA). Early research using iDNA techniques often had an epidemiological focus; for example, screening insect vector blood meals to identify hosts (review by Kent, 2009). More recently, these molecular techniques have been applied to biodiversity monitoring, including species of conservation concern (Schnell *et al.* 2012). A small number of studies have identified vertebrates from DNA contained within the blood-meals of haematophagous (blood-feeding) leeches (Weiskopf *et al.* 2017) while others have targeted blood or wound feeding arthropods including blowflies (Calvignac-Spencer *et al.* 2013b; Lee *et al.* 2015), mosquitoes and sand-flies (Kocher *et al.* 2017c). Sources of iDNA are not only restricted to blood; indeed, host DNA can be also recovered from invertebrate taxa that feed on faeces (Gómez & Kolokotronis 2016), and potentially from other excreta (see review by Calvignac-Spencer *et al.* 2013a).

Mounting interest in the potential use of invertebrates as samplers stems in part from falling sequencing costs in addition to a number of perceived advantages over more traditional methods (also see Weiskopf *et al.* 2017). For example, field sampling of invertebrates is often logistically easier, cheaper and tends to result in a greater number of individuals than direct or indirect sampling of vertebrates, including live- (Wells *et al.* 2004) or camera-trapping (Wearn *et al.* 2013). Sampling of vertebrates is also more tightly regulated than that of invertebrates, with stricter laws governing ethical handling and the transport of material across

international borders (Sikes & Gannon 2011). Furthermore, using iDNA removes the need for field-based taxonomic expertise as species identification can be achieved after sequencing with bioinformatics (Wheeler *et al.* 2004).

Despite the interest among ecologists in using iDNA for sampling, this approach still requires development and many aspects regarding its utility have not been fully addressed. For iDNA monitoring programmes to be successful we need a deeper understanding of how the choice of an invertebrate sampler might influence biodiversity estimates for a given ecosystem at a local scale. Multiple aspects of the invertebrate's biology will likely affect vertebrate detection probabilities (Calvignac-Spencer *et al.* 2013a). For example, variation in dispersal behaviour, habitat-use, feeding ecology, and rate of digestion should all ideally be taken into account when choosing a sampler species (Schnell *et al.* 2015a). Other biases are known to arise from the laboratory protocols, although these have been examined previously, and are better understood than those biases introduced by the invertebrate sampler (see Alberdi *et al.* 2017; Elbrecht *et al.* 2017).

Terrestrial leeches (family Haemadipsidae, phylum Annelida) are free-living blood-feeders, with ~50 species, distributed across the paleo-tropics (Sket & Trontelj 2008). Apart from being highly abundant and easy to collect, haemadipsid leeches may be particularly useful as iDNA sources because of their large body size and gut capacity compared with most blood feeding arthropods (Schnell *et al.* 2015a). In addition, the use of leech iDNA has been shown to be a complementary method to camera trapping (Weiskopf *et al.* 2017). Most common haemadipsid leeches have been suggested to be generalist feeders, opportunistically attaching to passing mammalian hosts (Govedich *et al.* 2004), and this has received recent support from wide-scale sampling. Schnell *et al.* (2018) compared haemadipsid leeches across five geographical regions and found evidence that across the family, species feed on a broad range of mammalian diversity, while Tessler *et al.* (2018) reported similar trends from three additional regions. However, finer-scale comparisons of iDNA samplers from the same site (e.g. Kocher *et al.* 2017c) have rarely been undertaken.

In this study, I perform a quantitative comparison of diet of two co-occurring terrestrial leeches *Haemadipsa sumatrana* (commonly known as the brown leech) and *H. picta* (the tiger leech) to test their relative usefulness as samplers of a diverse mammal fauna in lowland tropical forest in Borneo, Southeast Asia. Although these two species occupy the same forests, they appear to have different fine-scale habitat associations. *H. sumatrana is* found almost exclusively in the leaf litter while *H. picta* is found from the ground to around two metres in the understorey (Lai *et al.* 2011). *H. picta* also appears to be robust to microclimatic changes; they are more common than *H. sumatrana* in logged forest with open canopy (Kendall 2012), and are also more likely to occur on or near trails in the forest (Gąsiorek & Różycka 2017). Other unknown species-specific traits may also contribute to differences in their feeding ecology.

My specific aims were to (1) use metabarcoding techniques to ascertain the feeding ecology of both leech species, (2) compare the diets of spatially-matched leeches across a range of forest types, and (3) evaluate the suitability of using each of the leech species as an iDNA sampler for biodiversity monitoring. Being able to rapidly identify and understand differences in mammalian diversity in Bornean forests is especially pertinent in the light of ongoing land-use change. Specifically, forest outside of protected areas is often highly degraded due to timber extraction and conversion for agriculture (Gaveau *et al.* 2014). Furthermore, large vertebrates, such as charismatic and rare large mammals, are key considerations in the formulation of policy or conservation actions and decision-making will rest on the assumptions of data reliability.

2.3. Materials and methods

2.3.1. Study site and sample collection

I collected all samples at the Stability of Altered Forest Ecosystems (SAFE) site in the Kalabakan Forest Reserve, Sabah (4° 33′ N, 117° 16′ E) in Malaysian Borneo, a large-scale forest fragmentation experiment covering logged secondary forest (Ewers *et al.* 2011). There are 8 sites (3km^2 radius) that I have broadly classified as either twice- (n=4) or heavily logged forest (n=4) (Figure 2.1). The twice-logged sites are located within a large contiguous tract of managed forest and the heavily logged sites are with the SAFE experimental area and consist of degraded and fragmentated forest (Ewers *et al.* 2011). Each of these blocks contains between 8 and 16 permanent forest plots (25m^2) (for details see Ewers *et al.* 2011). I collected samples of leeches from 59 forest plots in heavily logged sites (Figure 1; blocks B, D, F & VJR) and 29 plots in twice-logged sites (Figure 1; blocks LF1-3 & LFE). Forest plots were selected based on accessibility and microclimatic conditions that support leech populations.

Within each plot, I sampled leeches by searching the forest floor and understorey for 20 minutes, and I re-sampled each plot four times between February and June 2015. All leeches encountered of both species, *H. picta* and *H. sumatrana*, were collected and placed into individual tubes containing RNA later. The samples were stored on ice packs in cool boxes until returning to the main camp, normally within 12 hours but for some remote sites the delay was 3-4 days. *H. picta* and *H. sumatrana* were the only leech species I encountered in the field and were easily identifiable based on their markings (Figure 1.3).

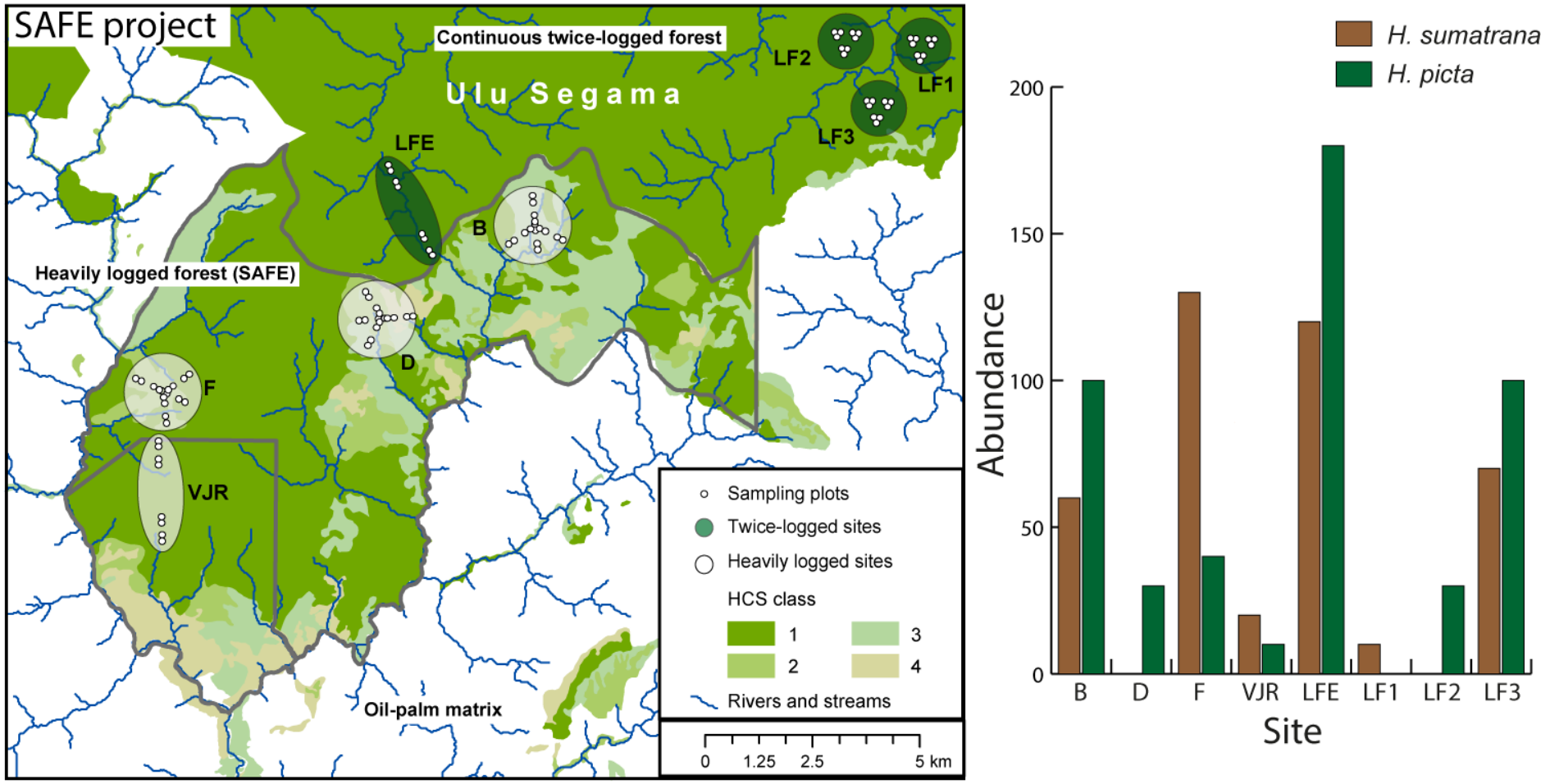


Figure 2.1. Schematic map of the sampling design used in this chapter. The twice-logged habitats are shaded in dark green within the continuous logged forest and the heavily logged habitats are shaded in white within the SAFE project experimental area. The location of the 25 m^2 vegetation plots where the sampling for leeches took place are shown as small white circles. Blue lines represent the rivers and streams, and the green shows the different classes of high carbon stock forest (HCS). Bar chart shows the number of individuals collected at each of the sites for *H. picta* and *H. sumatrana*.

2.3.2. DNA extraction and PCR amplification

I performed sequencing of pooled leeches following the protocol set out by Schnell *et al.* (2018) (Figure 2.2). Briefly, I performed tissue digestions on individual leeches using the tissue digestion buffer with enough buffer per leech to equal a volume approximately five times the leech body, and then incubated the samples overnight at 50°C while gently shaking. Following this incubation, I pooled 100µl of ten individual digests, ensuring that each 1000µl pool contained ten leeches collected from the same site (Figure 2.1). In total 490 *H. sumatrana* were pooled into 49 pools and 520 *H. picta* were pooled into 52 pools (Table 2.1). I purified the DNA from each leech digest pool using the QiaQuick DNA kit (Qiagen, UK) following the manufacturer's protocols but with a modified centrifugation procedure (1 min at 6,000g, 1 min at 10,000g followed by an additional 3 min at full speed, and 1 min at 12000g) and eluted in 50µl EB buffer. To ensure consistency, for each batch of extractions, I quantified the DNA from a subsample of pools using the Qubit dsDNA HS Assay Kit (Invitrogen).

Table 2.1. A summary of leech samples used in this study. For each site within a habitat type, the number of leech pools for which DNA was extracted and the subsequent number of these pools which were then used for 16S amplicon sequencing is shown for both species (*Haemadipsa sumatrana* and *H. picta*). The corresponding number of individuals in the pool is given for the samples sequenced. Before sequencing all samples were amplified using three replicate PCRs (except where specified)

Site	**Habitat type**	**Samples extracted (pools)**		**Samples sequenced (pools, individuals)**	
		H. sumatrana	*H. picta*	*H. sumatrana*	*H. picta*
B	Heavily logged	8	11	6, 60	10, 100
D	Heavily logged	0	4	0, 0	3, 30
F	Heavily logged	18	4	13, 130	4, 40†
VJR	Heavily logged	2	1	2, 20	1, 10
LFE	Twice-logged	13	19	12, 120	18, 180
LF1	Twice-logged	1	0	1, 10†	0, 0
LF2	Twice-logged	0	3	0, 0	3, 30†
LF3	Twice-logged	7	10	7, 70†	10, 100†

† *indicates samples which only had duplicate PCR replicates*

I amplified a 95bp fragment of the mammalian 16s mitochondrial gene using the primers "16Smam1" forward 5'-CGG TTG GGG TGA CCT CGGA-3' and "16Smam2" reverse 5'-GCT GTT ATC CCT AGG GTA ACT-3' primers (Taylor 1996). I conducted PCRs in triplicate, with the exception of 26 pools (from LF1-3 & F) which were conducted in duplicate during a preliminary experiment (Table 2.1). DNA was amplified using 5' nucleotide tagged primers (6-8bp) (Binladen *et al.* 2007), with identical tags on both forward and reverse primers to be able to identify possible errors due to "tag jumping" (Schnell *et al.* 2015b). All tags were designed to have two mismatches between each pair, to allow for identification in the case of sequencing error (Binladen *et al.* 2007). The 25µl PCR reactions consisted of 0.2mM of 10x buffer, 2.5mM MgCl2, 1unit DNA polymerase (AmpliTaq Gold, Applied Biosystems), 0.2mM dNTP mix (Invitrogen), 0.5mg/mL BSA, 0.6µM of each primer and 1µL of DNA template and with a thermal cycling profile of 95^0C for 5 minutes, then 40 cycles of 95^0C for 12 seconds, 59^0C for 30 seconds and 70^0C for 20 seconds with a final extension time of 7 minutes at 70^0C. Negative extraction, PCR and positive controls (giraffe DNA) were included in every run.

All PCR products (including controls) were visualised on 2% agarose gels and those reactions which contained DNA were pooled into libraries. PCR success rate was high, with 94% of *H. picta* pools and 83% of *H. sumatrana* pools seen to contain vertebrate DNA. Using these successful PCR replicates (including controls) I prepared indexed amplicon libraries for sequencing using the BEST v2.0 library build protocol (Carøe *et al.* 2017). All amplicon libraries were checked pre- and post-indexing using the 2100 Bioanalyzer, DNA high sensitivity kit (Agilent, Denmark). I pooled all indexed amplicon libraries at equimolar concentrations for sequencing on an Illumina MiSeq. Most libraries were sequenced (250bp paired-end) at the National High-throughput DNA Sequencing Centre (University of Copenhagen) with a smaller number sequenced (150bp paired-end) at the Bart's and the London Genome Centre (Queen Mary University London).

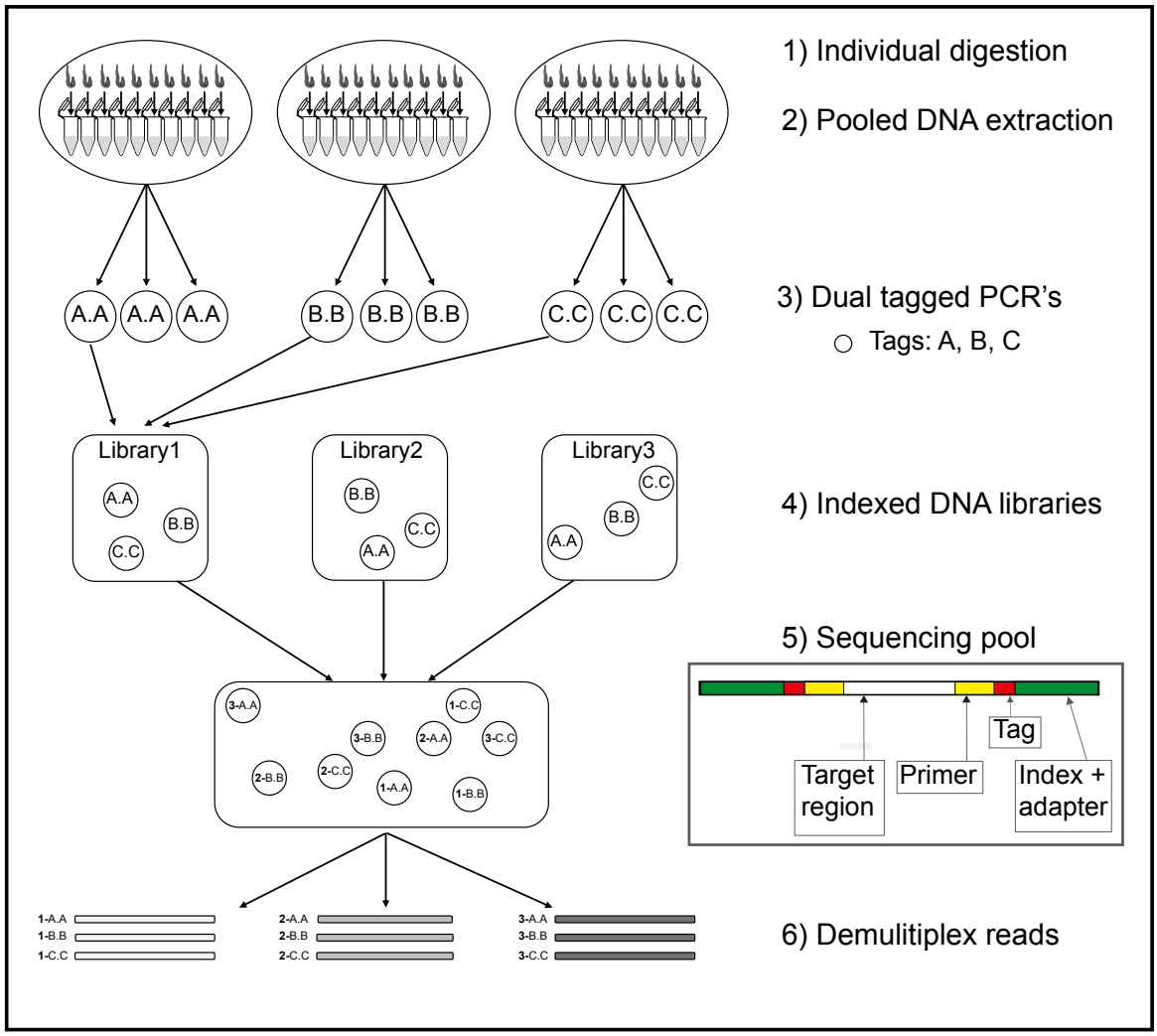


Figure 2.2. Workflow diagram for pooled terrestrial leech iDNA sequencing; (1) Individual leeches undergo tissue digestion, (2) Tissue digests are pooled and DNA is extracted from batches of ten individuals, (3) PCR replicates are uniquely dual-tagged, (e.g. A.A), (4) PCR replicates are pooled for DNA library build with unique tag/index combinations, (5) Libraries are mixed in equimolar concentration and sequenced, (6) Sequences are demultiplexed by index and sorted by PCR replicate. Inset shows magnification of the structure of an amplicon in the sequencing pool with the unique tag and index unique to each pooled DNA extract.

2.3.3. Bioinformatics and statistical analyses

Using AdapterRemoval v2 (Schubert *et al.* 2016) I merged the demultiplexed forward and reverse reads, with default parameters except for minalignment 100, minlength 50 and shift 5. On the merged reads, I used a modified version of DAMe (https://github.com/shyamsg/DAMe, Zepeda-Mendoza *et al.* 2016) to assign each sequence to the original sample based on the correct primer and nucleotide tag combination. With DAMe I retained only those sequences that were detected in a minimum of two replicates, clustered the filtered reads at 97% similarity using

sumaclust v1.3 (Mercier *et al.* 2013) and normalised the reads per sample to 50000 with DAMe to allow cross-sample comparisons. To identify potential sequencing errors, a post-clustering filtering procedure was then applied to the original OTU table using LULU, which removes erroneous rare OTUs based on both sequence similarity thresholds and within-sample patterns of co-occurrence (Frøslev *et al.* 2017). There was some evidence of contamination in the negative controls with sequences matching two OTUs, corresponding to *Rusa unicolor* and *Sus barbatus,* respectively. For the remaining samples, I therefore only considered these species to be present where numbers or reads exceeded those found in the controls.

2.3.4. Compiling the reference database

Using local knowledge of the field site, field guides (Payne *et al.* 1985; Phillipps & Phillipps 2016) and the IUCN red list distribution maps (www.iucnredlist.org), I complied a reference database of all mammals likely to occur at the study site. From NCBI GenBank nucleotide database, I retrieved all published 16S, aiming for five records per species (Supplementary Figure S2.1). I trimmed and aligned the selected sequences to the primer target region using AliView (Larsson 2014). To augment my database, new 16S sequences were generated for the following species: Common treeshrew (*Tupaia glis*), Small-toothed palm civet (*Arctogalidia trivirgata*), Banded civet (*Hemigalus derbyanus*), Common palm civet (*Paradoxurus hermaphroditus*), Hose's civet (*Diplogale hosei*), Malay civet (*Viverra tangalunga)*, Banded linsang (*Prionodon linsang*), Short-tailed mongoose (*Urva brachyura*), Collared mongoose (*Urva semitorquata*), Chinese ferret-badger (*Melogale moschata*), Malay weasel (*Mustela nudipes*) and Clouded leopard (*Neofelis nebulosa*) (sequences provided by G. Veron, Supplementary Table S2.1). To be able to identify contamination, I also included 16S sequence records from NCBI GenBank database for human, giraffe (positive control) and domestic/human-associated species (Supplementary Table S2.2). Representative 16S sequences for reptiles, amphibians and birds were also included in the reference database (Supplementary Table S2.2), sourced from NCBI GenBank, as detection of these taxa with iDNA by haemadipsids is known (Schnell *et al.* 2018; Tessler *et al.* 2018).

2.3.5. Taxonomic assignment

To assign each OTU to a mammalian taxon, I performed a BLAST search against a custom sequence reference database for Bornean taxa. With the MEGAN lowest common ancestor (LCA) algorithm, I assigned mammalian taxon to the OTUs from the top BLAST results with >90% similarity. The MEGAN parameters I used were minimum bit score = 150, top percent = 2, min support = 1, and weighted LCA with 90% coverage (Huson *et al.* 2007). I only considered assignments at the genus-level, as this has been shown to increase reliability of identifications in other ribosomal markers when reference databases are incomplete (Kocher *et al.* 2017b). In a small number of cases where the best matching species (>90% similarity) has no known congeners in Borneo, I was able to assign to the species-level (e.g. *Echinosorex gymnura*). Where there was no match to a reference sequence, the OTU remained unassigned and was removed from the analysis. I then filtered our results by removing any OTUs with a match to human or our positive control. Final taxonomic assignments are presented in Table 2.2.

2.3.6. Estimation of biodiversity determined by leech samplers

To determine the relative utility of using the two focal leech species for iDNA sampling, I produced sample-size based diversity accumulation curves using all samples together and compared these to leech species-specific curves. To give a deeper understand of the effects of rare and abundant taxa, I produced these curves by estimating three orders of Hill numbers of diversity (Hill 1973). These are the most commonly used Hill numbers and are equivalent to species richness ($q = 0$), the exponential of Shannon-Weiner index ($q = 1$) and the Simpson diversity ($q = 2$) (Chao *et al.* 2014). One of the benefits of using Hill numbers compared to other diversity indices is that they can be expressed in the terms of effective number of species, thus allowing different communities to be directly compared (Chao *et al.* 2014). In practice, this means a mammal community sampled by *H. picta* is comparable to a community sampled by *H. sumatrana*. For the accumulation curves, using rarefaction, I constructed 84% confidence intervals (CIs) that equate to an α-level of 0.05 for overlapping distributions, rather than 95% CIs that equate to an α-level of 0.01 and are thus considered overly conservative in such comparisons (MacGregor-Fors & Payton 2013). Diversity

accumulation curves with the standard 95% CI are shown in Supplementary Figure S2.2.

To test whether leech species differ in their utility as iDNA samplers I applied two approaches. First, I fitted a GLM in which I modelled the number of detections per leech pool as the response variable with Poisson error, and fitted leech species (*H. picta* and *H. sumatrana*), forest type (heavily and twice-logged) and block identity (B, D, F, VJR, LF1, LF2, LF3, LFE) as explanatory variables. I started with a full model containing all variables, and compared its fit based on AIC to seven reduced models (Supplementary Table S2.3). Models showed no overdispersion ($\theta < 2$). Second, to test whether the diets of either leech species differ with respect to composition of mammals, I examined patterns of beta diversity among pools. I calculated pairwise Bray-Curtis dissimilarity indices and visualised community composition using non-metric multidimensional scaling (NMDS). To test for greater dissimilarity between leech species, and different habitats, I applied a PERMANOVA analysis as a robust test of ecological community structure (Anderson & Walsh 2013). I conducted all analyses in R (R Core Team 2018) using Vegan (Oksanen *et al.* 2017) and iNEXT packages (Hsieh *et al.* 2016).

2.4. Results

2.4.1. Generation of reference database

The final reference database of sequences compiled for the field site contained 256 records of the 16S target sequence, from 28 mammalian families across ten orders. For 40 mammal species for which 16S sequences were not available, I either obtained published sequences from a related member of the same taxonomic family (30 cases), or I used newly generated sequences for Bornean native species (8 cases), or an Asian sister species (two cases; clouded leopard and Chinese ferret badger), via Sanger sequencing (Supplementary Table S2.1). In this latter case, sequences ranged from 90 to 101bp (new GenBank accession numbers MG996889 - MG996900) (Supplementary Figure S2.1).

2.4.2. Taxonomic assignment

By curating the total number of OTUs with the post-clustering algorithm (LULU) and filtering out contaminants, I reduced the number of clustered OTUs from 65 to 17 (26% retained) but with no loss of taxonomic diversity. All OTUs matched to native Bornean mammalian taxa, and there were no unexpected taxa in my results. Of the 17 OTUs, 14 matched with high similarity to the reference sequences with >90% similarity (Table 2.2). Two of the remaining OTUs (OTU49 and OTU21) matched less well to a reference sequence (both at 79%) but were consistently assigned to langur (Colobinae) and gibbon (Hylobatidae), respectively. I also found one OTU with a match that could not be resolved beyond the subfamily Cervinae, matching equally to both the cervid genera that occur at the site (*Muntiacus* and *Rusa*).

Table 2.2. Taxonomic identity assigned to the unique Operational Taxonomic Units (OTUs) which were generated from MiSeq amplicon sequencing and bioinformatic filtering. The level of confidence in each assignment is shown by % identity match given by BLAST and the bit score from MEGAN. If two different OTUs share the same taxonomic identity, values for both are given separated by a /

Common name	Order	Family (subfamily)	Taxa assigned	OTU	% Identity	Bit Score
Unknown deer	Cetartiodactyla	Cervidae (Cervinae)	Cervinae	OTU18	97	143
Sambar deer	Cetartiodactyla	Cervidae	*Rusa unicolor*	OTU4	100	171
Muntjac	Cetartiodactyla	Cervidae	*Muntiacus sp*	OTU5/ OTU7	91/90	159/154
Bearded pig	Cetartiodactyla	Suidae	*Sus barbatus*	OTU2	100	174
Mousedeer	Cetartiodactyla	Tragulidae	*Tragulus sp*	OTU6	99	171
Banded civet	Carnivora	Viverridae (Hemigalinae)	*Hemigalus derbyanus*	OTU8	100	178
Malay civet	Carnivora	Viverridae (Viverrinae)	*Viverra tangalunga*	OTU12	100	178
Moonrat	Eulipotyphla	Erinaceidae	*Echinosorex gymnura*	OTU10	100	171
Macaque	Primate	Cercopithecidae (Cercopithecinae)	*Macaca sp*	OTU9/ OTU29	95/97	141/154
Leaf monkey	Primate	Cercopithecidae (Colobinae)	*Trachypithecus sp*	OTU49	79	60
Gibbon	Primate	Hylobatidae	*Hylobates sp*	OTU21	79	60
Porcupine	Rodentia	Hystricidae	*Hystrix sp*	OTU13/OTU71	91/97	167/148
Long-tailed porcupine	Rodentia	Hystricidae	*Trichys fasciculata*	OTU15	90	161
Rat	Rodentia	Muridae	*Rattus sp*	OTU86	99	163

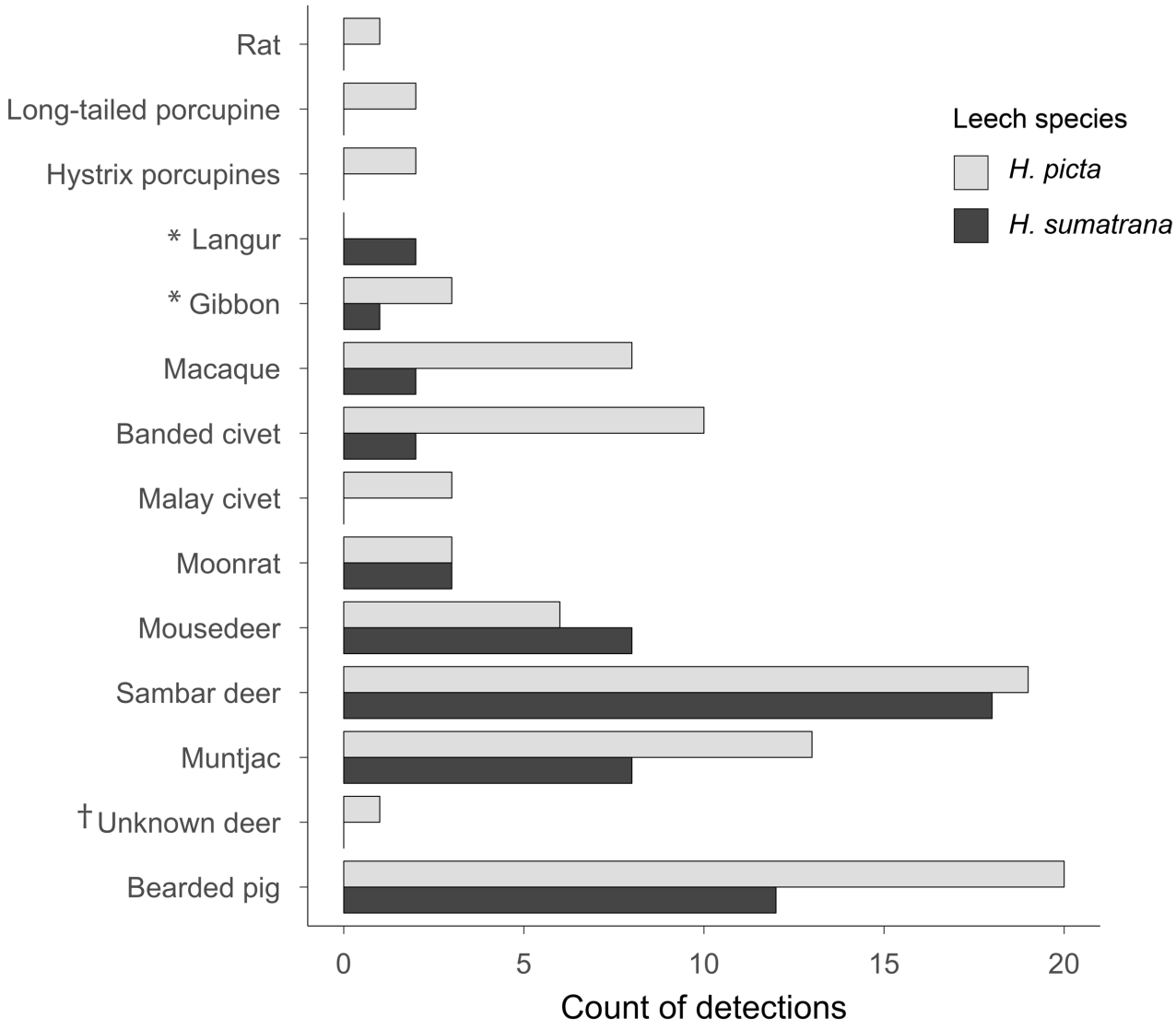


Figure 2.3. A comparison of the abundance of detections for different mammal taxa identified using tiger leech samplers (*H. picta*, light grey) compared with brown leech samplers (*H. sumatrana*, dark grey). * indicates the two cases where the sequence has a poor but consistent match to the reference database (<90% identity). † indicates the sequence with the best match to both cervid genera (*Rusa* and *Muntiacus*) therefore cannot be identified to species.

Eight mammalian taxa were common to both leech species (Figure 2.3), of which the most prevalent were the Bornean sambar deer (*Rusa unicolor*) and bearded pig (*Sus barbatus*), followed by the muntjac (*Muntiacus sp.*) and the mousedeer (*Tragulus sp.*) (Figure 2.3). Other taxa were detected in both leech species but with fewer detections in the brown leech (*H. sumatrana*) than in the tiger leech (*H. picta*): banded civet (*Hemigalus derbyanus*), moonrat (*Echinosorex gymnura*), macaque (*Macaca sp.*) and gibbon (*Hylobates sp.*). Additionally, I found four taxa in

the tiger leech that were not found in the brown leech: the Malay civet (*Viverra tangalunga*) and three rodents (two porcupine genera, *Hystrix* and *Trichys* as well as one *Rattus* sp). Finally, the langur (Colobinae) was only detected in the brown leech.

2.4.3. Mammal diversity in leech diets

There was a greater total number of detections in twice-logged forest in *H. picta* than *H. sumatrana* but very similar detection levels for both leech species in the heavily logged forest (Figure 2.4A). These trends were also reflected in most of the individuals blocks sampled (Figure 2.4B). The results of the GLM indicated that the number of mammal detections per pool was determined by leech species, with more detections in *H. picta*, but not by either habitat type or block (Table 2.3, alternative model summaries in Supplementary Table S2.3). Model comparisons suggested that the two best-fitting models, each with similar AIC values, contained leech species alone ($F_{2,88}$ =20.86, $p < 0.05$) and leech plus habitat type ($F_{1,88}$ = 30.28, p = 0.202). However, while the latter model was associated with the best fit (adj-R^2 = 0.31), leech species was the only significant single predictor (Table 3). Considering taxonomic representation, *H. picta* samples a greater proportion of orders (5/5), families (8/9) and genera (12/14) detected in this study compared with *H. sumatrana* (orders = 4/5, families = 7/9 and genera = 9/14). Of the species which could be identified, *H. picta* detects all six representatives while *H. sumatrana* detects four.

Table 2.3. Summary output of the Poisson generalised linear models with the lowest AIC values out of the possible candidate models. The models include the effect of the variables of the leech sampler species (brown or tiger) and habitat type on the total number of mammal detections found in the leech blood-meal. The parameter estimate is given with the corresponding standard error. The F-value and the p-value are given. * shows significance where $\alpha = 0.05$

	Model1			Model2		
	Estimate (± SE)	**F value**	**p-value**	**Estimate (± SE)**	**F value**	**p-value**
Intercept	0.77 (0.086)	79.62	>0.05*	0.83 (0.070)	138.58	>0.05*
Species	0.22 (0.092)	5.43	0.022	0.20 (0.091)	4.69	0.032
Habitat	0.12 (0.092)	1.65	0.202	-	-	-
AIC	295.65			294.42		
Adj-R^2	0.31			0.25		

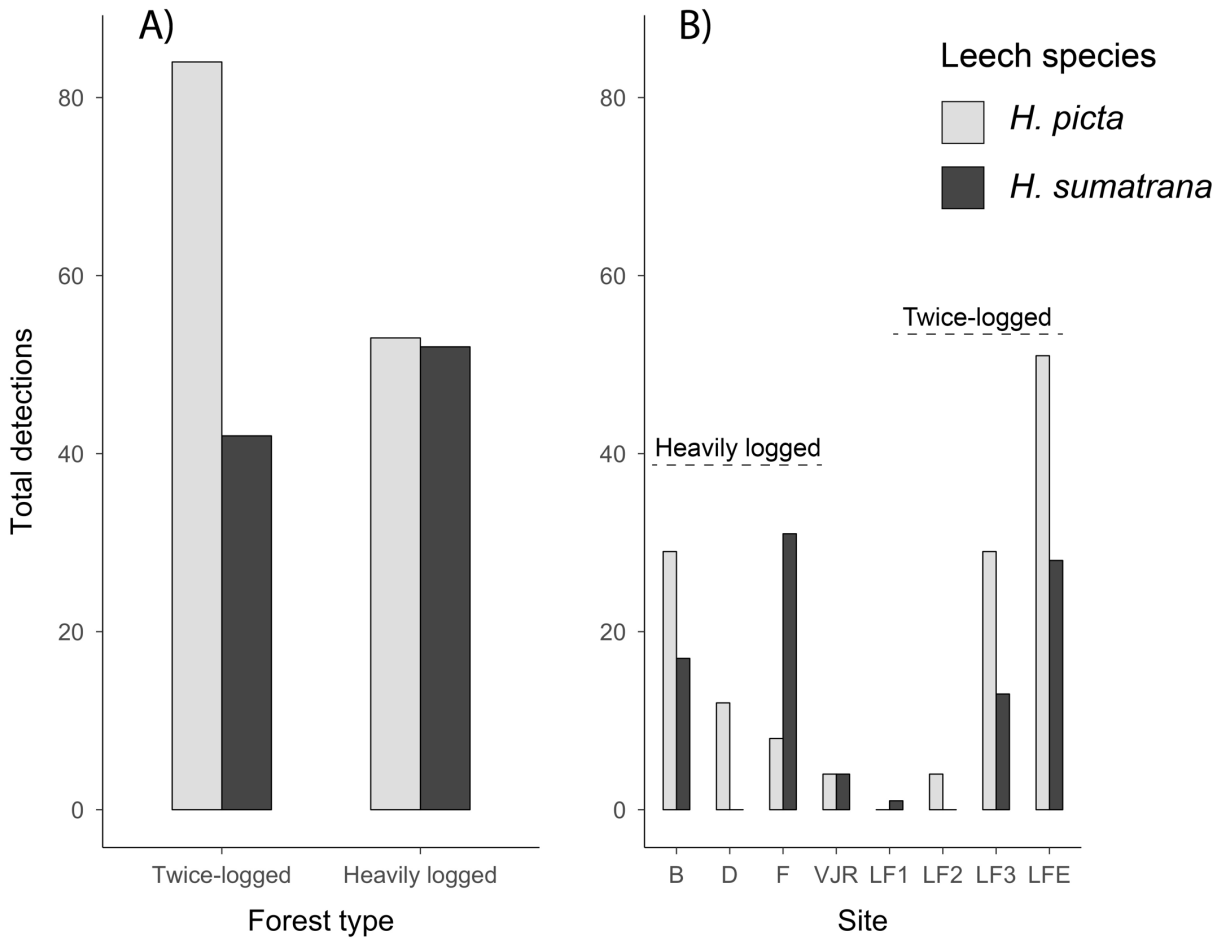


Figure 2.4. Total count of mammal detections found in the blood meals of each leech species, *H. picta* = light grey bars and *H. sumatrana*, dark grey bars. (A) the total detections depending on each forest type, either twice-logged or heavily logged and (B) shows the total detections split by each block from within the two forest types

2.4.4. Accumulation of taxonomic richness

Accumulation curves based on three metrics of mammalian diversity (equivalent to raw species richness, Shannon-Weiner index and Simpson diversity) showed consistent differences among the leech species. In each case, tiger leeches consistently sampled around 40% more diversity than did the brown leeches (Figure 2.5). Comparing these accumulation curves to corresponding curves based on pooled leeches suggested that *H. sumatrana* contained a subset of taxa of H. *picta*, with almost no additional diversity obtained by combining data from both leeches over that recorded for *H. picta* alone.

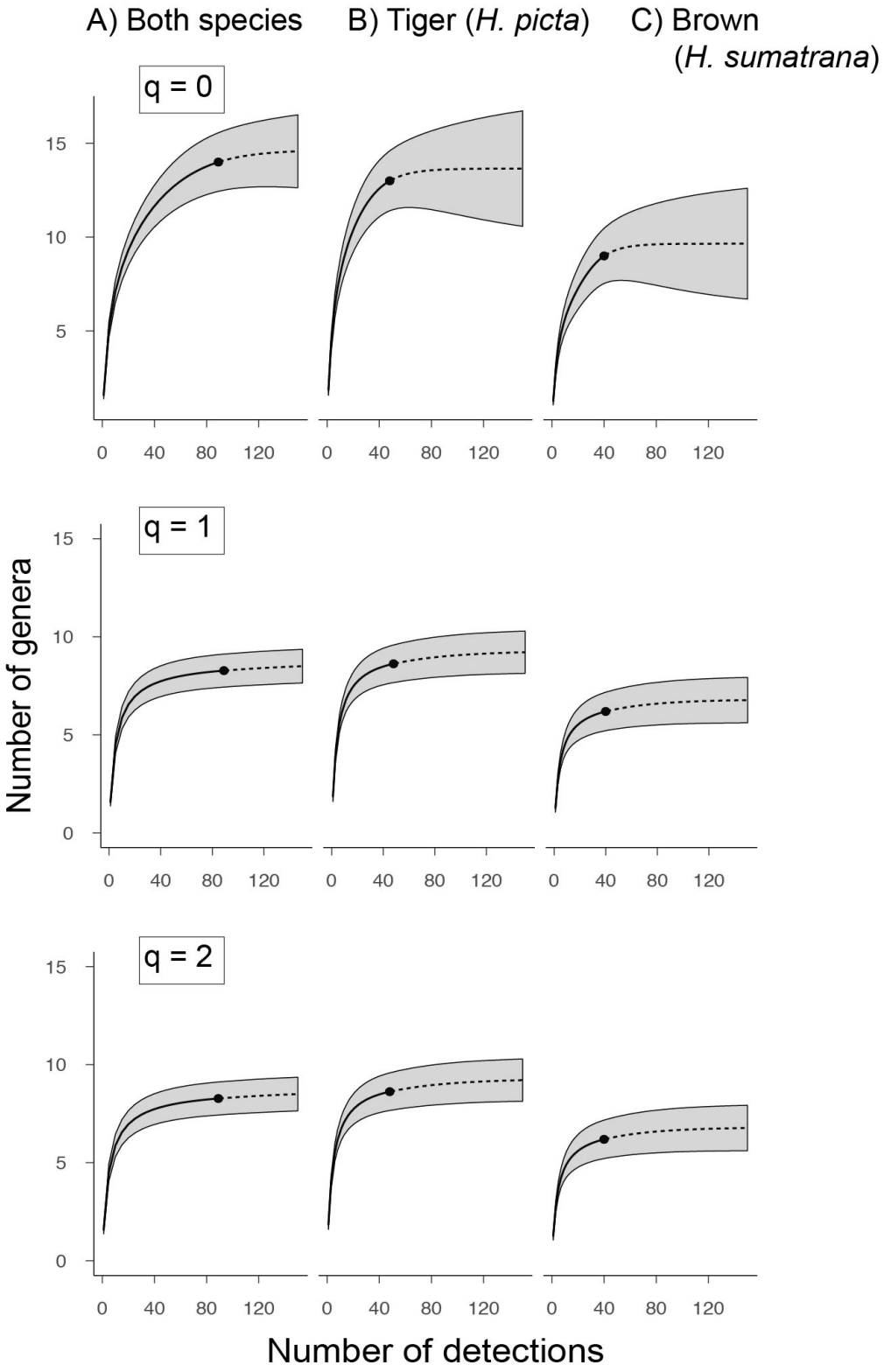


Figure 2.5. Diversity accumulation curves for: (A) both leech species together, (B) tiger leeches only (*H. picta*) and (C) brown leeches only (*H. sumatrana*). Each row shows the accumulation of mammal genera with increasing detections using the three hill numbers, q = 0, 1, 2 (corresponding to species richness, Shannon diversity index and Simpson index respectively). Curves are presented with 84% confidence intervals, which is equivalent to significance where $\alpha = 0.05$ (MacGregor-Fors & Payton 2013) and extrapolated to 125 samples (dashed lines) following Chao *et al.* (2014)

2.4.5. Estimates of local biodiversity between samplers

Visualising the differences in beta-diversity using NMDS showed some separation between the community of mammals detected in the two habitat types, twice- and heavily logged forest (Figure 2.6A). When the data points were grouped by leech species I found considerable overlap in the mammal communities sampled (Figure 2.6B). Despite some apparent separation with habitat, the PERMANOVA analysis found no significant difference between either of the factors or their interaction (Figure 2.6).

