MRS. IDA BÆRHOLM SCHNELL (Orcid ID : 0000-0002-0436-785X)

DR. KRISTINE BOHMANN (Orcid ID : 0000-0001-7907-064X)

MR. MIKKEL-HOLGER S. SINDING (Orcid ID : 0000-0003-1371-219X)

MR. MASON J CAMPBELL (Orcid ID : 0000-0001-6803-271X)

Article type : Resource Article

Debugging diversity - a pan-continental exploration of the potential of terrestrial blood-feeding leeches as a vertebrate monitoring tool

Ida Bærholm Schnell[1,2*], Kristine Bohmann[1,3], Sebastian E. Schultze[1], Stine R. Richter[1], Dáithí C. Murray[4], Mikkel-Holger S. Sinding[1,5], David Bass[6,7], John E. Cadle[8], Mason J. Campbell[9], Rainer Dolch[10], David P. Edwards[9,11], Thomas N. E. Gray[12], Teis Hansen[1], Anh Nguyen Quang Hoa[13], Christina Lehmkuhl Noer[1,2], Sigrid Heise-Pavlov[14], Adam F. Sander Pedersen[15], Juliot Carl Ramamonjisoa[16], Mark E. Siddall[17], Andrew Tilker[18,19], Carl Traeholt[2], Nicholas Wilkinson[20], Paul Woodcock[21], Douglas W. Yu[3,22], Mads Frost Bertelsen[2], Michael Bunce[4], M. Thomas P. Gilbert[1,4,23].

[1] Section for EvoGenomics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark,

[2] Center for Zoo and Wild Animal Health, Copenhagen Zoo, Frederiksberg, Denmark.

[3] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, UK,

[4] Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia,

[5] Greenland Institute of Natural Resources, Nuuk, Greenland,

[6] Department of Life Sciences, the Natural History Museum, London, UK,

[7] Cefas, Barrack Road, The Nothe, Weymouth, Dorset DT4 8UB, UK,

[8] Centre ValBio, Ranomafana, Ifanadiana, Fianarantsoa, Madagascar,

[9] Centre for Tropical Environmental and Sustainability Science (TESS) and College of Science and Engineering, James Cook University, Cairns, Queensland 4878, Australia,

[10] Association Mitsinjo, Andasibe, Madagascar,

[11] Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, UK,

[12] Wildlife Alliance, Phnom Penh, Cambodia,

[13] WWF Vietnam, Hue City, Thua Thien Hue, Vietnam,

[14] Centre for Rainforest Studies at the School for Field Studies, Yungaburra, Queensland, Australia,

[15] Centre for Medical Parasitology, Department of Immunology and Microbiology, University of Copenhagen, Copenhagen, Denmark,

[16] The Peregrine Fund, Antananarivo, Madagascar,

[17] Sackler Institute of Comparative Genomics and Division of Invertebrate Zoology, American Museum of Natural History, New York, USA,

[18] Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany,

[19] Global Wildlife Conservation, Austin, Texas, USA

[20] Department of Geography, University of Cambridge, Cambridge, UK,

[21] School of Biology, University of Leeds, Leeds, UK,

[22] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[23] NTNU University Museum, N-7491 Trondheim, Norway

*Corresponding Author: Ida Bærholm Schnell, Øster Voldgade 5-7, DK-1350 Copenhagen K, ibschnell@snm.ku.dk