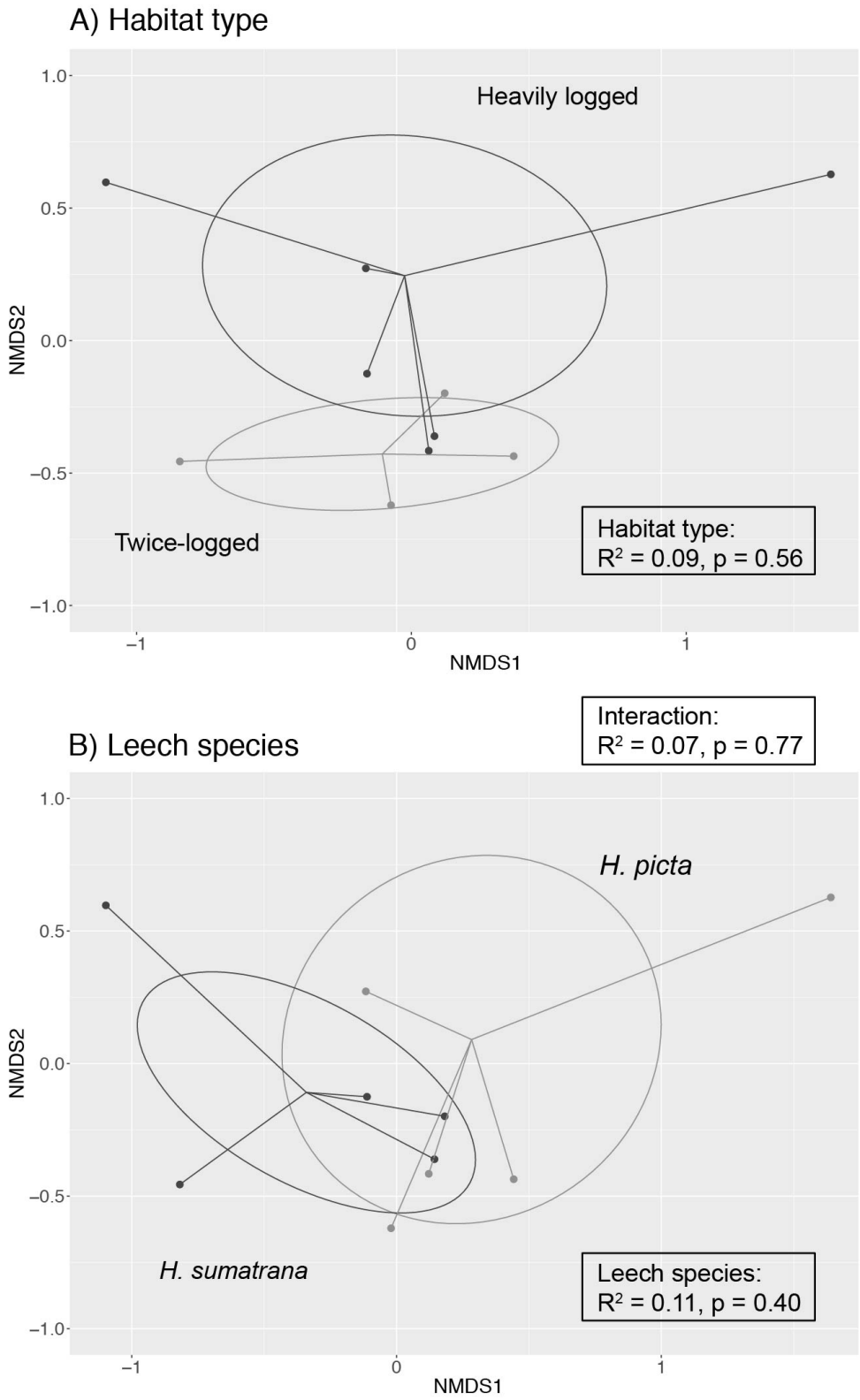


Figure 2.6. Non-metric Multi-Dimensional Scaling (NMDS) ordination plots showing the Bray Curtis dissimilarity between the mammal communities identified in the leech blood-meals depending on (A) habitat types and (B) leech species. Stress of the NMDS = 0.05. Ellipses represent the standard error of the ordination. Inset values show the parameter estimates from the corresponding PERMANOVA test showing the non-significance between the groupings and the interaction where $\alpha = 0.05$

2.5. Discussion

The use of invertebrates as iDNA samplers of vertebrates is gaining interest, and in this study, I systematically assessed the relative utility of two congeneric haemadipsid species, tiger (*Haemadipsa picta*) and brown (*H. sumatrana*) leech for detecting local mammal diversity. Using spatially matched samples for the detection of mammals in a degraded forest habitat in North Borneo, Southeast Asia, I analysed over 1000 individual leeches from the two species, sampled at 88 sites across the landscape which revealed the presence of terrestrial mammal species from nine families spanning five orders.

2.5.1. Leech-derived iDNA from Haemadipsa picta versus H. sumatrana

I found a high degree of overlap in the mammalian species richness detected in both *H. picta* and *H. sumatrana* diets. The most abundant detections correspond to the large common species found in the area such as sambar deer and bearded pig. However, of these leech species, *H. picta* has a significantly higher detection rate compared to *H. sumatrana*. There are nine overlapping mammal taxa in both leech species, but four taxa were only specific to *H. picta*, including all of the rodents detected. Sampling with *H. picta* results in a greater coverage of the total mammal community. By directly comparing the accumulation curves for a fixed sampling effort, i.e. same number of equally sized leech pools, I found the detection of a greater diversity of effective numbers of species using *H. picta* compared to *H. sumatrana*. However, while there is a greater abundance of detections, I did not detect a significant difference in the community composition using either leech sampler.

Feeding strategies have been suggested to affect iDNA detection (Schnell *et al.* 2015a) and *H. picta* and *H. sumatrana* show clear differences in their searching and feeding behaviour; *H. sumatrana* is almost exclusively found at ground level and is camouflaged in the leaf litter, whereas *H. picta* tends to wait on leaves in the undergrowth and thus, together with its more striking markings is easier to see and collect during sampling (Fogden & Proctor 1985). Taking these points together, I suggest that of the two species examined, *H. picta* represents the more

suitable iDNA sampler in our study area, due to the greater abundance of positive detections coupled with favourable behavioural traits for rapid sampling.

2.5.2. Detection of mammalian diversity

Although medium to large mammals, especially ungulates, were well represented in the sequence data, it was noticeable that very few small mammals were detected. In particular, non-volant mammals from three families, Tupaiidae (treeshrews), Scuiridae (squirrels) and Muridae (mice and rats) were not detected in any of the leech samples and yet are known to occur in the study area (Wearn *et al.* 2017). While my study is based on presence-only data, and thus non-detection cannot be used to infer absence from the habitat, it is noteworthy that a similar lack of Bornean small mammals was also recently reported by Schnell *et al.* (2018). These authors compared iDNA from leech blood meals sampled from a broad geographical scale, and were able to detect treeshrews, squirrels and murid rodents in mainland Southeast Asia, Madagascar and Australia, but not in leech samples from Borneo. In addition, representatives of rodents and treeshrews were detected in a study of 200 leeches from Bangladesh based on Sanger sequencing of the 16S rRNA marker (Weiskopf *et al.* 2017). A recent study by Tessler *et al*. (2018) also confirmed detections of these small mammal groups in China and Bangladesh. The absence of these taxa from Bornean leeches might indicate that haemadipsid leeches in Borneo are behaving differently to their congeners in other parts of Asia.

As small mammals form a large part of the mammalian biomass in Borneo and a rich diversity of species are known to occupy the forest, the lack of detection specifically in Borneo is intriguing. The reference database contained several representative sequences from all the small mammal families, so this is unlikely to be a consequence of missing reference data but in fact could reflect size-related feeding preference shown by Bornean haemadipsids in particular. Whether or not leeches actively prey on large mammals or are more easily detected and ejected by small mammals, is not known. Moreover, the underrepresentation of nocturnal mammals, such as murid rodents, cannot be explained by the timing of my surveys which were conducted during the day, since both of the focal leech species are

active during both day and night. Regardless of the underlying causes of the observed patterns of detection, my data would indicate that for the purposes of biodiversity surveys, Bornean leeches appear not to passively sample all non-volant mammals in their environment, as has been previously suggested. Thus, I recommend that future iDNA studies should include assessments of how the ecology and behaviour of the chosen invertebrate sampler might influence any results.

2.5.3. Imperfect detections and temporal resolution

A major issue with the use of iDNA for biodiversity monitoring is imperfect species detection resulting from problems of false detections, both positive and negative. This is discussed in the wider literature concerning environmental DNA (eDNA) (Roussel *et al.* 2015; Deiner *et al.* 2017) and also for iDNA studies (Schnell *et al.* 2015a). I took many steps to optimise the trade-off between false positives and negatives; for example using technical replicates to increase detection rates and reduce false negatives (Ficetola *et al.* 2016) while removing spurious tag combinations (Schnell *et al.* 2015b) and using conservative bioinformatic filtering to reduce false positives (Alberdi *et al.* 2017).

Imperfect detection of iDNA is also likely to be influenced by aspects of the biology of the sampler. Indeed, the rate of digestion of the blood-meal and intervals between feeding events will affect the window of DNA detection (Schnell *et al.* 2015a). For the medicinal leech (*Hirudo medicinalis*) the detection window has been empirically tested and shown to be at least 120 days for mammal DNA (Schnell *et al.* 2012), and up to 50 days for mammalian viral DNA (Kampmann *et al.* 2017). However, *H. medicinalis* is a larger-bodied taxon than *Haemadipsa* spp., with different ecology and behaviours, and the detection window remains unknown for the focal leech species. Therefore, I do not know the temporal resolution of the mammal detections, and, while this may be shorter than 120 days for *Haemadipsa* spp., I have only retained the most abundant OTUs per sample (removal of singletons etc) this is likely going to limit vertebrate detection to the most recent blood-meal only and potentially, standardise the time-frame of our detections between leech pools.

Previous studies have used several different molecular techniques for identifying iDNA, including PCR-only (Lee *et al.* 2015), qPCR (Kampmann *et al.* 2017), Sanger sequencing (Schnell *et al.* 2012) or shot-gun sequencing of individuals (Gómez & Kolokotronis 2016) and high throughput sequencing (HTS) of amplicon pools (Calvignac-Spencer *et al.* 2013b). Here my results independently confirm the observation by Schnell *et al.* (2018) that sample throughput can be successfully maximised by pooling individual DNA extracts before screening for iDNA, a technique that allowed me to conduct such a comprehensive investigation of the area. By only sequencing leech pools that contained successfully-amplified DNA, I was able to maximise cost-effectiveness in the study. At the same time, however, pooling represents a trade-off; by not determining the feeding behaviour of individual leeches within the pool, the importance of some mammalian prey may be under estimated. Discerning the consequences of pooling for iDNA detection in leeches is an important consideration when applying these technologies in conservation monitoring programmes.

2.5.4. Leeches in human-modified forests

I sampled leeches in a degraded human-modified landscape, which is becoming a typical ecosystem in Southeast Asia and has been associated with a different mammalian composition compared to primary rainforest (Wearn *et al.* 2017). Using iDNA, there was a 30-40% overlap in genera detected compared to two camera trapping studies which were conducted at the same field site (Deere *et al.* 2017; Wearn *et al.* 2017). With a much shorter field sampling campaign compared to the comprehensive camera trapping of these two studies, this highlights the potential of the iDNA method as a rapid and complementary sampling tool. Detecting diversity in leech diets is affected by many factors. One such factor being the restriction of terrestrial leeches to areas with high humidity as a consequence of their evolutionary history (Borda & Siddall 2004) and as I found in this study, some heavily degraded and open forest plots yielded no leeches on multiple visits. While it is unknown how leech populations will be affected with increasing land use change, temperature increases and humidity decreases as forests are fragmented (Hardwick *et al.* 2015) and logging has already been shown to affect a wide range of invertebrates (Ewers *et al.* 2015). As such, it is likely that land-use

change will have a detrimental effect on terrestrial leech populations. It might be beneficial therefore to test alternative invertebrate samplers, such as blow-flies, which are found in a greater variety of habitats (Calvignac-Spencer *et al.* 2013b).

One observable consequence of sampling leeches in logged forests was the high proportion of human DNA detected in my samples (45% of samples with >10% of total copy number). Human activity is high in degraded forests (also see Weiskopf *et al.* 2017) especially around the SAFE project field site, where there are semi-permanent forestry and oil-palm settlements scattered throughout the landscape. Thus, I would assume human blood-meals are sustaining leech populations in degraded landscapes.

2.6. Conclusion

It is important to understand how invertebrate behaviours will introduce biases and affect the biodiversity estimates from iDNA monitoring. By exploring the diets of two leech species, I found that they are not equal in their ability to detect mammals. Therefore, I would recommend that in the forests, where the focal leeches are co-occurring, *H. picta* is the more effective iDNA sampler for both molecular and behavioural reasons. However, in habitats where only *H. sumatrana* is found, iDNA recovered from this species should be sufficient to detect common mammals. The lack of small mammal detections from their diets, shows how little we know about terrestrial leech behaviours. As such I emphasise the need to for a greater ecological understanding invertebrate sampler of choice and how the species interacts with the environment. I would recommend more studies, such as this one, especially if the ultimate goal is for conservation monitoring. For these particular leech species, this study adds to the understanding of their feeding ecology for which previously there was little known and puts us a step closer to utilising iDNA in future monitoring programmes. Finally, few iDNA studies have considered the effects of over-harvesting invertebrate sampler species, and the conservation implications of this are not known (see Schnell *et al.* 2015a). In general, the role of leeches in the ecosystem is poorly understood; thus, I advocate to reduce the numbers extracted, in areas of Borneo where *H. picta* and *H. sumatrana* co-occur, only *H. picta* is needed to sample the community of mammals.

2.7 Supplementary Information

Additional OTU sequences and OTU tables can be found online at the NERC Environmental Information Data Centre, with the associated DOI: 10.5285/3affed0d-fe6f-4916-89e3-e672639191e5

Supplementary tables

Table S2.1. Description of the novel sequences generated for this study. Table shows the tissue origin and accession number from NCBI GenBank. Institution ID's are MNHM = Muséum National d'Histoire Naturelle, Paris, FR; USNM = Smithsonian National Museum of Natural History, Washington, USA; BZM = Museum fur Naturkunde, Berlin, GER; ROM = Royal Ontario Museum, Toronto, CA; NHM = Natural History Museum, London, UK; FMHN = Field Museum, Chicago, USA. The Chinese ferret badger sequence, though not native in Borneo, was used as an alternative closely related sister species to the Bornean ferret badger for which there was no sequence available.

Common name	Family	Genus	Species	Origin	Institution	Accession number
Clouded leopard	Felidae	*Neofelis*	*nebulosa*	Ménagerie, MNHN	MNHN	MG996889
Short tailed mongoose	Herpestidae	*Urva*	*brachyura*	Malaysia, Borneo	USNM	MG996890
Collared mongoose	Herpestidae	*Urva*	*semitorquata*	Borneo	BZM	MG996891
Chinese ferret badger	Mustelidae	*Melogale*	*moschata*	Vietnam	ROM	MG996892
Malay weasel	Mustelidae	*Mustela*	*nudipes*	Malaysia	MNHN	MG996893
Banded linsang	Prionodontidae	*Prionodon*	*linsang*	Cincinnati Zoo	MNHN	MG996894
Small-toothed palm civet	Viverridae	*Arctogalidia*	*trivirgata*	Ménagerie, MNHN	MNHN	MG996895
Hose's civet	Viverridae	*Diplogale*	*hosei*	Borneo	NHM	MG996896
Banded civet	Viverridae	*Hemigalus*	*derbyanus*	Indonesia, Borneo	MNHN	MG996897
Common palm civet	Viverridae	*Paradoxurus*	*hermaphroditus*	Indonesia, Borneo	ROM	MG996898
Malay civet	Viverridae	*Viverra*	*tangalunga*	Philippines	FMNH	MG996899
Common treeshrew	Tupaiidae	*Tupaia*	*glis*	Thailand	MNHN	MG996900

Table S2.2. List of species which were included in the 16S reference database used to taxonomically identify OUT sequences, these species are additional to the native Bornean mammals, and representative of human associated, invasive and non-mammalian species

Common name	Species name	Reason for inclusion
African civet	*Civettictis civetta*	Only one record for Malay civet, sister taxa
Indian crested porcupine	*Hystrix indica*	Only one record for Malay porcupine, no records for thick-spined porcupine, sister taxa
Giraffe	*Giraffa camelopardalis*	Used as positive control, divergent to Bornean species
Domestic cattle	*Bos taurus*	Potentially present in the area, human associated
Malaysian field rat	*Rattus tiomanicus*	Potentially present in the area, human associated
Brown rat	*Rattus norvegicus*	High probability species is present, human associated
House rat	*Rattus rattus*	High probability species is present, human associated
House mouse	*Mus musculus*	High probability species is present, human associated
Domestic dog	*Canis lupus familiaris*	Presence confirmed at the field site
Domestic cat	*Felis sylvestris*	Presence confirmed at the field site
Bent toe gecko	*Gehyra mutilata*	Representative common gecko sequence; potential blood meal
Asian house gecko	*Hemidactylus frenatus*	Representative common gecko sequence; potential blood meal
Chicken	*Gallus gallus*	Representative common bird sequence; presence confirmed at the field site; potential blood meal
Monitor lizard	*Varanus salvator*	Presence confirmed at the field site; potential blood meal
Freshwater fish	*Leprobarbus hovenii*	Representative common fish sequence

Table S2.3. Model structures for the candidate Poisson generalised linear models for testing the effects of habitat type, leech species, site, and two interactions (habitat type and leech species, and site and leech species) on the overall count of mammal detections in leech blood-meals. The values of AIC, degrees of freedom (DF), the delta AIC between the best fitting model and the model set (ΔAIC), the residual deviance (Res. dev) and the adjusted R^2 is given for each model in the candidate set

Model structure	AIC	DF	ΔAIC	Res. dev	Adj - R^2
Habitat + Leech + Site + Habitat:Leech + Site:Leech	307.25	11	12.8	37.5	0.55
Habitat + Leech + Site + Site:Leech	307.25	11	12.8	37.5	0.55
Habitat + Leech + Site + Habitat:Leech	302.60	8	8.2	38.8	0.51
Habitat + Leech + Site	301.25	7	6.8	39.5	0. 50
Leech + Habitat	295.65	3	0	42.6	0.31
Leech + Site	301.25	7	6.8	39.5	0.50
Leech	294.41	2	1.2	41.9	0.25

Supplementary figures

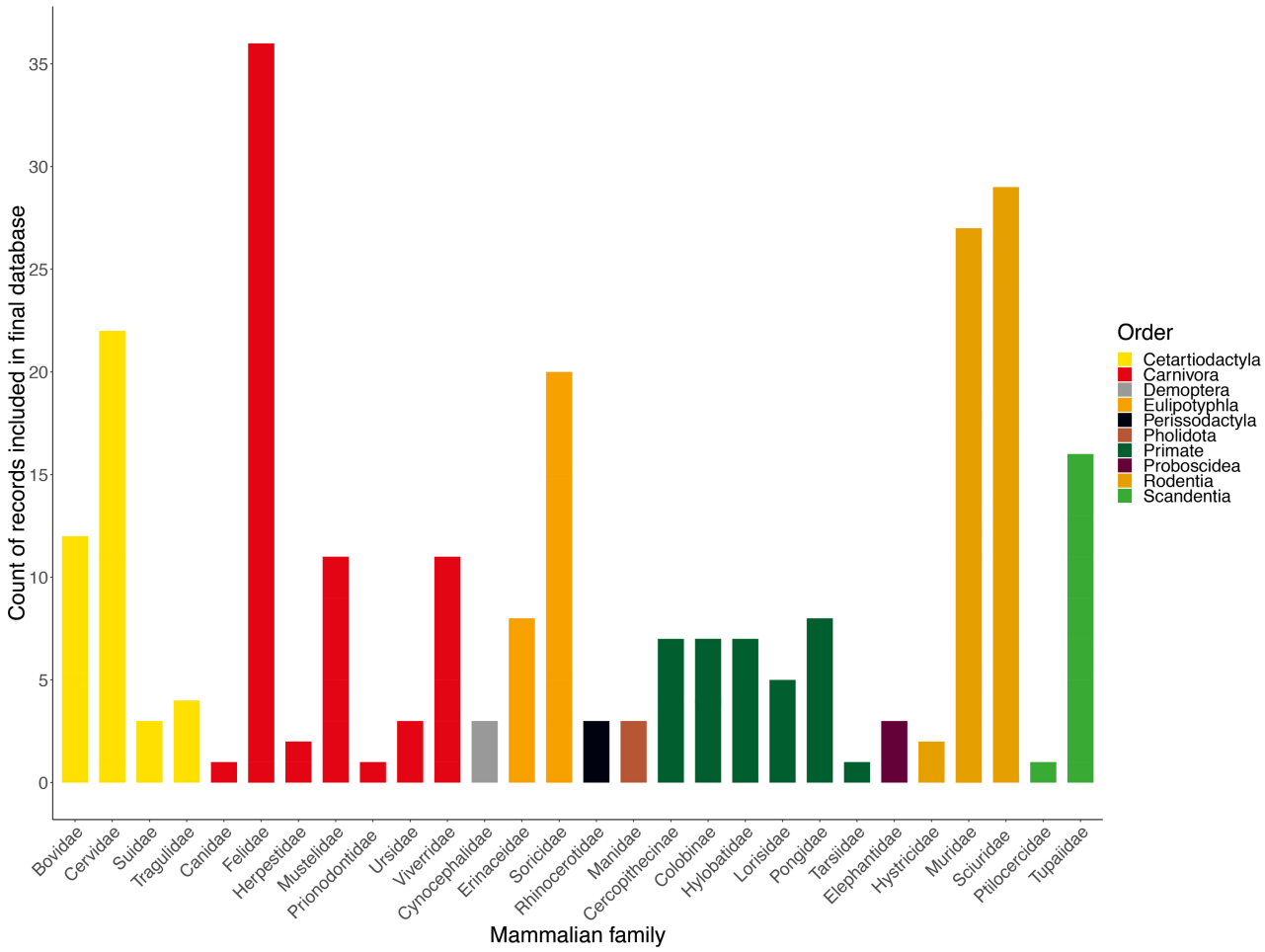


Figure S2.1. Summary of number of records included in the 16S rRNA reference database used to identify taxa from leech blood meals. The height of the bar shows the number of individual records for that mammalian family which were included in the database, gathered from the NCBI database or generated for this study. The bars are coloured according to mammalian order

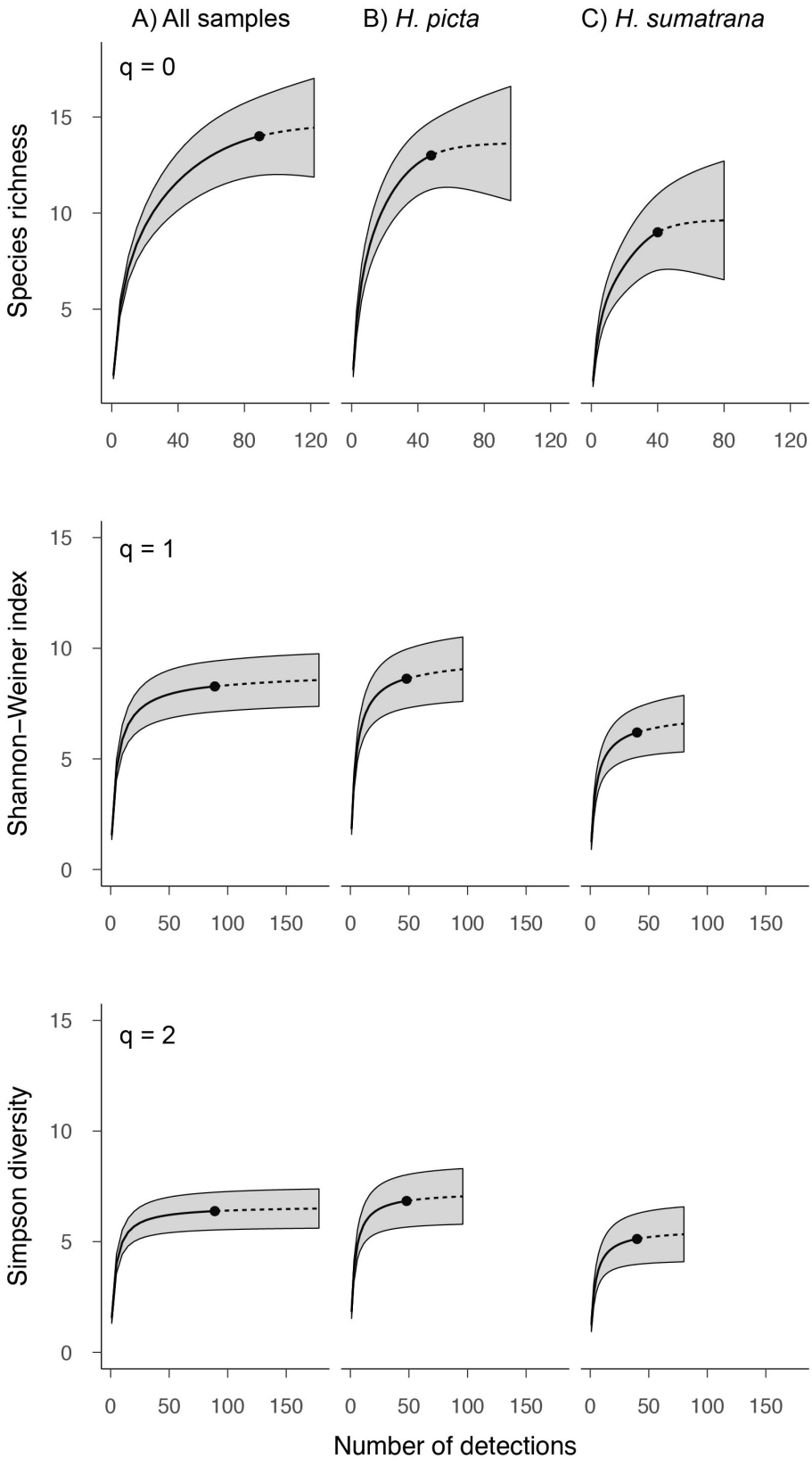


Figure S2.2. Diversity accumulation curves with 95% confidence intervals for: (A) both leech species together, (B) tiger leeches only (*H. picta*) and (C) brown leeches only (*H. sumatrana*). Each row shows the accumulation of mammal genera with increasing detections using the three hill numbers, q = 0, 1, 2 (corresponding to species richness, Shannon diversity index and Simpson index respectively). Curves are extrapolated to 125 samples (dashed lines) (Chao *et al.* 2014)

Chapter 3: Spatio-temporal changes in mammal diversity in a degraded human-modified tropical landscape, an iDNA approach

3.1. Abstract

Degraded forests dominate tropical landscapes outside of protected areas, and there is now an urgent need to understand the biodiversity and conservation value of these habitats. While both traditional survey methods and newer approaches such as camera trapping have been invaluable for conservation, many faunal groups in the tropics can still be a challenge to monitor. This is especially true of forest mammals, many of which are elusive and nocturnal. Emerging non-invasive molecular sampling methods, including those based on invertebrate-derived DNA, offer new opportunities for monitoring biodiversity, and have the potential to become valuable additions to the ecologist's conservation toolbox. Here I use the blood-feeding terrestrial leech *Haemadipsa picta* as a sampler of forest mammals to assess changes in diversity across a forest degradation gradient in Sabah, Malaysian Borneo. I screened over 1700 individual leeches collected in a dry and wet season, for mammal DNA, by targeting a small (~95bp) fragment of 16S rRNA gene, which was amplified and sequenced for pools of leeches. I identified mammals to genus-level using a reference database of Bornean mammals and detected a total of 17 genera from 181 pooled samples collected across the habitat gradient. The proportional phylogenetic richness of both orders and families was higher in the wet season compared to the dry season. I also found a significant effect of season on community composition. Despite this, I found no significant effect of vegetation variables such as canopy height and aboveground biomass on mammal species richness. By detecting broad spatial and temporal differences in the mammal community across a degradation gradient, my findings show that iDNA sequenced from terrestrial leeches can detect local-scale changes in diversity and additional taxa compared to camera traps. In practice, leech-based surveys have the potential to complement other conservation monitoring techniques to detect fine-scale diversity responses to human-mediated tropical forest degradation.

3.2. Introduction

3.2.1. Degraded forest landscapes

Across the wet tropics, recent years have seen a trend of increasing anthropogenic deforestation (Hansen *et al.* 2013) and associated activities leading to forest degradation (Lewis *et al.* 2015). The threats to tropical forests are numerous, and these ecosystems are vulnerable to both local and climatic stressors (Barlow *et al.* 2018). Forest degradation is a particularly grave concern across Southeast Asia which harbours a high diversity of endemic species yet a low percentage of forest is protected (Sodhi *et al.* 2010). Predictions for the future look bleak for this region with estimations of vertebrate extinctions as high as 85% by 2100 based on simple species-area relationships (Sodhi *et al.* 2010). Both oil palm plantation expansion and unsustainable logging practices are the biggest contemporary drivers of forest loss in the region (Stibig *et al.* 2014), with degraded forests being vulnerable to conversion to oil palm plantations (Edwards *et al.* 2011). Yet despite these trends, there is that these habitats can still support biodiversity, and have greater conservation value in comparison to agricultural landscapes (Edwards *et al.* 2011)

Degradation leads to a detrimental change in the natural characteristics of a forest habitat, and a decline in its ability to provide ecological goods and services (Food and Agriculture Organization 2011). Both natural and/or anthropogenic disturbances, such as tree fall from wind or logging, is a mechanism which can lead to forests becoming degraded (however, not all disturbance results in forest degradation) (Food and Agriculture Organization 2011). While the process of forest degradation is known to have particularly negative effects on biodiversity, the specific responses are varied and complex. Indeed Gibson *et al.* (2011) performed a global-scale meta-analysis of the effects of tropical forest degradation, and found that while biodiversity was lower in degraded forests compared to primary forests, responses varied across both taxa and regions. Moreover, degraded forests are not all homogeneous and the activities leading to this degradation can also vary widely. For example, in the context of logging, impacts on biodiversity can depend on the intensity at which the timber is extracted, the geographical region where the logging is taking place (Burivalova *et al.* 2014), and

the method of extraction, with clear differences between, for example, reduced impact logging (RIL) (Edwards *et al.* 2012b) versus salvage logging (Thorn *et al.* 2018). Furthermore, many analyses into the effects of forest degradation have revealed idiosyncratic effects, with biodiversity responses dependent on the species (e.g. Lawton *et al.* 1998), guild (e.g. Wearn *et al.* 2017) or region studied (e.g. Gibson *et al.* 2011).

In combination with habitat loss, a major consequence of land-use change is the fragmentation of once continuous forests (Fletcher *et al.* 2018). The negative effects of fragmentation have been shown in numerous studies across the tropics, including long-term ecological experiments (e.g. BDFFP, Laurance *et al.* 2002; and the SAFE project, Ewers *et al.* 2011). Studies show that increasing fragmentation and low connectivity can cause barriers to gene-flow (Scriven *et al.* 2015) and changes to ecosystem functions such as seed dispersal (Bovo *et al.* 2018) and community composition (Laurance *et al.* 2002). Habitat corridors play a vital role in increasing the connectivity of these isolated forest patches. Within plantations riparian forests surrounding watercourses are set aside to reduce the negative effects of land-use change, and are often protected by law and requirements of sustainable certification e.g. Roundtable on Sustainable Palm Oil (RSPO) (Luke *et al.* 2017). These riparian reserves often serve to connect forest fragments in a mosaic landscape, and there is growing support in their role for supporting biodiversity levels (Gray *et al.* 2014).

3.2.2. Bornean diversity under threat

Borneo, the largest island in Southeast Asia, has long been classified as a biodiversity hotspot due to its high level of endemism, resulting from the region's complex geological history (Sodhi *et al.* 2004) coupled with its high level of habitat loss (Myers *et al.* 2000). The biggest current threats to Bornean biodiversity are rampant commercial logging and conversion of forest for agriculture (Sodhi *et al.* 2004). Currently most agricultural land is used for oil palm plantations, and it is noteworthy that Malaysia and Indonesia - the two countries governing the largest areas of Borneo - are the world's principal producers of palm oil (Food and Agriculture Organisation UN 2017). In total, a 30% reduction in forest cover

largely related to oil palm expansion was reported for Borneo between 1973 and 2010 (Gaveau *et al.* 2014). This figure is partly explained by the realisation that most agricultural expansion for oil palm has taken place in logged-over forests, rather than on old cropland, as was earlier asserted (see Koh & Wilcove 2008).

Currently, most tracts of continuous forest in Malaysian Borneo are found within protected areas (Bryan *et al.* 2013). However, within the agricultural landscape, areas deemed too steep to log are designated as Virgin Jungle Reserves (VJRs) which afford some protected status. As such the forests within these small fragments tend to be of relatively high quality and potentially buffered from the negative effects of land-conversions (Ewers *et al.* 2011). VJRs can be connected by riparian forest reserves, again indicating the potential importance of these riparian habitat corridors (Gray *et al.* 2014; Mitchell *et al.* 2017).

The negative impacts of forest loss on Borneo for biodiversity might be especially serious given that a large proportion of the island's taxa are recognised as being of conservation concern on a global scale. For example, in a study by Brodie *et al.* (2014a) half of the mammals in Sabah that were detected on camera traps are classified by the IUCN as Endangered or Vulnerable. Moreover, these threats might be underestimated due to a lack of knowledge regarding the population status of numerous species, many of which are both threatened and data deficient (Schipper *et al.* 2008). As such, it is likely that species thought to be at risk of population decline or extinction will exclude taxa that are poorly known, and for which their conservation status is unclear.

3.2.3. The importance of mammals in conservation

The success of most environmental and conservation policies can be measured by their impact on either biodiversity or carbon stocks, or increasingly, both (Sollmann *et al.* 2017). Vertebrates such as mammals and birds, frequently feature in policy, and are used as baselines by which to evaluate the effectiveness of decisions. For example, major United Nations programmes such as REDD+ and the Convention on Biological Diversity's Aichi Targets both incorporate biodiversity at their core. The effectiveness of such policies on carbon and biodiversity

conservation needs to be evaluated (Deere *et al.* 2017). In general, mammals show less sensitivity to logging than birds and arthropods (Gibson *et al.* 2011). However, as a proportion of their total diversity, terrestrial mammals show severe threatened statuses based on IUCN assessments (Costantini *et al.* 2016). In an analysis by Costantini *et al.* (2016) 129 Bornean mammals considered 46% have a threatened status, including Vulnerable, Near Threatened or Endangered (IUCN *et al.* 2008) and of the 676 bird species considered, 35% were considered vulnerable. Recently, other groups, such as amphibians, are under threat from widespread disease outbreaks and overexploitation (Stuart *et al.* 2004). Unfortunately, studies have also shown low-levels of cross-taxon congruence in biodiversity 'hotspots' for threatened species of birds, mammals and amphibians (Grenyer *et al.* 2006). However, as a charismatic group, mammals tend to be well studied and thus can be beneficial for conservation monitoring.

Aside from their inherent biodiversity value, mammals are an important group for healthy ecosystem functioning, as both predators and prey species (Moreno *et al.* 2006) but also as pollinators (Ratto *et al.* 2018), seed dispersers (Wright *et al.* 2000), and keystone species (Sinclair 2003). As is frequently highlighted by conservation policy; biodiversity monitoring and research is integral to improving our predictions of the effects of human activity and industry on tropical ecosystems, and thus our ability to stem any negative impacts (Sodhi *et al.* 2010).

3.2.4. Biodiversity monitoring for mammals

At local spatial scales, mammals may show differences in the way that they respond to forest degradation (Wearn *et al.* 2018). In Sabah, some groups of mammals seem more resilient than others to increasing forest degradation (Wearn *et al.* 2017). Small mammals, and more generalist feeding guilds such as insectivores and omnivores, show increases in relative abundance along a habitat gradient from primary to logged forest compared to, respectively, large-bodied mammals and feeding guilds such as carnivores and frugivores (Wearn *et al.* 2017). Other studies have also recorded herbivorous mammals persisting in logged forests where hunting pressures are low, benefitting from pioneer tree growth in forest gaps (Brodie *et al.* 2014a; Granados *et al.* 2016).

Systematically establishing the abundance and diversity of human-modified landscapes enables us to understand the extent to which species persist in these degraded forests. Yet obtaining reliable information on the contribution of mammals to communities in degraded forest ecosystems can be difficult, and several methods have been used. One of the most popular survey techniques for mammals is camera trapping and has been applied to many situations (Ahumada *et al.* 2011; Samejima *et al.* 2012; Mazzolli *et al.* 2017). Remote camera trapping is now a standard technique for detecting the presence of otherwise difficult-to-observe mammals (Meek *et al.* 2014) as well as for describing community diversity and inferring species occupancy (Brodie *et al.* 2014a). Camera traps tend to be biased toward the detection of larger-bodied terrestrial mammals compared to, for example, live trapping methods, which typically record greater numbers of small mammals. Thus a comprehensive understanding of local level patterns and dynamics requires a combination of survey techniques (Wearn *et al.* 2017) but this can be financially and logistically constraining.

Alongside more traditional sampling techniques conservation monitoring using invertebrate-derived DNA (iDNA) could be feasible in the near future. With the reducing costs of metabarcoding, species can now be detected with relative ease by sequencing DNA from environmental, faecal and gut samples, including bloodmeal DNA from invertebrates (Schnell *et al.* 2015a). Several studies have demonstrated the use of such invertebrates as biodiversity monitoring tools, especially using blood feeding terrestrial leeches (Haemadipsidae), which seem to be popular due to their large body size and relative ease of sampling (Schnell *et al.* 2015a). On a broad geographic scale the utility of leeches as vertebrate samplers has been shown (Schnell *et al.* 2018; Tessler *et al.* 2018) but few studies have used iDNA as a tool to quantify local mammal (or vertebrate) diversity (Weiskopf *et al.* 2017). Field work to collect terrestrial leeches requires very little equipment and sites can be surveyed very quickly. This means that leech-based iDNA studies have the potential to complement and help to direct additional monitoring campaigns.

For this study, my main objective was to apply leech-based iDNA to assess and quantify differences in mammal diversity across a gradient of habitat degradation

in Sabah, North Borneo. The Stability of Altered Forest Ecosystems Project (SAFE) is subject to ongoing degradation, with logging activities taking place between the two years of sampling (Ewers *et al.* 2011). I tested whether differences in mammal community diversity could be detected using metabarcoding of leeches, and whether these differences in overall diversity reflect habitat quality over time and space. For this study I focus on an abundant species found in the lowland dipterocarp forests of Sabah, *Haemadipsa picta*, commonly called the tiger leech. By assigning DNA sequences from pooled leech blood meals to mammal taxa, I test the hypothesis that mammalian diversity decreases with habitat degradation. Many studies, on various taxa have demonstrated a negative effect of forest degradation on levels of overall biodiversity (e.g. Gibson *et al.* 2011; Edwards *et al.* 2014). To test this, I construct Generalised Linear Mixed Effects Models (GLMMs) in which I model richness and Shannon's diversity index against habitat quality metrics and seasonal differences. Finally, to test the hypothesis that the dissimilarity in community composition increases across the habitat gradient, and between years, I use non-metric multi-dimensional scaling and permuted ANOVAs.

3.3. Materials and methods

3.3.1. Study design and sample collection

To obtain information on temporal and spatial changes in mammal diversity, I undertook sampling at sites across the SAFE landscape in a dry season (February-June) and a wet season (September-December 2016). I aimed to visit the same vegetation plots within these sites in both seasons, but this was not possible in all cases. For example, several sites had become inaccessible in the intervening year due to road degradation and tree extraction (e.g. F). In some cases, I also found sites with no or too few *H. picta* present (e.g. E, C, LF1), which is likely to be due to changes in microclimate resulting from the removal of trees. The field sampling regime was exactly the same as described in Chapter 2, searching the leaf litter and understory for leeches within the boundaries of each vegetation plot, for twenty minutes. I only had one opportunity to sample in the primary rainforest, at the Danum Valley Conservation Area (DVCA). In DVCA the same sampling regime as in SAFE was followed, setting up 25m^2 plots and hand-searching for 20 minutes. A summary of the samples for this chapter can be found in Table 3.1 and a detailed schematic of the sampling design can be found in Figure 3.1.

To measure the effects of habitat degradation, I used two methods of quantifying habitat quality. First, based on land-use history, I used a broad categorical system to classify each site into one of four forest types along a gradient of degradation: primary, riparian, twice-logged and heavily logged forest. Primary forest at DVCA was of the highest quality, followed by the twice-logged forest which had undergone two rounds of selective logging but remained within continuous forest. The sites of the lowest quality were the heavily logged forest sites which were all within the SAFE experimental area and were eventually to be isolated within an oil palm matrix (Figure 2.1). Previous studies have used a similar land-use intensity gradient, showing primary forest has higher maximum canopy heights and greater PAI than increasingly logged forests (Jucker *et al.* 2018). Finally, there are the riparian sites, three of which were along rivers in the heavily logged forest in SAFE. The riparian forest at within the SAFE area was in the process of being experimentally fragmented to different widths of 0 m, 5 m, 30 m and 120 m as part

of the wider SAFE project (Ewers *et al.* 2011). Riparian forest width is an important factor in determining levels of biodiversity at the site, however, the optimum width is location and species dependent (Luke *et al.* 2018; Mitchell *et al.* 2018). The quality of the forests along these rivers is highly variable and consists of a high proportion of forest edge, thus they tend to be of similar quality to the surrounding heavily logged forests. One river (RLFE) which was located within the twice-logged forest (LFE) which meant that the riparian forest surrounding this river, was of higher quality than the other three sites. Second, across all of these sites, I obtained several metrics describing vegetation structure, which were obtained from remote sensing data (for details see below). Analyses based on these two measures are hereafter referred to as 'forest type' and 'vegetation', respectively.

Table 3.1. Summary of number of leech pools used in this study, which site the samples were collected from and from within which forest type. The number of leech pools sequenced in the dry and wet season for each site is given, along with the total number of individuals that this corresponds to in brackets. Numbers of individual leeches per pool ranged between 4-12 individuals (mean = 9, median = 10)

Site	Forest type	Dry season pools - 2015	Wet season pools - 2016
OG	Primary	0	23 (222)
VJR	Twice-logged	6 (63)	5 (37)
LF1	Twice-logged	1 (6)	0
LF2	Twice-logged	3 (30)	0
LF3	Twice-logged	9 (92)	12 (117)
LFE	Twice-logged	18 (180)	14 (133)
B	Heavily logged	14 (136)	7 (70)
D	Heavily logged	8 (73)	10 (87)
E	Heavily logged	2 (18)	3 (28)
F	Heavily logged	5 (44)	0
R0	Riparian	6 (60)	1 (8)
R30	Riparian	7 (70)	0
R5	Riparian	7 (61)	0
RLFE	Riparian	15 (139)	5 (50)

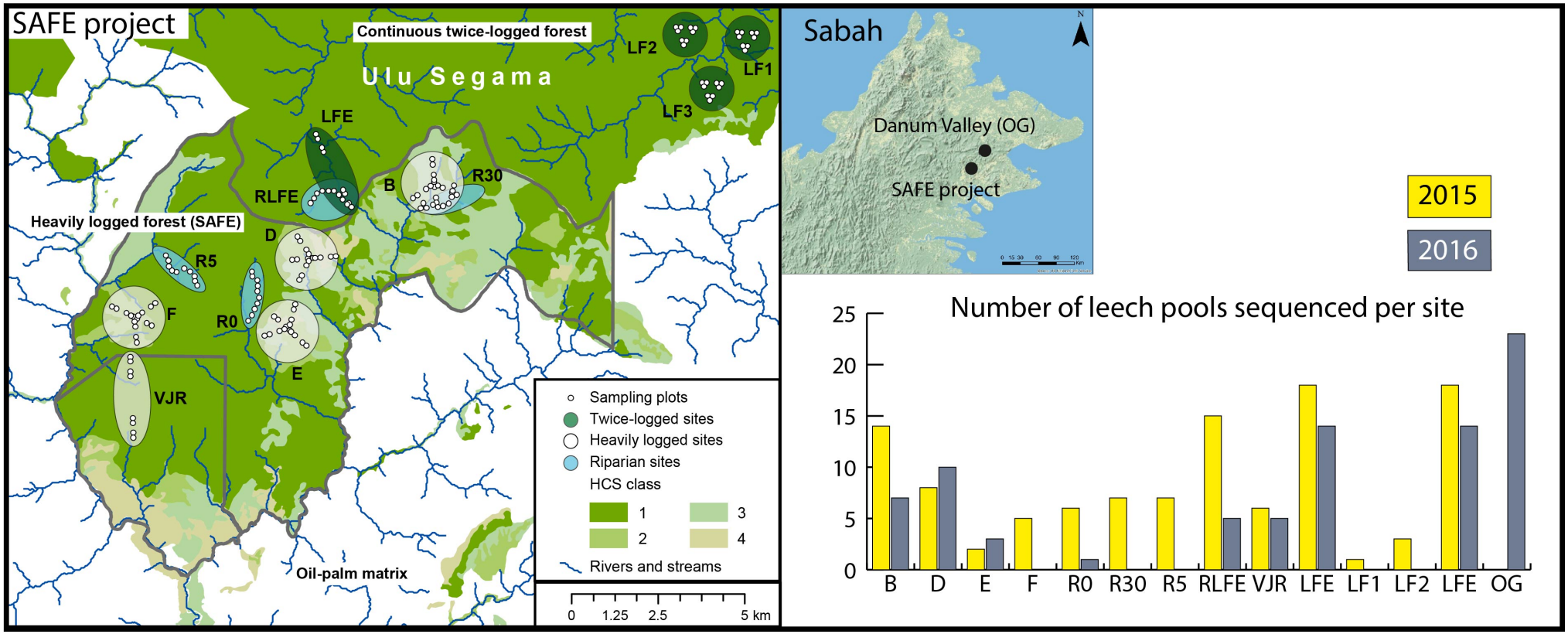


Figure 3.1. Schematic map of the sampling design used in this chapter. The twice-logged habitats are shaded in dark green within the continuous logged forest, the heavily logged habitats are shaded in white and riparian sites are shaded in blue within the SAFE project experimental area. The location of the vegetation plots where the sampling took place are shown as small white circles. Blue lines represent the rivers and streams, and the green shows the different classes of high carbon stock forest (HCS). Bar chart shows the number of pools sequenced at each of the sites for *H. picta* in 2015 in yellow and 2016 in grey. Inset map shows the location of the SAFE project and Danum Valley for the primary forest sites (old growth - OG) in the state of Sabah, Malaysian Borneo.

3.3.2. DNA extraction, PCR amplification and library pooling

For this study I generated new sequence data from leeches collected in 2015 and 2016. For a subset of leeches from 2015 from sites B, D, F (heavily logged forest sites) and LFE, LF2 and LF3 (twice-logged forest sites) I re-extracted the DNA (using a modified protocol) from the stored digests. I took this approach, rather than re-use the same extracts, to increase DNA yield, and also to avoid batch effects from different sequencing runs.

All steps for DNA extraction followed those described in Chapter 2, with the addition of an extra lysis step following the initial incubation with proteinase K and pooling of individuals. At this point, I added 200µL of buffer AL, from the DNeasy Blood and Tissue kit (Qiagen), to a 200µL subsample of each pooled digest then incubated for 15 minutes at 56°C. I then mixed in an additional 200µL of 100% ethanol and added the samples to QiaQuick spin columns (Qiagen) and centrifuged at 6000g for one minute. The application of the modified method was based on preliminary results that indicated a small but non-significant increase in the number of OTUs and taxa recovered (see Supplementary Figure S3.1). The DNA was then purified and resuspended, again following the same protocol as in Chapter 2.

As described in Chapter 2, alongside each batch of the extractions I also conducted at least one extraction control (i.e. a blank sample that contained all of the reagents minus the tissue). A subsample of the extracts from each batch of extractions was quantified using the Qubit dsDNA HS Assay Kit (Invitrogen), to check the success of the extractions and the DNA concentration of the negative controls. PCR amplification steps followed the protocol in Chapter 2. Successful PCR replicates were mixed into amplicon pools for a single-tube library build (Carøe *et al.* 2017). The amplicon pools were sequenced in two batches: batch1 = 11 amplicon pools, and batch 2 = 22 amplicon pools. All samples were sequenced at The Genome Centre (Queen Mary University of London) using the Illumina MiSeq platform for a target of 150bp paired end reads.

3.3.3. Taxonomic assignment

Full details of the taxonomic assignment process can be found in Chapter 2. Briefly, I merged read pairs together, and then matched the nucleotide tags on the ends of the amplicons to original leech samples using the DAMe pipeline, retaining only those sequences that appeared in a minimum of two PCR replicates. Chimeric sequences were removed using mothur (Schloss *et al.* 2009) before clustering at 97% similarity with sumaclust (Mercier *et al.* 2013) and post-clustering filtering algorithm with LULU (Frøslev *et al.* 2017).

All OTUs were identified to genus, except for those 11 species for which no congeners occur in Sabah; these could thus confidently be identified to species-level. Genus-level identification is preferred when using potentially incomplete databases (Kocher *et al.* 2017b) and there is evidence to suggest this represents the best approach when working with ribosomal markers due to the limited taxonomic resolution achieved (Hillis & Dixon 1991; Axtner *et al.* 2018). As the highest taxonomic level common to all assignments, all analysis is conducted at genus-level. The majority of OTUs matched well to the database, with the exception of two OTUs that had poor quality matches; one of these matched poorly to a leaf monkey sequence, and the other to an amphibian which were removed. Two further OTUs were removed as they were suspected to be errors, matching only to nuclear DNA. The same threshold filtering approach was used as in Chapter 2 to screen the samples, based contamination found in the controls.

3.3.4. Vegetation structure

To quantify habitat quality, I used vegetation data from a Leica ALS50-II LiDAR sensor flown by NERC's Airborne Research Facility which covered the SAFE landscape in 2014 (details in Jucker *et al.* 2018, data provided by T. Swinfield and D. Coombes). The habitat had undergone experimental logging between the LiDAR flight and the surveys, however, as the sites within SAFE should remain untouched, this data were deemed appropriate. Vegetation metrics which could differentiate between the variable structure of the forest sites were extracted at the site level. Using a 1 km buffer around the centroid of the site ensured that there was no overlap in the metrics used for each site while still being appropriate to the size of

potential home-ranges of the mammal species. These metrics were canopy height (range = 8.66 m to 33.79 m, mean = 20.06), above ground biomass (range = 21.46 kg to 199.76 kg, mean = 91.81 kg), gap fraction (range = 0.02 to 0.62, mean = 0.22) (the inverse of which is forest cover) and a measure of habitat heterogeneity (Moran's I) (range = 0.38 to 0.74, mean = 0.57). Habitat heterogeneity values ranges from -1 to 1, representing a gradient from evenly dispersed canopies up to perfect clustering, while a value of zero represents perfectly random canopy dispersion. In practice, values approaching 1 indicate greater clustering of the canopy and thus represent strong contrasts in habitat availability such as gaps or very large trees within a matrix of intermediate canopy heights. Negative values are rare in natural forests and intact, homogenous canopies would have values closer to zero. Due to the relatively small sample size, I constructed a principal component analysis (PCA) to include information from all the metrics in a singular value, in the subsequent models.

3.3.5. Statistical analysis

Species accumulation

For all analyses, the leech pool is considered the sampling unit. Using taxonomic assignments generated from the OTU data, I estimated alpha diversity of mammals at each of the forest types using the Chao2 species richness estimator (Chao 1987; Gotelli & Colwell 2011). This metric can account for 'under-sampling' and thus allows estimation of the taxonomic richness of the actual pool.

To obtain richness estimates for each forest type as well as the entire sample, and to check whether the community of mammals had been fully sampled, I generated accumulation curves using sample-based rarefaction and extrapolation. Each curve was recalculated for three Hill numbers (q = 0,1,2), which represent bias-corrected estimates of, respectively, species richness, the exponential of the Shannon-Wiener index, and the Simpson diversity metric (see Chao et al., 2014). The use of Hill numbers is recommended in cases of incomplete sampling, for example due to a high abundance of rare species (Chao *et al.* 2014). In this study, estimates of alpha diversity based on invertebrate samplers of mammals will certainly suffer from incomplete sampling. This is likely given that mammal DNA is rare in the mixed

DNA sample, in part due to the potentially long intervals between leech feeding events, as well as degradation of the mammal DNA over time within the blood meal. Extrapolation was conducted using Chao2 to estimate the undetected diversity in the reference sample (i.e. the observed value). I extrapolated to double the sample size of the reference for the reason that where q = 0 (i.e. species richness), extrapolations beyond this point have been shown to become unreliable whereas estimates for q = 1 and q = 2 remain relatively unbiased (see Chao *et al.* 2014). For each curve, the confidence intervals were generated using the bootstrap method proposed in Chao *et al.* (2014) for 1000 replicates. To compare curves, I used 84% confidence intervals (CI) rather than 95% CIs, which have been shown to be more conservative than an alpha level of 0.05 (see Chapter 2). Curves with 95% CIs are provided for information (Supplementary Figure S3.2). All curves were produced in the iNEXT package (Hsieh *et al.* 2016) in R (R Core Team 2018).

Effects of habitat quality and vegetation variables on diversity

To identify the factors determining the diversity of mammals detected among pools from across the habitat gradient, I used generalised linear mixed effect models (GLMM). I repeated the models for two response variables based on, respectively, taxon richness (Hill = 0) using a Poisson error distribution, and Shannon's diversity (Hill = 1) using a normal error distribution. In each model, I included several fixed effects: forest type, year of sampling, number of leeches sequenced per pool, and principal component 1 from the PCA of vegetation structure. I used site identity as a random effect to account for potential spatial autocorrelation, as I would expect pools of leeches collected from within one site to be more similar than between sites.

For each response variable, I first generated a 'global model' with all variables and two interactions: year and forest type, and year and PC1. With likelihood ratio tests, I simplified the model by removing non-significant terms. Then I identified the best-fitting model using AIC (delta AIC <4 and/or an AIC weight >0.09) which are the commonly used thresholds discussed in Burnham & Anderson (2002). Detections from the primary forest were excluded from this analysis due problems

with introducing bias from only one replicate. I used the packages lme4 for GLMMs (Bates *et al.* 2015) in R.

Community composition across a habitat gradient

To visualise the differences between mammal community composition between the different forest types and years, I used non-metric multidimensional scaling (NMDS) based on Chao's dissimilarity metric. This metric (like the richness estimator for alpha diversity) shows the effects undetected species on the whole species pool (Chao *et al.* 2005). Chao *et al.* (2005) show that classic measures i.e. Jaccard/Sørensen indices perform poorly when a sample consists of a high number of rare species. I checked for relationships between the communities at each site and the vegetation metrics (canopy height, aboveground biomass, habitat clustering, and gap fraction) by fitting environmental vectors to the ordination. To test for differences in variance between the factors of forest type and year, I used a permuted analysis of variance (PERMANOVA). All models were run for 9999 permutations and constrained by site identity to reflect the study design. All these analyses were conducted in the vegan package (Oksanen *et al.* 2017) in R.

3.4. Results

3.4.1. Sequence summary

I retrieved a total of 13,311,805 forward and reverse reads from the two sequencing runs (batch1 = 6,114,937 reads and batch2 = 7,196,868 reads). There was an overall success rate for merging the paired reads of 93.30%, with slightly higher success for the batch1 reads (96.81%) compared to the batch2 reads (90.31%) due to the low sequencing yield for two of the amplicon pools. In total, there were 12,419,253 successfully merged reads. Of these, 81.99% contained the correct primer and tag combination, 11.37% had no primer sequence, and 6.11% of the reads were tag jumps. Only 12,746 reads (0.1% of successful reads) were identified as chimeras and subsequently removed. The initial clustering resulted in 128 OTUs which was reduced to 43 OTUs after applying the post-clustering filter. When I removed the contamination and collapsed OTUs with matching taxonomic assignments, I was left with a final set of 17 OTUs found in 181 samples.

3.4.2. Identity of mammals

All OTUs could be identified to at least genus-level. Overall, I found evidence of mammals from 17 genera, 12 families, and six orders. All OTUs matched to the reference database with a percent similarity greater than 90%, with the one exception of OTU27, which consistently matched to the gibbon reference sequences but with lower confidence (79% similarity). I was able to assign ten OTUs to species-level with confidence based on the knowledge that only a single member of the genus occurred in the area (Table 3.2). I found a maximum of four mammals in one leech pool (three samples; F.2015, D.2016 and LFE.2016) but the average detection per pool was one and out of 181 samples, 72 contained no amplifiable (non-human) mammalian DNA.

Table 3.2. Taxonomic assignments of unique OTUs given with the classification of order, family, genus and species level identity (in cases where there were no closely related congeneric species in the study region). Bit score and % similarity are reported from BLAST and where there two OTUs were assigned to the same taxonomic group, two values are given for IUCN status, and confidence values

Common name (IUCN status) *	Order	Family	Genus	Species	Bit score	%Similarity
Cow (LC/EN)	Cetartiodactyla	Bovidae	*Bos*	*Bos* sp	171	100%
Muntjac (LC/NT)	Cetartiodactyla	Cervidae	*Muntiacus*	*Muntiacus* sp	154/159	90%/91%
Sambar deer (VU)	Cetartiodactyla	Cervidae	*Rusa*	*Rusa unicolor*	171	100%
Bearded pig (VU)	Cetartiodactyla	Suidae	*Sus*	*Sus barbatus*	174	100%
Mousedeer (LC/LC)	Cetartiodactyla	Tragulidae	*Tragulus*	*Tragulus* sp	141/171	94%/99%
Cat (LC/EN)	Carnivora	Felidae	*Prionailurus*	*Prionailurus* sp	169	100%
Sun bear (VU)	Carnivora	Ursidae	*Helarctos*	*Helarctos malayanus*	167	99%
Small toothed palm civet (LC)	Carnivora	Viverridae	*Arctogalidia*	*Arctogalidia trivirgata*	150	96%
Banded civet (NT)	Carnivora	Viverridae	*Hemigalus*	*Hemigalus derbyanus*	176	100%
Masked palm civet (LC)	Carnivora	Viverridae	*Paguma*	*Paguma larvata*	174	100%
Malay civet (LC)	Carnivora	Viverridae	*Viverra*	*Viverra tangalunga*	178	100%
Sunda pangolin (CR)	Pholidota	Manidae	*Manis*	*Manis javanica*	159	100%
Gibbon (EN)	Primate	Hylobatidae	*Hylobates*	*Hylobates* sp	60	79%
Macaque (LC/VU)	Primate	Cercopithecidae	*Macaca*	*Macaca* sp	141	95%
Elephant (EN)	Proboscidea	Elephantidae	*Elephas*	*Elephas maximus*	172	92%
Porcupine (LC/LC)	Rodentia	Hystricidae	*Hystrix*	*Hystrix* sp	159/167	90%/91%
Long-tailed porcupine (LC)	Rodentia	Hystricidae	*Trichys*	*Trichys fasciculata*	161	90%

**IUCN classifications are as follows: LC = Least Concern, NT = Near Threatened, VU = Vulnerable, EN = Endangered, CR = Critically Endangered*

Overall the number of mammalian orders and families detected was greater in 2016 compared to 2015 and was highest in the twice-logged forest sites (Figure 3.2). In all forest types, terrestrial members of the order Cetartiodactyla were the most frequently detected in both years and in all of the four forest types. In 2016 there were more detections of mammals from the Rodentia with additional detections in twice-logged forest compared to 2015. Primates were detected at low numbers across all forest types in 2015, but only in the higher quality forests in 2016. In addition, elephants (Proboscidea; Elephantidae) and Sunda pangolins (Pholidota; Manidae) were only detected in 2016, in both cases in the twice logged forest.

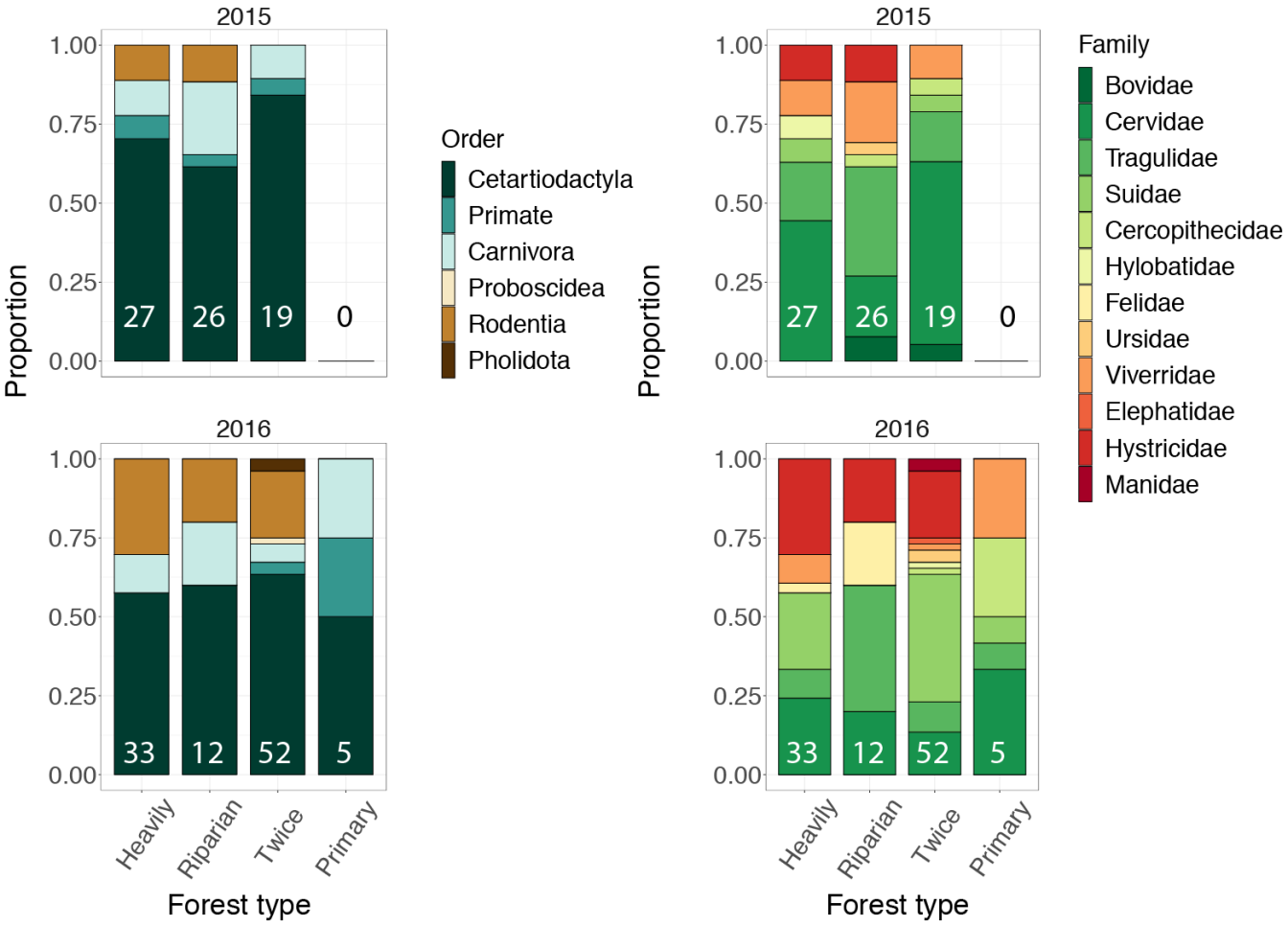


Figure 3.2. Proportion of mammalian taxonomic groups recorded in the leech blood-meals from 2015 and 2016. based on taxonomic order (left) and family (right) for each forest type. Sample number shown in white at the base of each bar.

3.4.3. Taxon diversity

I observed 17 genera across all samples, but using the Chao2 non-parametric estimator, richness was calculated as 19.24 (±3.38) genera (Table 3.3). The alpha diversity I observed was greatest in the twice-logged forests and lowest in the primary forest. Based on the richness estimator richness values, however, I found that for heavily logged forest had the lowest diversity of the four forest types, and closest to the observed richness value. The estimated richness for the primary forest sites increases by over 50%, potentially as a consequence of the small sample size in this forest type, although with a large standard error the value.

Table 3.3. Observed and estimated taxon richness at the genus level. Estimated richness calculated with the Chao2 estimator (Chao, 1987), for all samples combined and for each forest type separately

Forest type	Observed richness	Estimated richness (±S.E)	Sample size
All samples	17	19.24 (3.38)	181
Twice-logged	14	17.07 (3.60)	57
Heavily logged	10	11.97 (3.68)	60
Riparian	10	13.90 (5.17)	41
Primary	7	14.65 (11.17)	23

The greatest differences in accumulation of genera along the habitat gradient is seen for genera richness (q = 0). Twice-logged forest sites were associated with greater diversity and more rapid accumulation of mammal genera compared to the primary, heavily logged and riparian forests (Figures 3.3B-E). The curve for twice-logged forest also most closely matches that of all samples combined (Figures 3.3A & C). Sampling in heavily logged forest has almost reached the asymptote for richness, indicating near-complete detection of the full assemblage of mammals that are fed on by leeches in this disturbed forest type (Figure 3.3D). Looking at the accumulation of diversity at the two orders of Hill numbers, exponential of Shannon's diversity index (q = 1) and the inverse of Simpsons diversity, the habitat specific differences become minimal (Figures 3.3F-O). The same pattern of diversity accumulation is seen for the heavily-logged and twice-logged forest sites when analysing the Simpson's diversity metric (Figures 3.3M & N). However, all

three diversity curves for primary forest habitat indicate under sampling (Figures 3.3B, G & L), where even after the recommended extrapolation the curves do not reach a stable plateau

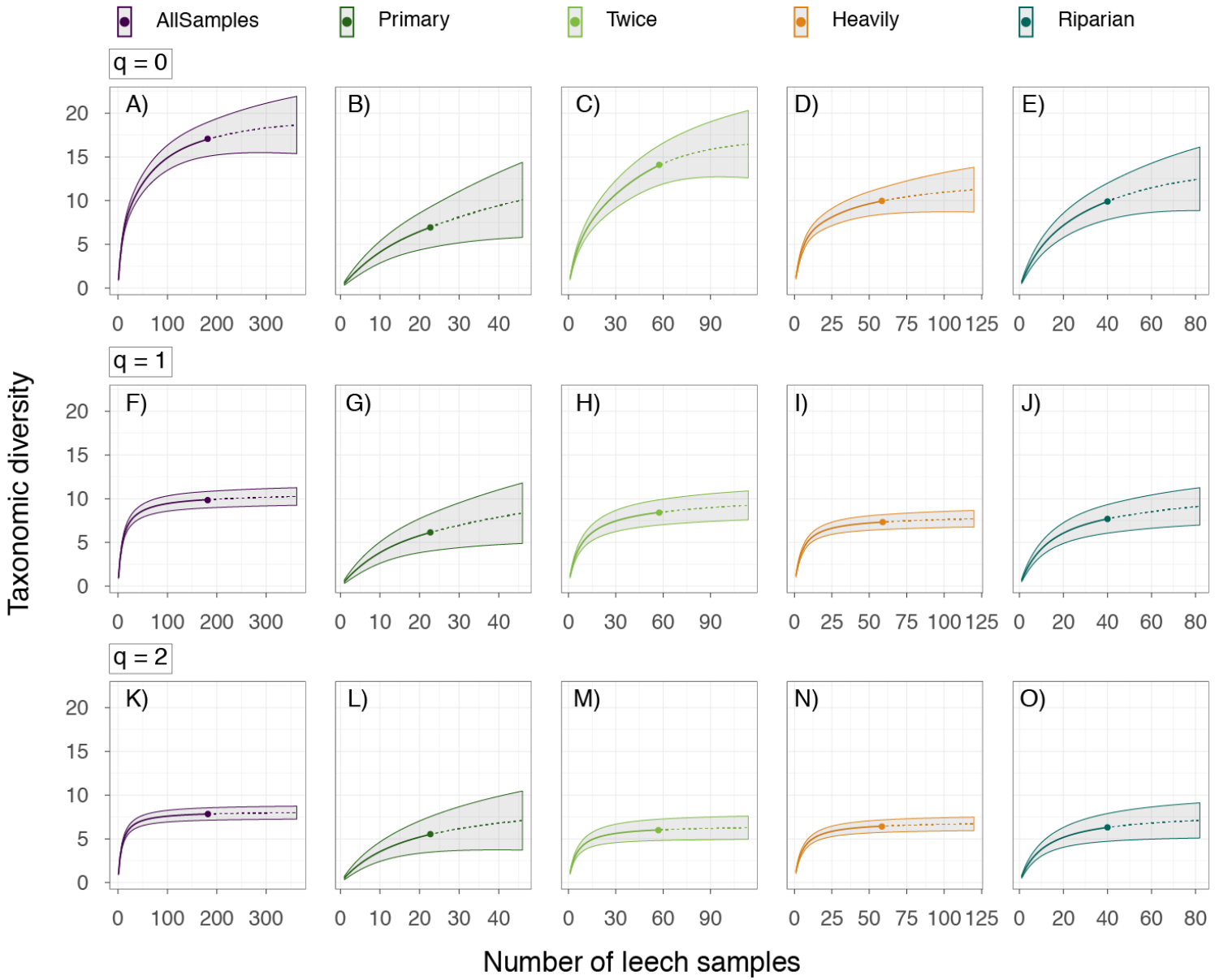


Figure 3.3. Diversity accumulation curves (at the genus level) comparing the effect of increasing leech samples on the taxonomic diversity. The curves are calculated using three orders of hill numbers, q = 0, 1 & 2 which are equivalent to species richness, Shannon diversity index and Simpson index, respectively. The diversity of all samples combined (A, F & K) is compared to the accumulation of diversity in the four different habitat types: primary (B, G & L), twice-logged (C, H & M), heavily-logged (D, I & N) and riparian (E, J & O). The x-axis varies depending on the number of samples. The solid line represents the rarefied values, and the dashed line represents the extrapolated values and is extended to double the reference sample (empirical value, solid circle). The accumulation curves are presented with 84% confidence interval which has been shown to be equivalent to a significance value of α = 0.05 (MacGregor-Fors & Payton, 2013)

3.4.4. Effects of habitat quality on mammal diversity

To obtain a single measure of vegetation structure for use in models of mammal diversity, I first performed a principal component analysis (PCA) of the four measures of vegetation structure across the study sites. In this analysis, PC1 explained 86.9% of the variation, and PC2 and PC3 explained 8.4% and 4.7%, respectively (Figure 3.4, Supplementary Figure S3.3) The PCA showed separation of sites classified as logged forest from those classified primary forest, with further separation of sites based of logging intensity. Sites from heavily logged and riparian forest formed broadly overlapping clusters, with the exception of the least disturbed riparian site (RLFE) that was more similar to other sites sampled from within the same continuous twice-logged forest. Examination of the vector loadings revealed that primary forest and twice-logged forest are correlated with greater above-ground biomass, canopy height and proportion forest cover. Conversely, heavily logged and riparian forest are more closely associated with habitat clustering, indicative of canopy height heterogeneity, and gap fraction (inverse of forest cover) (Figure 3.4). When plotting PC1 against PC3, the same patterns are seen but with larger confidence ellipses (Supplementary Figure S3.3).

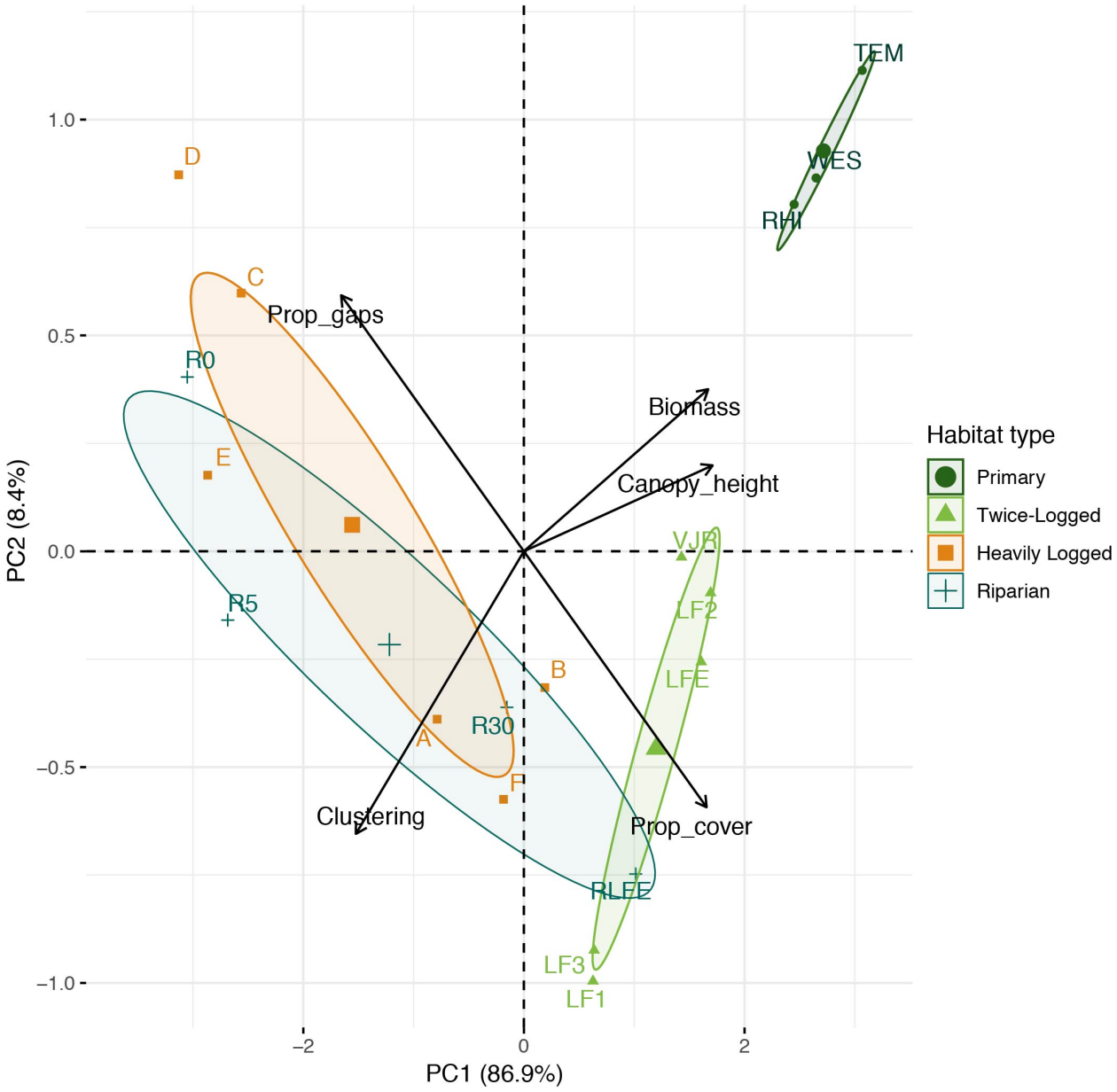


Figure 3.4. Principal component analysis (PCA) showing the relationship between components of vegetation structure (as calculated from LiDAR data) and the sampling sites where surveys took place. The first (PC1) and second (PC2) axis, which explain the most total variation, are shown with the respective percentages. Points are labelled with the site ID and the mean centroid point, while habitat types are denoted by different symbols and colours. Ellipses show the 95% confidence intervals around the four habitat types. The vegetation metrics shown are habitat heterogeneity (Clustering), forest cover (Prop_cover), gap fraction (Prop_gaps), aboveground biomass (Biomass) and canopy height (Canopy_height), and the direction of arrows shows the relationship between the metrics and sites.

To determine the impact of habitat quality on mammal diversity (Richness and Shannon diversity index) I constructed GLMMs including continuous metrics of vegetation structure (PC1 from my PCA) and the categorical forest type. For the models in which richness was the response variable, I found three models, each associated with a ΔAIC of <4 and with a corresponding AIC weight of >0.09. All three models included the fixed effects of both year and forest type (Table 3.4; all model summary in Supplementary Table S3.1). Greater richness was detected in the wet season (2016) compared to the dry season (2015). Additionally, higher richness was detected in twice-logged forest types compared to heavily logged forest (Figure 3.5) but this difference was not significant (Table 3.5). These two variables were the only significant predictors of richness in the leech pools; differences in annual richness are shown in Figure 3.5. There was no significant effect of either the continuous vegetation metrics (PC1) or the number of leeches sequenced, and there was only weak and non-significant support for an interaction between forest type and year. The fixed effects in the best-fitting models explained approximately 20% of the variance in species richness ($R^2 = 0.23$).

For the models with the Shannon diversity index as the response, model simplification using likelihood ratio testing resulted in a single model with only year remaining as a fixed effect (Table 3.4; model comparisons in Supplementary Table S3.2). However, year was not a significant predictor of Shannon diversity index and given that this metric is more heavily weighted toward common species; these model results imply common species show no clear response to habitat quality for either year.

Table 3.4. Best fitting Poisson GLMM models as determined by ΔAIC. The model residual deviance, the conditional R^2 and weight is given. The response variable for each model is either *Richness* or *Shannon-diversity* index. The fixed effects structure for each model is shown where *'Type'* = forest type, *'Year'* = sampling year, *'Leeches'* = number of leeches sequenced and '*Year:Type'* = interaction between year and forest type

Model	Response	Fixed effects	DF	AIC	ΔAIC	Weight	Resid. dev.	Cond R^2
M4.SR	Richness	Year + Type +Year:Type	7	397.6	0.0	0.45	149.0	0.23
M3.SR	Richness	Year + PCA + Type + Year:Type	8	398.4	0.8	0.30	149.3	0.23
M2.SR	Richness	Year + PCA + Type + Leeches + Year:Type	9	399.9	2.3	0.14	149.8	0.23
M7.sh	Shannon	Year	4	153.8	0.0	0.98	125.7	0.13

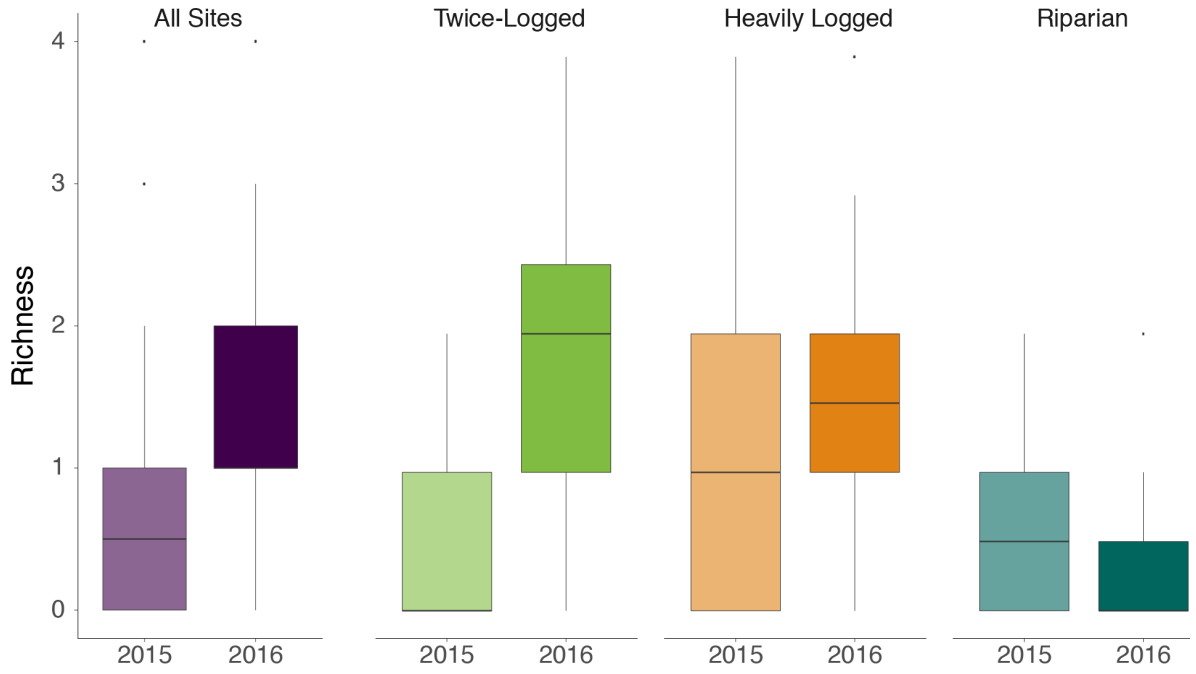


Figure 3.5. Comparison of the number of genera detected from within the leech pools sampled across the habitat gradient, from twice-logged, heavily-logged and riparian forest, in 2015 (dry season) compared to 2016 (wet season). Primary forest was not included as it was not sampled in 2015. Richness is measured in number of unique genera detected in each forest type

Table 3.5. Model summary table for each of the best-fitting GLMM Poisson models from the candidate model set. Three models with Richness as the response variable (M4.SR, M3.SR, & M2.SR) and one model with the Shannon diversity index as the response variable (M7.sh). Estimates for each parameter is given with the corresponding standard error and * indicates the parameter is significant at 0.05

Parameter (±S.E)	**M4.SR**	**M3.SR**	**M2.SR**	**M7.sh**
Intercept	-0.13 (0.16)	0.02 (0.18)	-0.04 (0.51)	0.15 (0.04)
Wet (2016)	0.68* (0.16)	0.71* (0.16)	0.71* (0.16)	0.26 (0.06)
PCA		-0.02 (0.02)	-0.02 (0.02)	
Heavily logged	0.14 (0.17)	0.27 (0.19)	0.27 (0.19)	
Logged	-0.14 (0.17)	-0.27 (0.19)	-0.27 (0.19)	
Riparian	-0.44 (0.24)	-0.50* (0.24)	-0.50* (0.24)	
Leeches			0.01 (0.05)	

3.4.5. Spatial and temporal changes in community

In the NMDS using the Chao dissimilarity index (Figure 3.6), I found no separation between the mammal communities identified from leeches among the sites regardless of forest type. Similarly, I also found no significant relationship between the NMDS and continuous vegetation metrics when fitting the environmental vectors to the ordination. However, I found a clear separation in the communities sampled between 2015 and 2016. This result was supported by the PERMANOVA analysis. The overall model explained a total of 40% of the variance in mammal communities among sites ($R^2 = 0.41$). Of this total variance, year explained 15% ($R^2 = 0.17$, $p > 0.05$) and the interaction between year and habitat explained a further 15% ($R^2 = 0.17$, $p = 0.07$) (Table 3.6).

Table 3.6. PERMANOVA model summary showing the explained variation for each parameter R^2 after 9999 iterations calculated using the Chao dissimilarity index. * indicates the significance of the parameter in the model at 0.05

	DF	**R^2**	**F value**	**P (of model)**
Overall	6	0.42	1.75	0.054
Year	1	0.18	4.31	<0.05*
Habitat	3	0.09	0.75	0.68
Interaction	2	0.16	1.98	0.08
Residual	14	0.57		

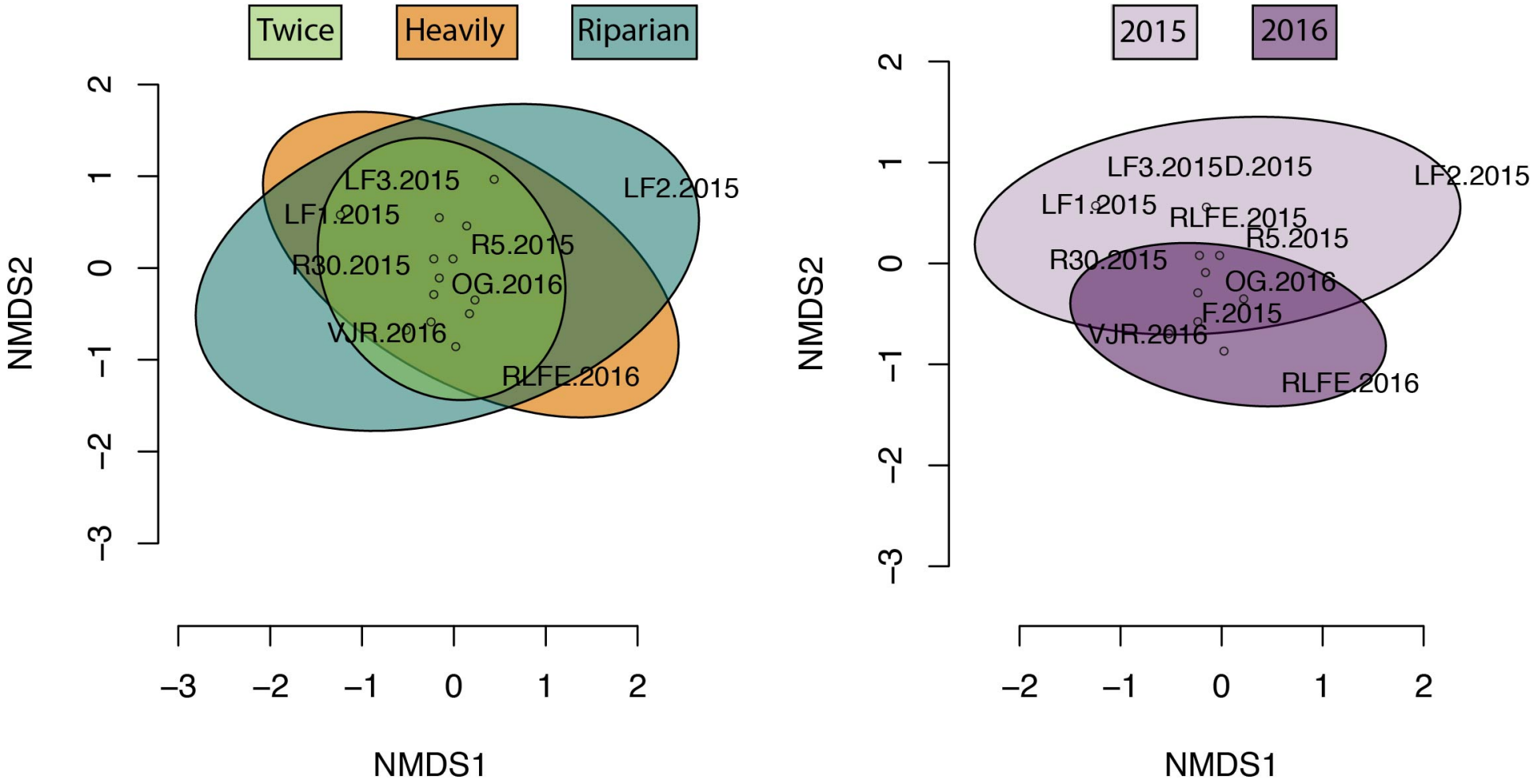


Figure 3.6. Non-metric multidimensional scaling's (NMDS) ordinations using the Chao dissimilarity index to show the difference in community structure between sampling sites. Ellipses represent the 95% confidence interval when sites are grouped by (A) forest type which indicates that the communities of all three habitat types are overlapping and (B) year where 2015 was the dry season and 2016 was the wet season. Stress value for the ordination = 0.13.

3.5. Discussion

In this study I aimed to test whether leech-based iDNA surveys could be used to assess mammalian diversity across a habitat degradation gradient in Borneo. Overall, I identified 17 mammalian genera from 181 pools of leeches (*Haemadipsa picta*) representing 1,724 individual leeches. The majority of the mammals detected were larger bodied members of the orders Cetartiodactyla and Carnivora, with the greatest number of genera detected from the Viverridae (civets) family (four genera). In total, this study revealed seven mammal taxa that were not previously recorded in the comparison of two leeches (see Chapter 2). These additional taxa comprised *Bos sp.*, *Prionailurus sp.*, *Manis javanica*, *Helarctos malayanus*, *Arctogalidia trivirgata*, *Paguma larvata* and *Elephas maximus*. Conversely, three taxa that were previously reported (Chapter 2) were not found in this dataset: *Echinosorex gymnura*, *Rattus sp.* and *Trachypithecus sp.*. In generating the sequences for this chapter, I used a modified protocol to re-extract DNA from samples collected from 2015 that were previously used in Chapter 2 and this discrepancy is likely to be reflecting the stochastic nature of PCR-based sampling.

3.5.1. Estimating mammalian richness with leeches

Using the bias-corrected Hill numbers to generate diversity accumulation curves showed near complete sampling when all samples were combined, from across the habitat gradient. In contrast, considering the forest types separately, the accumulation curve of diversity did not plateau, implying that leech-based sampling of individual habitats was incomplete. This was particularly evident in the riparian habitats that were the least intensively surveyed. In general, these trends were most pronounced where $q = 0$ (species richness). At the higher order Hill numbers $q = 1$ (corresponding to the exponential of Shannon's diversity index) and $q = 2$ (inverse Simpson concentration index), which accounts for species evenness and dominance respectively, there was less difference between forest types. These findings imply that variation in simple species diversity ($q = 0$) among forest types is driven by the presence of particularly rare or abundant species.

Overall the results from my generalised linear mixed effects models, which accounted for the nested nature of the sampling design, showed a greater diversity of mammals in 2016 than in 2015. In terms of spatial differences in mammal diversity across the degradation gradient, I found that taxon richness was lower in the heavily logged forest than in the twice-logged forest but that there was no difference in Shannon diversity index. Other studies of Bornean diversity have also demonstrated high levels of diversity in logged forest. For example, Wearn *et al.* (2017) found an increase in mammal abundance along a gradient from primary to logged forest, although there was a sharp decrease in oil palm habitats, which were not sampled in my study due to the absence of leeches. Although the finding that logged forests retain much of the diversity of primary forests has been reported from wide geographical regions and taxonomic groups (Putz *et al.* 2012), relatively few studies have examined the impacts of multiple rounds of logging. In one exception Edwards *et al.* (2011) (Edwards *et al.* 2011) showed that while the initial logging event can have negative effects for some taxa, subsequent rounds of logging have greater negative effects on biodiversity. Nonetheless, this study also showed that considerable biodiversity persisted in even the most degraded forest. The results from my own study support these trends, with the greatest richness detected in heavily degraded sites. Indeed, my analysis of community dissimilarity showed no clear differences in composition between any of the logged forest sites. Comparisons of multiple land-use classes show that logged forest can support a greater number of species than the relatively depauperate landscapes of plantations or pastures (Edwards *et al.* 2014). My results would support the notion that logged forests retain species and therefore hold conservation value, and thus also indicate that some mammal species are resilient to recent logging events.

Despite the differences in mammal richness between twice- and heavily logged forest, I found surprisingly no effect of vegetation structure on the relative diversity of detections. This is contrary to the findings of Wearn *et al.* (2017) who found a strong relationships between mammalian abundance, and above-ground biomass and forest cover. Yet, the results of the principal component analysis (Figure 3.4) showed that the coarse-scale forest type groupings showed separation for the continuous metrics, except the riparian versus the heavily logged sites. One

plausible explanation for the apparent lack of influence of vegetation structure on mammals is there could be a mis-match in spatial scale between the vegetation data and the response of the mammal to the environment. Indeed, the spatial scale of the covariates should be tailored to the individual focal species using averaged home-range sizes or simulations (Niedballa *et al.* 2015).

3.5.2. Seasonality and detection of diversity

A principal result of my analyses was a strong effect of year on mammalian diversity detected across the habitat gradient, which was more pronounced than for any other metric. There were also differences in community composition revealed by my analyses of Chao dissimilarity. This might be reflecting seasonal differences as the samples were collected once in the rainy season (2016) and once in the dry season (2015). Thus, the differences in diversity could be a consequence of changes in food availability or optimal conditions for forest mammals determining which species were available for the leeches to feed upon. For example, tree phenology will affect the distribution (and thus the detection) of frugivores (Fleming *et al.* 1987). Indeed, my results show an increased diversity of frugivorous species, such as civets, during the rainy season.