Abstract

The use of environmental DNA (eDNA) has become an applicable non-invasive tool with which to obtain information about biodiversity. A sub-discipline of eDNA is iDNA (invertebrate-derived DNA), where genetic material ingested by invertebrates is used to characterise the biodiversity of the species that served as hosts. While promising, these techniques are still in their infancy, as they have only been explored on limited numbers of samples from only a single or a few different locations. In this study, we investigate the suitability of iDNA extracted from more than 3,000 haematophagous terrestrial leeches as a tool for detecting a wide range of terrestrial vertebrates across five different geographical regions on three different continents. These regions cover almost the full geographical range of haematophagous terrestrial leeches, thus representing all parts of the world where this method might apply. We identify host taxa through metabarcoding coupled with high-throughput sequencing on Illumina and IonTorrent sequencing platforms to decrease economic costs and workload and thereby make the approach attractive for practitioners in conservation management. We identified hosts in four different taxonomic vertebrate classes: mammals, birds, reptiles, and amphibians, belonging to at least 42 different taxonomic families. We find that vertebrate blood ingested by haematophagous terrestrial leeches throughout their distribution is a viable source of DNA with which to examine a wide range of vertebrates. Thus, this study provides encouraging support for the potential of haematophagous terrestrial leeches as a tool for detecting and monitoring terrestrial vertebrate biodiversity.

Keywords

iDNA, metabarcoding, high throughput sequencing, terrestrial haematophagous leeches, vertebrate diversity, vertebrate monitoring.

Introduction

Estimating abundance and distribution of vertebrate species is one of the fundamental challenges for conservation biologists, and an essential input to track progress (or lack thereof) toward the Aichi Biodiversity Targets (Convention on Biological Diversity 2010), a worldwide strategic plan for biodiversity and its protection. Even well-studied groups such as mammals require continual monitoring efforts to document temporal and spatial changes (Jones & Safi 2011) and appropriately inform conservation efforts. Thus, any methodological innovations in species monitoring approaches are potentially valuable for biomonitoring in a conservation framework, especially if they make data collection easier or more efficient, lower the impact on the target species, and/or reduce costs.

Recently, the use of environmental DNA (eDNA) has been touted as a widely applicable non-invasive tool with which to obtain information about biodiversity (e.g. Bohmann *et al.* 2014). Furthermore, if coupled to metabarcoding (i.e. high throughput sequencing of taxonomically informative PCR amplicons), eDNA enables identification of multiple species within single bulk samples (Taberlet *et al.* 2012). An emerging sub-discipline within eDNA is 'iDNA' (invertebrate-derived DNA), in which genetic material ingested by invertebrates is used to characterise the biodiversity of the species that served as hosts (Calvignac-Spencer *et al.* 2013a). Examples include vertebrate DNA extracted from blood-feeding midges (Lassen *et al.* 2012), ticks (Gariepy *et al.* 2012), terrestrial leeches (Schnell *et al.* 2012), carrion-feeding blowflies (Calvignac-Spencer *et al.* 2013b), blood-feeding sand flies and mosquitoes (Kocher *et al.* 2017). These methods are promising as they enable collection of "blood/tissue samples" that can be difficult to obtain, such as from vertebrate species that are challenging to monitor because they are shy, nocturnal, dangerous, or live in inaccessible habitats (Bohmann *et al.* 2013). However, while promising, the use of iDNA to detect vertebrates is still in its infancy, and prior to the application of iDNA as a biodiversity monitoring tool, there is a need to verify its potential and to define the limits (both biological and geographical) within which it may be useful (Schnell *et al.* 2015b). In addition, improving labour and financial cost-effectiveness is essential if iDNA is to become a useful biodiversity-monitoring tool.

We present here a large expansion on our initial study that reported the potential of leech-derived iDNA as a source of vertebrate DNA (Schnell *et al.* 2012). More specifically, we focus on the haematophagous (blood-feeding) terrestrial leeches that belong to the family Haemadipsidae, within the suborder Hirudiniformes (Borda & Siddall 2004; Borda *et al.* 2008; Phillips & Siddall 2009).

Members of this family occupy large parts of Asia, Australasia, and Madagascar (Sawyer 1986; Borda *et al.* 2008), areas that are known for their extensive tropical or subtropical rainforests, rich biodiversity, and high number of endemic and/or threatened vertebrate species (Myers *et al.* 2000; Ceballos & Ehrlich 2006; Schipper *et al.* 2008). Generally, leeches only feed a few times annually and possess an ability to retard the digestion rate of ingested blood (Sawyer 1986; Schnell *et al.* 2012). As a result of this, preliminary observations record high detection rates of well-preserved vertebrate DNA in leech blood meals (Schnell *et al.* 2012).