Yet due to the complex and heterogeneous nature of this human-modified tropical landscape there are additional factors which could be intensifying these differences, such as logging. Although logging at the wider SAFE field site was ongoing throughout the sampling periods, such that the heavily logged sites were being degraded over the life-time of the project (Ewers *et al.* 2011), logging intensity was lower during 2016. Thus, even though the sites in 2016 had suffered more timber extraction, there was a reduction of mechanical noise. Mammals have been shown to be sensitive to the intensity of logging (Burivalova *et al.* 2014), so the reduction of human activity could be linked to the higher diversity in 2016. Also noteworthy is that sites in 2016 showed considerable regrowth of grasses and pioneer tree species, and several studies have reported the re-growth of pioneer trees is linked to increased abundance of food sources for herbivores and frugivores, including browsing species, such as the sambar deer (Brodie *et al.* 2014a; Granados *et al.* 2016). Large ungulates, which make up the bulk of my

detections, are also primary targets for poaching. With the reduced human activity, associated poaching may have decreased. The pressures on mammals from poaching is thought to be a greater concern that the direct effects of logging itself (Brodie *et al.* 2014b).

An additional potential cause of the observed inter-annual variation that needs to be considered is the occurrence of the large El Niño/Southern Oscillation (ENSO) in 2015. The sampling in 2015 took place during the El Niño, which in Borneo typically manifests as prolonged dry seasons and increased fire risks (Chen *et al.* 2016). The impacts of dry weather associated with the El Niño on biodiversity are known to be especially severe in fragmented landscapes, such as SAFE (Pfeifer *et al.* 2017), due to edge effects (Fletcher *et al.* 2018).

It must also be considered that conducting slightly different extractions on the samples may have led to differences in the detected diversity. Samples which were re-extracted from a lysate (and not an original tissue) have a higher risk of being degraded through increased freeze-thaw cycles and stability of sample in the extraction buffers. These were samples collected in the dry season of 2015. Different extractions protocols have been shown to result in different levels of detections (Deiner *et al.* 2015), however, initial DNA quantification showed the concentrations to be comparable and the ability to use these samples allowed for an important comparison.

3.5.3. Comparisons with other studies

Previous studies using terrestrial leeches have reported higher rates of detection than those reported in this current study. For example, Schnell *et al.* (2012) detected mammal DNA in 21 out of 25 individual leeches, although this was a small-scale study in which, unlike my study, the leeches were not pooled. It is therefore possible that that the process of pooling, while allowing high throughput screening of large sample sizes, risks the obscuring of rare DNA as it is outcompeted by more common sequences during PCR (Pompanon *et al.* 2012). Aside from methodological differences, variation among the detection rates could also reflect regional or habitat differences in the communities being studied. In a

recent study of leeches samples from across the Palaeotropics, Schnell *et al.* (2018) applied a pooling method and detected multiple vertebrate classes including birds and reptiles (using the same primers), however, only mammals were identified from the Bornean samples that were included in this broad-scale study. Other studies have also reported detection from birds in leeches collect from different regions of Southeast Asia (Tessler *et al.* 2018).

Comparing the results from my study to those from the Bornean leeches screened by Schnell *et al.* (2018), I found the same mammalian families, but with additional detections for Manidae, and Elephantidae. One of the claimed benefits of using leech iDNA is the speed at which samples can be collected. Indeed, Weiskopf *et al.* (2017) analysed leeches collected over a four-day period in Bangladesh and found 12 mammal species, which represented half of the species identified from the same site from over 1300 camera trap nights. Weiskopf *et al.* (2017) also compared costs of sampling using leeches and camera traps, and found that leech-based sampling with pooling and high throughput sequencing was by far the most cost effective due to drasatic reductions in the costs of field work.

Despite this, camera trapping remains the most successful and comprehensive in terms of sampling completeness for terrestrial mammals. Comparing my results to those of a camera trapping study performed at the same site I found that of the 25 mammal genera detected at least once by Deere *et al.* (2017), 17 (68%) were also recorded here. However, I also detected *Bos* sp. and the Bornean gibbon (*Hylobates muelleri*), neither of which were reported by Deere *et al.* (2017), demonstrating the potential of iDNA as a complementary technique to camera trapping In particular, the addition of the arboreal gibbon is intriguing. While *H. picta* is known to be able to climb upwards in the understory, little else is known as to whether it forages in the canopy. Thus, understanding more about the behaviour and space-use of *H. picta* could help us to tailor iDNA studies towards particular mammals. Uncertainty around the independent dispersal ability of terrestrial leeches when not attached to a host remains a potentially significant cause of error in data interpretation. Although I (as with all iDNA studies of leeches to date) have

assumed that a leech has not travelled unless being transported by a mammal (or other vertebrate host) more work is needed to test this assumption.

3.6. Conclusion

My results show that the sequencing of leech iDNA can be used to determine differences in relative diversity across different forest types, as well as over time. I have shown that while leeches cannot provide an exhaustive catalogue of the mammals present at a given site, leech iDNA is nevertheless still capable of assaying a representative mammalian community, with detection richness close to that of camera traps for Sabah. By comparing the relative diversity across the habitat gradient of degradation, I was able to pinpoint areas with greater richness and diversity. Working towards monitoring with iDNA, technical aspects of this method would benefit from further development for individual identifications for example and a deeper understanding leech foraging behaviour. My findings showcase the potential for using iDNA based sampling methods for biodiversity surveys in degraded and pristine tropical forests.

3.7 Supplementary information

Supplementary tables

Table S3.1. Candidate model set with genus richness as the response variable. Model comparison is based on ΔAIC

Model	Response	Fixed effects	DF	AIC	ΔAIC	weight	Res. Dev.
M1.SR	Richness	Year + Habitat + PC1 + Leech + Year:PC1 + Year:Habitat	10	401	3.8	0.06	151.0
M2.SR	Richness	Year + Habitat + PC1 + Leech + Year:Habitat	9	399	1.8	0.16	148.7
M3.SR	Richness	Year + Habitat + PC1 + Year:Habitat	8	397	0.0	0.40	148.4
M4.SR	Richness	Year + Habitat + Year:Habitat	7	397	0.3	0.34	148.7
M5.SR	Richness	Year + Habitat	5	402	4.9	0.04	159.6
Null.SR	Richness	Only random effect	2	421	24.4	<0.001	180.9

Table S3.2. Candidate model set with Shannon-diversity index as the response variable. Model comparison is based on ΔAIC

Model	Response	Fixed effects	DF	AIC	ΔAIC	weight	Res. Dev.
M1.sh	Shannon	Year + Habitat + PC1 + Leech + Year:PC1 + Year:Habitat	11	185	32.0	<0.001	154.3
M2.sh	Shannon	Year + Habitat + PC1 + Leech + Year:Habitat	10	176	23.2	<0.001	135.8
M4.sh	Shannon	Year + Habitat + Leech + Year:Habitat	9	169	16.2	<0.001	128.9
M5.sh	Shannon	Year + Habitat + Year:Habitat	8	164	10.3	0.005	122.1
M6.sh	Shannon	Year + Habitat	6	162	8.4	0.01	122.1
M7.sh	Shannon	Year	4	153	0.0	0.98	121.9
Null.sh	Shannon	Only random effect	3	165	11.3	0.003	126.1.

Supplementary figures

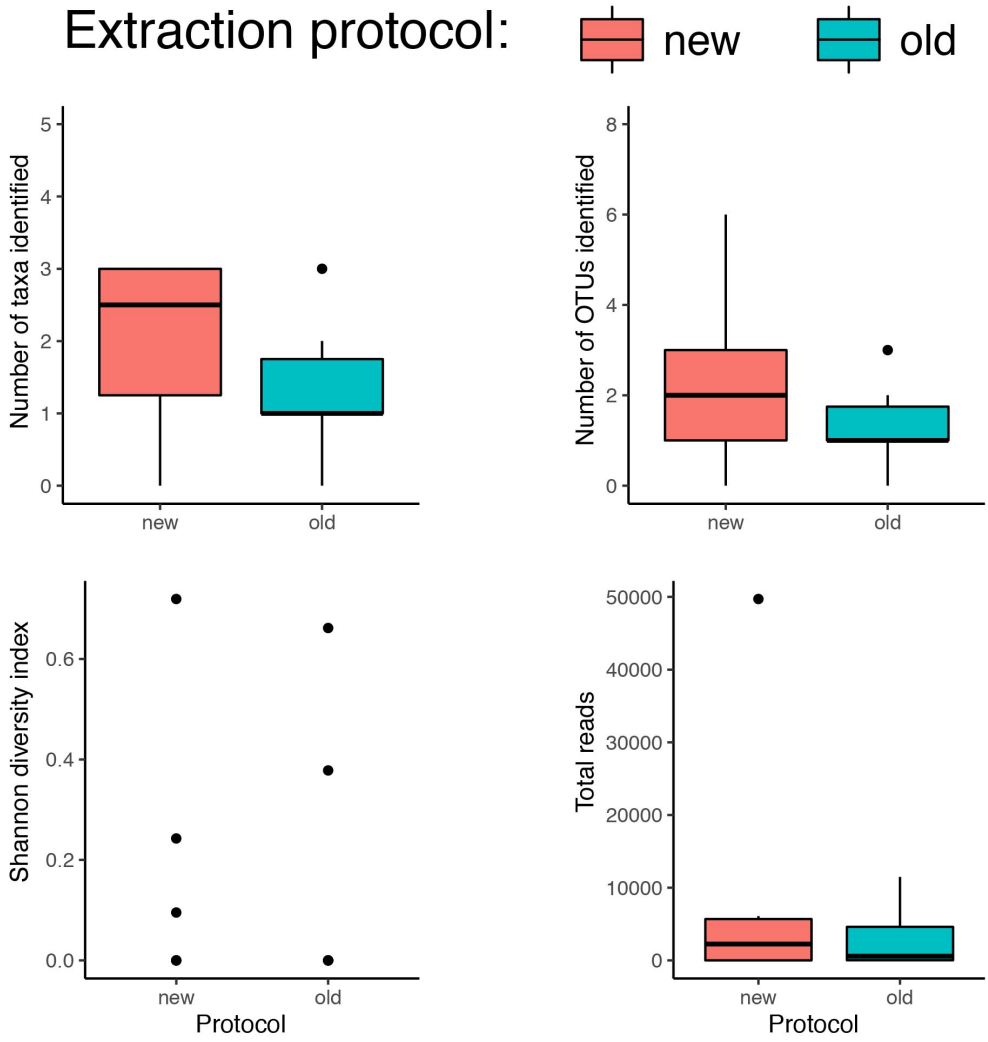


Figure S3.1. The results of a preliminary experiment comparing two DNA extraction protocols used in Chapter two and Chapter three - *new* = is the protocol used in Chapter 3 with the additional lysis buffer (AL) and *old* = used in Chapter 2. Four metrics are shown calculated for all the samples (1) number of taxa assigned to OTUs, (2) number of OTUs initially identified, (3) Shannon diversity index and (4) the total number of reads generated.

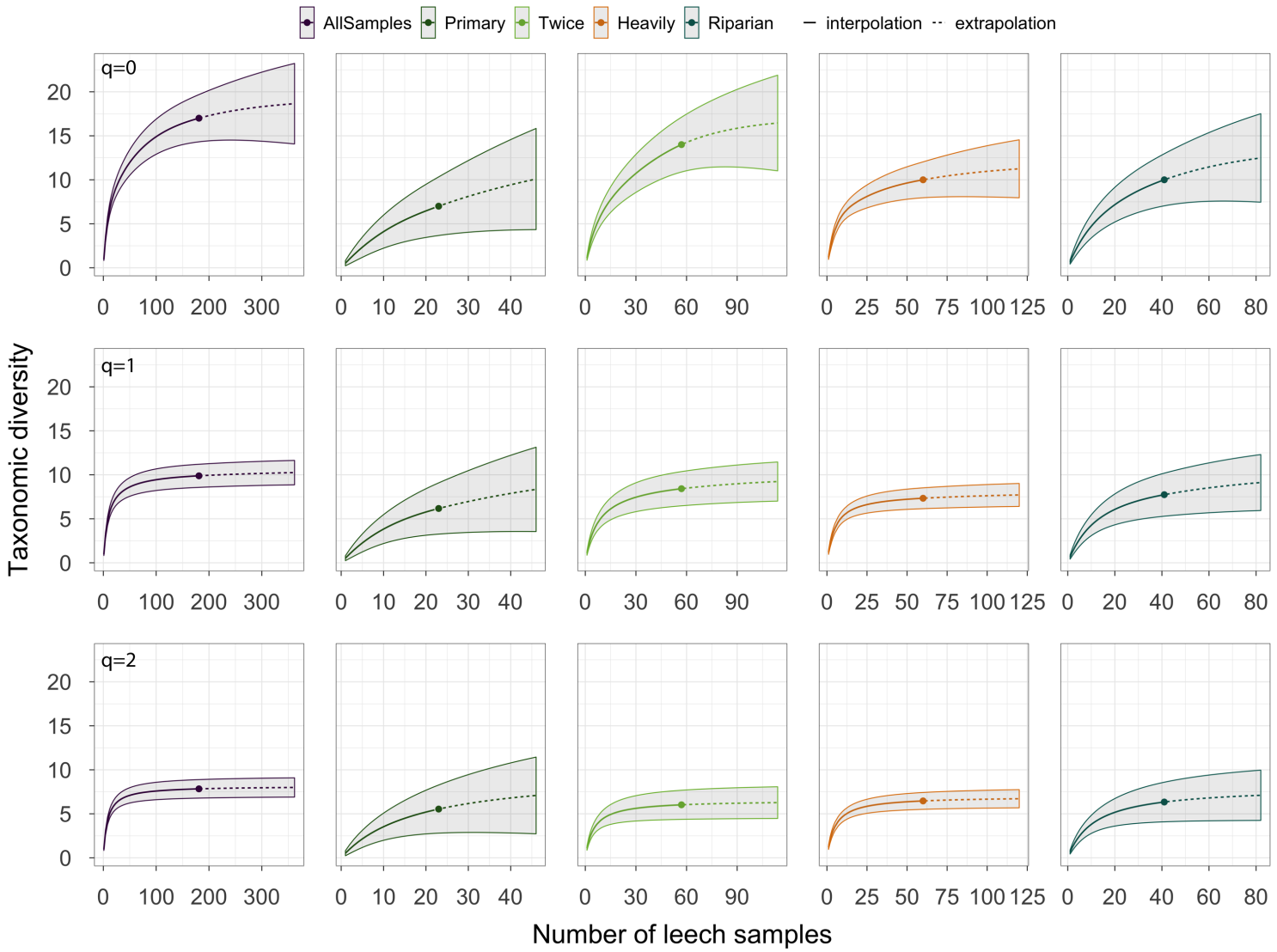


Figure S3.2. Diversity accumulation curves (at the genus level) comparing the effect of increasing leech samples on the taxonomic diversity. The curves are calculated using three orders of hill numbers, q = 0, 1 & 2 which are equivalent to species richness, Shannon diversity index and Simpson index, respectively. The diversity of all samples combined (A, F & K) is compared to the accumulation of diversity in the four different habitat types: primary (B, G & L), twice-logged (C, H & M), heavily-logged (D, I & N) and riparian (E, J & O). The x-axis varies depending on the number of samples. The solid line represents the rarefied values, and the dashed line represents the extrapolated values and is extended to double the reference sample (empirical value, solid circle). The accumulation curves are presented with 95% confidence intervals

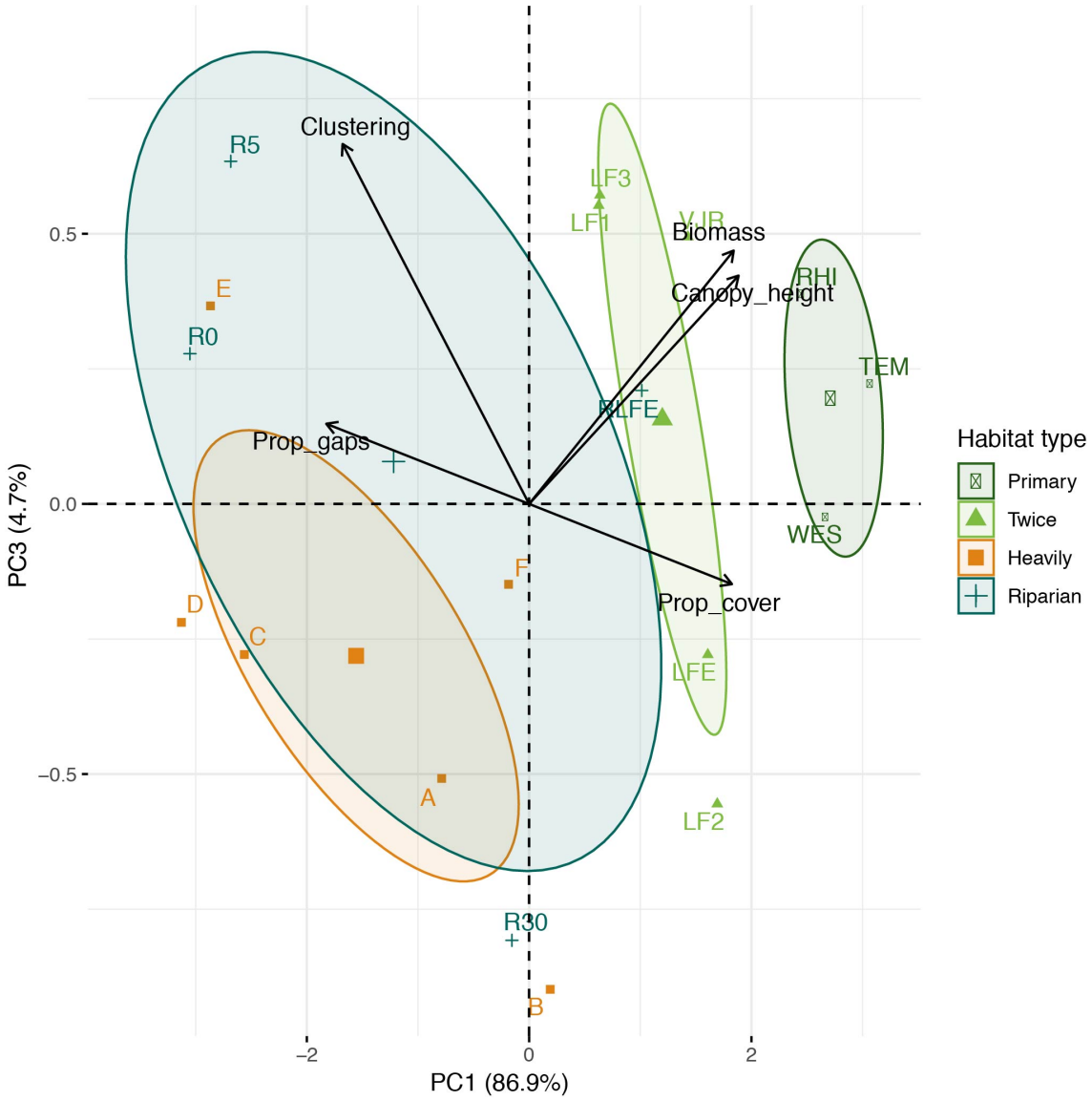


Figure S3.3. Principal component analysis (PCA) showing the relationship between components of vegetation structure (as calculated from LiDAR data) and the sampling sites where surveys took place. The first (PC1) and third (PC3) axis are shown with the respective variation they explain. Points are labelled with the site ID and the mean centroid point, while habitat types are denoted by different symbols and colours. Ellipses show the 95% confidence intervals around the four habitat types. The vegetation metrics shown are habitat heterogeneity (Clustering), forest cover (Prop_cover), gap fraction (Prop_gaps), aboveground biomass (Biomass) and canopy height (Canopy_height), and the direction of arrows shows the relationship between the metrics and sites.

Chapter 4: Modelling imperfect detections with multiscale occupancy models

4.1. Abstract

Molecular surveys are rising in popularity as a way of detecting species that are shy, elusive or otherwise difficult to track. Yet these approaches remain prone to the inherent issues of imperfect detection. In tropical landscapes, rapid forest loss means that there is a pressing need for reliable species monitoring. Occupancy modelling is a powerful statistical tool for accounting for imperfect detections that has been widely adopted for analysing biodiversity survey data, but which has rarely been applied to molecular data. Despite this, occupancy models offer the potential to improve invertebrate-derived DNA (iDNA) approaches so these methods move beyond cataloguing the presence of species in an area. Specifically, occupancy models can help address problems relating to invertebrates feeding behaviour as well as stochasticity in the results of PCR and DNA sequencing. Here I apply occupancy modelling to ten mammalian taxa for which detection data were generated from the blood meals of leeches collected in primary and logged forest. Using multiscale occupancy models, I estimated probabilities of occupancy and availability, as well as detectability, and compared these values across forest types and samples. My findings show that, overall, iDNA occupancy increases with measures of habitat quality. I also found that the availability parameter (the probability that mammal DNA was available for detection) was affected to a small extent by sampling effort. In terms of technical determinants, on average, iDNA detection probability increased with the concentration of DNA in the extract. Yet despite these trends, there were species-specific exceptions. When estimating minimum numbers of samples and PCR replicates needed, I found that these were strongly influenced by year and species, with up to ten leech pool samples and 10 to 20 PCR replicates needed for >80% probabilities of detection. This study demonstrates the usefulness of combining occupancy models with iDNA approaches. At the same time, I show that there are still knowledge gaps and limitations that need to be addressed in using invertebrate samplers.

4.2. Introduction

4.2.1. General occupancy modelling

In recent years several studies have demonstrated that vertebrate DNA can be obtained from blood-feeding invertebrates (Weiskopf *et al.* 2017; Schnell *et al.* 2018; Tessler *et al.* 2018), and this has led to growing interest in the utility and potential of invertebrate samplers for biodiversity assessments (Calvignac-Spencer *et al.* 2013b; Schnell *et al.* 2015a; Kocher *et al.* 2017c). To date, most invertebrate-derived DNA (iDNA) studies have focused almost exclusively on proving the presence of vertebrate species, with little or no consideration of the extent to which such methods detect, or fail to detect, species that are present. However, like all survey methods, iDNA surveys are still prone to the inherent issues of imperfect detection, and there is a need to understand how these methods can be improved to account for sampling biases.

Occupancy modelling is a powerful statistical tool for accounting for imperfect detections that in recent years has been widely adopted for analysing biodiversity survey data. In particular, the site-occupancy, or proportion of sites occupied by a species, is commonly used variable for biodiversity monitoring (Guillera-Arroita *et al.* 2010). The application of occupancy modelling to biodiversity surveys attempts to address the issues that no technique will record all individuals without error, and that some species are more likely than others to be missed by a given survey (MacKenzie *et al.* 2002). Imperfect detections can result from false positives, such as through the misidentification of a species, or from false negatives, such as through failing to record a species when it is actually present. Missing the detection of a species during a survey can have a large impact on the estimates of occupancy, resulting in underestimates of sites occupied (Mackenzie & Royle 2005). To counter this problem, replicate surveys needs to be conducted and the detection probabilities can be estimated (Guillera-Arroita *et al.* 2010). Additionally, environmental covariates are routinely included to evaluate how a species' occupancy and detectability varies with particular habitat characteristics (MacKenzie *et al.* 2006). There is a large amount of theoretical and applied

literature demonstrating the flexibility and versatility of these models (MacKenzie *et al.* 2002; Royle & Link 2006; Nichols *et al.* 2008).

4.2.2. Occupancy models for molecular survey data

Despite their potential utility and relevance, occupancy models have rarely been applied to the findings of molecular surveys (e.g. Schmidt *et al.* 2013; Hunter *et al.* 2015). Yet molecular data from invertebrate samplers (i.e. iDNA) typically consist of detections and non-detections of prey vertebrate species occurrence across sites, from which it is possible to infer species occupancy. In particular, occupancy modelling represents a statistical framework that enables the simultaneous estimation of both occupancy and detection probabilities from biodiversity survey data. These methods have previously been used to analyse survey data obtained by a range of methods, including camera traps (Brodie *et al.* 2014b; Rich *et al.* 2016) and point counts (Royle & Nichols 2003). Recently, occupancy models and multiscale (or hierarchical) extensions have been applied to species occurrence data detected using environmental DNA (eDNA), that is, DNA extracted from environmental samples (e.g. water) (Schmidt *et al.* 2013; Hunter *et al.* 2015; Dorazio & Erickson 2017b).

Biodiversity surveys based on sampling blood-feeding leeches for iDNA have been gaining interest and, while still relatively new, appear to offer several potential benefits including speed and reduced field costs and logistics (Weiskopf *et al.* 2017). Although there are currently no published studies of leech-based iDNA that have integrated occupancy modelling, the potential for doing so has been discussed previously (Schnell *et al.* 2015a). In this system, there are at least two general pathways by which imperfect detections could be introduced into occupancy probability. First, it is possible that the leeches themselves demonstrate feeding preferences. The invertebrate sampler of choice for my studies has been the haemadipsid leeches (*Haemadipsa picta* and *H. sumatrana*). These species have the potential to introduce imperfect detections if they actively avoid feeding on one or more particular species, even when that species is present in the environment. This would result in an erroneous non-detection causing occupancy to be underestimated. Unfortunately, the general behaviour and ecology of these

leech species, and indeed of terrestrial leeches in general, remains understudied. While it is difficult to account for any feeding biases, my two detailed studies described in Chapters 2 and 3 demonstrate broad diets and my finding of generalist feeding behaviour agrees with other studies of the Haemadipsidae across a greater extent of their range (Schnell *et al.* 2018; Tessler *et al.* 2018). At the same time, however, I also highlight missing groups (e.g. small and arboreal mammals), and I find some evidence of differences between the both leech species examined (see Chapter 2).

The second main cause of imperfect detections may arise via technical bias, whereby imperfect detections are introduced during PCR and sequencing of the leech samples. Following leech collection, there are many technical aspects of PCR-based metabarcoding which can cause a species DNA to be missed. These aspects include, but are not limited to, primer bias (Elbrecht & Leese 2015), inhibition (Goldberg *et al.* 2016) and bioinformatic parameters (Alberdi *et al.* 2017). Successful PCR amplification also becomes more variable as the rarity of DNA in a sample increases (Barnes & Turner 2015). To account for many of these problems, metabarcoding studies routinely use replicate PCR reactions to increase the confidence in a detection. Understanding the detectability of particular species from both leech feeding and PCR amplification are important for implementing an adequate number of replicates.

4.2.3. Hierarchical occupancy models and DNA

A classic two-level model contains the detection probability (p) and the occupancy (ψ) (Mackenzie *et al.* 2003). In contrast, in a multiscale model, detection probability is partitioned into two parameters, termed the availability parameter (θ) and detection probability (p) (Kéry & Royle 2016). Thus, in its basic form, the three-level multi-scale occupancy model estimates three parameters, as follows:

1. **Occupancy probability (ψ)** - probability of site occupancy by target species
2. **Availability probability (θ)** - probability of site used by the target species during the survey
3. **Detection probability (p)** - probability of detecting the target species during the survey

Multi-scale occupancy models have been used to estimate the occupancy of mobile species (Mordecai *et al.* 2011). Here, the availability parameter (θ) aims to estimate how a species could remain undetected if it moves from an occupied patch and is therefore "unavailable" for detection during a sampling event.

In well-designed iDNA surveys, taking multiple samples from within a site during a survey is analogous to spatial replication in a traditional survey, whereas PCRs from the same DNA extract can be considered equivalent to temporal replicates. This is because subsample of DNA used for the PCR reaction is a repeated measure of the same sample, like returning to the same site on a different survey (Dorazio & Erickson 2017b). Thus these nested levels of replication can be incorporated into a multi-scale ('hierarchical') occupancy model (analogous to the approaches of Nichols *et al.* 2008 & Mordecai *et al.* 2011). Following the framework described in Hunter *et al.* (2015), the three nested levels of a multi-scale model for iDNA would then correspond to:

1. **iDNA Occupancy probability** = the probability of a species' DNA occurring at a site (ψ)
2. **iDNA Availability probability** = the conditional probability that the target DNA occurs at a site, given that the species is presence at the site (θ)
3. **iDNA Detection probability** = the conditional probability that DNA is detected from a subsample (PCR replicate), given that the DNA is present in a sample from a given site (p)

The first example where these models were applied to eDNA sampling found that previous studies had underestimated the occurrence of the amphibian pathogen (*Batrachochytrium dendrobatidis*, Bd) from pond samples (Schmidt *et al.* 2013). Multi-scale occupancy models have also been used to track the invasive but elusive Burmese python in the Florida Everglades using eDNA (Hunter *et al.* 2015). Using occupancy models to understand imperfect detection in iDNA studies will be particularly beneficial in developing monitoring schemes using invertebrates (Schnell *et al.* 2015a). This is especially pertinent in study regions, such as Borneo,

where forest degradation and land-conversion is wide-spread (Sodhi *et al.* 2004) and the need to accurately monitor species without bias is crucial for directing policy and conservation actions.

In this chapter I use Bayesian multiscale occupancy models following the broad approach of (Hunter *et al.* 2015). I apply multiscale occupancy models to detection histories for ten mammals generated from sequence data, from two years of iDNA sampling. To understand species responses to forest degradation, I model the three parameters (ψ, θ & p) as functions of covariates. I predict that estimates of occupancy will change as species-habitat associations will be affected by vegetation structure and habitat quality. Additionally, due to continued logging across the landscape, occupancy probabilities could decrease between 2015 and 2016 as the habitats become more degraded and modified. Previous studies have shown that while logged forests can support mammal diversity, there is threshold by which the landscape becomes too degraded (Burivalova *et al.* 2014; Wearn *et al.* 2017). Finally, I estimate the minimum amount of sampling and PCR replication needed to confidently detect these species, which is an important financial and logistical constraint on metabarcoding studies.

4.3. Materials and methods

4.3.1. Data generation

All sequence data used for this study were generated and described previously in Chapter 3. To make the data suitable for analysis using multiscale hierarchical occupancy models, I modified the steps of the bioinformatics pipeline. Briefly, for this chapter I used sequences from pools of tiger leeches *Haemadipsa picta* collected at the SAFE project, Sabah, in the dry season (February – June 2015) and a wet season (September – December 2016). All field protocols were the same as the previous two chapters; where individual leeches were collected from within the boundaries of 25m^2 vegetation plots for 20 minutes and stored in RNA later. As detailed previously, at a larger scale vegetation plots are grouped into sites, consisting of 8-16 plots. I used leech pools from all four forest types sampled; primary, twice-logged, heavily logged and riparian forest.

From the pooled leech samples, I amplified and sequenced a fragment of the 16s rRNA gene, which had been extracted using the modified protocol detailed in Chapter 3. Pools of leeches ranged from 4-13 individuals, with an average of 9.53 and median of 10. Leeches within pools were always from the same site and identifiable by the addition of unique tags to the 5' end of the primers during PCR. Amplicon libraries were then combined into an equimolar sequencing pool, for 150bp paired-end sequencing on an Illumina MiSeq at the Bart's and the London Genome Centre, Queen Mary University London.

4.3.2. Bioinformatics

Following the initial steps of the bioinformatics pipeline in Chapter 2, raw demultiplex reads were merged in AdapterRemoval v2 (Schubert *et al.* 2016) using the same parameters. I used the *sort.py* script in the modified version of DAMe (https://github.com/shyamsg/DAMe, Zepeda-Mendoza *et al.* 2016) to sort the merged reads into the original samples.

Sequence reads were filtered based on length and quality (see Chapter 2). Unique sequences were then retained on condition that they were present in at least one

of the three replicates in each sample (as opposed to two of three replicates, as applied in previous chapters) using the *filter.py* script in DAMe. This retained information on the presence of each sequence per PCR replicate was used in generating the detection histories needed to calculate occupancy (see below). I then clustered the filtered reads at 97% similarity with sumaclust v1.3 (Mercier *et al.* 2013) and removed chimeras with mothur (Schloss *et al.* 2009). I then applied the LULU post-clustering algorithm (Frøslev *et al.* 2017). The remaining OTUs were then assigned to taxa following the steps in Chapter 3 (e.g. using BLAST and MEGAN). The ten taxa with the most abundant detections were selected (Table 4.1), of which six which could be identified to species and four to genus.

Table 4.1. Four-letter codes used to identify the mammal taxa in the occupancy models

Code	Scientific name	Common name
HEDE	*Hemigalus derbyanus*	Banded civet
HYMU	*Hylobates muelleri*	Bornean gibbon
HYSP	*Hystrix sp*	Porcupine
MASP	*Macaca sp*	Macaque
MUSP	*Muntiacus sp*	Muntjac
RUUN	*Rusa unicolor*	Sambar deer
SUBA	*Sus barbatus*	Bearded pig
TRFA	*Trichys fasciculata*	Long-tailed porcupine
TRSP	*Tragulus sp*	Mousedeer
VITA	*Viverra tangalunga*	Malay civet

From the files generated by DAMe, I extracted the read frequencies corresponding to the OTUs assigned to the target taxa. For each taxon, if its corresponding OTU sequence was found in a PCR replicate, this replicate was coded as 1 for a detection, whereas if no sequence reads were found the replicate was coded as 0 for a non-detection. By doing this for all three PCR replicates, for each OTU, this generated the detection history for each species. Additionally, I applied stringent filtering thresholds, first by removing any reads with only singletons or doubletons, then filtering out reads based on any contamination in the negative

controls. Finally, I removed the lowest 1% of reads for each OTU (if not already removed during the previous steps). I applied this strict filtering to reduce any potential erroneous detections introduced by accepting all three PCRs, as opposed to the standard minimum threshold of two out of three that was applied in my previous studies (see Chapter 3).

4.3.3. Study design and occupancy model assumptions

For multiscale-occupancy modelling, I used three nested levels of sampling and constructed the models separately for each season

1. The 13 sites across the human-modified forest landscape, encompassing primary, twice-logged, heavily logged and riparian forest (as described in Chapter 3).
2. The spatially replicated leech pools collected within each of these sites
3. The technical replication through PCR replicates, where DNA is subsampled from the pooled leech DNA extracts

I generated the detection histories for each species, for each PCR replicate, within each leech pool, from each site. For example, for a taxon at a given site, if there were three pools of leeches, each of which was amplified in triplicate, a hypothetical detection history could be expressed as {101 111 100} based on conventional notation in which PCR replicates are nested within pools. Thus, in this example, there is a detection in the first and third PCR replicates of the first pool, detections in all three replicates of the second pool, and detection in only the first replicate of the third pool. There are several potential reasons for the occurrence of a non-detection: (i) the target species is not present at the site (true absence), (ii) the target is present, but the leeches have not fed on it, or (iii) the leeches have fed on it, but it was not amplified during PCR. Both of these latter scenarios are examples of imperfect detections.

The classical occupancy framework makes several assumptions: (1) independent detections, (2) detection and occupancy are constant in space (or accounted for in the covariates), (3) the sites do not change occupancy state during the time of the survey (i.e. within a closed season) and (4) there are no false positives (Bailey *et al.*

2007). For this study I am considering each leech pool as an independent detection event; that is, the detection of a mammal in one pool does not have any influence on the probability that it is detected in another pool. Covariates are added to the model to account for any variation in habitat preferences or sampling that would change the occupancy and detectability of the species. Also, I am considering the detections to be within a closed season, because the field sampling was conducted over 3-4 months, and it is likely that only the last blood meal is being detected for each leech. Finally, I am assuming that there are no false positives as a result of the strict filtering and only accepting confident taxonomic assignments (> 90% similarity). However, there are some caveats to these assumptions (see Discussion).

4.3.4. Occupancy models

I used the single species multi-scale occupancy models described in Dorazio & Erickson (2017c) and Hunter *et al.* (2015), generating separate models for the dry and wet season data. This model comprises three levels of Bernoulli trials to describe the processes leading to a DNA detection. My data for each taxon is binary, taking the form of detection or non-detection, at the *i*th site (*i* = 1-13, field sites at SAFE), for the *j*th leech pool (*j* = 1 up to 23 samples taken from a site), in the *k*th PCR replicate (*k* = 1-3, triplicate PCR reactions). The levels of the hierarchical model are therefore:

1) *Occupancy - the probability the DNA from the target taxon is occupying site i where terrestrial leeches are present.* The variable Z describes the presence (*Z* = 1) or absence (*Z* = 0) of the target DNA at site *i*. This is modelled as a function of the occupancy parameter at the site ψ_i

$$Z_i \sim \text{Bernoulli}\ (\psi_i)$$

2) *Availability - the probability that the leeches collected for this study had fed on the blood of the target taxon given that it was present was present at the site.* This is the probability that the DNA was available to be sequenced. Here the availability parameter θ_{ij} is conditional on the value of z_i. A is the probability the target DNA occurs in the sample *j*, given the species was present (*z* = 1) at

site *i*. If the site is unoccupied i.e. $z_i = 0$, A_{ij} has to equal zero meaning the target mammal had not been present at the site, and the DNA is not available for detection. This restricts the inclusion of false positives.

$$A_{ij}|z_i \sim \text{Bernoulli}\ (z_i\theta_{ij})$$

3) *Detection - The probability that the DNA of the taxon is detected during PCR given that the DNA was present in the leech pool.* The final equation describes the probability that DNA will be detected in a PCR replicate Y_{ijk} (detection = 1, non-detection = 0) conditional on the occurrence of the DNA within the leech sample a_{ij} (the realised value of A_{ij}). p_{ijk} is the detection probability, which is conditional on DNA being amplified in the *k*th PCR replicate of the *j*th leech pool, collected at the *i*th site. Again, this equation does not allow for false positives as Y_{ijk} will equal to zero if there is no DNA in the *j*th leech sample from the *i*th site.

$$Y_{ijk}|a_{ij} \sim \text{Bernoulli}\ (a_{ij}p_{ijk})$$

For each level of the model, specific covariates can be added depending on how they are assumed *a priori* to affect the target mammal's occupancy at the site (ψ), the availability of the DNA in the sample (θ) or the detection of DNA in the PCR (p) as the parameters β, α and δ, respectively.

I fitted eight models for each species including a different site, sample or replicate covariate using the logit link function, for Bernoulli distributions. Models were run using the eDNAoccupancy package (Dorazio & Erickson 2017b) in R (R Core Team 2018), which uses Bayesian inference to estimate the parameters of ψ, θ and p, and assumes uniform prior distributions (for more details on the MCMC algorithm see the supporting information for the R package, (Dorazio & Erickson 2017a). I set the MCMC chain up to run for a total of 80,000 iterations but updating the chain four times with 20,000 iterations, using the *updateOccModel* function. I used trace plots to assess model convergence and I compared goodness-of-fit for each of the models using two commonly applied criteria for Bayesian models: Watanabe's AIC (WAIC, Watanabe 2010), and posterior-predictive loss (PPLC, Gelfand & Ghosh 1998). For nested models, a lower value of these models indicates a better fit.

4.3.5. Covariate selection and model construction

Initially, I selected four site-level covariates (β) which represented the heterogeneity across the SAFE project and which would reflect differences in habitat use by mammals. These covariates were canopy height, aboveground biomass, habitat heterogeneity and forest cover, and were calculated from a Leica ALS50-II LiDAR sensor flown by NERC's Airborne Research Facility which covered the SAFE landscape in 2014 (unpublished data, T. Swinfield and D. Coombes). As with the previous chapter, these covariates were extracted at the block level (with a 1 km buffer around the centroid) which was deemed more appropriate to the home-ranges of the mammal species than the 25 m^2 vegetation plot. However, when I tested these metrics for collinearity, I found that all four covariates were highly correlated (Table S4.1). Therefore, I decided to proceed using the habitat heterogeneity covariate in order to capture the variability across the human-modified landscape. Habitat heterogeneity (measured using Moran's I) is a value which ranges from -1 to 1 and represents a gradient of forest canopy clustering. Values approaching 1 indicate greater clustering of the canopy and thus represent strong contrasts in habitat availability such as gaps or very large trees within a matrix of intermediate canopy heights, while a value of 0 represents perfectly random canopy dispersion.

As a sample-level covariate (α), I used the number of pools per site, as an indicator of sampling effort. This covariate is likely to affect the probability of DNA being detected in a sample of leeches (θ). Finally, I also used two replicate level covariates (δ) which will affect the detection parameter, p. First, I used the concentration of the DNA extract for each pool, as measured by the Nanodrop (Thermo-Scientific) for the 2015 samples and Qubit (Invitrogen) for the 2016 samples. DNA concentration can affect the amplification success of the PCR reaction and thus the detectability by either causing inhibition, if there is too much DNA in the sample, or by increasing the stochasticity of amplifying the target species DNA from a mixed sample if there is too little DNA. I also used the summed relative weight of the leech individuals in the pool. I could only weigh the individuals after they had been stored in RNA later and as such this is not the true

weight. I included this covariate to test whether there was an effect of the large individual interspecific size difference (ranging from 0.01 g to 2 g).

For each of the ten mammals (Table 4.1) and for each of the two years of sampling (2015 and 2016), I constructed five models. These comprised four univariate models, each including one covariate, and the null model that contained no covariates. I was not able to investigate the effect of interactions between covariates because the dataset was not large enough. Indeed, model testing showed that MCMC chains did not converge for more complex models, and thus these results would be prone to error in their parameter estimates. Details of the model structure can be found in Table 4.2.

Table 4.2. The five multiscale occupancy models fitted to the detection histories each species repeated for 2015 and 2016. Site- (ψ), sample (θ) - and replicate-level (p) covariates are in brackets and (.) indicates no covariate

Model	Covariate structure	Description
Model 1	ψ (.) θ (.) p (.)	No covariates
Model 2	ψ (heterogeneity) θ (.) p (.)	Habitat heterogeneity
Model 3	ψ (.) θ (pools) p (.)	Survey effort
Model 4	ψ (.) θ (.) p (conc)	iDNA concentration
Model 5	ψ (.) θ (.) p (weight)	Leech pool total weight

4.3.6. Cumulative probabilities

Finally, to test the effects of sampling and PCR replication on the probability that the DNA is in the sample given the target was at the site (availability, θ) and the probability of detection of the DNA given it was in the sample (detectability, p), I calculated the cumulative probability scores for these parameters. This is an important consideration for molecular sampling as increasing the number of samples and PCR replicates to be sequenced will increases costs. Any increase in cost needs to be traded off against the increasing probability of detection with greater sample size, and the benefits of increased confidence in the detections by confirming a detection in multiple PCR reactions. This is of particular concern when designing molecular surveys for conservation monitoring schemes where

confidence in the presence of a rare target species is crucial. As in Hunter *et al.* (2015), I calculated the cumulative availability probability as:

$$1 - (1 - \theta)^{k}$$

And the cumulative detection probability as:

$$1 - (1 - p)^{k}$$

Where θ is the availability parameter, p is the detection probability and k described the number of leech samples or PCR replicates. I calculated this parameter for k = 1 up to k = 20. In both cases, θ and p, the starting value for k = 1 was derived from the mean posterior probability taken from model 1 (which included no covariates).

4.4. Results

4.4.1. Detection histories

To generate the detection history for each taxon, in each of the three PCR replicates, I set the lowest minimum PCR threshold to one, which retains all sequences passing quality filtering. Initially this resulted in 485 OTUs, of which 45 (9%) were identified as chimeric sequences and so were removed. Of the remaining 440 OTUs, only 40 (9%) were identified as true sequences by LULU (post-clustering filtering) and retained for further analysis. The ten taxa that were selected were those species with the highest number of detections (Table 4.1). However, after the strict filtering of the read frequencies, the number of times a species was detected in a minimum of one of the PCR replicates (per pool) ranged from 3 to 66: *Sus barbatus* = 66, *Muntiacus sp* = 33, *Hylobates muelleri* = 25, *Rusa unicolor* = 22, *Hystrix sp* = 14, *Tragulus sp* = 12, *Hemigalus derbyanus* = 9, *Viverra tangalunga* = 9, *Macaca sp* = 8, and *Trichys fasciculata* = 3. Although initially seeming to have enough detections, after read frequency filtering the long-tailed porcupine, *Trichys fasciculata*, was removed from further analysis, as the number of detections was too low to allow accurate parameter estimations (MacKenzie *et al.* 2002).

4.4.2. Model selection

The MCMC traces for each model showed chain convergence for all but two models, using 80,000 iterations. The two non-converging models corresponded to the mousedeer models for 2016 which included the replicate level covariates (models seven and eight for mousedeer) and these models were removed from further analysis.

I compared the five models constructed for each taxon (within each year) and found that these showed little difference in fit, based on two separate goodness-of-fit criteria (WAIC and PPL) (Supplementary Table S4.2). Two exceptions were the models for Bornean gibbon and macaque, where the replicate-level models (DNA concentration and leech weight covariates) in 2016 showed lower criterion values. Thus, indicating that these covariates are having a greater effect on the detection

of these species compared to those at the levels of site and sample. For the remaining species, there appeared to be no difference in the relative impact that each of the covariates had on the respective parameters. Using the median values estimated from the posterior distribution, I was able to identify taxa-specific trends in their responses to the five covariates in both years and plot the mean trend line. Due to the low number of overall detections for each species, the resulting Bayesian posterior credible intervals (CI) were large, and thus for clarity and to show general trends, the CI's are not included in the main Figures. However, I present individual taxa responses in Supplementary Figure S4.1 and individual responses split by season in Supplementary Figure S4.1.

4.4.3 iDNA responses to covariates

For each of the nine taxa (not including long tailed porcupine), I tested how the site-occupancy, sample-availability and PCR detection probability of the target species iDNA varied with the habitat quality, sampling effort and technical replication (variables β, α, δ).

Site effects on DNA occupancy

I found that the iDNA occupancy probability varied among taxa, seasons and across a range of canopy heterogeneity values. In general, during the dry season (2015) values of ψ decreased with increasing levels of habitat heterogeneity. This indicates that probability of species occupancy declines as the habitat becomes more clustered, with increased gaps in the canopy (Figure 4.1). The species with the highest probability of occupancy and the smallest change in occupancy across the range of habitats in 2015 was bearded pig and muntjac. I found that the species with the lowest probability of occupancy was banded civet, and both the banded civet and sambar deer showed the largest decrease in occupancy across the range. Only the Malay civet shows the opposite trend, where the probability of site occupancy has a positive relationship with habitat heterogeneity. On the contrary, estimates of occupancy from the wet season (2016) did not reveal similar trends, and there was no clear relationship between habitat heterogeneity across the gradient and iDNA occupancy probability (Figure 4.1).

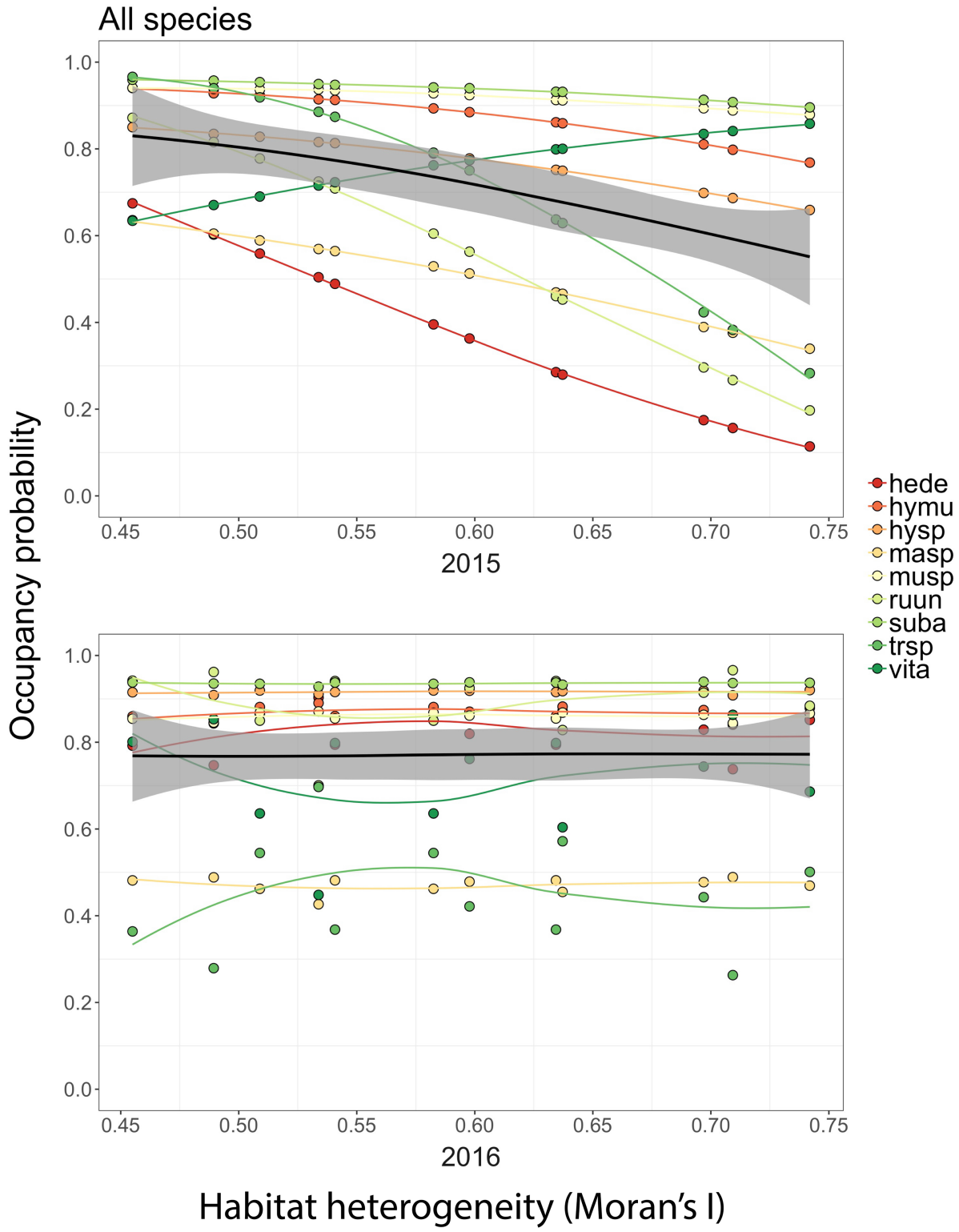


Figure 4.1. Changes in probability of site-occupancy (ψ) for each taxon in response to changes in habitat heterogeneity. Increases in heterogeneity represent increases in gaps and clustering of the canopy at that site. Each taxa's response is shown in a different colour. The responses for the dry season (2015) are in the top panel and for the wet season (2016) in the bottom panel. The four-letter code is used for each taxon, for full taxonomic names refer to Table 4.1. The black line shows the mean occupancy probability (ψ), across all taxa with the shaded area showing the 95% confidence interval.

Sampling effects on DNA availability

Examination of the effects of with sampling effort (i.e. number of pools sequenced) revealed no strong average response among all species looked at (Figure 4.2). During the dry season of 2015, increasing the sampling effort was related to an increase in the probability of iDNA availability for some taxa (muntjac, gibbon, mousedeer and macaque), and a decrease in probability for other taxa (banded civet, porcupine and sambar deer). The availability of bearded pig and Malay civet DNA was relatively consistent with varying amounts of leech pools, at ~80% and ~50% respectively, from 0 – 20 pools. Considering the wet season of 2016, there was less taxon-level variation in responses, and the probability of the DNA occurring in the leech samples remained low (< 50%) and most species showed no (or very little) effect of increasing the number of pools. Bearded pig and banded civet DNA showed small increases and macaque DNA showed small decreases in availability in availability probability with increasing sampling effort (i.e. number of pools).

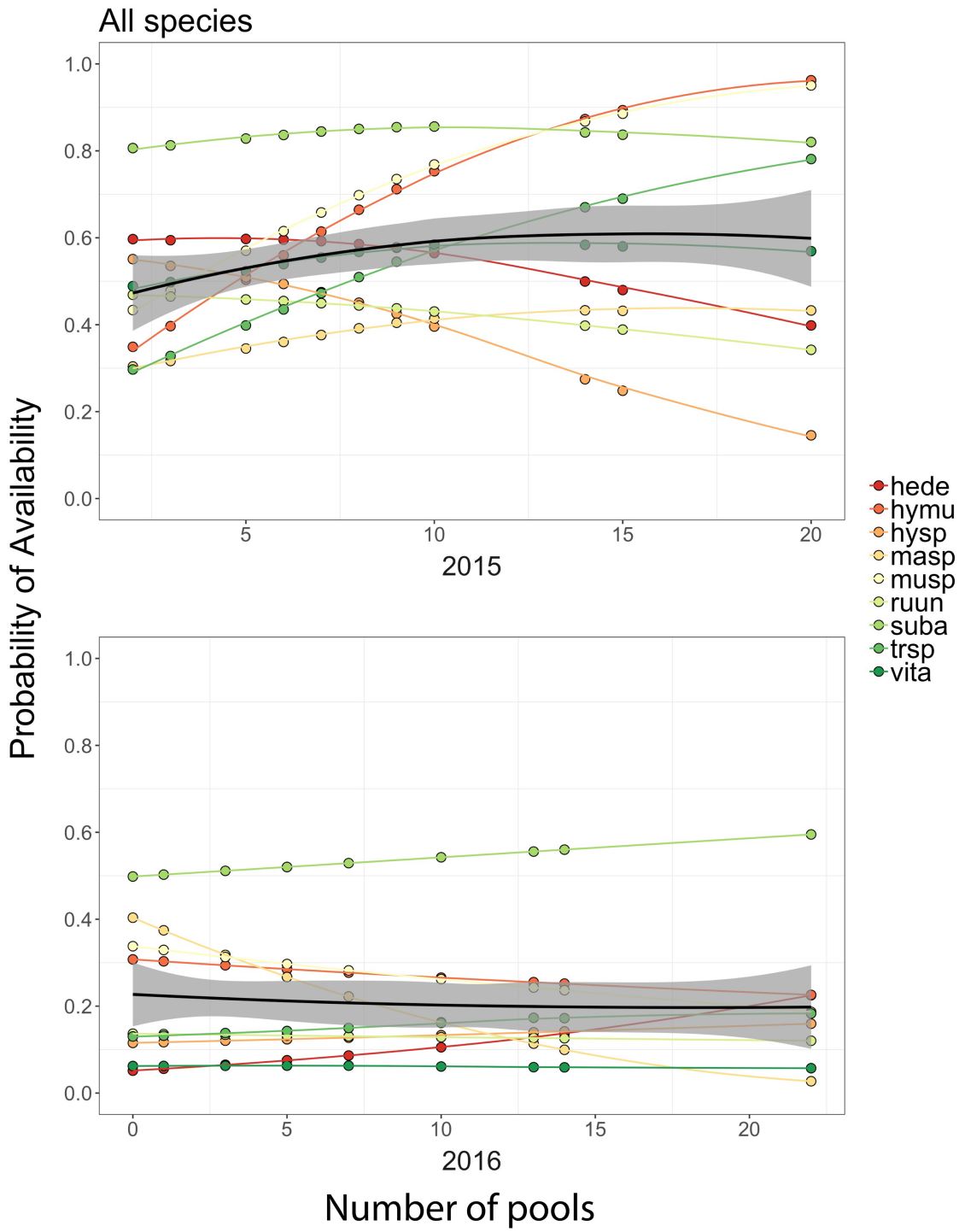


Figure 4.2. Changes in iDNA sample-availability probability (θ) for each taxon in response to changes in sampling effort, i.e. the number of leech pools sequenced. Each taxa's response is shown in a different colour. The responses for the dry season (2015) are in the top panel and for the wet season (2016) in the bottom panel. The four-letter code is used for each taxon, for full taxonomic names refer to Table 4.1. The black line shows the mean availability probability (θ) across all taxa, with the shaded area showing the 95% confidence interval.

Replicate covariate effect on detection probability

In both years, the total weight of the leech pool had no effect on the detectability of DNA in the PCR replicate, with values of p remaining consistent within species (Figure 4.3). In 2015, probabilities ranged downwards from 0.8, for muntjac and bearded pig, to 0.3 for Hystrix porcupine and macaque. In 2016, the range was between 0.6 for bearded pig and <0.1 for Malay civet.

Considering the effect of DNA concentration, in 2015, I found that sambar deer showed the greatest positive response, with the steepest increase in detection probability, rising from <0.2 up to 0.5 (Figure 4.3). Four taxa, banded civet, gibbon, bearded pig, and muntjac, showed small decreases in detection probability with increasing DNA concentration. In 2016, the probability of detecting DNA from macaque, Hystrix porcupine, sambar deer, bearded pig and gibbon increased with DNA concentration, and for macaques there was a considerable increase, from a detection probability of zero to 90% of replicates as DNA concentration increases two-fold. DNA concentration had no effect on the detectability of banded civet or muntjac DNA in the PCR replicates and the probability of detecting Malay civet DNA in the PCR replicate decreases as total DNA concentration increases.

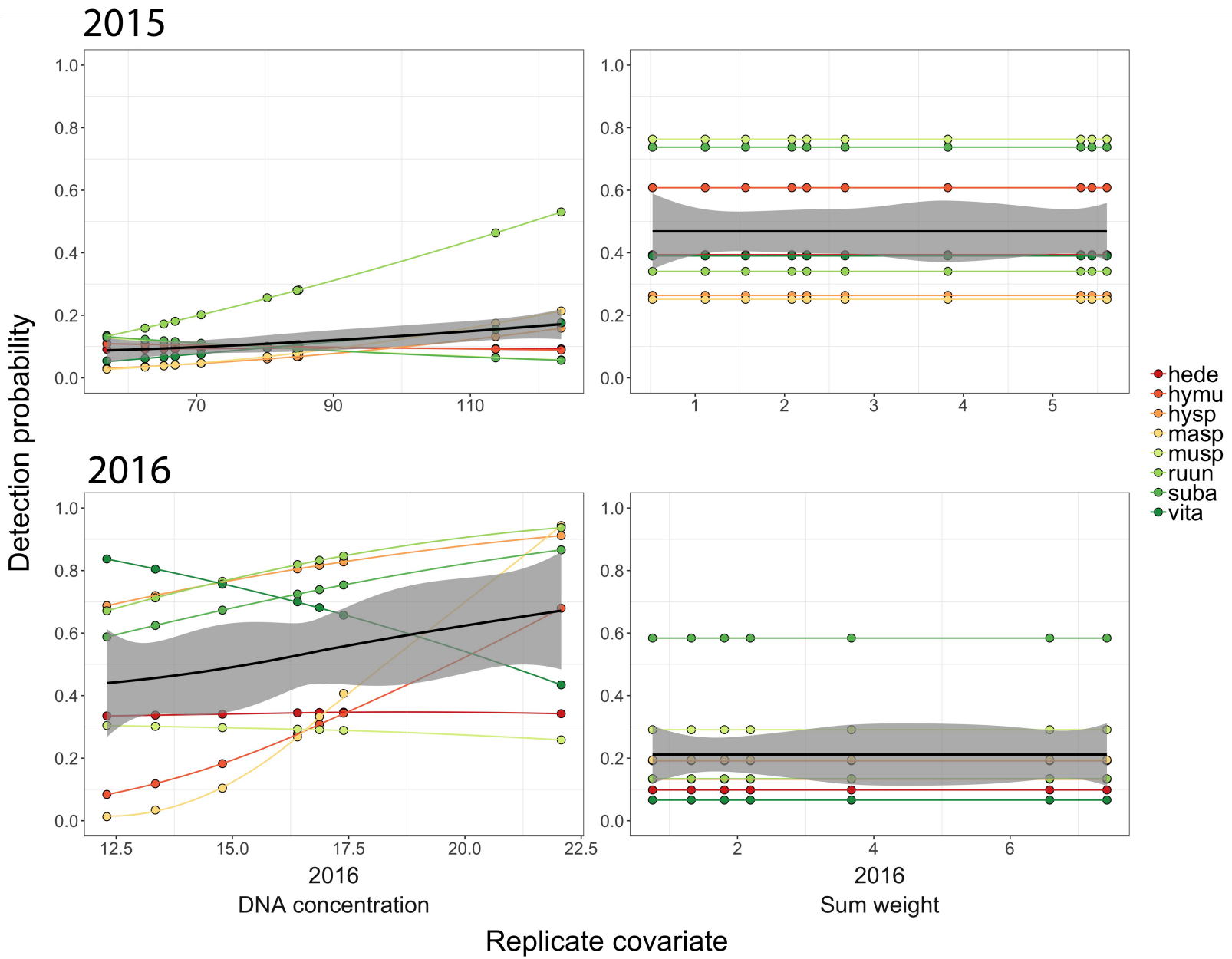


Figure 4.3. Changes in detection probability (p) for each taxon in response to changes in two replicate covariates – *left* = DNA concentration of the pooled extract, this is measured using the Nanodrop for the dry season samples in 2015 and the Qubit for the wet season samples in 2016. *Right* = total relative weights of the individuals pooled together for the DNA extraction. Each taxa's response is shown in a different colour. The responses for the dry season (2015) are in the top panel and for the wet season (2016) in the bottom panel. The four-letter code is used for each taxon, for full taxonomic names refer to Table 4.1. The black line shows the mean detection probability (p) across all taxa, with the shaded area showing the 95% confidence interval.

4.4.4. How much replication is needed?

For leech pools, my findings from 2015 show that for each taxon, and for all taxa combined, I would have needed around five pools per site to achieve an 80% probability of the DNA being available in the pool given that the site is occupied (i.e. ψ = 1) (Figure 4.4). All taxa follow the same trend, with cumulative availability rising to 1.0 in <10 samples. In contrast, in 2016, the probability of availability was much lower and more species-specific, so to reach a 50% and 80% probability of occurrence, I would have needed five pools and ten pools respectively per site, reaching 100% at over 20 samples. For Malay civet, this curve showed no evidence of reaching a plateau even at 20 leech pools.

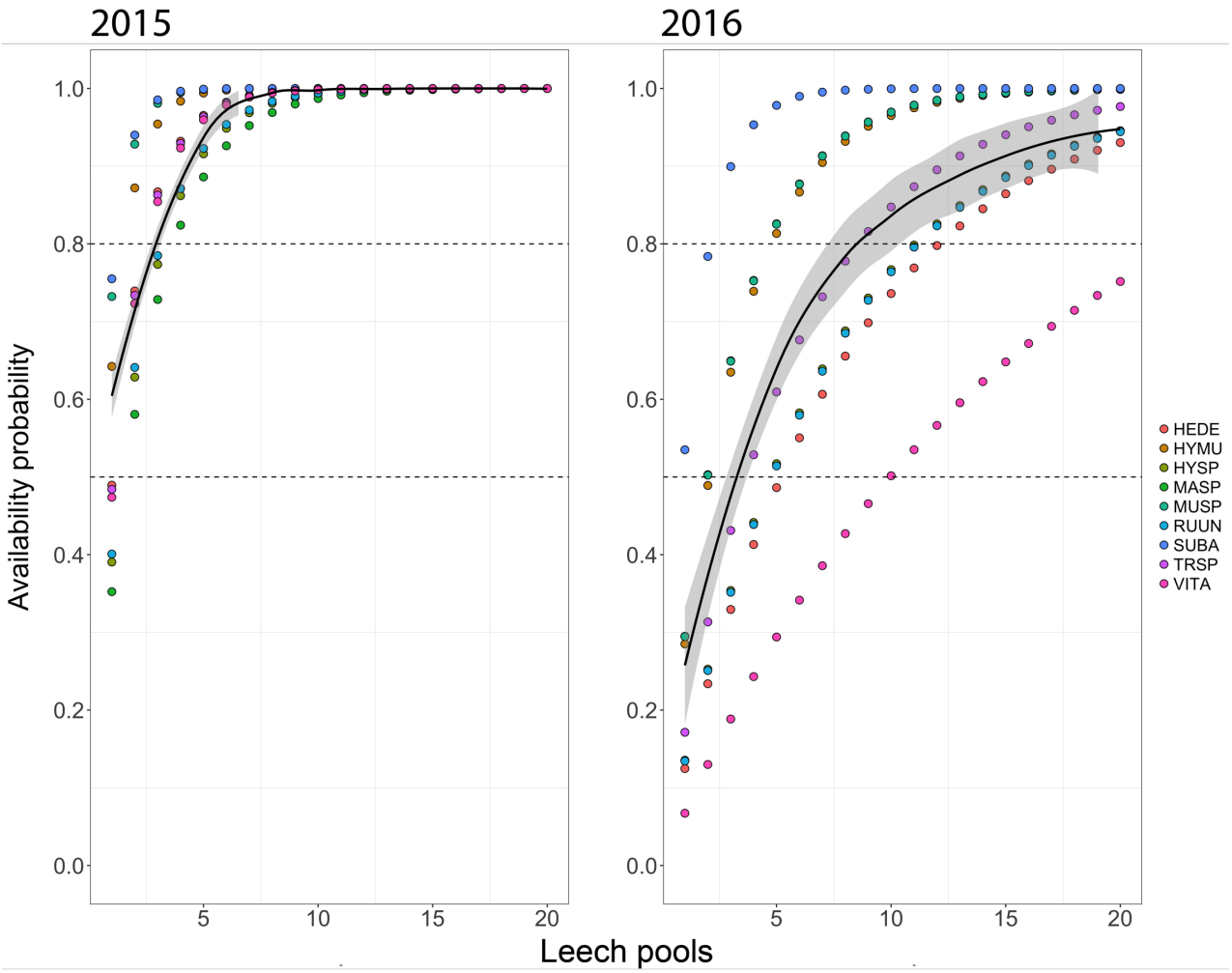


Figure 4.4. The effect of increasing the number of leech pools sequenced on the cumulative availability probability (θ) for mammal iDNA in the leech pools. The response in the dry season (2015) is shown on the left and the response in the wet season (2016) is shown on the right. Each taxon is represented by a different colour and its unique four-letter code (full taxonomic names in Table 4.1). The mean response across all taxa is shown in black with 95% confidence interval. The two dashed lines represent the 50% and 80% probability.

In contrast to availability based on pools, the change in detection probability based on PCR replicates showed the opposite pattern between years (Figure 4.5). Given that the DNA was in the sample (i.e. θ = 1), for 2015, I would have needed at least six PCR replicates to reach an average detection probability of 50%. This rises steeply to 14 PCR replicates required to achieve an average 80% detectability, and >20 replicates required for 100% detectability. For 2016, however, the average detection probability was much higher, with four PCR replicates required for an average of 50% detectability. I found that around eight replicates were needed to reach an average detection probability of 100%, and 17 replicates were needed to ensure 100% detection probability of every species.

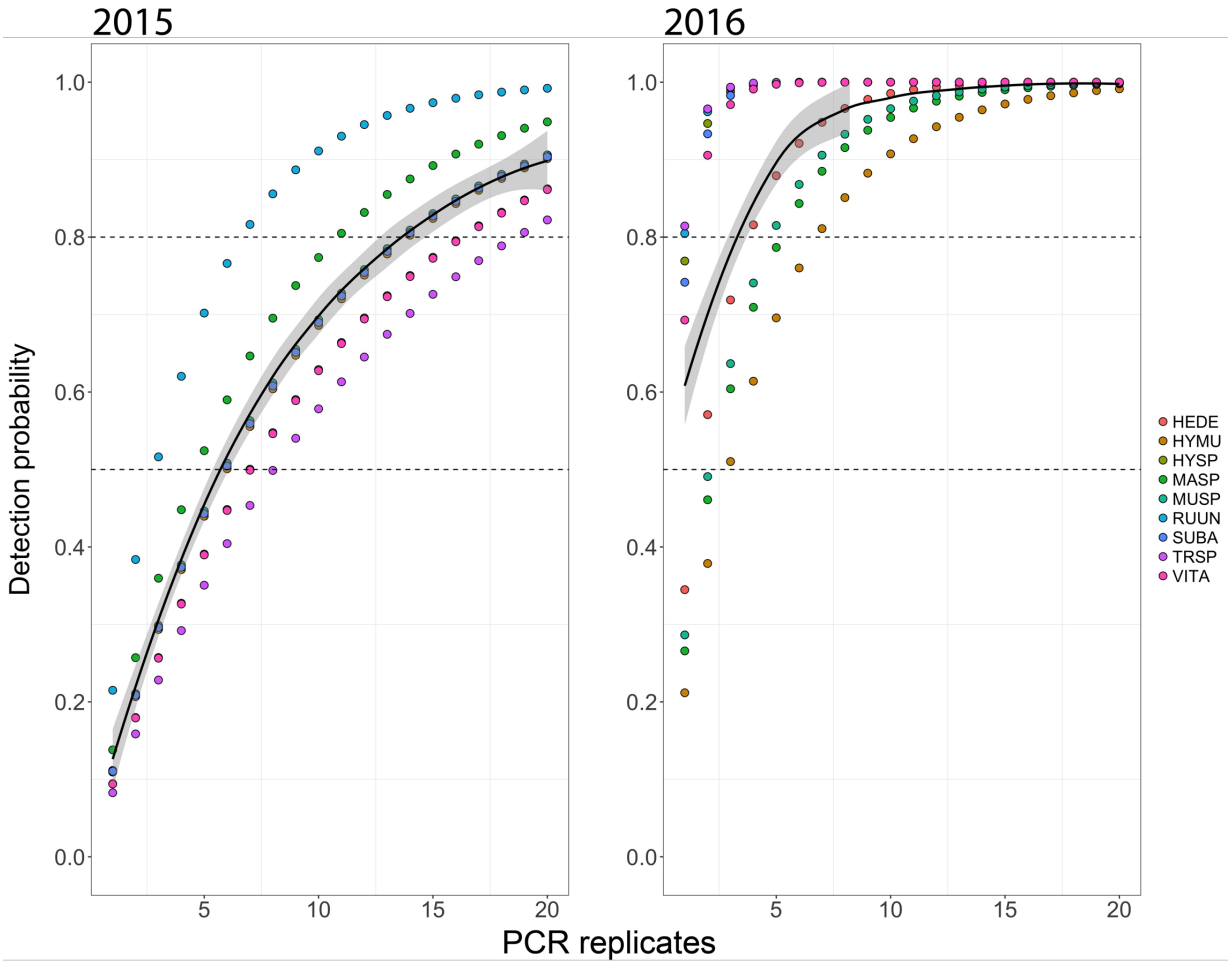


Figure 4.5. The effect of increasing the number of PCR replicates on the cumulative detection probability (p) for mammal iDNA in the leech pools. The response in the dry season (2015) is shown on the left and the response in the wet season (2016) is shown on the right. Each taxon is represented by a different colour and its unique four-letter code (full taxonomic names in Table 4.1). The mean response across all taxa is shown in black with 95% confidence interval. The two dashed lines represent the 50% and 80% probability.

4.5. Discussion

In this study I applied an occupancy modelling statistical framework to molecular detection data for mammals that were derived from the sequencing of leech blood-meals from Borneo. By using a multiscale extension of the model, I have demonstrated that it is possible to incorporate imperfect detection at two-levels; first, using the availability parameter θ to estimate the probability that the DNA is available for sequencing (i.e. it is in the leech sample), and second, the detection parameter p to estimate the probability DNA detected in a PCR replicate. I was also able to include covariates to model how these probabilities change with varying habitat quality, sampling effort and molecular factors. To my knowledge, this is the first time that iDNA data has been analysed using multiscale occupancy models and the results provide a deeper insight into the impacts of habitat degradation in a human-modified landscape using a relatively novel molecular sampling technique.

4.5.1. Habitat effects and occupancy

Using an occupancy modelling approach, my results provide evidence that the occurrence of mammal DNA is related to habitat quality, even after accounting for imperfect detections. Specifically, habitat quality had a greater impact on occupancy probability during the dry season (2015) than in the wet season (2016). In general, decreasing habitat quality (i.e. increasing clustering) was related to decreased probability of occupancy. As well as supporting my original prediction, which is similar to the findings of Deere *et al.* (2017) based on camera traps, who recorded increased community occupancy probabilities in higher quality logged forest. My results showed a strong impact of season on the species-specific patterns of occupancy with only Malay civet showing a different trend. It is unclear why the two years of sampling gave different results. One possibility is that the environment was more stable during the wet season in 2016 compared to 2015. The dry season in 2015 was characterised by a long drought period, coupled with large amounts of timber extraction across the field site. During periods of increased disturbance higher quality habitats (such as the managed forest at SAFE) might provide refugia for species, and thus increase the probability that their DNA

occurs at these sites. The response shown by Malay civet iDNA in the dry season (2015) could reflect this species' relatively high tolerance to disturbance, as it is widespread across primary and secondary forests. Although there is some evidence that it is negatively affected by logging and forest degradation (Brodie *et al.* 2014a).

4.5.2. Sample effects and availability

I found no consistent strong effect of sampling effort on the average probability of availability over the two seasons of sampling. There appears to be a positive effect of sampling in the dry season and no effect in the wet season. This is likely to be related to the movement patterns of mammals during the dry season. When mammalian prey are rare, leeches may have a large inter-feeding interval; this will allow for more degradation in the blood meal, and thus increase the impact of sampling effort on iDNA availability probability. We know that this window should have an effect on iDNA but very little is known about the feeding behaviour of these leeches (see Schnell *et al.* 2015 & Drinkwater *et al.* 2018).

4.5.3. PCR effects and detectability

DNA detection probability was not found to be affected by the total weight of leeches in the pool in either season (2015 or 2016). This is perhaps unfortunate because, had an effect been found, it would suggest that increasing the biomass of leeches in the pool would be an easy way for practitioners to enhance detection rates. These results also run counter to earlier findings suggesting that longer leeches may increase detections, although these were not based on occupancy modelling (Weiskopf *et al.* 2017). These findings from Weiskopf *et al.* (2017) are based on different leech species and another source of the discrepancy might be the method of measuring biomass. In my case, biomass could only be quantified after the leeches had been preserved in the field, and thus it is possible that leeches had become heavier through the absorption of the preservation reagent.

When considering the impact of DNA concentration of the pooled extract, my results suggest that concentration is positively associated with probability of detection. This effect was stronger for samples from the wet season (2016) than

for the dry season (2015). However, the two seasons cannot be strictly compared resulting from the differences in the sensitivity of the two different platforms that were used to quantify DNA (Qubit and Nanodrop, respectively) in these two years. This overall positive effect of DNA concentration was unexpected given that high concentrations are regularly thought to reduce PCR efficiency due to the presence of inhibitors. On the other hand, low starting concentrations of DNA are known to cause PCR stochasticity, which may be alleviated by increasing the concentration (Alberdi *et al.* 2017). My findings thus highlight the importance of using and developing protocols to increase the yield of degraded DNA, such as those developed for ancient DNA (Schnell *et al.* 2015a). In practice, concentration thresholds could be set before sequencing to save time and financial resources.

4.5.4. How much replication is needed?

By calculating the cumulative availability probability, I was able to estimate the number of leech samples needed to achieve different chances of detection. For the samples, in both years, I estimated that five to ten leech pools per site would be required to achieve between 50% to 80% probability of availability. This number would be reasonable in practice and calculating these values prior to a study could help to optimise field sampling and reduce problems of overharvesting. On the other hand, these values are averages and, for some species (e.g. Malay civet), many more than 20 samples appear to be needed to reach the same high levels of confidence. In metabarcoding studies, it is typical to conduct replication in triplicate without much consideration given to the impacts of this parameter choice. However, recently studies have shown that the amount of replication and the type of replication thresholds applied (e.g. additive or restrictive) can affect the diversity detected (Alberdi *et al.* 2017). My results show that even the rule of thumb of three replicates would not be sufficient for many species, highlighting the need to consider these parameters on a case by case basis.

When estimating cumulative detection probabilities for PCR replicates, however, the curves showed a far greater number of replicates per sample would be needed to reach 50% detection probability in the dry season (2015) than in the wet season (2016). In the 2015 data, the requirement for over ten PCR replicates (in order to

be confident of the detection results) is currently unlikely to be feasible for most laboratory studies, due to financial and time constraints. Yet there is a growing need for reliable eDNA testing, especially given the popularity of sampling for species of conservation concern. For example, water sampling for eDNA has been used by the British government (The Department for Environment, Food and Rural Affairs, DEFRA) in surveys for the EU-protected great crested newt (*Triturus cristatus*) (Bigg *et al.* 2014). To obtain high confidence in detections, Bigg *et al.* (2014) used 12 replicates; however, this original study used an approach based on qPCR, which is less costly than one based on high throughput sequencing. Terrestrial leeches are also currently being used by the World Wildlife Fund (WWF) to search for the critically endangered saola (*Pseudoryx nghentinhensis*) (WWF 2013). This represents a case where certainty of detection is extremely important due to limited conservation funds, and the potential consequences of a false positive detection are high. The species-specific variation for detection probability also shows the importance of preliminary studies so that parameters can be tailored to particular taxa.

One way to improve the reliability of detections in eDNA (and iDNA) studies is by combining information over multiple loci to obtain better resolution in species assignments. For example, in my study, several taxa could not be assigned to species-level, and instead were only resolved at the level of genus (e.g. *Hystrix* was used for *H. brachyura* and *H. crassispinis*). Behavioural and ecological differences among congeners could potentially drive erroneous estimates of occupancy and detectability. In some cases, congeneric taxa may also differ in conservation status. For example, within the genus *Muntiacus, M. muntjac* a listed by the IUCN a Least Concern species whereas its congener *M. atherodes* is listed as Near Threatened. A similar case is seen in the macaques, with *Macaca nemestrina* considered Vulnerable but *M. fasciculata* is of Least Concern. Consequently, the conservation decision and actions implemented after detecting these species could be different, especially if it were found that the more common species was driving estimates of occupancy. Apart from using just 16S rRNA gene, other studies have also combined this with the 12S rRNA marker to increase resolution for other blood-feeding insects (Kocher *et al.* 2017c). However, the usefulness of any genetic marker is

related to the availability of reference sequences, and indeed the completeness and quality of the reference database are major constraints. Currently 16S has the best reference database for Bornean mammals in part due to previous studies (e.g. Mohd Salleh *et al.* 2017), although the COI marker has been adopted as the barcoding marker more generally, and has a very comprehensive and well-curated COI barcode reference database (BOLD, the Barcode of Life Database), (Ratnasingham & Hebert 2013).

4.5.5. Caveats of occupancy modelling iDNA

Problems with false positives

To my knowledge this is the first iDNA study to apply occupancy models and more research is needed to understand the sources of error in such cases. To date, occupancy model studies have focused on the problem of false negative detections (failing to record a species that is present). For survey methods such camera traps, false negatives may have a more detrimental effect on parameter estimates than species misidentifications (Moilanen 2002). However, more so than other survey methods, metabarcoding approaches, such as used in my study, are likely to be more prone to higher levels of false positives through ambiguous or erroneous taxonomic assignments, contamination (from the field or laboratory) and incomplete reference sequences. Dealing with false positives is considered statistically difficult, and, compared to false negatives, less work has been done on developing models that account for them. Ficetola *et al.* (2016) advocated zeroing single detection histories assuming these to be artefacts, while others have considered this practice to introduce bias (Lahoz-Monfort *et al.* 2016).

In this study, I adopted several field and laboratory practices to reduce the risk of generating false positives. These included the use of protocols specifically designed for degraded DNA (e.g. ancient DNA), running and sequencing negative controls (PCR and extractions), matched primer tagging to identify tag-jumps, using lower numbers of PCR cycles to reduce non-target DNA amplification, and using PCR free laboratories. I also implemented bioinformatic techniques to reduce contamination, such as using post-clustering filtering (LULU), strict filtering thresholds and removing singletons, but this greatly reduced the number of

detections. Ideally, it would be best to calibrate the models using unambiguous detections such as detections from camera traps which were collected alongside the eDNA samples (Lahoz-Monfort *et al.* 2016). Despite these measures, however, it is still possible that false positives were introduced in the field. For example, I extracted DNA from the whole leech and there is a risk that any mammal DNA found on the outside of the leech could be amplified too. This type of problem is probably more serious for eDNA studies that use water samples from flowing water bodies; for example, where DNA can move downstream causing a positive detection in a site where the species was not present (Deiner & Altermatt 2014).

Closed season sampling

Another caveat of the occupancy model used is the assumption that sampling takes place within a closed season, where the true occupancy state of the species does not change. Due to lack of knowledge on the dispersal ecology of the leeches I assumed that individuals did not move large distances without being attached to a mammalian host (Schnell *et al.* 2015a). Although not yet recorded, it is possible that terrestrial leeches do move independently of the host, and there is potential that they could change the occupancy state and thus violate model assumptions. A deeper understanding of the ecology of the invertebrate sampler needs to be known to be able to account for leech movement in the models.

Independent leeches and herding animals

The final caveat concerns the assumption of independent detections. In this study, I pooled individuals from the same site. While it is assumed that these leeches represent independent samples (i.e. the detection of a mammal in the blood meal of one does not influence the detection of another in the blood meal), it is possible that this is not the case. For example, some of the mammals studied are known to aggregate and travel in groups, such as sambar deer (*Rusa unicolor*) and bearded pig (*Sus barbatus*). If the leeches were actually feeding on individuals of a family group or if multiple individuals were on one animal, this would artificially increase the estimates site occupancy for that species. Using the leech iDNA for individual identification would help to estimate how many individual mammals are being fed

on at a site, I would need microsatellite markers to achieve this, but it has been successful for other invertebrate samplers (Schubert *et al.* 2015).

4.6. Conclusion

In this study I have shown that powerful and versatile occupancy models can be combined with molecular detection data from iDNA studies. This allows the estimation of two parameters which describe imperfect detections and uncertainty in the presence of a species at a particular site. By detecting species-specific differences in occupancy and detectability regarding several habitat quality and technical metrics, my findings demonstrate that leech iDNA occupancy models can be usefully applied to monitoring in a conservation context. Yet caveats exist, and I highlight some aspects of this approach, both in the laboratory, and in understanding of the invertebrate ecology, where ongoing research and improvements are needed.

4.7 Supplementary information

Supplementary tables

Table S4.1 Pearson's correlation coefficients between the LiDAR vegetation covariates used in occupancy models, * shows the significance where $\alpha = 0.05$

	Canopy height	Habitat heterogeneity	Above ground biomass	Forest cover
Canopy height	1.00			
Habitat heterogeneity	-0.72*	1.00		
Aboveground biomass	1.00*	-0.72*	1.00	
Forest cover	0.91*	-0.72*	0.87*	1.00

Table S4.2 Table of model comparison criteria for all models (m1-m5), for all mammalian taxa (full taxonomic names in Table 4.1) in both seasons, dry (2015) and wet (2016). The covariates used in the models are: **m1** = null (none), **M2** = habitat heterogeneity (morans) **M3** = number of pools per site, (pools) **M4** = DNA concentration (conc), **M5** = total pool weight (weight). The two criteria used for model comparison are Watanabe's AIC (WAIC) and posterior-predictive loss (PPL)

			2015		2016	
Species	**Model**	**Covars**	**WAIC**	**PPL**	**WAIC**	**PPL**
HEDE	m1	none	0.13	8.86	0.16	10.50
HEDE	m2	moran	0.14	8.94	0.16	10.43
HEDE	m3	pools	0.14	9.23	0.17	10.97
HEDE	m4	conc	0.14	8.83	0.16	10.51
HEDE	m5	weight	0.14	8.82	0.16	10.59
HYMU	m1	none	0.41	31.28	0.28	18.60
HYMU	m2	moran	0.41	31.14	0.29	18.63
HYMU	m3	pools	0.42	31.92	0.31	19.62
HYMU	m4	conc	0.41	31.08	0.24	12.98
HYMU	m5	weight	0.41	30.98	0.27	15.32
HYSP	m1	none	0.13	7.57	0.20	15.10
HYSP	m2	moran	0.13	7.65	0.20	15.14
HYSP	m3	pools	0.13	7.88	0.20	15.24
HYSP	m4	conc	0.13	7.52	0.21	15.27
HYSP	m5	weight	0.12	7.44	0.18	14.11
MASP	m1	none	0.06	3.93	0.21	14.51
MASP	m2	moran	0.07	3.99	0.22	14.57
MASP	m3	pools	0.07	4.17	0.23	15.19
MASP	m4	conc	0.06	3.90	0.13	7.39
MASP	m5	weight	0.07	3.91	0.13	7.64
MUSP	m1	none	0.50	38.83	0.35	25.84
MUSP	m2	moran	0.50	38.95	0.35	25.81
MUSP	m3	pools	0.49	38.54	0.37	26.92
MUSP	m4	conc	0.50	38.81	0.35	25.48
MUSP	m5	weight	0.50	38.72	0.36	25.55
RUUN	m1	none	0.30	26.17	0.17	12.25
RUUN	m2	moran	0.29	26.08	0.17	12.29
RUUN	m3	pools	0.32	27.77	0.17	12.30
RUUN	m4	conc	0.28	23.96	0.18	13.03

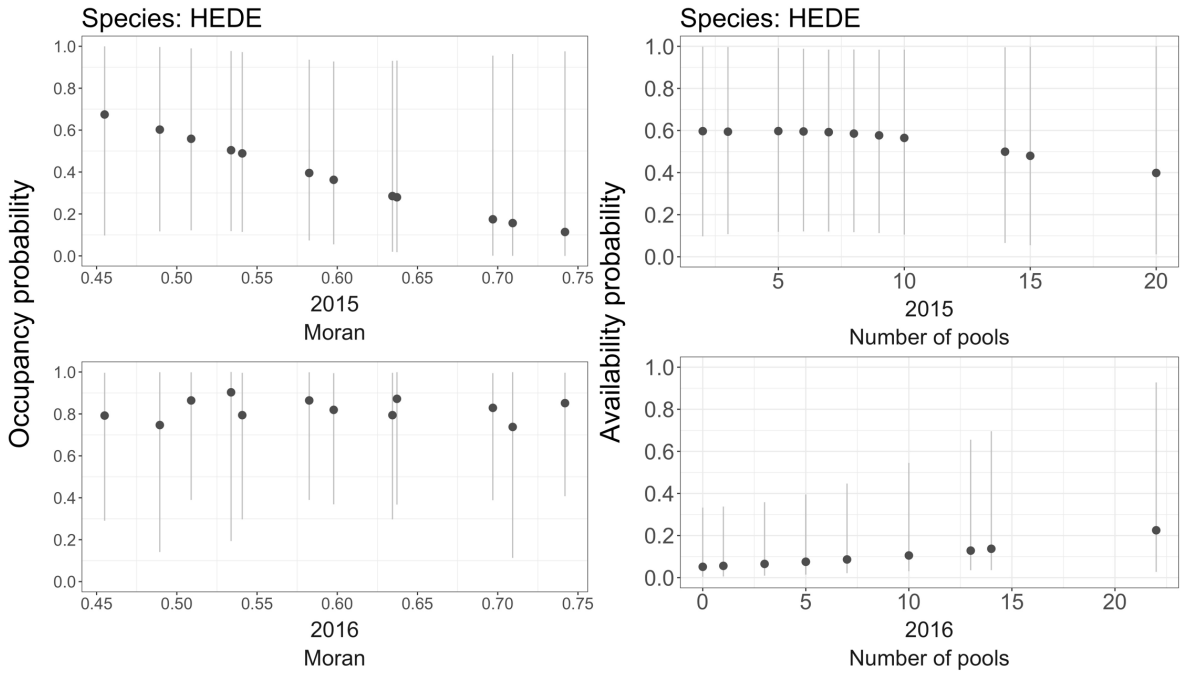
RUUN	m5	weight	0.28	22.99	0.19	13.25
SUBA	m1	none	0.52	40.54	0.79	62.31
SUBA	m2	moran	0.51	40.43	0.79	62.27
SUBA	m3	pools	0.53	41.21	0.80	62.45
SUBA	m4	conc	0.52	40.26	0.83	63.47
SUBA	m5	weight	0.51	40.00	0.83	63.06
TRSP	m1	none	0.20	12.77	0.09	6.26
TRSP	m2	moran	0.20	12.70	0.09	6.28
TRSP	m3	pools	0.21	13.24	0.09	6.26
TRSP	m4	conc	0.20	12.68	0.10	6.40
TRSP	M5	weight	0.20	12.68	0.10	5.87
VITA	m1	none	0.17	10.87	0.09	5.95
VITA	m2	moran	0.17	10.81	0.09	5.88
VITA	m3	pools	0.18	11.23	0.09	6.00
VITA	m4	conc	0.16	10.46	0.09	6.03
VITA	m5	weight	0.17	10.66	0.09	6.08

Supplementary figures

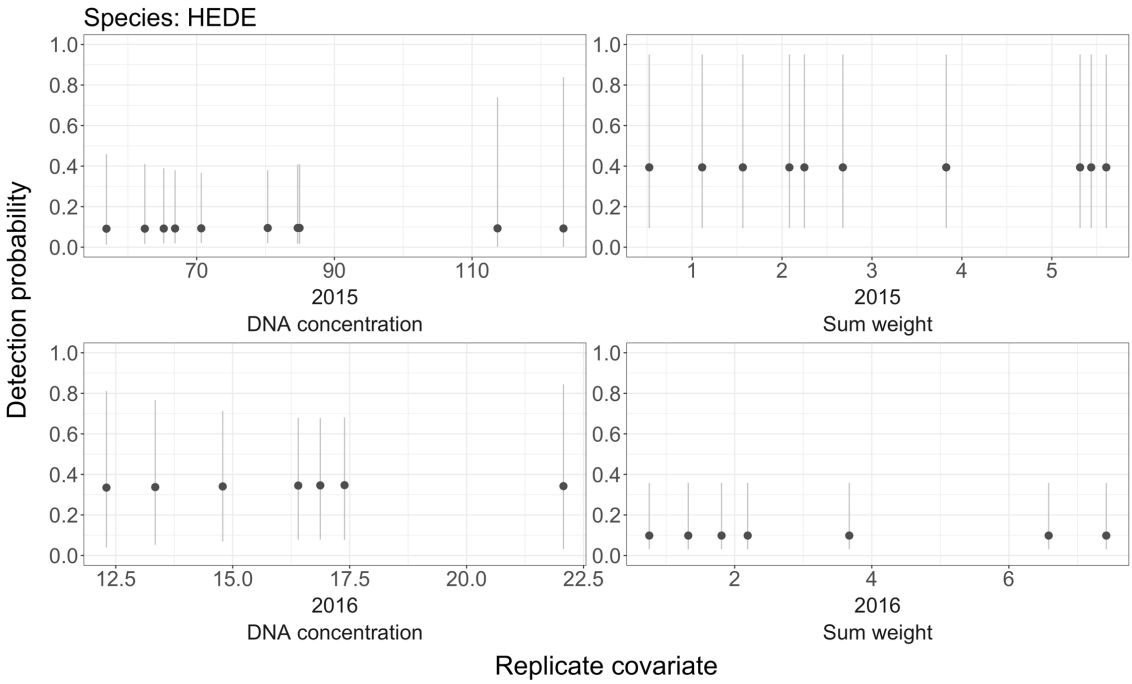
Figures S4.1. Individual taxa response plots to the site, sample and replicate covariates with their respective 95% Bayesian credible intervals. For each taxa the top panel is the response in the dry season (2015) and the bottom panel is the wet season (2016) response. Plot (1) shows the response of occupancy probability to changes in habitat heterogeneity (Moran) on the left side and the response to changes in sampling effort on the right side. Plot (2) shows the response of detection probability to the two replicate-level covariates, DNA concentration of the extract on the left and the total weight of the individual leeches on the right.