Our initial study (Schnell *et al.* 2012) reported detection of vertebrate host DNA in blood meals from just 25 individuals of terrestrial haematophagous leeches collected within a single forest complex in the Annamite Mountains of Vietnam. Thus, the number of leech species studied and mammals identified were equally small. The present study investigates detection of iDNA from a wide range of terrestrial vertebrates, using haematophagous terrestrial leeches collected over almost their full geographical range. Specifically, this includes the two morphologically distinct groups of terrestrial leeches, the duognathous (two-jawed) and trignathous (three jawed) leeches in South-East Asia, Madagascar, and Australia.

We demonstrate the feasibility of replacing traditional molecular cloning and Sanger sequencing-based characterisation with metabarcoding coupled to high throughput sequencing. Furthermore, to increase data output and decrease the "hands-on" sample processing workload, iDNA was generated from pools of leeches as opposed to single leeches to more accurately reflect how leech iDNA assessment might be practically implemented in a conservation setting.

Increasing throughput by moving from single leech extractions and Sanger sequencing to pools of leeches sequenced on next generation sequencing platforms requires careful methodological considerations, with decisions on how to balance detection of biodiversity with error removal, workload and costs (see Alberdi *et al.* (2018) for an exploration of some key decisions in metabarcoding studies). Here we evaluate and discuss a number of these decisions, including the use of human blocking probes to reduce amplification of human DNA, and how/if PCR replicates and their treatment affect the final result.

Materials and Methods

Sample sites

In total, 3,427 individual terrestrial leeches were collected from 21 sites in five geographical regions, namely Madagascar, Laos/Vietnam, Peninsular Malaysia, Malaysian Borneo, and mainland Australia and Tasmania (Table 1 & Fig. 1). Leeches from Madagascar were collected in the Analamazaotra Forest Station (Mitsinjo Reserve) in Andasibe (Périnet) and Ranomafana National Park, both placed in the mid-eastern part of the island. In Peninsular Malaysia, leeches were collected in Krau Wildlife Reserve in the central state of Pahang. Malaysian Bornean leeches were from four sites in the state of Sabah: Danum Valley Conservation Area, Maliau Basin Conservation Area, Sandakan Rainforest Discovery Centre, and Kinabalu National Park. Leeches were collected from multiple locations in Vietnam and Laos, focussing on the Annamite Mountains with sites including Thua Thien Hue and Quang Nam Saola Reserves in Vietnam, and Nam Kading and Xe Sap National Protected Areas in Laos. The vertebrate biodiversity identified from Vietnam and Laos are presented together due to the small spatial distance between the collection sites in the two countries. In Australia, leeches were collected at multiple sites in, or near, a number of national parks along the east coast in both the states of New South Wales (NSW) and Queensland (QLD). In Tasmania (TAS), leeches were collected in Cradle Mountain National Park. The samples collected in mainland Australia and Tasmania are henceforth collectively referred to as Australian samples.

Sample collection & storage

All leeches were collected by hand as they approached the collectors as part of their natural foraging behaviour. Leeches were immediately put into tubes containing RNAlater (Life Technologies, Carlsbad, CA), which killed them within minutes, or were killed by freezing, and subsequently thawed and submerged in either RNAlater or 96 % ethanol. Multiple leeches from the same location and collection date were put in the same tube. Where possible, the collectors wore latex gloves to minimise contamination with human DNA. Leeches were subsequently incubated at ca. 5°C for two days to allow full RNAlater or ethanol impregnation and then stored at ca. -20°C until analysis.

Sample preparation & DNA extraction

Prior to digestion, leeches stored in ethanol were placed in clean tubes with open lids at 55°C for up to 30 minutes to facilitate evaporation of ethanol. Leeches stored in RNAlater were removed from the RNAlater and digested directly.

Digesting leeches in pools has been tested and found possible (data not included). However in this study, leeches were digested individually, despite the increase in costs and workload. This made it possible to use any individual leech DNA extracts for other studies. Each leech was digested in approximately 5x their volume of digestion buffer, principally following the method described in Schnell *et al.* (2012), although with a reduction of proteinase K from 10% to 1% as we have observed this to be equally efficient at digesting the leeches (I.B. Schnell, unpublished observations).

Prior to DNA purification, 30-50 µL leech digest from 1-17 (average 7.6) individual leeches collected at the same site and date were pooled to decrease workload and costs of DNA extraction and sequencing. The leech digest pools were vortexed and centrifuged, and the supernatant was purified using a Qiaquick PCR purification kit (Qiagen, Valencia, CA), principally following the manufacturer's guidelines, although with a reduced speed of 6000 g during binding. DNA was eluted in 50 µL 10 mM Tris HCl buffer (pH 8) after 10 min incubation at 37°C. One negative extraction control was included for every 5-20 leech-pool extractions, covering new reagents, days of extraction and collection sites.

DNA amplification and sequencing

Asian and Malagasy leeches:

Prior to sequencing on an Illumina MiSeq platform, a ca. 95 bp (excluding primers) subfragment of the vertebrate mtDNA 16S locus was PCR amplified using mammalian-generic primers 16Smam1/16Smam2 (Taylor 1996) (Table 2). The primers were modified by 5´-labelling with unique nucleotide tags (with at least two differences between tags) (Binladen *et al.* 2007), creating 59 differently tagged forward and reverse primers. All leech-pool DNA extracts were PCR amplified in two replicates with different combinations of tags. PCRs were carried out in 25 µL reactions using the Amplitaq Gold enzyme system (Applied Biosystems, Foster City, CA) consisting of 2.5 mM $MgCl_2$, 1x Gold PCR Buffer, 0.4 µM each primer, 0.1 mM dNTPs, 0.5U Amplitaq Gold, and 1 µL purified DNA. In addition, 4.0 µM of a human blocking probe was added to each reaction to ensure minimum amplification of human DNA (Vestheim and Jarman 2008, Boessenkool *et al.* 2012) (Table 2). PCR conditions were as follows: 95°C for 5 min of enzyme activation followed by 40 cycles of 95°C for 12 s, 59°C for 30 s, and 70°C for 25 s, and one cycle for final extension of 7 min at 70°C. Post PCR, amplicons were visualised on 2% agarose gels stained with GelRed (Biotium, Hayward, CA). PCR products with different tag combinations, from leech digest pools within the same geographic region, were pooled in approximately equimolar ratios based on gel band strength and known concentrations on a subset of leech digest pools measured with a Qubit fluorometer (dsDNA high sensitivity kit, Invitrogen) (data not shown). Extraction blanks were included in a subset of the pools in an amount similar to the other PCR products (Supporting Information Table S2). The pooled PCR products were purified using Qiaquick PCR purification kits following the manufacturer's guidelines, although a 10 min incubation time at 37°C was included prior to elution of the DNA. The purified PCR pools were size-selected (150 bp +/−10%) on a LabChip XT (Caliper Life Sciences, Hopkinton, MA).

The purified PCR pools were then converted into Illumina MiSeq libraries using the NEBNext DNA library Prep Master Mix Set for 454 (#E6070) (NEB, Ipswich, MA, USA) with blunt end Illumina adapters (Meyer & Kircher 2010) in place of Roche/454 FLX adaptors and subsequently subjected to index PCR as described in Hope *et al.* (2014), although the index PCRs were run in three separate reactions instead of five, with each containing 10 μL library instead of 5 μL.

The libraries and a library blank for each round of library preparation were visualised on a 2100 Bioanalyser Chip (Agilent Technologies, Santa Clara, CA), then pooled in equimolar ratio (a subset of the library blanks were added to the pools in a volume similar to that of the other libraries) prior to sequencing using 150 bp paired-end chemistry on the Illumina MiSeq platform (Fig. 2).