HEDE - Banded civet (*Hemigalus derbyanus*)

1. Occupancy probability with habitat covariates and availability probability with sample-level covariates

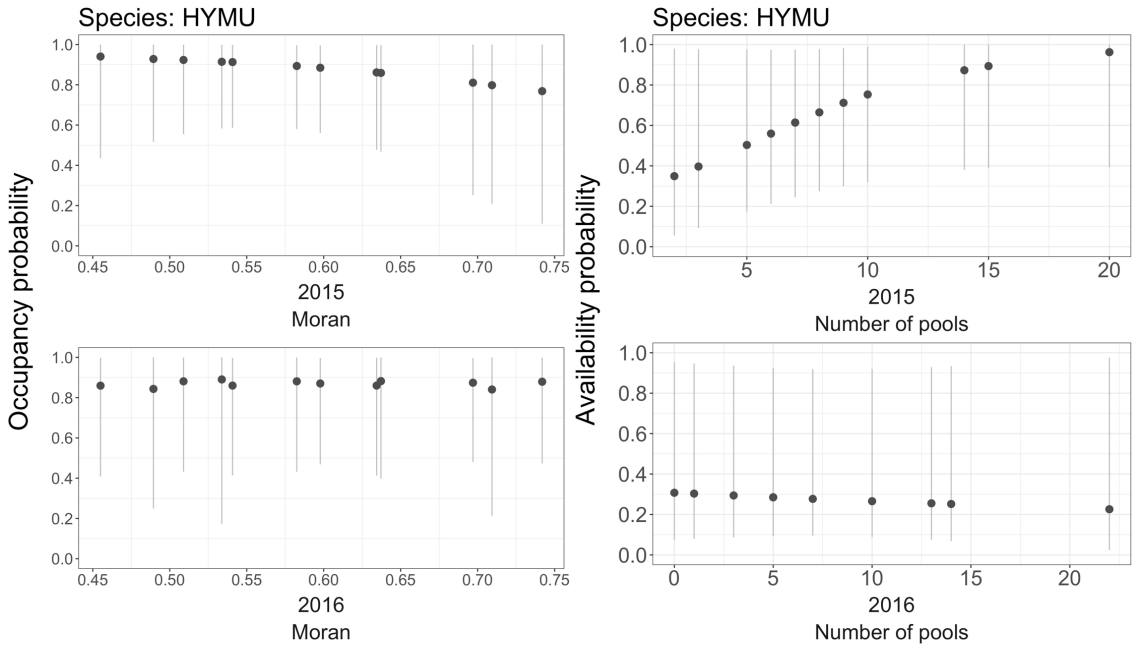


2. Detection probability with replicate-level covariates

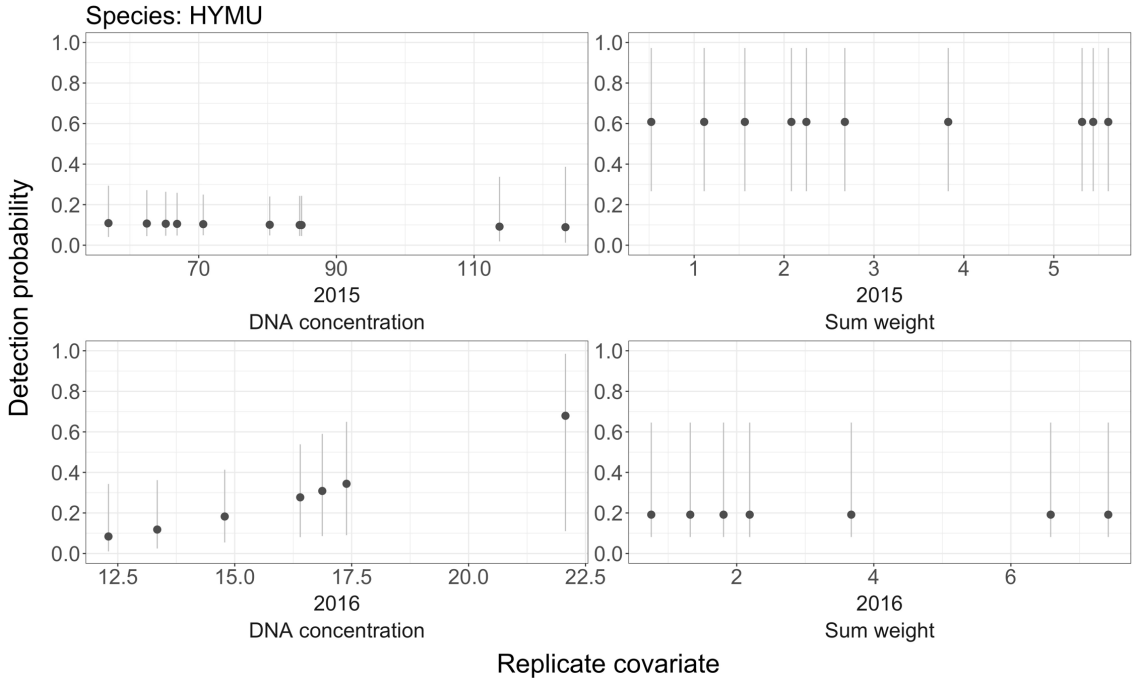


HYMU - Bornean gibbon (*Hylobates muelleri*)

1. <u>Occupancy probability with habitat covariates and availability probability with sample-level covariates</u>

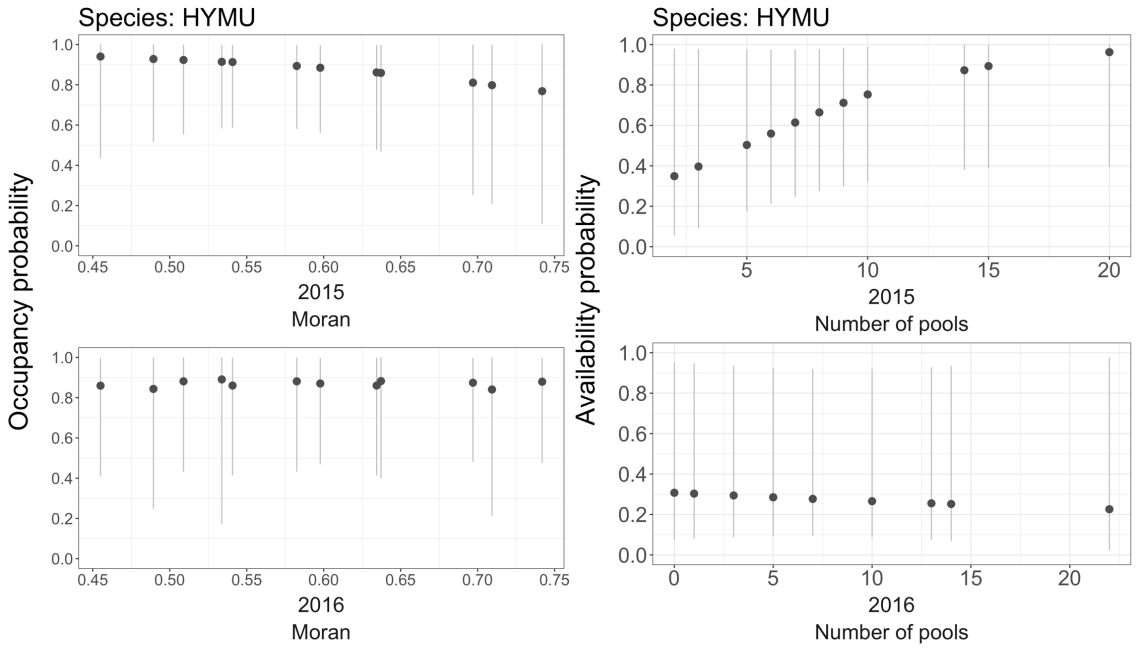


2. <u>Detection probability with replicate-level covariate</u>

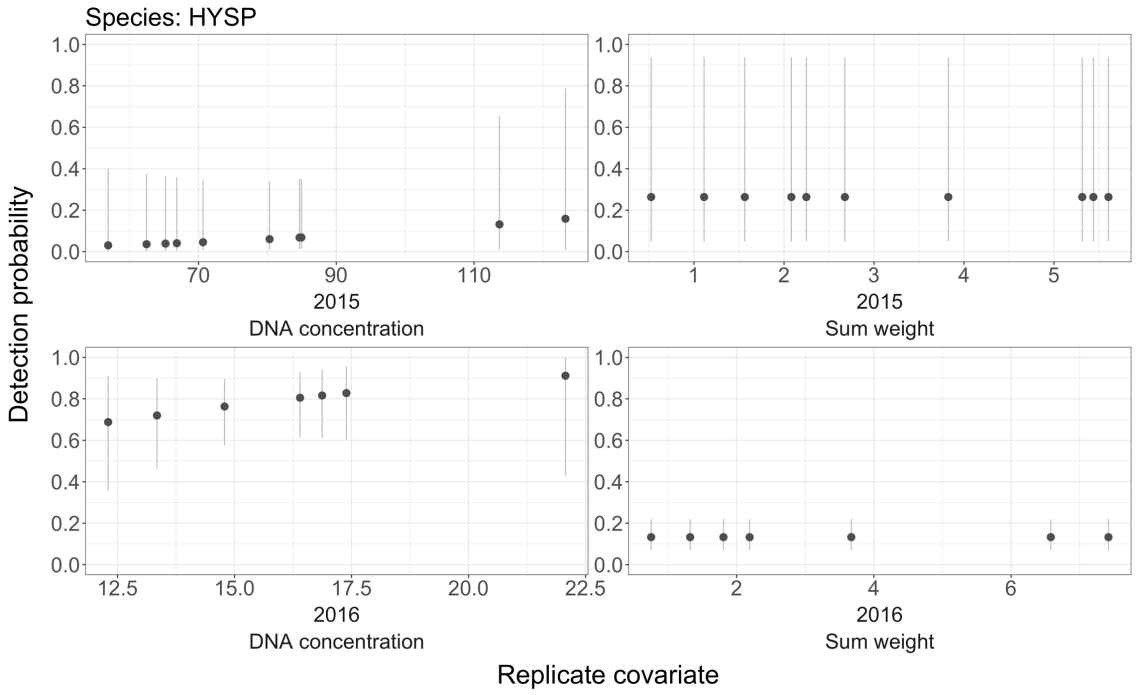


HYSP - porcupine (*Hystrix sp*)

1. <u>Occupancy probability with habitat covariates and availability probability with sample-level covariates</u>

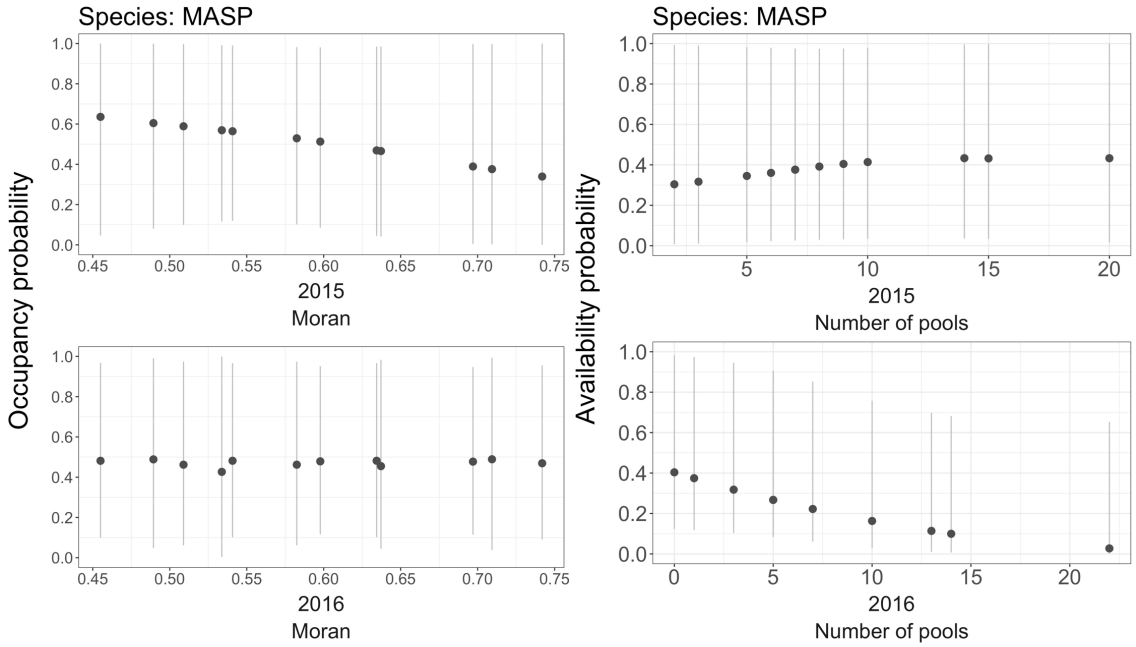


2. <u>Detection probability with replicate-level covariates</u>

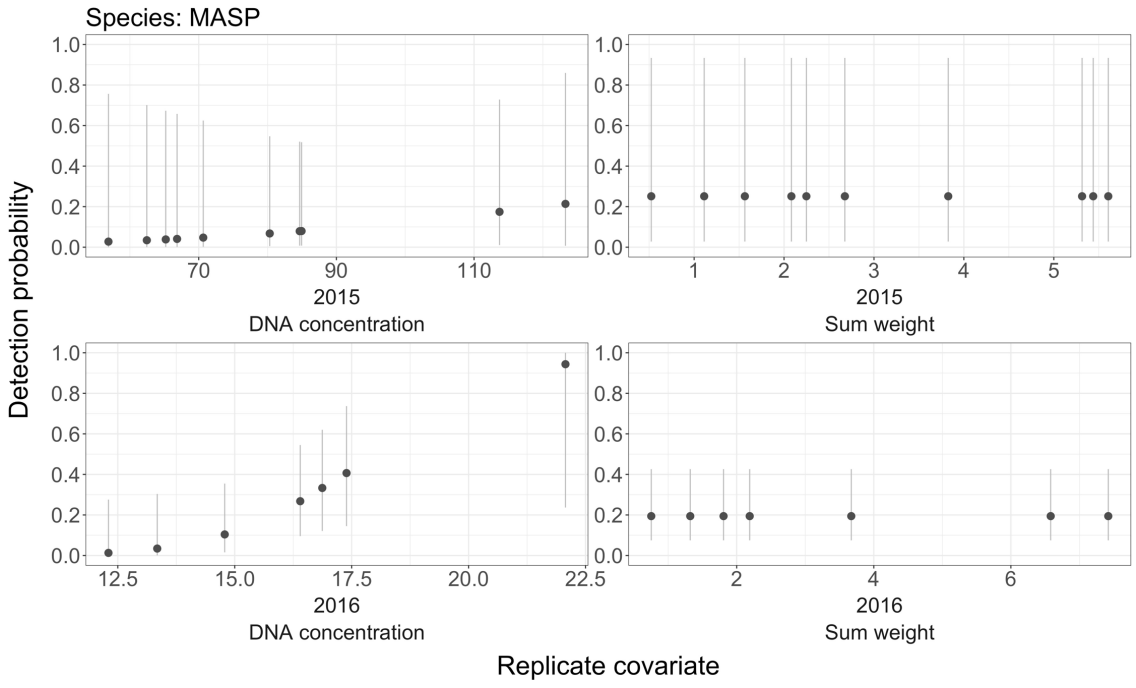


MASP - Macaque (*Macaca sp)*

1. <u>Occupancy probability with habitat covariates and availability probability with sample-level covariates</u>

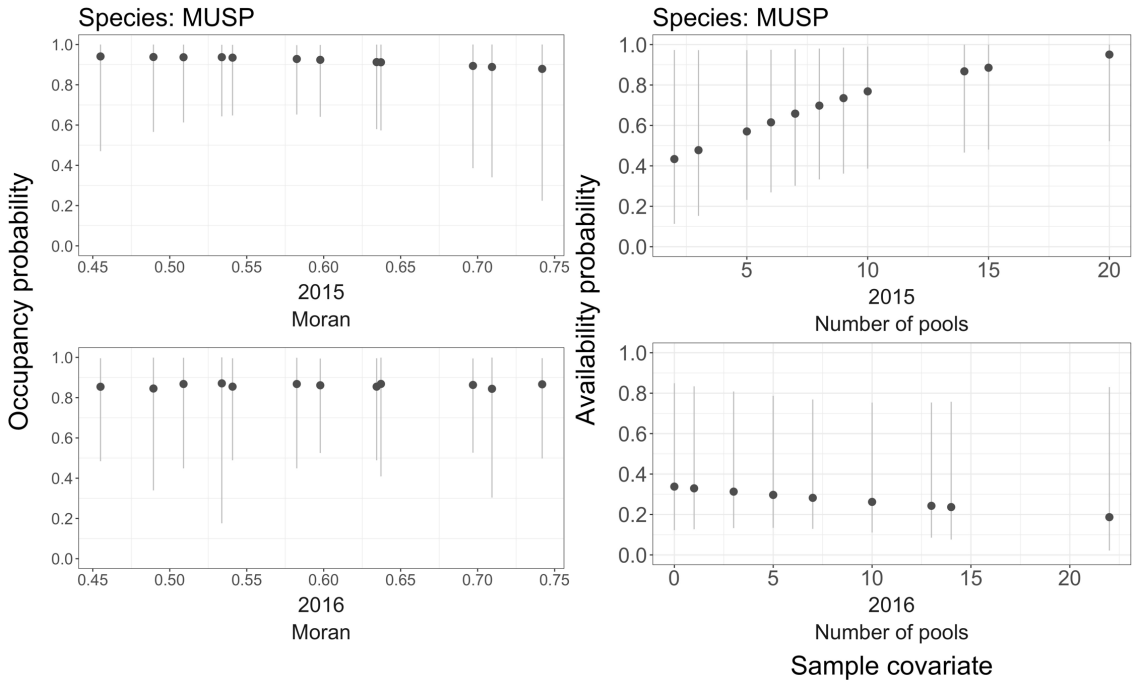


2. <u>Detection probability with replicate-level covariates</u>

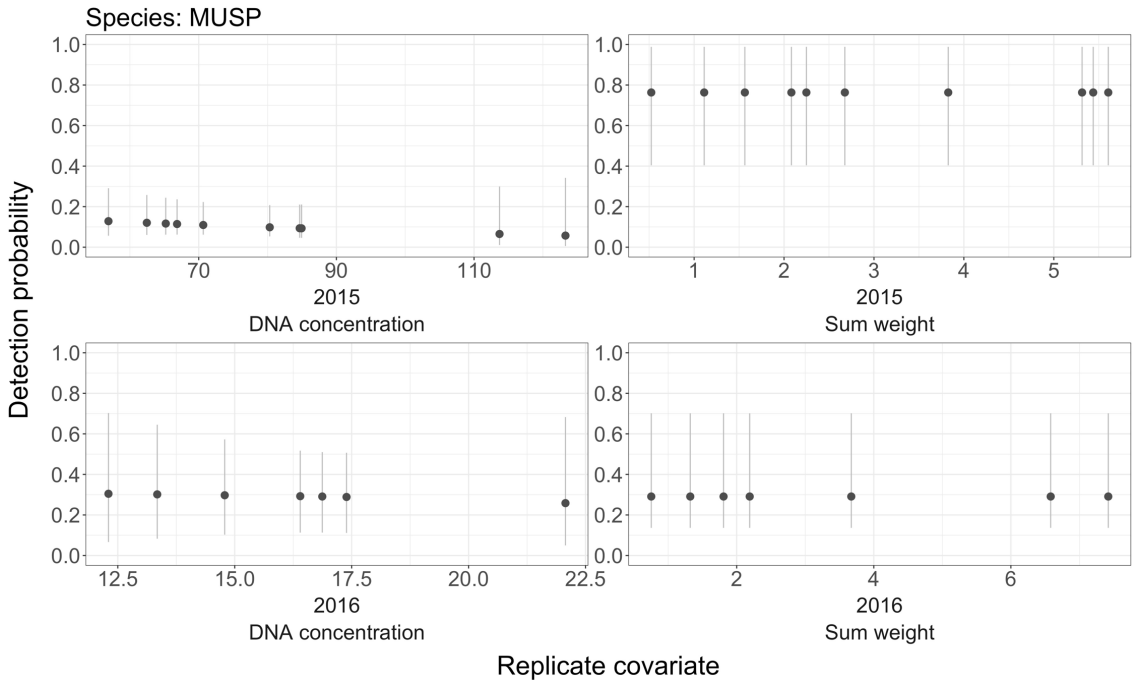


MUSP - Muntjac (*Muntiacus sp*)

1. Occupancy probability with habitat covariates and availability probability with sample-level covariates

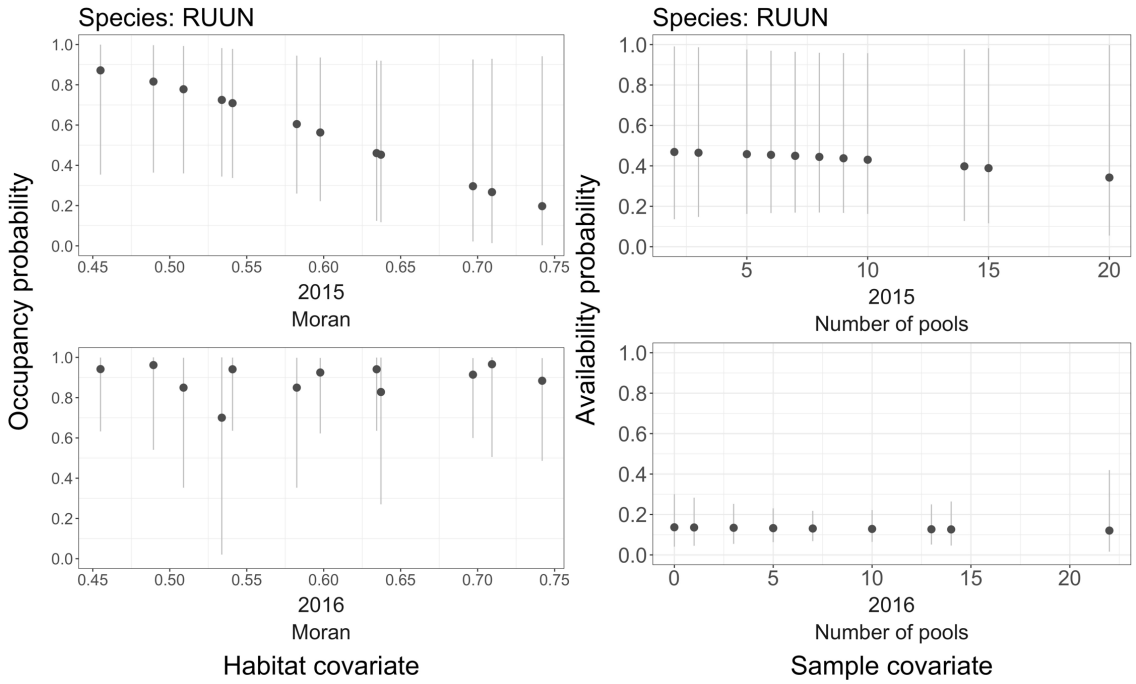


2. Detection probability with replicate-level covariates

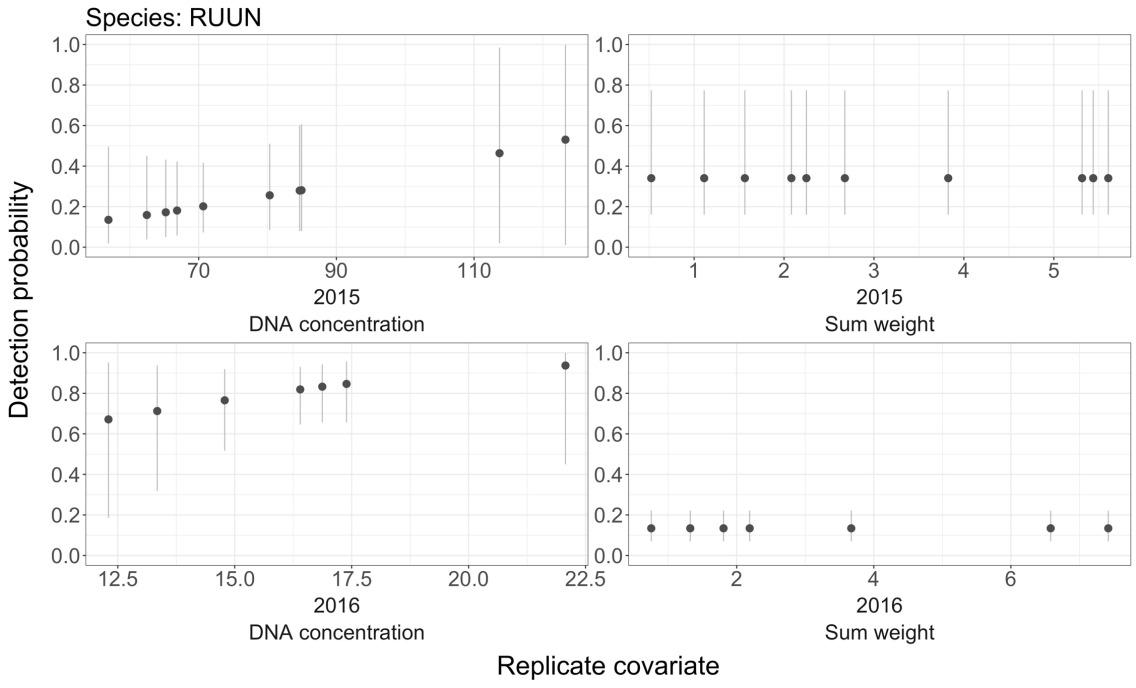


RUUN - Sambar deer (*Rusa unicolor*)

1. Occupancy probability with habitat covariates and availability probability with sample-level covariates

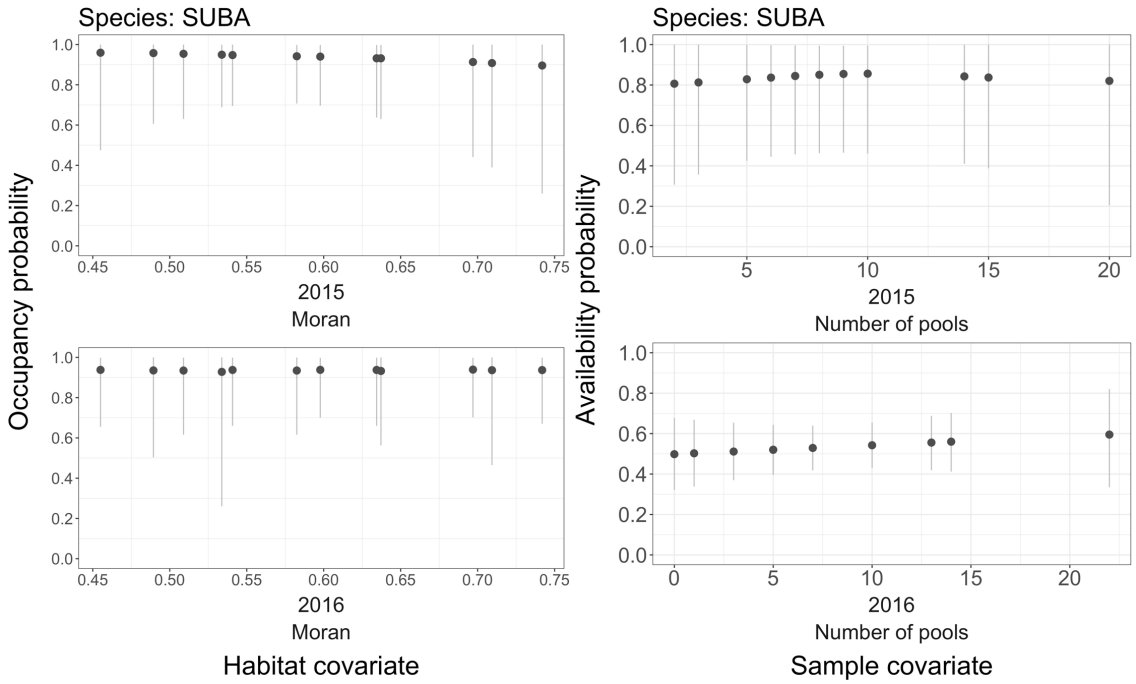


2. Detection probability with replicate-level covariates

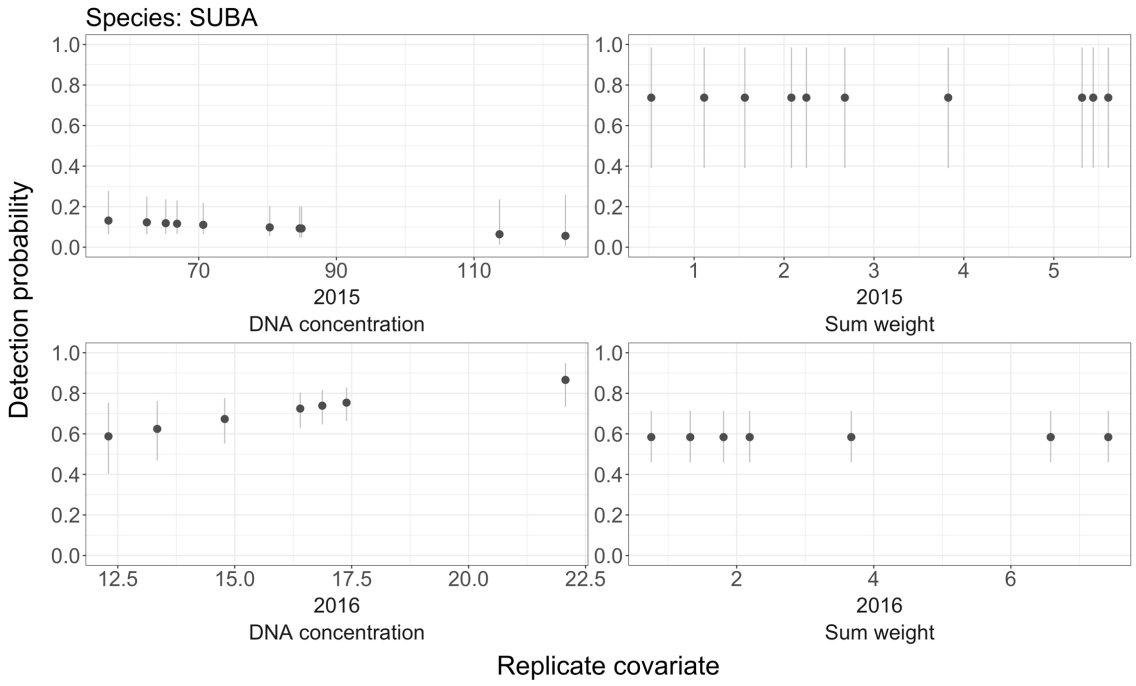


SUBA - Bearded pig (*Sus barbatus*)

1. <u>Occupancy probability with habitat covariates and availability probability with sample-level covariates</u>

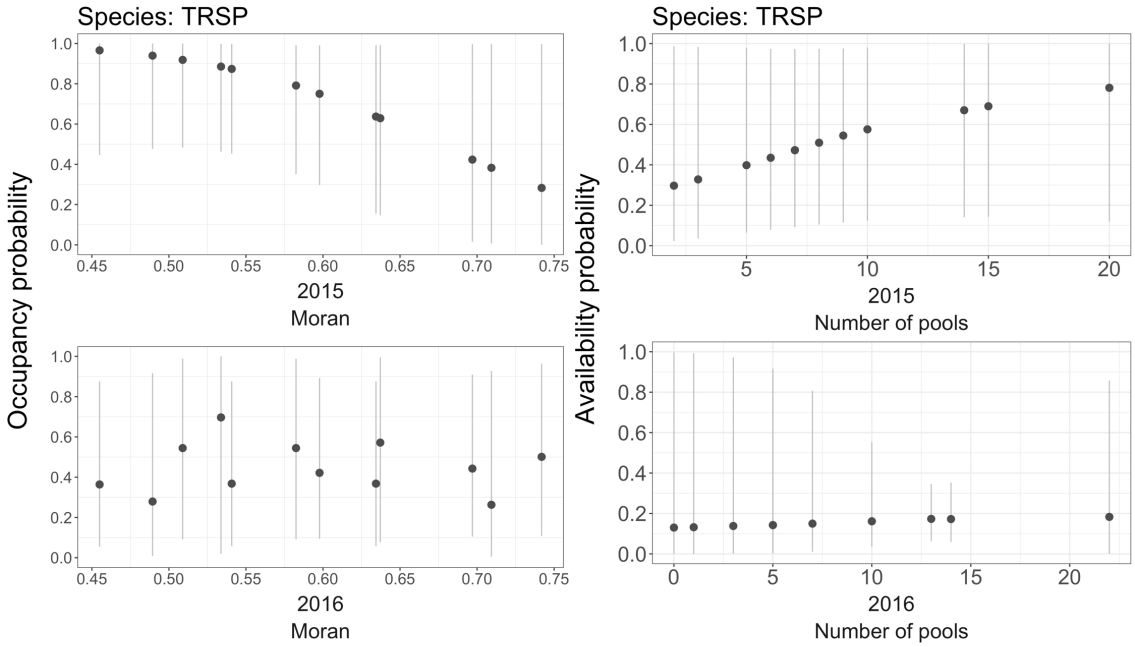


2. <u>Detection probability with replicate-level covariates</u>

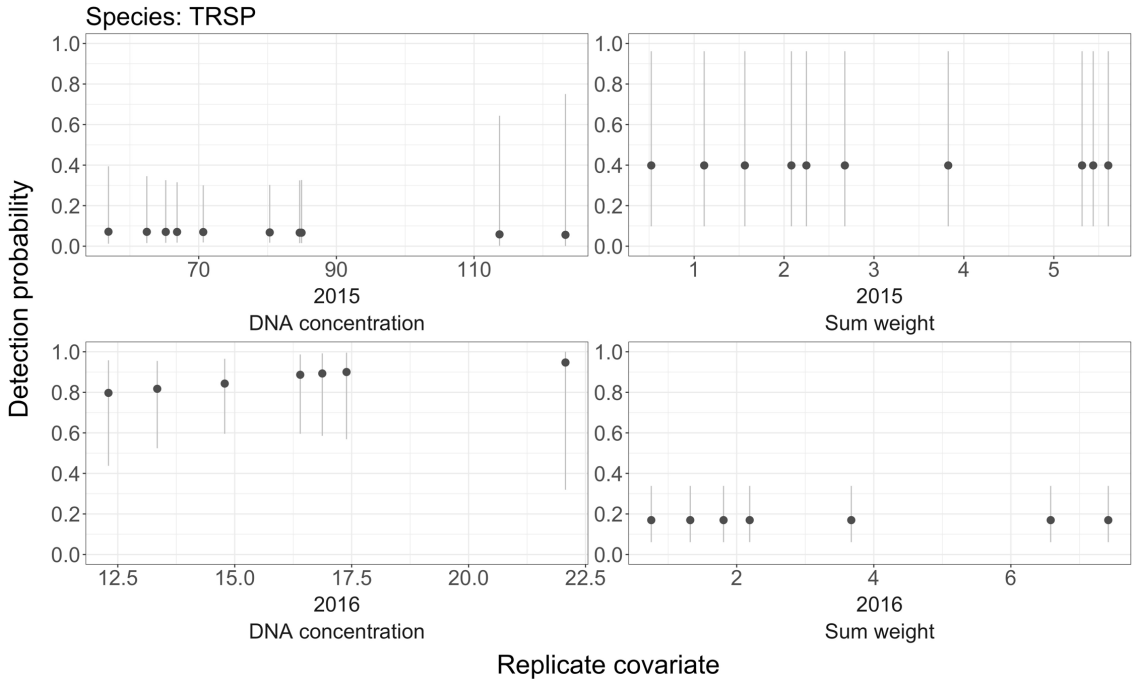


TRSP - Mousedeer (*Tragulus sp*)

1. <u>Occupancy probability with habitat covariates and availability probability with sample-level covariates</u>

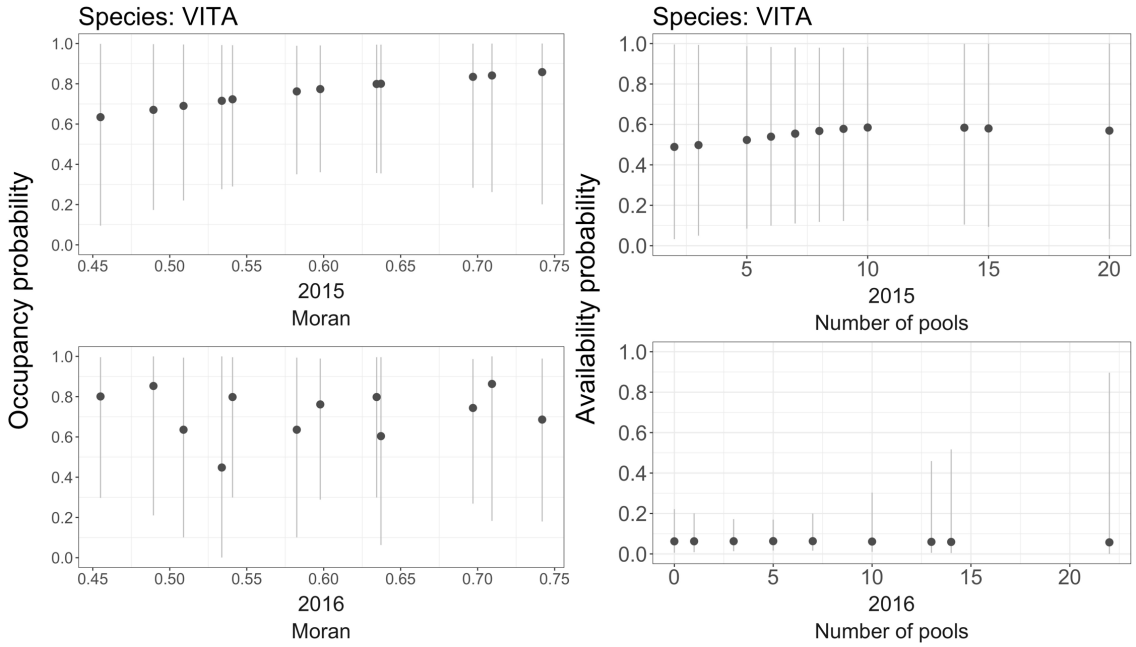


2. <u>Detection probability with replicate-level covariates</u>

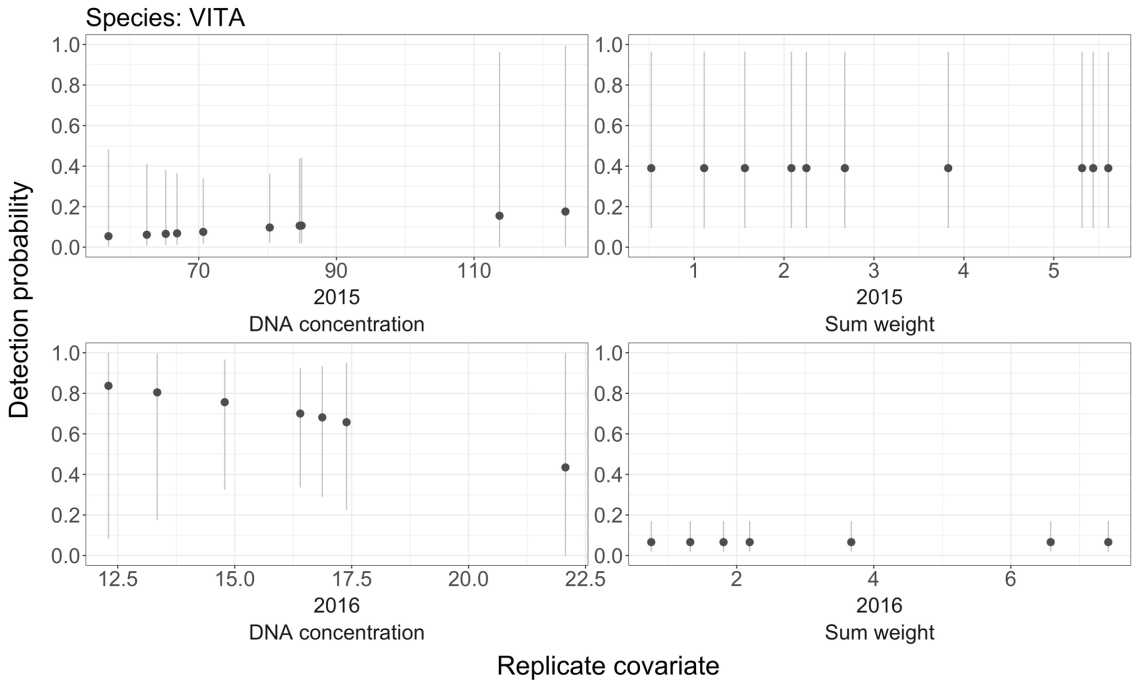


VITA - Malay civet (*Viverra tangalunga*)

1. Occupancy probability with habitat covariates and availability probability with sample-level covariates

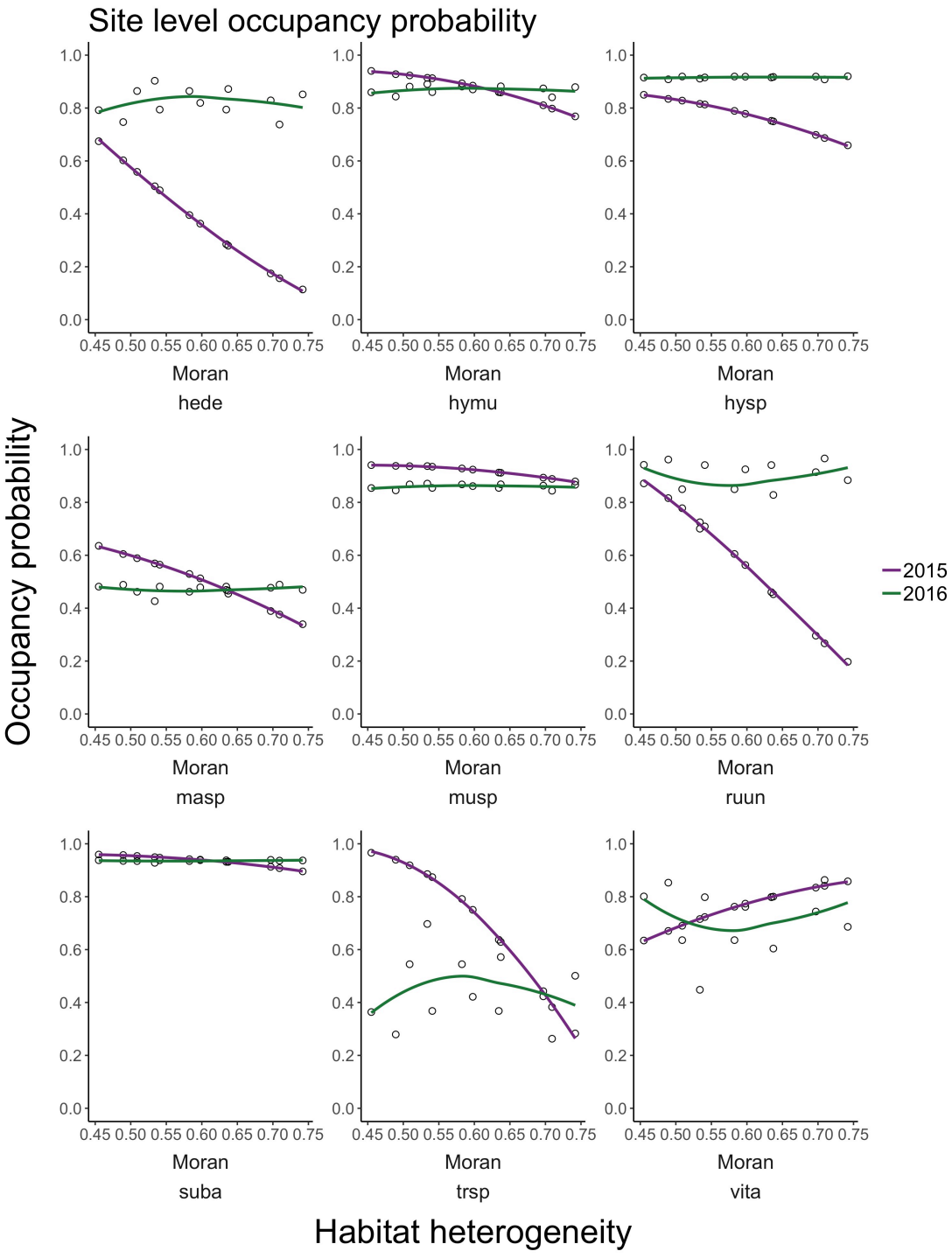


2. Detection probability with replicate-level covariates

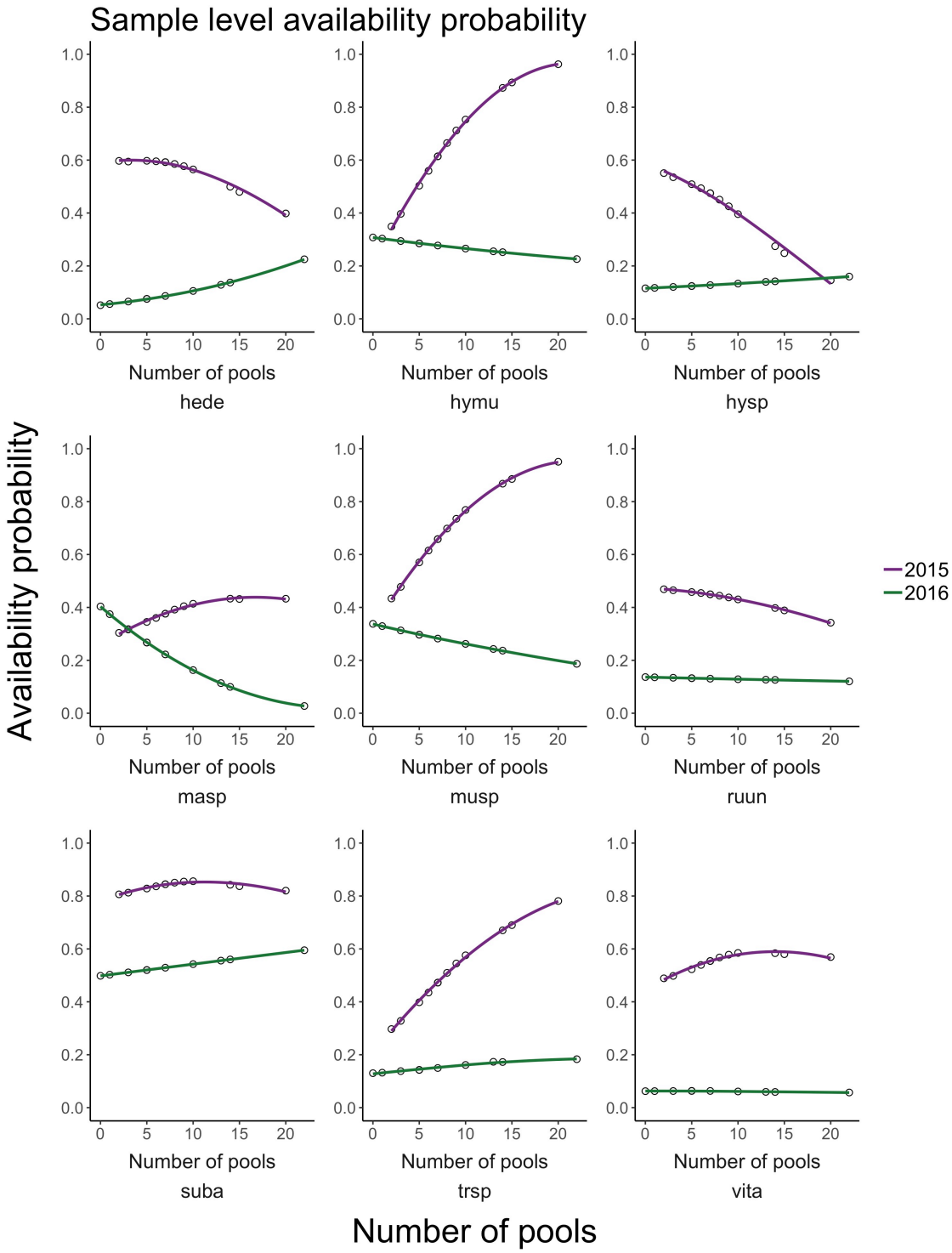


Figures S4.2. Shows the taxa specific seasonal responses to the (1) site-, (2) sample- and (3) replicate-level covariates. Each panel is a different taxon and all taxa names are given as their four-letter code, full taxonomic names can be found in Table 4.1. The purple lines represent probabilities calculated in the dry season of 2015 and the green lines represent probabilities calculated in the wet season of 2016

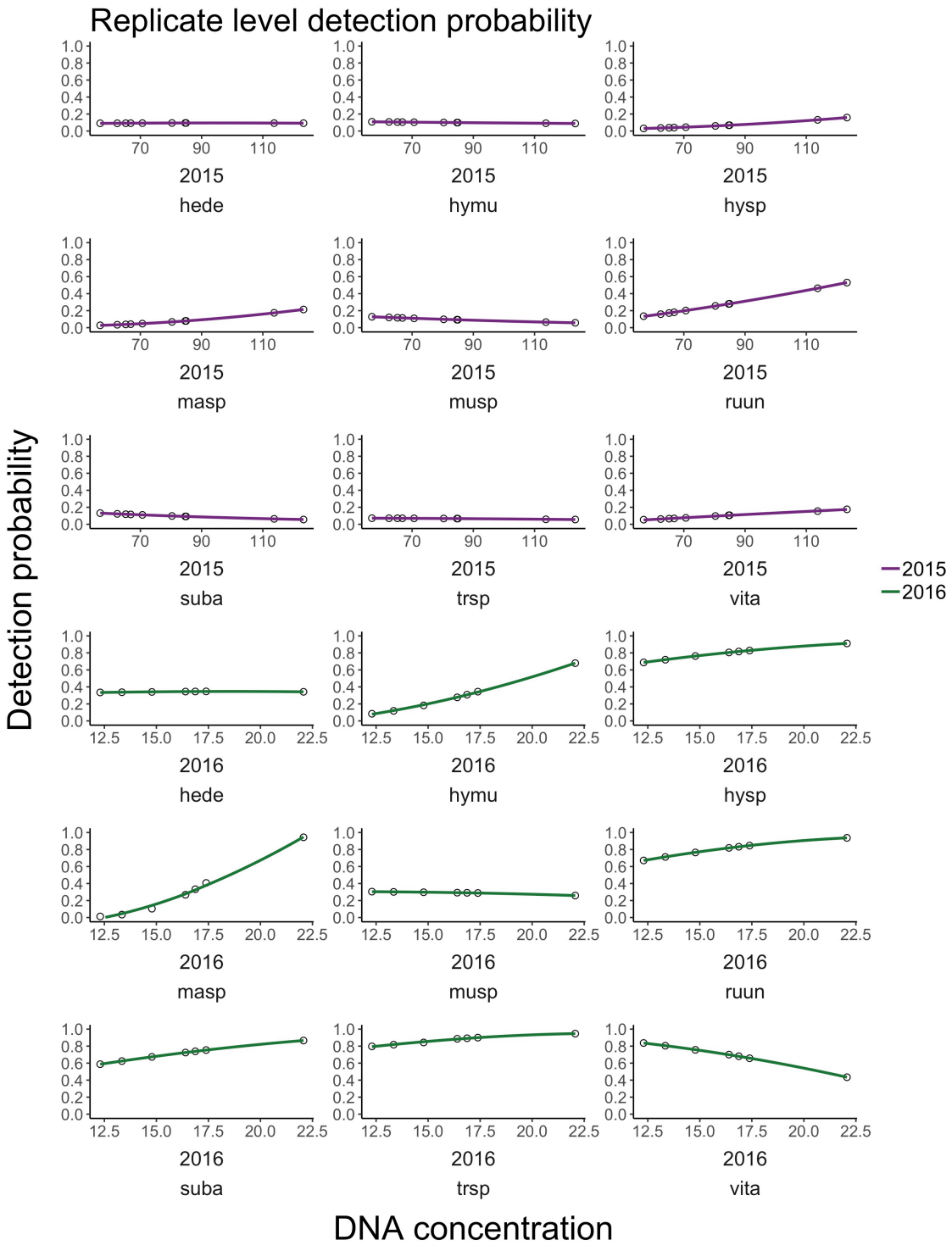
(1) Seasonal differences in site-occupancy probability at sites with different levels of habitat heterogeneity



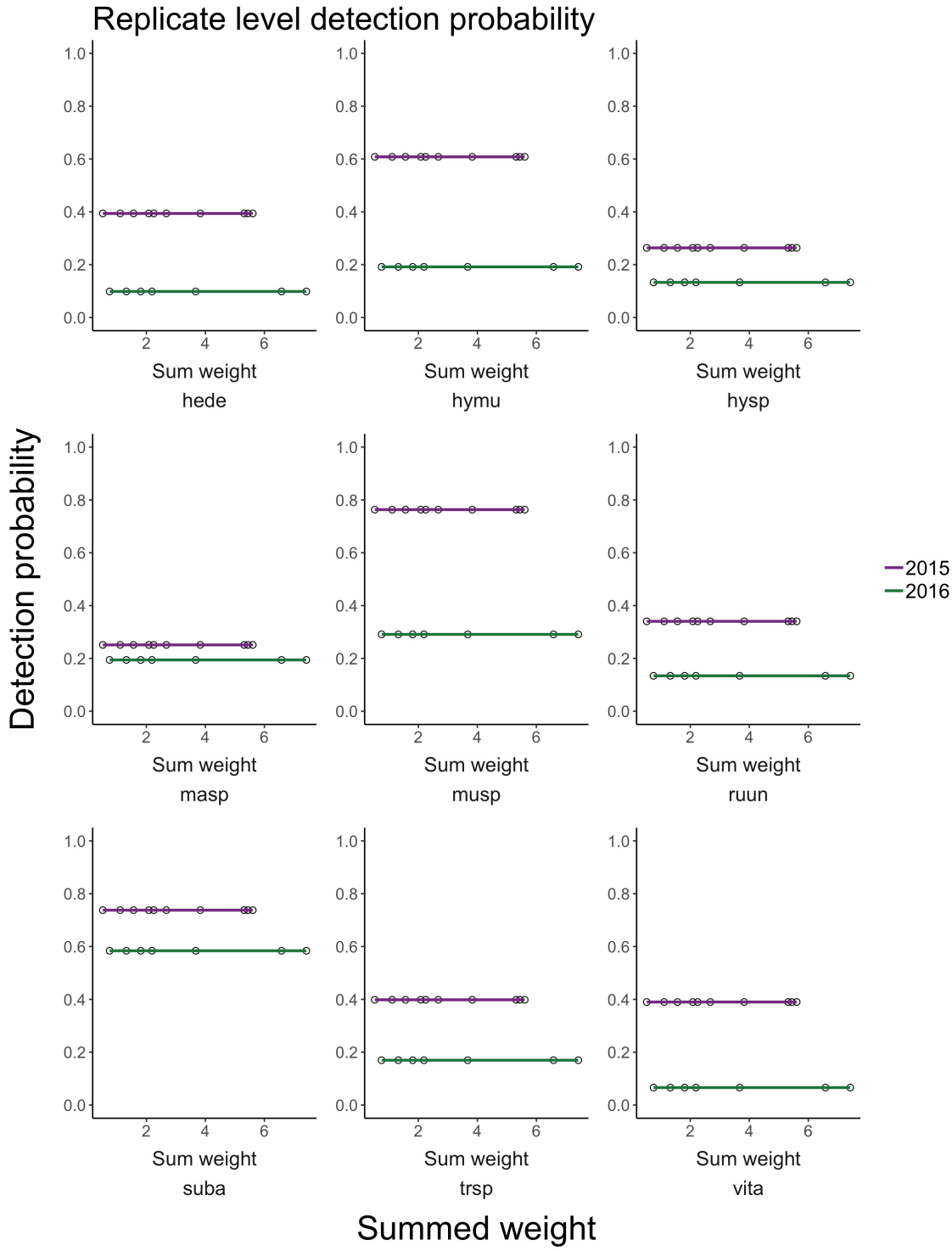
(2) Seasonal differences in sample-availability probability with different levels of sampling effort as determined by the number of pools sequenced



(3) Seasonal differences in detection probability with the concentration of the DNA extract. The seasons are shown on separate panels as the concentration was measured using two different techniques, thus not strictly comparable (Nanodrop in 2015 and Qubit in 2016).



(4) Seasonal differences in detection probability with the total relative weight of the individual leeches included in the DNA extract



Chapter 5: General discussion

5.1. Main findings - biodiversity in a modified landscape

In this thesis I have explored the use of leech-derived iDNA to detect mammalian diversity across a human-modified landscape. In light of the rapid and increasing modification taking place across the tropics, especially in Southeast Asia (Wilcove *et al.* 2013), new technologies are needed to understand the impacts that this has on biodiversity (Corlett 2016). Metabarcoding, the sequencing of short loci from mixed samples (e.g. bulk, dietary and environmental), is becoming increasingly popular for conservation and monitoring, especially in the area of aquatic monitoring for rare or invasive species (Jerde *et al.* 2011; Rees *et al.* 2014). Examples of this approach also include examining the DNA from the blood meals of invertebrates to identify their previous prey (iDNA), which has potential for use as a method to survey wildlife (Weiskopf *et al.* 2017). Yet not all invertebrates are equal as biodiversity samplers, based on differences in digestion rates, dispersal distances and host preferences (Calvignac-Spencer *et al.* 2013a). To date, two of the most frequently used invertebrate groups for this application are carrion flies (Calvignac-Spencer *et al.* 2013b) and haematophagous terrestrial leeches (Schnell *et al.* 2018). In this thesis I have isolated and studied iDNA sourced from two species of terrestrial leech and have shown that this technique can detect relatively high levels of mammalian diversity. Focusing on a tropical landscape that has been subject to intense pressure, I was also able to detect differences in mammal diversity among forest types, as well as show a strong interannual difference. I also present one of the first applications of occupancy models to iDNA data, to understand occurrence of mammal DNA while accounting for detectability.

5.1.1. Dietary differences in two blood-feeding leeches

Previous assessments of iDNA, including those based on leeches, have analysed multiple species together and have not considered the potential biases introduced by potential differences in their diets or host preference (Schnell *et al.* 2012; Weiskopf *et al.* 2017). For this chapter I analysed the differences between two congeneric and co-distributed terrestrial leeches that are abundant in Sabah: the

tiger leech (*Haemadipsa picta*) and the brown leech (*Haemadipsa sumatrana*). Very little is known on the ecology of these species, and by sequencing the blood meals of pools of both species I found that in fact there were differences in their feeding behaviours. From my results, *H. sumatrana* appears to feed on a nested subset of mammals that are fed on by *H. picta*. While the diversity of small-bodied mammals detected was low in general, the only detections of rodents were made using *H. picta*. Community composition did not differ significantly based on the two species or in different habitats (heavily vs. twice-logged logged forest). Considering diversity detected and behavioural traits, I concluded that *H. picta* appears to be a better sampler of mammals. Preliminary studies on the effect of invertebrate behaviour is crucial to understanding the biases they introduce into sampling schemes (Schnell *et al.* 2015a). Additionally, as the taxonomy and phylogenetic relationships of the haemadipsid leeches are an active area of research (Trontelj & Utevsky 2005; Borda *et al.* 2008; Borda & Siddall 2010), studies including sequencing of the leeches themselves are also greatly important (Schnell *et al.* 2018).

5.1.2. Spatial and temporal changes in mammalian diversity within a human-modified landscape

Building on the results of Chapter 2, in Chapter 3 I then used *H. picta* to perform an in-depth analysis of mammal diversity across different habitat types, from primary forest to logged forest of different degrees of degradation. I found that while finer-scale vegetation metrics did not significantly explain the levels of alpha diversity detected, broader classifications of habitat type did. Higher levels of diversity were found in logged forest habitats which is in agreement with previous research (Wearn *et al.* 2017). However, due to the almost complete absence of terrestrial leeches in extremely degraded and/or dry habitats, including oil palm plantations, these could not be surveyed for mammals using this method.

Overall, I show that the detected mammalian diversity from 2016 was greater than that detected in 2015. The underlying reason for these differences is complex and could be related to logging disturbance (Burivalova *et al.* 2014) and climatic changes, mass fruiting, and droughts associated with an ENSO event (Curran &

Leighton 2000). My findings in Chapter 3 show that logged forest habitats harbour diversity and have conservation value which need to be protected from land-conversion (Berry *et al.* 2010; Edwards *et al.* 2011). The results also demonstrate the terrestrial leech iDNA can be used to detect local-scale differences in diversity.

5.1.3. Accounting for imperfect detections

In Chapter 4, I apply an occupancy modelling framework to my data and model iDNA detections for ten taxa as a function of site-, sample- and PCR replicate-level covariates. I accounted for imperfect detections by using hierarchical models to partition the data into the availability and detectability parameters (Mordecai *et al.* 2011). The results from these models again show a strong difference between occupancy and detectability between years. In general, I found that higher DNA occupancy is associated with higher quality sites, however, this is very species-dependent. On the other hand, the majority of species showed a greater chance of DNA detection with higher DNA extract concentration. This is the first time, to my knowledge, that occupancy modelling has been applied to iDNA data, and as such there are some caveats to the assumptions of these models, concerning false positives and closed season sampling. False positives and negatives are inherent in iDNA sampling, and can arise from misidentifications and missing reference data. For the reason that failing to account for detectability of a species will result in biased estimates (MacKenzie *et al.* 2002), it is essential that if monitoring with iDNA is to be applicable and useful in conservation scenarios, then the development of these statistical frameworks is a key milestone (Schnell *et al.* 2015a).

5.2. Mammalian diversity detected with leeches

This field of biodiversity monitoring with iDNA is very much in its infancy. For example, very few iDNA studies have been published on terrestrial leeches of the family (Haemadipsidae) (Schnell *et al.* 2012, 2018; Weiskopf *et al.* 2017; Tessler *et al.* 2018). My findings from a single landscape build on the results of these earlier broader-scale studies and confirm that terrestrial leeches sample a wide-range of mammalian diversity from different habitats. I also found rare species, and species of high conservation concern such as the Sunda pangolin (*Manis javanica*), which

was one of the initially posited benefits of using iDNA (Schnell *et al.* 2012). Even though I was not able to compare my results directly with temporally and spatially matched camera traps, one of the advantages of sampling at a large collaborative project such as SAFE is that my results can later be compared to other datasets that have been, or are currently being, generated by others working at the same site. However, preliminary comparisons suggest that the maximum species richness detected with iDNA is lower than that identified using camera- and live-traps (Deere *et al.* 2017; Wearn *et al.* 2017), with my detections representing a nested subset of detections obtained by these other trapping methods. For example, using camera traps (Deere *et al.* 2017) found 24 mammalian genera across 16 families in 2015, with which I find a 41% overlap. In an earlier study (2011-2014) also conducted at the same field site, Wearn *et al.* (2017) found 31 genera of medium- to large-bodied mammals using a combination of live and camera trapping. In this thesis, I was able to detect a third of these mammals using rapid surveys of leeches, and, importantly I recorded three extra genera using iDNA (two arboreal primates and a murid rodent) that were not identified using cameras (see Deere *et al.* 2017).

Eleven of the taxa detected in my study (including potential sister species) had an IUCN red list classification putting them in a vulnerable category e.g. NT = Near Threatened, VU = Vulnerable, EN = Endangered, CR = Critically Endangered. These are mammals with vulnerable and declining populations and are of serious conservation concern. This included the Bornean gibbon (*Hylobates muelleri,* EN), Asian elephant (*Elephas maximus*, EN) and the particularly imperilled Sunda pangolin (*Manis javanica*, CR). Logging has been shown to reduce abundance of species within these vulnerable classifications (Costantini *et al.* 2016). Even so, like other studies, by detecting such species in this landscape, this supports the conservation value of logged and disturbed forests (Edwards *et al.* 2014). My results were dominated by medium- to large-bodied ungulates, in particular sambar deer (*Rusa unicolor*) and bearded pig (*Sus barbatus*). Even though these species are perceived to be common, they are both classified as Vulnerable meaning they are experiencing population declines and range reductions (IUCN 2001). Monitoring these populations, along with other large herbivores, is a crucial

part of conservation. Indeed, these species are disproportionately targeted for bushmeat, and their species-specific responses to logging are not well understood (Meijaard & Sheil 2008). Considering those species of 'Least Concern', it is still of utmost importance to include these in conservation efforts as 52% of species classified as least concern are declining globally (Schipper *et al.* 2008).

5.3. Limitations of sampling with iDNA

5.3.1. Missing mammalian groups

Despite the utility of leech-based iDNA shown in this study, there were some species and groups of mammals that I did not detect, but which are known to occur at the study site. These included detections of wild cats (Felidae), in particular. Borneo has five cat species, only one of which does not have a vulnerable classification by the IUCN (leopard cat, *Prionailurus bengalensis*, LC) (www.iucnredlist.org). All species of cat, including the extremely rare bay cat (*Pardofelis badia*, EN) have been detected before at SAFE by camera trapping (Wearn *et al.* 2013). Additionally, *H. picta* and *H. sumatrana* did not appear to feed commonly on small or arboreal mammals, or on any bats, all of which show high diversity in Borneo (Wells *et al.* 2004; Struebig *et al.* 2013; Chapman *et al.* 2018). These limitations are important to understand if iDNA approaches are to be used for targeted surveys in conservation.

5.3.2. Restrictions with iDNA sampler choice

As mentioned previously, terrestrial leeches are restricted to habitats with high humidity, particularly in damp tropical forests (Borda *et al.* 2008) and while they are often found in degraded forests (e.g. this thesis; Kendall 2012; Gąsiorek & Różycka 2017), they are not found in the exposed agricultural plantations. This reduces the ability to use leeches in non-forested or temperate environments, where blowflies (Hoffmann *et al.* 2018) or dung beetles (Kerley *et al.* 2018) may be served as better potential iDNA samplers. Finally, it is important to note that, compared to some groups, it is not possible to set unattended traps for leeches. This contrasts to groups such as dung beetles or carrion flies that can be collected in large numbers using traps baited with dung or meat, respectively. While methods of trapping may influence the species captured, they also offer the means

to replicate trapping effort, as well as remove the potential for any observer biases during leech collections.

5.3.3. Invertebrate overharvesting impacts

Finally, nothing is known of the impacts of extracting invertebrates from the environment for iDNA studies, and this is an area that needs more work. In areas which are already at risk from extended dry seasons and land-conversion, overharvesting of individuals could place extra stressors on population viability of invertebrate groups, with potentially knock-on consequences for interacting taxa such as predators. For terrestrial leeches in particular, their ecology is not well understood (Borda & Siddall 2010) yet recent studies have highlighted the cascading impact of invertebrate removal. For example, the experimental suppression of termites in Borneo led to a loss of ecosystem functioning, and reduced resilience to drought during the 2015-2016 El Nino drought (Ashton *et al.* 2019). While the ecological roles of terrestrial leeches may not be as critical as those of termites, anecdotally leeches are thought to be a prey species for some of the native birds and fishes, and thus their removal may directly impact food webs and trophic interactions. Steps to reduce sampling in iDNA studies could include preliminary testing of primers, and well-planned and executed study designs. Limits on the numbers of individuals taken should also be considered. Such steps to mitigate potential negative impacts on such populations are likely to become more important as interest in leeches as iDNA samplers of biodiversity increases. In general, more research is needed on the taxonomy and behaviour of this understudied group (e.g. Chapter 2; Weiskopf *et al.* 2017; Schnell *et al.* 2018; Tessler *et al.* 2018).

5.4. Future developments for iDNA sampling

While my study has demonstrated the utility of iDNA for monitoring, several key future developments and protocols might further improve this approach. First, iDNA studies will benefit from "ground-truthing" using camera-traps, to provide accurate comparisons between sampling methods (e.g. Lee *et al.* 2016; Weiskopf *et al.* 2017). Once we know where the discrepancies in detection lie, iDNA sampling can be deployed as a complementary method alongside large-scale or long-term

camera trapping campaigns. Using these two methods in conjunction with each other might offer benefits for sampling across seasons, when a single method is less effective on its own. Indeed, as my research has shown, surveying biodiversity with leeches is successful during rainy seasons, when camera trapping can be impacted by water damage, condensation, and flooding. Second, the ability to quantify biomass from iDNA would add enormous additional benefit for conservation monitoring. Attempting to quantify biomass from relative read abundance is currently a controversial area of research but one that is gaining traction (Deagle *et al.* 2018). Finally, the identification of individuals of host species will be a crucial development (Schubert *et al.* 2015). Individual identities would allow for estimation of abundances and the use of capture-mark-release analysis techniques (Schnell *et al.* 2015a). Being able to quantify intraspecific variation would allow for more detailed assessments of the vulnerability of a population, e.g. sex-specific space use in clouded leopards (Wearn *et al.* 2013). Aside from such methodological steps, the ongoing falling costs of sequencing will continue to increase the scope of these iDNA based screening programmes.

In practice, I have shown that using iDNA has potential as a biodiversity monitoring technique in tropical forests. Like the use of eDNA for freshwater, iDNA can bring benefits to biodiversity monitoring, such as remote detection of species that are difficult to spot, and the ability to achieve greater geographic coverage with reduced costs in the field. In the face of rapid habitat loss and climate change, we need to be innovative and use all new technologies we have access to, to gain the most accurate biodiversity information. For conservation monitoring, using leeches (or another invertebrate sampler) is a realistic approach for gaining rapid information. However, to advance iDNA sampling further, a greater appreciation and understanding for the ecology of terrestrial leeches is needed, including host preferences and dispersal distances. Understanding this would allow biodiversity surveys to be tailored to the research question or target species. Perhaps the most important deciding factor in the future uptake of this method as a standard component in the "ecologist's toolbox" is the continuing reduction in sequencing costs and improvements in reference databases. Regardless of the developments that need to be made, my findings show that the vulnerable and degraded forests

of Sabah have retained mammalian biodiversity that appears resilient to serious disturbance events (e.g. ENSO and logging). Therefore, it is in these forests, within human-modified landscapes that must remain at the forefront of conservation action, if we are going to slow the biodiversity crisis.

References

Ahumada, J. a, Silva, C.E.F., Gajapersad, K., Hallam, C., Hurtado, J., Martin, E., *et al.* (2011). Community structure and diversity of tropical forest mammals: data from a global camera trap network. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 366, 2703–11.

Aizpurua, O., Budinski, I., Georgiakakis, P., Gopalakrishnan, S., Ibañez, C., Mata, V., *et al.* (2018). Agriculture shapes the trophic niche of a bat preying on multiple pest arthropods across Europe: Evidence from DNA metabarcoding. *Mol. Ecol.*, 27, 815–825.

Alberdi, A., Aizpurua, O., Gilbert, M.T.P. & Bohmann, K. (2017). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.*, 9, 134–147.

Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K.H., *et al.* (2012). Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.*, 21, 1966–79.

Anderson, M.J. & Walsh, D.C.I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecol. Monogr.*, 83, 557–574.

Arrizabalaga-Escudero, A., Clare, E.L., Salsamendi, E., Alberdi, A., Garin, I., Aihartza, J., *et al.* (2018). Assessing niche partitioning of co-occurring sibling bat species by DNA metabarcoding. *Mol. Ecol.*, 27, 1273–1283.

Ashton, L.A., Griffiths, H.M., Parr, C.L., Evans, T.A., Didham, R.K., Hasan, F., *et al.* (2019). Termites mitigate the effects of drought in tropical rainforest, *Science*, *363*(6423), 174-177.

Axtner, J., Crampton-Platt, A., Hoerig, L.A., Xu, C.C.Y., Yu, D.W. & Wilting, A. (2018). An efficient and improved laboratory workflow and tetrapod database for larger scale eDNA studies. *bioRxiv*.

Bailey, L.L., Hines, J.E., Nichols, J.D. & MacKenzie, D.I. (2007). Sampling design trade-offs in occupancy studies with imperfect detection: Examples and software. *Ecol. Appl.*, 17, 281–290.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., *et al.* (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Mol. Ecol. Resour.*, 14, 306–323.

Barber, C.P., Cochrane, M.A., Souza, C.M. & Laurance, W.F. (2014). Roads, deforestation, and the mitigating effect of protected areas in the Amazon. *Biol. Conserv.*, 177, 203–209.

Barlow, J., França, F., Gardner, T.A., Hicks, C.C., Lennox, G.D., Berenguer, E., *et al.* (2018). The future of hyperdiverse tropical ecosystems. *Nature*, 559, 517–526.

Barnes, M.A. & Turner, C.R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conserv. Genet.*, 17, 1–17.

Basset, Y., Cizek, L., Cuenoud, P., Didham, R.K., Guilhaumon, F., Missa, O., *et al.* (2012). Arthropod Diversity in a Tropical Forest. *Science,* 338(6113), 1481-1484.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67, 1–48.

Beastall, C., Shepherd, C.R., Hadiprakarsa, Y. & Martyr, D. (2016). Trade in the Helmeted Hornbill Rhinoplax vigil: The "ivory hornbill." *Bird Conserv. Int.*, 26, 137–146.

Bell, K.L., Fowler, J., Burgess, K.S., Dobbs, E.K., Gruenewald, D., Lawley, B., *et al.* (2017). Applying Pollen DNA Metabarcoding to the Study of Plant–Pollinator Interactions. *Appl. Plant Sci.*, 5, 1–10.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., *et al.* (2013). GenBank. *Nucleic Acids Res.*, 41, 36–42.

Berry, N.J., Phillips, O.L., Lewis, S.L., Hill, J.K., Edwards, D.P., Tawatao, N.B., *et al.* (2010). The high value of logged tropical forests: lessons from northern Borneo. *Biodivers. Conserv.*, 19, 985–997.

Bicknell, J.E., Struebig, M.J., Edwards, D.P. & Davies, Z.G. (2014). Improved timber harvest techniques maintain biodiversity in tropical forests. *Curr. Biol.*, 24, R1119–R1120.

Bierregaard, R.O., Lovejoy, T.E., Kapos, V., dos Santos, A.A. & Hutchings, R.W. (1992). The biological dynamics of tropical rainforest fragments: a prospective comparison of fragments and continuous forest. *Bioscience*, 42, 859–866.

Bigg, J., Ewald, N., Valentini, A., Gaboriaud, C., Griffths, R., Wilkinson, J., *et al.* (2014). *Analytical and methodological development for improved surveillance of the Great Crested Newt Final Report*. Oxford.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R., *et al.* (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, 2, 1–9.

Blaxter, M. (2016). Imagining sisyphus happy: DNA barcoding and the unnamed majority. *Philos. Trans. R. Soc. B Biol. Sci.*, 371, 20150329.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., *et al.* (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.*, 29, 358–67.

Borda, E., Oceguera-Figueroa, A. & Siddall, M.E. (2008). On the classification, evolution and biogeography of terrestrial haemadipsoid leeches (Hirudinida:

Arhynchobdellida: Hirudiniformes). *Mol. Phylogenet. Evol.*, 46, 142–54.

Borda, E. & Siddall, M.E. (2004). Arhynchobdellida (Annelida: Oligochaeta: Hirudinida): phylogenetic relationships and evolution. *Mol. Phylogenet. Evol.*, 30, 213–225.

Borda, E. & Siddall, M.E. (2010). Insights into the evolutionary history of Indo-Pacific bloodfeeding terrestrial leeches (Hirudinida: Arhynchobdellida: Haemadipisdae). *Invertebr. Syst.*, 24, 456–472.

Bovo, A.A.A., Ferraz, K.M.P.M.B., Magioli, M., Alexandrino, E.R., Hasui, É., Ribeiro, M.C., *et al.* (2018). Habitat fragmentation narrows the distribution of avian functional traits associated with seed dispersal in tropical forest. *Perspect. Ecol. Conserv.*, 16, 90–96.

Brodie, J., Post, E. & Laurance, W.F. (2012). Climate change and tropical biodiversity: A new focus. *Trends Ecol. Evol.*, 27, 145–150.

Brodie, J.F., Giordano, A.J. & Ambu, L. (2014a). Differential responses of large mammals to logging and edge effects. *Mamm. Biol.*, 80, 7–13.

Brodie, J.F., Giordano, A.J., Zipkin, E.F., Bernard, H., Mohd-Azlan, J. & Ambu, L. (2014b). Correlation and persistence of hunting and logging impacts on tropical rainforest mammals. *Conserv. Biol.*, 29, 110–121.

Brodie, J.F., Helmy, O.E., Brockelman, W.Y. & Maron, J.L. (2009). Bushmeat Poaching Reduces the Seed Dispersal and Population Growth Rate of a Mammal-Dispersed Tree. *Ecol. Appl.*, 19, 854–863.

Bryan, J.E., Shearman, P.L., Asner, G.P., Knapp, D.E., Aoro, G. & Lokes, B. (2013). Extreme Differences in Forest Degradation in Borneo: Comparing Practices in Sarawak, Sabah, and Brunei. *PLoS One*, 8, e69679.

Burivalova, Z., Şekercioğlu, Ç.H. & Koh, L.P. (2014). Thresholds of logging intensity to maintain tropical forest biodiversity. *Curr. Biol.*, 24, 1893–1898.

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed)*. *Ecol. Modell.*

Calvignac-Spencer, S., Leendertz, F.H., Gilbert, M.T.P. & Schubert, G. (2013a). An invertebrate stomach's view on vertebrate ecology. *BioEssays*, 35, 1004–13.

Calvignac-Spencer, S., Merkel, K., Kutzner, N., Kühl, H., Boesch, C., Kappeler, P.M., *et al.* (2013b). Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. *Mol. Ecol.*, 22, 915–24.

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., *et al.* (2017). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.*, 9, 410–419.

Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability Author. *Biometrics*, 43, 783–791.

Chao, A., Chazdon, R., Colwell, R.K. & Chen, T.-J. (2005). A new statistical approach for assessing similarity of species compostition with incidence and abundance data. *Ecol. Lett.*, 40, 1705–1708.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., *et al.* (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.*, 84, 45–67.

Chapman, P.M., Wearn, O.R., Riutta, T., Carbone, C., Rowcliffe, J.M., Bernard, H., *et al.* (2018). Inter-annual dynamics and persistence of small mammal communities in a selectively logged tropical forest in Borneo. *Biodivers. Conserv.*, 27, 3155–3169.

Chen, C.C., Lin, H.W., Yu, J.Y. & Lo, M.H. (2016). The 2015 Borneo fires: What have we learned from the 1997 and 2006 El Niños? *Environ. Res. Lett.*, 11, 104003.

Clare, E.L. (2014). Molecular detection of trophic interactions: Emerging trends, distinct advantages, significant considerations and conservation applications. *Evol. Appl.*, 7, 1144–1157.

Clare, E.L., Chain, F.J.J., Littlefair, J.E. & Cristescu, M.E. (2016). The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome*, 59, 981–990.

Comtet, T., Sandionigi, A., Viard, F. & Casiraghi, M. (2015). DNA (meta)barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens. *Biol. Invasions*, 17, 905–922.

Corlett, R.T. (2007). What's so special about Asian tropical forests? *Curr. Sci.*, 93, 1551–1557.

Corlett, R.T. (2016). A Bigger Toolbox: Biotechnology in Biodiversity Conservation. *Trends Biotechnol.*, 35, 1–11.

Costantini, D., Edwards, D.P. & Simons, M.J.P. (2016). Life after logging in tropical forests of Borneo: A meta-analysis. *Biol. Conserv.*, 196, 182–188.

Curran, L.M. & Leighton, M. (2000). Vertebrate Responses to Spatiotemporal Variation in Seed Production of mast-fruiting Dipterocarpaceae. *Ecol. Monogr.*, 70, 101–128.

Deagle, B.E., Eveson, J.P. & Jarman, S.N. (2006). Quantification of damage in DNA recovered from highly degraded samples - A case study on DNA in faeces. *Front. Zool.*, 3, 1–10.

Deagle, B.E., Thomas, A.C., McInnes, J.C., Clarke, L.J., Vesterinen, E.J., Clare, E.L., *et al.* (2018). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Mol. Ecol.*, 1–16.

Deere, N.J., Guillera - Arroita, G., Baking, E.L., Bernard, H., Pfeifer, M., Reynolds, G., *et al.* (2017). High Carbon Stock forests provide co-benefits for tropical biodiversity. *J Appl Ecol*, 55, 1–12.

Deiner, K. & Altermatt, F. (2014). Transport distance of invertebrate environmental DNA in a natural river. *PLoS One*, 9, e88786.

Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., *et al.* (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.*, 26, 5872–5895.

Deiner, K., Fronhofer, E.A., Mächler, E., Walser, J.-C. & Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nat. Commun.*, 7, 1–9.

Deiner, K., Walser, J.C., Mächler, E. & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biol. Conserv.*, 183, 53–63.

Dorazio, R.M. & Erickson, R.A. (2017a). eDNAoccupancy: An R Package for Multi-scale Occupancy Modeling of Environmental DNA Data. Appendix: Markov chain Monte Carlo algorithm. *Mol. Ecol. Resour.*, 18.

Dorazio, R.M. & Erickson, R.A. (2017b). EDNAOCCUPANCY: An Rpackage for multiscale occupancy modelling of environmental DNA data. *Mol. Ecol. Resour.*, 18, 368–380.

Edwards, D.P., Fisher, B. & Wilcove, D.S. (2012a). High conservation value or high confusion value? Sustainable agriculture and biodiversity conservation in the tropics. *Conserv. Lett.*, 5, 20–27.

Edwards, D.P., Hodgson, J. a., Hamer, K.C., Mitchell, S.L., Ahmad, A.H., Cornell, S.J., *et al.* (2010). Wildlife-friendly oil palm plantations fail to protect biodiversity effectively. *Conserv. Lett.*, 3, 236–242.

Edwards, D.P., Larsen, T.H., Docherty, T.D.S., Ansell, F. a, Hsu, W.W., Derhé, M. a, *et al.* (2011). Degraded lands worth protecting: the biological importance of Southeast Asia's repeatedly logged forests. *Proc. Biol. Sci.*, 278, 82–90.

Edwards, D.P., Tobias, J.A., Sheil, D., Meijaard, E. & Laurance, W.F. (2014). Maintaining ecosystem function and services in logged tropical forests. *Trends Ecol. Evol.*, 29, 511–520.

Edwards, D.P., Woodcock, P., Edwards, F. a., Larsen, T.H., Hsu, W.W., Benedick, S., *et al.* (2012b). Reduced-impact logging and biodiversity conservation: A case study from Borneo. *Ecol. Appl.*, 22, 561–571.

Edwards, D.P., Woodcock, P., Newton, R.J., Edwards, F. a., Andrews, D.J.R., Docherty, T.D.S., *et al.* (2013). Trophic flexibility and the persistence of understory birds in intensively logged rainforest. *Conserv. Biol.*, 27, 1079–1086.

Elbrecht, V. & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLoS One*, 10, e0130324.

Elbrecht, V., Peinert, B. & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecol. Evol.*, 7, 6918–6926.

Ewers, R.M., Boyle, M.J.W., Gleave, R.A., Plowman, N.S., Benedick, S., Bernard, H., *et al.* (2015). Logging cuts the functional importance of invertebrates in tropical rainforest. *Nat. Commun.*, 6, 6836.

Ewers, R.M., Didham, R.K., Fahrig, L., Ferraz, G., Hector, A., Holt, R.D., *et al.* (2011). A large-scale forest fragmentation experiment: the Stability of Altered Forest Ecosystems Project. *Philos. Trans. R. Soc. B Biol. Sci.*, 366, 3292–3302.

Ficetola, G.F., Miaud, C., Pompanon, F. & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biol. Lett.*, 4, 423–425.

Ficetola, G.F., Taberlet, P. & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Mol. Ecol. Resour.*, 16, 604–607.

Fleming, T.H., Breitwisch, R. & Whitesides, G.H. (1987). Patterns of tropical vertebrate frugivore diveristy. *Annu. Rev. Ecol. Syst.*, 18, 91–109.

Fletcher, R.J., Didham, R.K., Banks-Leite, C., Barlow, J., Ewers, R.M., Rosindell, J., *et al.* (2018). Is habitat fragmentation good for biodiversity? *Biol. Conserv.*, 226, 9–15.

Floyd, R., Abebe, E., Papert, A. & Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Mol. Ecol.*, 11, 839–850.

Fogden, S.C.L. & Proctor, J. (1985). Notes on the feeding of land leeches (Haemadipsa zeylanica Moore and H . picta Moore) in Gunung Mulu National Park, Sarawak. *Biotropica*, 17, 172–174.

Food and Agriculture Organisation UN. (2017). *FAO - oil palm production ranking by country*. Available at: http://www.fao.org/faostat/en/#rankings/countries_by_commodity. Last accessed 20 October 2018.

Food and Agriculture Organization. (2011). Assessing forest degradation: Towards the development of globally applicable guidlines. *Food Agric. Organ.*, 99.

Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A.K., Pietroni, C., *et al.* (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.*, 8, 1188.

Gardner, T.A., Barlow, J., Chazdon, R., Ewers, R.M., Harvey, C.A., Peres, C.A., *et al.* (2009). Prospects for tropical forest biodiversity in a human-modified world. *Ecol. Lett.*, 12, 561–582.

Gąsiorek, P. & Różycka, H. (2017). Feeding strategies and competition between terrestrial Haemadipsa leeches (Euhirudinea:Arhynchobdellida) in Danum Valley rainforest (Borneo,Sabah). *Folia Parasitol. (Praha).*, 64.

Gaveau, D.L. a., Sloan, S., Molidena, E., Yaen, H., Sheil, D., Abram, N.K., *et al.* (2014).

Four Decades of Forest Persistence, Clearance and Logging on Borneo. *PLoS One*, 9, e101654.

Gaveau, D.L.A., Sheil, D., Husnayaen, Salim, M.A., Arjasakusuma, S., Ancrenaz, M., *et al.* (2016). Rapid conversions and avoided deforestation: Examining four decades of industrial plantation expansion in Borneo. *Sci. Rep.*, 6, 1–13.

Gelfand, A.E. & Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85, 1–11.

Giam, X., Hadiaty, R.K., Tan, H.H., Parenti, L.R., Wowor, D., Sauri, S., *et al.* (2015). Mitigating the impact of oil-palm monoculture on freshwater fishes in Southeast Asia. *Conserv. Biol.*, 29, 1357–1367.

Gibson, L., Lee, T.M., Koh, L.P., Brook, B.W., Gardner, T.A., Barlow, J., *et al.* (2011). Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature*, 505, 710–710.

Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., *et al.* (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.*, 7, 1299–1307.

Gómez, A. & Kolokotronis, S.-O. (2016). Genetic identification of mammalian meal source in dung beetle gut contents. *Mitochondrial DNA Part A*, 28, 612–615.

Gotelli, N.J. & Colwell, R.K. (2011). Estimating species richness. In: *Biological Diversity* (eds. Magurran, A.A. & McGill, B.J.). pp. 39–54.

Govedich, F.R., Moser, W.E. & Davies, R.W. (2004). Annelida : Clitellata , Hirudinea , Euhirudinea. In: *Freshwater Invertebrates of the Malaysian Region* (eds. Yule, C.M. & Yong, H.S. (Eds)). Academy of Science Malaysia, pp. 175–190.

Granados, A., Crowther, K., Brodie, J.F. & Bernard, H. (2016). Persistence of mammals in a selectively logged forest in Malaysian Borneo. *Mamm. Biol.*, 81, 268–273.

Gray, C.L., Hill, S.L.L., Newbold, T., Hudson, L.N., Boïrger, L., Contu, S., *et al.* (2016). Local biodiversity is higher inside than outside terrestrial protected areas worldwide. *Nat. Commun.*, 7, 12306.

Gray, C.L., Slade, E.M., Mann, D.J. & Lewis, O.T. (2014). Do riparian reserves support dung beetle biodiversity and ecosystem services in oil palm-dominated tropical landscapes? *Ecol. Evol.*, 4, 1049–1060.

Grenyer, R., Orme, C.D.L., Jackson, S.F., Thomas, G.H., Davies, R.G., Davies, T.J., *et al.* (2006). Global distribution and conservation of rare and threatened vertebrates. *Nature*, 444, 93–96.

Guillera-Arroita, G., Ridout, M.S. & Morgan, B.J.T. (2010). Design of occupancy studies with imperfect detection. *Methods Ecol. Evol.*, 1, 131–139.

Haddad, N.M., Brudvig, L.A., Clobert, J., Davies, K.F., Gonzalez, A., Holt, R.D., *et al.*

(2015). Habitat fragmentation and its lasting impact on Earth ' s ecosystems. *Appl. Ecol.*, 1–9.

Hansen, M.C., Stehman, S. V. & Potapov, P. V. (2010). Quantification of global gross forest cover loss. *Proc. Natl. Acad. Sci.*, 107, 8650–8655.

Hansen, M.C.C., Potapov, P. V, Moore, R., Hancher, M., Turubanova, S.A. a, Tyukavina, A., *et al.* (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science,* 342(6160), 850-853.

Hardwick, S.R., Toumi, R., Pfeifer, M., Turner, E.C., Nilus, R. & Ewers, R.M. (2015). The relationship between leaf area index and microclimate in tropical forest and oil palm plantation: Forest disturbance drives changes in microclimate. *Agric. For. Meteorol.*, 201, 187–195.

Harvey, C.A., Dickson, B. & Kormos, C. (2010). Opportunities for achieving biodiversity conservation through REDD. *Conserv. Lett.*, 3, 53–61.

Havmøller, R.G., Payne, J., Ramono, W., Ellis, S., Yoganand, K., Long, B., *et al.* (2016). Will current conservation responses save the Critically Endangered Sumatran rhinoceros Dicerorhinus sumatrensis? *Oryx*, 50, 355–359.

Heinrich, S., Wittmann, T.A., Prowse, T.A.A., Ross, J. V., Delean, S., Shepherd, C.R., *et al.* (2016). Where did all the pangolins go? International CITES trade in pangolin species. *Glob. Ecol. Conserv.*, 8, 241–253.

Hill, M.O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54, 427–432.

Hillis, D.M. & Dixon, M.T. (1991). Ribosomal DNA: Molecular Evolution and Phylogenetic Inference. *Quarterky Rev. Biol.*, 66, 411–453.

Hoffmann, C., Merkel, K., Sachse, A., Rodríguez, P., Leendertz, F.H. & Calvignac-Spencer, S. (2018). Blow flies as urban wildlife sensors. *Mol. Ecol. Resour.*, 18, 502–510.

Hsieh, T.C., Ma, K.H. & Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.*, 7, 1451–1456.

Hunter, M.E., Oyler-McCance, S.J., Dorazio, R.M., Fike, J.A., Smith, B.J., Hunter, C.T., *et al.* (2015). Environmental DNA (eDNA) sampling improves occurrence and detection estimates of invasive Burmese pythons. *PLoS One*, 10, 1–17.

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007). MEGAN analysis of metagenomic data. *Genome Res.*, 17, 377–386.

IUCN. (2001). *2001 IUCN Red List categories and criteria version 3.1*. *http://www.iucnredlist.org/technical-documents/categories-and-criteria/2001-categories-criteria*.

IUCN, International, C., University, A.S., University, T.A., Rome, U. of, Virginia, U. of,

et al. (2008). *An Analysis of Mammals on the 2008 IUCN Red List*.

Jennings, A.P., Naim, M., Advento, A.D., Aryawan, A.A.K., Ps, S., Caliman, J.P., *et al.* (2015). Diversity and occupancy of small carnivores within oil palm plantations in central Sumatra, Indonesia. *Mammal Res.*, 60, 181–188.

Jerde, C.L., Chadderton, W.L., Mahon, A.R., Renshaw, M.A., Corush, J., Budny, M.L., *et al.* (2013). Detection of Asian carp DNA as part of a Great Lakes basin-wide surveillance program. *Can. J. Fish. Aquat. Sci.*, 70, 522–526.

Jerde, C.L., Mahon, A.R., Chadderton, W.L. & Lodge, D.M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conserv. Lett.*, 4, 150–157.

Jucker, T., Hardwick, S.R., Both, S., Elias, D.M.O., Ewers, R.M., Milodowski, D.T., *et al.* (2018). Canopy structure and topography jointly constrain the microclimate of human-modified tropical landscapes. *Glob. Chang. Biol.*, 24(11), 5243-5258.

Kampmann, M.L., Schnell, I.B., Jensen, R.H., Axtner, J., Sander, A.F., Hansen, A.J., *et al.* (2017). Leeches as a source of mammalian viral DNA and RNA—a study in medicinal leeches. *Eur. J. Wildl. Res.*, 63(2), 36.

Kendall, A. (2012). The effect of rainforest modification on two species of South-East Asian terrestrial leeches, Haemadipsa zeylanica and Haemadipsa picta. (MSc Thesis). Imperial College London.

Kent, R.J. (2009). Molecular methods for arthropod bloodmeal identification and applications to ecological and vector-borne disease studies. *Mol. Ecol. Resour.*, 9, 4–18.

Kerley, G.I.H., Landman, M., Ficetola, G.F., Boyer, F., Bonin, A., Rioux, D., *et al.* (2018). Diet shifts by adult flightless dung beetles Circellium bacchus, revealed using DNA metabarcoding, reflect complex life histories. *Oecologia*, 188, 107–115.

Kéry, M. & Royle, J.A. (2016). *Applied hierarchical modeling in ecology : analysis of distribution, abundance and species richness in R and BUGS. Volume 1, Prelude and static model*. *Appl. Hierarchical Model. Ecol.*

Kocher, A., Gantier, J.C., Gaborit, P., Zinger, L., Holota, H., Valiere, S., *et al.* (2017a). Vector soup: high-throughput identification of Neotropical phlebotomine sand flies using metabarcoding. *Mol. Ecol. Resour.*, 17, 172–182.

Kocher, A., de Thoisy, B., Catzeflis, F., Huguin, M., Valière, S., Zinger, L., *et al.* (2017b). Evaluation of short mitochondrial metabarcodes for the identification of Amazonian mammals. *Methods Ecol. Evol.*, 8, 1276–1283.

Kocher, A., de Thoisy, B., Catzeflis, F., Valière, S., Bañuls, A.L. & Murienne, J. (2017c). iDNA screening: Disease vectors as vertebrate samplers. *Mol. Ecol.*, 26, 6478–6486.

Koh, L.P. & Wilcove, D.S. (2008). Is oil palm agriculture really destroying tropical

biodiversity? *Conserv. Lett.*, 1, 60–64.

Konnai, S., Mekata, H., Odbileg, R., Simuunza, M., Chembensof, M., Witola, W.H., *et al.* (2008). Detection of *Trypanosoma brucei* in Field-Captured Tsetse Flies and Identification of Host Species Fed on by the Infected Flies. *Vector-Borne Zoonotic Dis.*, 8, 565–574.

Konopik, O., Steffan-Dewenter, I. & Grafe, T.U. (2015). Effects of Logging and Oil Palm Expansion on Stream Frog Communities on Borneo, Southeast Asia. *Biotropica*, 47, 636–643.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Mol. Ecol. Resour.*, 16, 673–685.

Lai, Y. Te, Nakano, T. & Chen, J.H. (2011). Three species of land leeches from Taiwan, Haemadipsa rjukjuana comb. n., a new record for Haemadipsa picta Moore, and an updated description of Tritetrabdella taiwana (Oka). *Zookeys*, 139, 1–22.

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30, 3276–3278.

Laurance, W.F., Goosem, M. & Laurance, S.G.W. (2009). Impacts of roads and linear clearings on tropical forests. *Trends Ecol. Evol.*, 24(12), 659–669.

Laurance, W.F., Lovejoy, T.E., Vasconcelos, H.L., Bruna, E.M., Didham, R.K., Stouffer, P.C., *et al.* (2002). Ecosystem decay of Amazonian forest fragments : a 22-years investigation. *Conserv. Biol.*, 16, 605–618.

Lawton, J.H., Bignell, D.E., Bolton, B., Bloemers, G.F., Eggleton, P., Hammond, P.M., *et al.* (1998). Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nat.*, 391, 72–76.

Lee, P.-S., Gan, H.M., Clements, G.R., Wilson, J.-J. & Adamowicz, S. (2016). Field calibration of blowfly-derived DNA against traditional methods for assessing mammal diversity in tropical forests. *Genome*, 59, 1008–1022.

Lee, P.S., Sing, K.W. & Wilson, J.J. (2015). Reading mammal diversity from flies: The persistence period of amplifiable mammal mtDNA in blowfly guts (Chrysomya megacephala) and a new DNA mini-barcode target. *PLoS One*, 10, e0123871.

Lewis, S.L., Edwards, D.P. & Galbraith, D. (2015). Increasing human dominance of Tropical Forests. *Science,* 349(6250), 827-832.

Littlefair, J.E., Zander, A., de Sena Costa, C. & Clare, E.L. (2018). DNA metabarcoding reveals changes in the contents of carnivorous plants along an elevation gradient. *Mol. Ecol.*, 0–2.

Logue, K., Keven, J.B., Cannon, M. V., Reimer, L., Siba, P., Walker, E.D., *et al.* (2016). Unbiased Characterization of Anopheles Mosquito Blood Meals by Targeted High-Throughput Sequencing. *PLoS Negl. Trop. Dis.*, 10, 1–18.

Luke, S.H., Barclay, H., Bidin, K., Chey, V.K., Ewers, R.M., Foster, W.A., *et al.* (2017). The effects of catchment and riparian forest quality on stream environmental conditions across a tropical rainforest and oil palm landscape in Malaysian Borneo. *Ecohydrology*, 10, 1–14.

Luke, S.H., Slade, E.M., Gray, C.L., Annammala, K. V., Drewer, J., Williamson, J., *et al.* (2018). Riparian buffers in tropical agriculture: scientific support, effectiveness, and directions for policy. *J. Appl. Ecol.*, 0–3.

MacGregor-Fors, I. & Payton, M.E. (2013). Contrasting Diversity Values: Statistical Inferences Based on Overlapping Confidence Intervals. *PLoS One*, 8, 8–11.

Mackenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, B., Franklin, A.B., *et al.* (2003). Estimating Site Occupancy, Colonization, and Local Extinction When a Species Is Detected Imperfectly. *Ecology*, 84, 2200–2207.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.A. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006). *Occupancy Estimation and Modelling*. Academic Press.

Mackenzie, D.I. & Royle, J.A. (2005). Designing occupancy studies: General advice and allocating survey effort. *J. Appl. Ecol.*, 42, 1105–1114.

Malhi, Y. (2012). The productivity, metabolism and carbon cycle of tropical forest vegetation. *J. Ecol.*, 100, 65–75.

Malhi, Y., Gardner, T. a., Goldsmith, G.R., Silman, M.R. & Zelazowski, P. (2013). Tropical Forests in the Anthropocene. *Annu. Rev. Environ. Resour.*, 39, 140906185140006.

Malmqvist, B., Strasevicius, D., Hellgren, O., Adler, P.H. & Bensch, S. (2004). Vertebrate host specificity of wild-caught blackflies revealed by mitochondrial DNA in blood. *Proc. R. Soc. B Biol. Sci.*, 271, S152–S155.

Mazzolli, M., Haag, T., Lippert, B.G., Eizirik, E., Hammer, M.L.A. & Al Hikmani, K. (2017). Multiple methods increase detection of large and medium-sized mammals: working with volunteers in south-eastern Oman. *Oryx*, 51, 290–297.

McInnes, J.C., Alderman, R., Lea, M.A., Raymond, B., Deagle, B.E., Phillips, R.A., *et al.* (2017). High occurrence of jellyfish predation by black-browed and Campbell albatross identified by DNA metabarcoding. *Mol. Ecol.*, 26, 4831–4845.

Meek, P.D., Ballard, G., Claridge, A., Kays, R., Moseby, K., O'Brien, T., *et al.* (2014). Recommended guiding principles for reporting on camera trapping research. *Biodivers. Conserv.*, 23, 2321–2343.

Meijaard, E. & Sheil, D. (2008). The persistence and conservation of Borneo's mammals in lowland rain forests managed for timber: Observations,

overviews and opportunities. *Ecol. Res.*, 23, 21–34.

Mercier, C., Boyer, F., Bonin, A. & Coissac, E. (2013). SUMATRA and SUMACLUST : fast and exact comparison and clustering of sequences. In: *Programs and Abstracts of the SeqBio workshop*. pp. 27–29.

Ministry of Natural Resources and Environment. (2016). *National policy on Biological Diversity (NPBD)*. Putrajaya, Malaysia.

Mioduchowska, M., Czyz, M.J., Gołdyn, B., Kur, J. & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? *PLoS One*, 13, 1–16.

Mitchell, S. L., Edwards, D. P., Bernard, H., Coomes, D., Jucker, T., Davies, Z. G., & Struebig, M. J. (2018). Riparian reserves help protect forest bird communities in oil palm dominated landscapes. *Journal of Applied Ecology*, *55*(6), 2744-2755.

Mohd Salleh, F., Ramos-Madrigal, J., Peñaloza, F., Liu, S., Mikkel-Holger, S.S., Riddhi, P.P., *et al.* (2017). An expanded mammal mitogenome dataset from Southeast Asia. *Gigascience*, 6, 1–8.

Moilanen, A. (2002). Implications of empirical data quality to metapopulation model parameter estimation and application. *Oikos*, 96, 516–530.

Mordecai, R.S., Mattsson, B.J., Tzilkowski, C.J. & Cooper, R.J. (2011). Addressing challenges when studying mobile or episodic species: Hierarchical Bayes estimation of occupancy and use. *J. Appl. Ecol.*, 48, 56–66.

Moreno, R.S., Kays, R.W. & Samudio JR, R. (2006). Competitive release in diets of ocelot (*Leopardus pardalis*) and puma (*Puma concolor*) decline. *J. Mammal.*, 87, 808–816.

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

Newbold, T., Hudson, L.N., Hill, S.L.L., Contu, S., Lysenko, I., Senior, R.A., *et al.* (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, 520, 45–50.

Nichols, J.D., Bailey, L.L., O'Connell, A.F., Talancy, N.W., Campbell Grant, E.H., Gilbert, A.T., *et al.* (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *J. Appl. Ecol.*, 45, 1321–1329.

Nicholson, E., Collen, B., Barausse, A., Blanchard, J.L., Costelloe, B.T., Sullivan, K.M.E., *et al.* (2012). Making robust policy decisions using global biodiversity indicators. *PLoS One*, 7(7), e41128.

Niedballa, J., Sollmann, R., Mohamed, A. Bin, Bender, J. & Wilting, A. (2015). Defining habitat covariates in camera-trap based occupancy studies. *Sci. Rep.*, 5, 1–10.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., *et al.* (2017). vegan: Community Ecology Package. *R Packag. ver. 2.4–3*.

Pan, Y., Birdsey, R.A., Fang, J., Houghton, R., Kauppi, P.E., Kurz, W.A., *et al.* (2011). A Large and Persistent Carbon Sink in the World's Forests. *Science,* 333(6045), 988-993.

Paoli, G.D., Wells, P.L., Meijaard, E., Struebig, M.J., Marshall, A.J., Obidzinski, K., *et al.* (2010). Biodiversity Conservation in the REDD. *Carbon Balance Manag.*, 5(1), 7.

Pardo, L.E., Campbell, M.J., Edwards, W., Clements, G.R. & Laurance, W.F. (2018). Terrestrial mammal responses to oil palm dominated landscapes in Colombia. *PLoS One*, 13, e0197539.

Payne, J., Francis, C.M. & Phillipps, K. (1985). *Field guide to the mammals of Borneo*. Sabah Society.

Pedersen, M.W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. Der, Haile, J., Hellstrom, M., *et al.* (2015). Ancient and modern environmental DNA. *Philos. Trans. B*, 370, 20130383.

Pfeifer, M., Lefebvre, V., Peres, C.. A., Banks-Leite, C., Wearn, O.R., Marsh, C.J., *et al.* (2017). Creation of forest edges has a global impact on forest vertebrates. *Nature*, 551, 187–191.

Pfeifer, M., Lefebvre, V., Turner, E., Cusack, J., Khoo, M., Chey, V.K., *et al.* (2015). Deadwood biomass: an underestimated carbon stock in degraded tropical forests? *Environ. Res. Lett.*, 10, 044019.

Phillipps, Q. & Phillipps, K. (2016). *Phillips's field guide to the mammals of Borneo and their ecology*. Kota Kinabalu: Natural History Publication.

Pochon, X., Zaiko, A., Fletcher, L.M., Laroche, O. & Wood, S.A. (2017). Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. *PLoS One*, 12, e0187636.

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N. & Taberlet, P. (2012). Who is eating what: Diet assessment using next generation sequencing. *Mol. Ecol.*, 21, 1931–1950.

Putz, F.E. & Redford, K.H. (2010). The Importance of Defining 'Forest': Tropical Forest Degradation, Deforestation, Long-term Phase Shifts, and Further Transitions. *Biotropica*, 42, 10–20.

Putz, F.E., Zuidema, P.A., Synnott, T., Peña-Claros, M., Pinard, M.A., Sheil, D., *et al.* (2012). Sustaining conservation values in selectively logged tropical forests: The attained and the attainable. *Conserv. Lett.*, 5, 296–303.

R Core Team. (2018). R Core Team (2018). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria. URL http//www.R-project.org/.*, R Foundation for Statistical Computing.

Ratnasingham, S. & Hebert, P.D.N. (2007). Barcoding BOLD : The Barcode of Life Data System (www.barcodinglife.org). *Mol. Ecol. Notes*, 7, 355–364.

Ratnasingham, S. & Hebert, P.D.N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS One*, 8(7), e66213.

Ratto, F., Simmons, B.I., Spake, R., Zamora-Gutierrez, V., MacDonald, M.A., Merriman, J.C., *et al.* (2018). Global importance of vertebrate pollinators for plant reproductive success: a meta-analysis. *Front. Ecol. Environ.*, 16, 82–90.

Rees, H.C., Bishop, K., Middleditch, D.J., Patmore, J.R.M., Maddison, B.C. & Gough, K.C. (2014). The application of eDNA for monitoring of the Great Crested Newt in the UK. *Ecol. Evol.*, 4, 4023–4032.

Rich, L.N., Miller, D.A.W., Robinson, H.S., Mcnutt, J.W. & Kelly, M.J. (2016). Using camera trapping and hierarchical occupancy modelling to evaluate the spatial ecology of an African mammal community. *J. Appl. Ecol.*, 1225–1235.

Riutta, T., Malhi, Y., Kho, L.K., Marthews, T.R., Huasco, W.H., Khoo, M.S., *et al.* (2017). Logging disturbance shifts net primary productivity and its allocation in Bornean tropical forests. *Glob. Chang. Biol.*, 12, 3218–3221.

Rodgers, T.W., Xu, C.C.Y., Giacalone, J., Kapheim, K.M., Saltonstall, K., Vargas, M., *et al.* (2017). Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community. *Mol. Ecol. Resour.*, 17, e133–e145.

Rosa, I.M.D., Smith, M.J., Wearn, O.R., Purves, D. & Ewers, R.M. (2016). The Environmental Legacy of Modern Tropical Deforestation. *Curr. Biol.*, 26, 2161–2166.

Roussel, J.M., Paillisson, J.M., Tréguier, A. & Petit, E. (2015). The downside of eDNA as a survey tool in water bodies. *J. Appl. Ecol.*, 52, 823–826.

Royle, J.A. & Link, W.A. (2006). Generalized Site Occupancy Models Allowing For False Positive and False Negative Errors. *Ecology*, 87, 835–841.

Royle, J.A. & Nichols, J.D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology*, 84, 777–790.

Salinas-Ramos, V.B., Herrera Montalvo, L.G., León-Regagnon, V., Arrizabalaga-Escudero, A. & Clare, E.L. (2015). Dietary overlap and seasonality in three species of mormoopid bats from a tropical dry forest. *Mol. Ecol.*, 24, 5296–5307.

Samejima, H., Ong, R., Lagan, P. & Kitayama, K. (2012). Camera-trapping rates of mammals and birds in a Bornean tropical rainforest under sustainable forest management. *For. Ecol. Manage.*, 270, 248–256.

Schipper, J., Chanson, J. S., Chiozza, F., Cox, N. A., Hoffmann, M., Katariya, V., ... & Baillie, J. (2008). The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science*, 322(5899), 225-230.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al.* (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75, 7537–7541.

Schmidt, B.R., Kéry, M., Ursenbacher, S., Hyman, O.J. & Collins, J.P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian pathogen. *Methods Ecol. Evol.*, 4, 646–653.

Schnell, I.B., Bohmann, K., Schultze, S.E., Richter, S.R., Murray, D.C., Sinding, M.-H.S., *et al.* (2018). Debugging diversity - a global scale exploration of the potential of terrestrial bloodfeeding leeches as a vertebrate monitoring tool. *Mol. Ecol. Resour.*, 18(6), 1282-1298.

Schnell, I.B., Sollmann, R., Calvignac-Spencer, S., Siddall, M.E., Yu, D.W., Wilting, A., *et al.* (2015a). iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool – prospects, pitfalls and avenues to be developed. *Front. Zool.*, 12, 24.

Schnell, I.B., Thomsen, P.F., Wilkinson, N., Rasmussen, M., Jensen, L.R.D., Willerslev, E., *et al.* (2012). Screening mammal biodiversity using dna from leeches. *Curr. Biol.*, 22, R262–R263.

Schnell, I.B.B., Bohmann, K. & Gilbert, M.T.P. (2015b). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.*, 15, 1289–1303.

Schubert, G., Stockhausen, M., Hoffmann, C., Merkel, K., Vigilant, L., Leendertz, F.H., *et al.* (2015). Targeted detection of mammalian species using carrion fly-derived DNA. *Mol. Ecol. Resour.*, 15, 285–294.

Schubert, M., Lindgreen, S. & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes*, 9, 88.

Scriven, S.A., Hodgson, J.A., McClean, C.J. & Hill, J.K. (2015). Protected areas in Borneo may fail to conserve tropical forest biodiversity under climate change. *Biol. Conserv.*, 184, 414–423.

Senior, M.J.M., Brown, E., Villalpando, P. & Hill, J.K. (2015). Increasing the Scientific Evidence Base in the "High Conservation Value" (HCV) Approach for Biodiversity Conservation in Managed Tropical Landscapes. *Conserv. Lett.*, 8, 361–367.

Sigsgaard, E.E., Broman, I., Henrik, N., Anders, M., Steen, K., Knudsen, W., *et al.* (2017). Seawater environmental DNA reflects seasonality of a coastal fish community. *Mar. Biol.*, 164, 1–15.

Sikes, R.S. & Gannon, W.L. (2011). Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *J. Mammal.*, 92, 235–253.

Sinclair, A.R.E. (2003). Mammal population regulation, keystone processes and ecosystem dynamics. *Philos. Trans. R. Soc. B Biol. Sci.*, 358, 1729–1740.

Sket, B. & Trontelj, P. (2008). Global diversity of leeches (Hirudinea) in freshwater. *Hydrobiologia*, 595, 129–137.

Slik, F.J., Alvarez-loayza, P., Alves, L.F., Ashton, P., Balvanera, P., Bastian, M.L., *et al.* (2015). An estimate of the number of tropical tree species. *Proc. Natl. Acad. Sci.*, 112, E4628–E4629.

Sodhi, N.S., Koh, L.P., Brook, B.W. & Ng, P.K.L. (2004). Southeast Asian biodiversity: an impending disaster. *Trends Ecol. Evol.*, 19, 654–60.

Sodhi, N.S., Posa, M.R.C., Lee, T.M., Bickford, D., Koh, L.P. & Brook, B.W. (2010). The state and conservation of Southeast Asian biodiversity. *Biodivers. Conserv.*, 19, 317–328.

Sollmann, R., Mohamed, A., Niedballa, J., Bender, J., Ambu, L., Lagan, P., *et al.* (2017). Quantifying mammal biodiversity co-benefits in certified tropical forests. *Divers. Distrib.*, 23, 317–328.

Stibig, H.J., Achard, F., Carboni, S., Raši, R. & Miettinen, J. (2014). Change in tropical forest cover of Southeast Asia from 1990 to 2010. *Biogeosciences*, 11, 247–258.

Struebig, M.J., Turner, A., Giles, E., Lasmana, F., Tollington, S., Bernard, H., *et al.* (2013). *Quantifying the Biodiversity Value of Repeatedly Logged Rainforests : Gradient and Comparative Approaches from Borneo*. *Glob. Chang. Multispecies Syst. Part III*. 1st edn. Elsevier Ltd.

Struebig, M.J., Wilting, A., Gaveau, D.L.A., Meijaard, E., Smith, R.J., Fischer, M., *et al.* (2015). Targeted Conservation to Safeguard a Biodiversity Hotspot from Climate and Land-Cover Change. *Curr. Biol.*, 25, 372–378.

Stuart, S.N., Chanson, J.S., Cox, N.A., Young, B.E., Rodrigues, A.S.L., Fischman, D.L., *et al.* (2004). Status and Trends of Amphibian Declines and Extinctions Worldwide, 306, 1783–1787.

Sullivan, M.J.P., Talbot, J., Lewis, S.L., Phillips, O.L., Qie, L., Begne, S.K., *et al.* (2017). Diversity and carbon storage across the tropical forest biome. *Sci. Rep.*, 7, 1–12.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012a). Environmental DNA. *Mol. Ecol.*, 21, 1789–1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012b). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.*, 21, 2045–2050.

Taylor, P.G. (1996). Reproducibility of Ancient DNA Sequences from Extinct Pleistocene Fauna. *Mol. Biol. Evol.*, 283–285.

Tessler, M., Weiskopf, S. R., Berniker, L., Hersch, R., Mccarthy, K. P., Yu, D. W., & Siddall, M. E. (2018). Bloodlines: Mammals, leeches, and conservation in southern Asia. *Systematics and Biodiversity*, *16*(5), 488-496.

Thomas, A.C., Howard, J., Nguyen, P.L., Seimon, T.A. & Goldberg, C.S. (2018). ANDe™: A fully integrated environmental DNA sampling system. *Methods Ecol. Evol.*, 2018, 1–7.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmussen, M., Gilbert, M.T.P., *et al.* (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Mol. Ecol.*, 21, 2565–73.

Thorn, S., Bässler, C., Brandl, R., Burton, P.J., Cahall, R., Campbell, J.L., *et al.* (2018). Impacts of salvage logging on biodiversity: A meta-analysis. *J. Appl. Ecol.*, 55, 279–289.

Toledo, V.M., Ortiz-Espejel, B., Cortés, L., Moguel, P. & Ordoñez, M. de J. (2003). The multiple use of tropical forests by indigenous peoples in Mexico: A case of adaptive management. *Conserv. Ecol.*, 7, 9.

Trontelj, P. & Utevsky, S.Y. (2005). Celebrity with a neglected taxonomy: Molecular systematics of the medicinal leech (genus Hirudo). *Mol. Phylogenet. Evol.*, 34, 616–624.

Turner, E.C., Abidin, Y.Z., Barlow, H., Fayle, T.M., Jaafar Mohd Hattah Hj., Khen, C.V., *et al.* (2012). The Stability of Altered Forest Ecosystems Project: Investigating the Design of Human-Modified Landscapes for Productivity and Conservation. *Plant.*, 88, 453–468.

United Nations. (2018). *Reducing Emissions from Deforestation and Degradation*.

US Fish and Wildlife. (2013). *QAPP: eDNA Monitoring of bighead and silver carps*.

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory, 11, 3571–3594.

Wearn, O.R., Carbone, C., Rowcliffe, J.M., Pfeifer, M., Bernard, H. & Ewers, R.M. (2018). Land-use change alters the mechanisms assembling tropical rainforest mammal communities in Borneo. *J. Anim. Ecol.*, 1–8.

Wearn, O.R., Rowcliffe, J.M., Carbone, C., Bernard, H. & Ewers, R.M. (2013). Assessing the Status of Wild Felids in a Highly-Disturbed Commercial Forest Reserve in Borneo and the Implications for Camera Trap Survey Design. *PLoS One*, 8, e77598.

Wearn, O.R., Rowcliffe, J.M., Carbone, C., Pfeifer, M., Bernard, H. & Ewers, R.M. (2017). Mammalian species abundance across a gradient of tropical land-use intensity: A hierarchical multi-species modelling approach. *Biol. Conserv.*, 212, 162–171.

Weiskopf, S.R., McCarthy, K.P., Tessler, M., Rahman, H.A., McCarthy, J.L., Hersch, R., *et al.* (2017). Using terrestrial haematophagous leeches to enhance tropical

biodiversity monitoring programmes in Bangladesh. *J. Appl. Ecol.*, 55, 2071–2081.

Wells, K., Pfeiffer, M., Lakim, M.B. & Linsenmair, K.E. (2004). Use of arboreal and terrestrial space by a small mammal community in a tropical rain forest in Borneo, Malaysia. *J. Biogeogr.*, 31, 641–652.

Wheeler, Q.D., Raven, P.H. & Wilson, E.O. (2004). Taxonomy: Impediment or Expedient? *Science,* 303, 285–285.

Wilcove, D.S., Giam, X., Edwards, D.P., Fisher, B. & Koh, L.P. (2013). Navjot's nightmare revisited: Logging, agriculture, and biodiversity in Southeast Asia. *Trends Ecol. Evol.*, 28, 531–540.

Wilkinson, C. L., Yeo, D. C., Tan, H. H., Fikri, A. H., & Ewers, R. M. (2018). Land-use change is associated with a significant loss of freshwater fish species and functional richness in Sabah, Malaysia. *Biological Conservation*, *222*, 164-171.

Willerslev, E., Hansen, A.J., Brand, T., Binladen, J., Gilbert, T.M.P., Shapiro, B.A., *et al.* (2003). Diverse plant and animal DNA from Holocene and Pleistocene sedimentary records. *Science,* 300(5620), *791-795.*

Wright, S.J., Zeballos, H., Dominguez, I., Gallardo, M.M., Moreno, M.C. & Ibáñez, R. (2000). Poachers alter mammal abundance, seed dispersal, and seed predation in a neotropical forest. *Conserv. Biol.*, 14, 227–239.

WWF. (2013). The Saola's Battle for Survival on The Ho Chi Minh Trail.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C., *et al.* (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.*, 3, 613–623.

Zepeda-Mendoza, M.L., Bohmann, K., Carmona Baez, A. & Gilbert, M.T.P. (2016). DAMe: A toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC Res. Notes*, 9, 1–13.

Appendix

Appendix 1

Accession numbers and taxonomic details for the 16S reference sequences included in the database used to assign mammal species to the unknow OTU sequences from leech iDNA. Accession numbers are from NCBI Genbank.

Genus	Species	Accession	Genus	Species	Accession
Aeromys	*tephromelas*	AY227482	*Muntiacus*	*muntjak*	NC4563
Aonyx	*cinerea*	KY117535	*Mus*	*musculus*	NC12387
Aonyx	*cinerea*	KY117536	*Mustela*	*nudipes*	MG996893
Arctictis	*binturong*	KY117560	*mutilata*	*isolate*	FJ613473
Arctogalidia	*trivirgata*	MG996895	*Nannosciurus*	*melanotis*	AY227461
Bos	*javanicus*	AB915322	*Nannosciurus*	*melanotis*	KT001463
Bos	*javanicus*	AB915322	*Nasalis*	*larvatus*	JF293094
Bos	*javanicus*	JN632605	*Nasalis*	*larvatus*	KM889667
Bos	*javanicus*	JN632606	*Nasalis*	*larvatus*	NC8216
Bos	*javanicus*	NC12706	*Nasalis*	*larvatus*	U39012
Bos	*taurus*	AB090990	*Neofelis*	*nebulosa*	MG996889
Bos	*taurus*	AB090991	*Niviventer*	*cremoriventer*	KY117572
Bos	*taurus*	AB090992	*Niviventer*	*cremoriventer*	KY117573
Bos	*taurus*	AB090993	*Nycticebus*	*coucang*	AF212952
Bos	*taurus*	AF492351	*Nycticebus*	*coucang*	AJ309867
Bos	*taurus*	KT375529	*Nycticebus*	*coucang*	AY773981
Bos	*taurus*	KT827208	*Nycticebus*	*coucang*	GQ253662
Callosciurus	*adamsi*	KR911800	*Nycticebus*	*coucang*	NC2765.1
Callosciurus	*adamsi*	NC30071	*Paguma*	*larvata*	KP233214
Callosciurus	*notatus*	AY227453	*Paguma*	*larvata*	KT191130
Callosciurus	*notatus*	KY117541	*Paguma*	*larvata*	NC29403
Callosciurus	*notatus*	KY117542	*Paradoxurus*	*hermaphroditus*	KJ698653
Callosciurus	*prevostii*	KY117543	*Paradoxurus*	*hermaphroditus*	MG996898
Canis	*lupus*	EU740414	*Pardofelis*	*marmorata*	AY499299
Catopuma	*badia*	AF006435	*Pardofelis*	*marmorata*	AY499300
Catopuma	*badia*	KP202256	*Pardofelis*	*marmorata*	KF754883
Catopuma	*badia*	KR135746	*Pardofelis*	*marmorata*	KF754884
Catopuma	*badia*	KX265094	*Pardofelis*	*marmorata*	KP202263
Catopuma	*badia*	KX265095	*Pardofelis*	*marmorata*	NC28303
Catopuma	*badia*	KX265096	*Petaurillus*	*kinlochii*	AY227490
Catopuma	*badia*	NC28300	*Petaurista*	*elegans*	KU579289
Cephalopachus	*bancanus*	AF348159	*Petinomys*	*setosus*	AY227492

Genus	Species	Accession	Genus	Species	Accession
Cervus	*unicolor*	AY391769	*Pongo*	*pygmaeus*	AY765084
Cervus	*unicolor*	AY391770	*Pongo*	*pygmaeus*	AY765085
Cervus	*unicolor*	DQ989636	*Pongo*	*pygmaeus*	AY765087
Cervus	*unicolor*	EF035448	*Pongo*	*pygmaeus*	AY765089
Chimarrogale	*himalayica*	DQ630345	*Pongo*	*pygmaeus*	AY765090
Chimarrogale	*himalayica*	GU981061	*Pongo*	*pygmaeus*	D38115
Chimarrogale	*himalayica*	GU981062	*Pongo*	*pygmaeus*	KU353723
Civettictis	*civetta*	KJ193027	*Pongo*	*pygmaeus*	NC1646
Civettictis	*civetta*	KJ193028	*Prionailurus*	*bengalensis*	HM185183
Civettictis	*civetta*	KJ193029	*Prionailurus*	*bengalensis*	JN392459
Civettictis	*civetta*	KJ193283	*Prionailurus*	*bengalensis*	KF754877
Civettictis	*civetta*	KJ193284	*Prionailurus*	*bengalensis*	KF754878
Crocidura	*foetida*	EF524857	*Prionailurus*	*bengalensis*	KF754879
Crocidura	*foetida*	EF524859	*Prionailurus*	*bengalensis*	KJ850247
Crocidura	*fuliginosa*	EF524812	*Prionailurus*	*bengalensis*	KJ850248
Crocidura	*fuliginosa*	EF524813	*Prionailurus*	*bengalensis*	KP202257
Crocidura	*fuliginosa*	EF524860	*Prionailurus*	*bengalensis*	KP202258
Crocidura	*fuliginosa*	EF524883	*Prionailurus*	*bengalensis*	KP202259
Crocidura	*fuliginosa*	GU981077	*Prionailurus*	*bengalensis*	KP202260
Cynocephalus	*variegatus*	AJ428849	*Prionailurus*	*bengalensis*	KP246843
Cynogale	*bennetti*	KY117544	*Prionailurus*	*bengalensis*	KR132586
Dendrogale	*melanura*	JF795293	*Prionailurus*	*bengalensis*	KR132587
Dicerorhinus	*sumatrensis*	FJ905816	*Prionailurus*	*bengalensis*	KX857784
Dicerorhinus	*sumatrensis*	KY117545	*Prionailurus*	*bengalensis*	NC16189
Dicerorhinus	*sumatrensis*	NC12684	*Prionailurus*	*bengalensis*	NC28301
Diplogale	*hosei*	MG996896	*Prionailurus*	*planiceps*	AF006407
Dremomys	*everetti*	KR911798	*Prionailurus*	*planiceps*	KP202280
Echinosorex	*gymnura*	AF348079	*Prionailurus*	*planiceps*	KR132592
Echinosorex	*gymnura*	NC2808	*Prionailurus*	*planiceps*	KR135743
Elephas	*maximus*	AJ428946	*Prionailurus*	*planiceps*	NC28312
Elephas	*maximus*	DQ316068	*Prionodon*	*linsang*	MG996894
Elephas	*maximus*	NC5129	*Pteromyscus*	*pulverulentus*	AY227491
Exilisciurus	*exilis*	AY227456	*Ptilocercus*	*lowii*	JF795291
Exilisciurus	*exilis*	KR911801	*Rattus*	*baluensis*	KY611363
Exilisciurus	*exilis*	KY117546	*Rattus*	*baluensis*	KY611388
Exilisciurus	*exilis*	NC30072	*Rattus*	*baluensis*	KY611390
Galeopterus	*variegatus*	JN800721	*Rattus*	*exulans*	NC12389
Galeopterus	*variegatus*	NC4031	*Rattus*	*rattus*	NC12374
Gallus	*gallus*	AP003319	*Rattus*	*tiomancus*	KY117580
Gehyra	*mutilata*	FJ613470	*Rattus*	*tiomanicus*	KP876560

Genus	Species	Accession	Genus	Species	Accession
Gehyra	*mutilata*	FJ613471	*Rattus*	*tiomanicus*	KY117579
Gehyra	*mutilata*	FJ613472	*Rattus*	*tiomanicus*	KY117581
Giraffa	*camelopardalis*	NC12100	*Rattus*	*tiomanicus*	NC29888
Giraffa	*camelopardalis*	NC24820	*Ratufa*	*affinis*	AY227495
Helarctos	*malayanus*	EF196664	*Rheithrosciurus*	*macrotis*	AY227498
Helarctos	*malayanus*	FM177765	*Rhinosciurus*	*laticaudatus*	AY227463
Helarctos	*malayanus*	NC9968	*Rusa*	*unicolor*	GQ411195
Hemidactylus	*frenatus*	GQ245970	*Rusa*	*unicolor*	GQ411196
Hemidactylus	*frenatus*	HM012691	*Rusa*	*unicolor*	JN714148
Hemidactylus	*frenatus*	HM192680	*Rusa*	*unicolor*	KX156946
Hemidactylus	*frenatus*	NC12902.2	*Rusa*	*unicolor*	KY117576
Hemigalus	*derbyanus*	MG996897	*Rusa*	*unicolor*	NC31835
Homo	*sapiens*	NR137295	*Rusa*	*unicolor*	NC8414
Homo	*sapiens*	NR137295	*Suncus*	*etruscus*	DQ630314
Hylobates	*muelleri*	AB050178	*Suncus*	*etruscus*	FJ486943
Hylobates	*muelleri*	AB050179	*Suncus*	*etruscus*	FJ486944
Hylobates	*muelleri*	EF152491	*Suncus*	*etruscus*	FJ716832
Hylobates	*muelleri*	HQ622778	*Suncus*	*etruscus*	JN556042
Hylobates	*muelleri*	HQ622779	*Suncus*	*murinus*	DQ630306
Hylobates	*muelleri*	HQ622780	*Suncus*	*murinus*	DQ630347
Hylobates	*muelleri*	HQ622781	*Suncus*	*murinus*	EF507191
Hylomys	*suillus*	AM905041	*Suncus*	*murinus*	FJ486952
Hylomys	*suillus*	AM905042	*Suncus*	*murinus*	FJ486953
Hylomys	*suillus*	AY121770	*Sundamys*	*infraluteus*	KY117583
Hylomys	*suillus*	DQ630368	*Sundamys*	*muelleri*	KY117584
Hylomys	*suillus*	NC10298	*Sundamys*	*muelleri*	KY117585
Hystrix	*brachyurus*	KR816507	*Sundasciurus*	*brookei*	AY227465
Muntiacus	*muntjak*	AF108038	*Sundasciurus*	*brookei*	KY117586
Hystrix	*indica*	DQ901405	*Sundasciurus*	*lowii*	KY117587
Hystrix	*indica*	JN714136	*Sus*	*barbatus*	GQ338944
Hystrix	*indica*	JN714143	*Sus*	*barbatus*	KP789021
Hystrix	*indica*	JN714144	*Sus*	*barbatus*	NC26992
Hystrix	*indica*	JN714145	*Tarsius*	*bancanus*	NC2811
Iomys	*horsfieldi*	AY227488	*Trachypithecus*	*cristatus*	KJ174503
Lariscus	*insignis*	AY227459	*Trachypithecus*	*cristatus*	KY117598
Lariscus	*insignis*	KR911799	*Trachypithecus*	*cristatus*	NC23971
Lariscus	*insignis*	KY117550	*Tragulus*	*kanchil*	JN632709
Lariscus	*insignis*	NC30070	*Tragulus*	*kanchil*	NC20753
Leopoldamys	*sabanus*	KY117551	*Tragulus*	*Napu*	KY117549
Leopoldamys	*sabanus*	KY117552	*Tragulus*	*napu*	M55539

Genus	Species	Accession	Genus	Species	Accession
Leopoldamys	*sabanus*	KY117553	*Trichys*	*fasciculata*	KY117590
Leopoldamys	*sabanus*	KY117554	*Tupaia*	*dorsalis*	JF795305
Leopoldamys	*sabanus*	KY117555	*Tupaia*	*glis*	JF795307
Leptobarbus	*hoevenii*	AP011286	*Tupaia*	*glis*	JF795308
Leptobarbus	*hoevenii*	NC15528	*Tupaia*	*glis*	MG996900
Lutra	*sumatrana*	KY117556	*Tupaia*	*gracilis*	JF795309
Lutrogale	*perspicillata*	KY117557	*Tupaia*	*longipes*	JF795311
Lutrogale	*perspicillata*	KY117558	*Tupaia*	*minor*	JF795313
Macaca	*fascicularis*	KM851002	*Tupaia*	*minor*	JF795314
Macaca	*fascicularis*	KM851003	*Tupaia*	*montana*	JF795315
Macaca	*fascicularis*	KM851004	*Tupaia*	*picta*	F795318
Macaca	*fascicularis*	KM851005	*Tupaia*	*splendidula*	JF795319
Macaca	*nemestrina*	KP765688	*Tupaia*	*splendidula*	JF795320
Macaca	*nemestrina*	KY117594	*Tupaia*	*tana*	AF203727
Macaca	*nemestrina*	NC26976	*Tupaia*	*tana*	JF795321
Manis	*javanica*	KP306515	*Tupaia*	*tana*	JF795322
Manis	*javanica*	KT445979	*Urva*	*brachyura*	MG996890
Manis	*javanica*	NC26781	*Urva*	*semitorquata*	MG996891
Martes	*flavigula*	FJ719367	*Varanus*	*salvator*	AB980995
Martes	*flavigula*	HM106326	*Viverra*	*tangalunga*	MG996899
Martes	*flavigula*	KM347744	*Muntiacus*	*muntjak*	EF523635
Martes	*flavigula*	NC12141	*Muntiacus*	*muntjak*	EF523636
Maxomys	*surifer*	KY117565	*Muntiacus*	*muntjak*	EF523637
Maxomys	*surifer*	KY117566	*Muntiacus*	*muntjak*	EF523638
Maxomys	*whiteheadii*	KY117568	*Muntiacus*	*muntjak*	EF523639
Maxomys	*whiteheadii*	KY117569	*Muntiacus*	*muntjak*	KY117560
Maxomys	*whiteheadii*	KY117570	*Muntiacus*	*muntjak*	AF108039
Maxomys	*whiteheadii*	KY117571	*Muntiacus*	*muntjak*	AY225986
Melogale	*moschata*	MG996892	*Muntiacus*	*atherodes*	KY117559

Appendix 2

OTU table for Chapter 3 – A species by site table showing a 1 where that species was detected and a 0 for a non-detection

Column headings

- **SITE:** B, D, E, F = heavily logged, LFE, LF1, LF2, LF3 = twice-logged, OG = primary, R0, R5, R30, RLFE = riparian
- **NO.**: Number of leeches per pool
- **Year**: 15 = 2015, 16 = 2016
- **Four letter codes** used for species detected in leech iDNA

SUBA	*Sus barbatus*	**RUUN**	*Rusa unicolor*	**MUSP**	*Muntiacus sp*
BOSP	*Bos sp*	**TRSP**	*Tragulus sp*	**MAJA**	*Manis javanica*
HYSP	*Hystrix sp*	**TRFA**	*Trichys fasciculata*	**PRSP**	*Prionailurus sp*
VITA	*Viverra tangalunga*	**HEDE**	*Hemigalus derbyanus*	**ARTR**	*Arctogalidia trivirgata*
PALA	*Paguma larvata*	**HEMA**	*Helarctos malayanus*	**MASP**	*Macaca sp*
HYMU	*Hylobates muelleri*	**ELMA**	*Elephas maximus*		

Site	NO.	Year	SUBA	RUUN	MUSP	BOSP	TRSP	MAJA	HYSP	TRFA	PRSP	VITA	HEDE	ARTR	PALA	HEMA	MASP	HYMU	ELMA
B	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
B	7	15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
B	10	15	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
B	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	9	15	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
B	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	16	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	10	16	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
B	10	16	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	10	16	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	10	16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	10	16	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
D	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	11	15	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
D	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Site	NO.	Year	SUBA	RUUN	MUSP	BOSP	TRSP	MAJA	HYSP	TRFA	PRSP	VITA	HEDE	ARTR	PALA	HEMA	MASP	HYMU	ELMA
D	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	6	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
D	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	6	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	10	16	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
D	4	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	10	16	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	11	16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	10	16	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	9	16	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
D	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	3	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
E	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
E	8	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	10	16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
E	10	16	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
E	8	16	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
F	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	10	15	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
F	10	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	4	15	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LF1	6	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LF2	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF2	10	15	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
LF2	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	12	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	8	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	16	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
LF3	10	16	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	16	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	16	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
LF3	10	16	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
LF3	10	16	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
LF3	9	16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LF3	10	16	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LF3	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	10	16	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
LF3	10	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
LFE	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Site	NO.	Year	SUBA	RUUN	MUSP	BOSP	TRSP	MAJA	HYSP	TRFA	PRSP	VITA	HEDE	ARTR	PALA	HEMA	MASP	HYMU	ELMA
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
LFE	10	15	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	3	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LFE	10	16	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
LFE	10	16	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
LFE	10	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
OG	8	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	7	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
OG	9	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
OG	12	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	5	16	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	10	16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Site	NO.	Year	SUBA	RUUN	MUSP	BOSP	TRSP	MAJA	HYSP	TRFA	PRSP	VITA	HEDE	ARTR	PALA	HEMA	MASP	HYMU	ELMA
OG	9	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OG	12	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
R0	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R0	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R0	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R0	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R0	10	15	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
R0	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R0	8	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R30	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
R5	7	15	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
R5	9	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R5	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
R5	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R5	4	15	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
R5	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R5	11	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	11	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	6	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	11	15	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
RLFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
RLFE	4	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	11	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	15	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RLFE	11	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	11	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	4	15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
RLFE	11	16	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
RLFE	9	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	16	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RLFE	10	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	10	15	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	10	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	13	15	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	10	15	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	10	15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	10	15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
VJR	6	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	4	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Site	NO.	Year	SUBA	RUUN	MUSP	BOSP	TRSP	MAJA	HYSP	TRFA	PRSP	VITA	HEDE	ARTR	PALA	HEMA	MASP	HYMU	ELMA
VJR	10	16	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
VJR	9	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VJR	8	16	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Appendix 3

Vegetation metrics calculated from LiDAR data provided by T. Swinfield and D. Coomes

Metrics used in the analysis for Chapter 3 and Chapter 4

Column headings

- **Site:** A, B, D, E, F, VJR = heavily logged, LFE, LF1, LF2, LF3 = twice-logged, RHI, TEM, WES = primary, R0, R5, R30, RLFE = riparian
- **Moran** = Habitat heterogeneity measured by Morans I
- **CanopyHeight =** top of canopy height measured in metres
- **CH_SD** = Standard deviation of canopy height
- **AGB =** Above ground biomass
- **ForestCov** = Proportion forest cover, the inverse is Gap fraction – used in Chapter 3

Site	Moran	CanopyHeight	CH_SD	GapFraction	AGB	ForestCov
A	0.59	14.61	6.34	0.23	50.53	0.77
B	0.51	17.08	6.48	0.13	65.32	0.87
C	0.62	10.31	6.34	0.53	28.54	0.47
D	0.63	8.66	5.95	0.62	21.46	0.38
E	0.71	10.89	7.89	0.53	31.25	0.47
F	0.60	18.08	8.06	0.16	71.70	0.84
LF1	0.64	22.92	7.14	0.06	105.73	0.94
LF2	0.45	24.40	5.65	0.02	117.20	0.98
LF3	0.64	23.16	7.32	0.07	107.61	0.93
LFE	0.49	24.79	5.87	0.02	120.28	0.98
VJR	0.54	27.37	11.49	0.09	141.43	0.91
R0	0.70	10.21	7.54	0.57	28.10	0.43
R30	0.53	15.98	6.25	0.16	58.57	0.84
R5	0.74	12.10	8.73	0.49	37.12	0.51
RLFE	0.58	23.52	6.65	0.04	110.37	0.96
RHI	0.44	31.86	13.74	0.08	181.42	0.92
TEM	0.38	33.79	12.41	0.04	199.76	0.96
WES	0.39	31.30	12.23	0.05	176.21	0.95