Australian leeches (collected in mainland Australia and Tasmania):
Due to export restrictions, leeches collected in Australia could not be analysed in Denmark alongside the other leeches. Therefore, the laboratory protocols and sequencing technologies available slightly differed from those of the Asian and Malagasy leeches. The DNA from Australian samples was real-time PCR (qPCR) amplified using two sets of primers. These sets were, the mammal generic primer set 16Smam1/16Smam2 (as described above (Taylor 1996)), and a vertebrate generic primer set 12SA/12SO (Poinar 1998) (Table 2). Both primer sets were modified into fusion-tagged primers containing the appropriate Ion Torrent adaptors, A and P1, followed by a unique nucleotide tag and the template specific primer at the 3′ end. For contamination control, in each primer set individual extracts were assigned forward and reverse tags in combinations that had never been amplified in the laboratory before. The uniquely tagged qPCR amplifications on all leech-pools, across both primer sets, were carried out using the Amplitaq Gold enzyme system in 25 µL volumes containing 2.5 mM $MgCl_2$, 1x buffer, 0.4 µM each primer, 0.2 mM mixed dntps, 1U Amplitaq Gold, 0.6 µL SYBR Green (SYBR® Green I Nucleic Acid Gel Stain - 10,000X concentrate in DMSO, Invitrogen) and 2 µL purified DNA. Reaction conditions varied depending on the primer set as described in Murray *et al.* (2013).

The fusion-tagged qPCR reactions were performed in duplicate for each leech-pool. Post PCR, successful replicates from each leech-pool were combined to reduce the effects of PCR stochasticity. To reduce costs associated with purification, PCR-amplicons from five leech-pools were pooled in approximately equal proportions (as determined by qPCR end-point values), prior to double purification using Agencourt AMPure XP Bead PCR Purification (Beckman Coulter Genomics, MA, USA). The purified PCR pools were visualised on 2% agarose gels to confirm the presence of appropriate length DNA and subsequently pooled in roughly equimolar ratios based on band intensity to form the sequencing libraries.

Sequencing libraries were diluted 1:1000 and double purified using Agencourt AMPure XP Bead PCR Purification and eluted in 40 µL EB buffer. The libraries were then quantified using qPCR against a dilution series of a standard library of known molarity to determine the appropriate input volume of library for emulsion PCR (emPCR) prior to amplicon sequencing on the Ion Torrent PGM. All amplicon sequencing was performed as per the manufacturer’s instructions using 314 chips and 200 bp reagents (Fig. 2).

For all leeches:

To check for cross- and background contamination, at least one negative PCR control was included for every five leech-pool extracts (all geographic regions) amplified. All extraction blanks were amplified, and a subset of both extraction blanks and negative PCR controls was included in sequencing.

Use of human blocking probes

Thirty-three individual leeches collected in Peninsular Malaysia were PCR amplified both with and without human blocking probes to explore differences in amplification of human DNA. PCR setup and conditions were the same as for the other leech pools from Peninsular Malaysia, however two PCR replicates were made with human blocking probes and two without. Post-amplification treatment was the same as used for the other samples that were sequenced on the Illumina platform. The proportion of sequences within OTUs assigned to humans of the total amount of sequences in assigned OTUs was calculated, and the human proportions were compared between amplifications with and without the addition of human blocking probes. A paired t-test was conducted to compare the averages of the two treatments.

Sequence analyses

The Illumina and Ion Torrent libraries were sequenced on several sequencing runs. After sequencing, reads from each of the Illumina libraries were trimmed and merged in AdapterRemoval (v. 1.5.4) (Lindgreen 2012), using the default settings, except for the following: mismatchrate=0.01, minlength=100, shift=5, minquality=28, minalignmentlength=50, trimns, trimqualities and collapse.

Trimmed and merged sequences from the Illumina MiSeq, and sequences obtained from the Ion Torrent PGM, were sorted and filtered in leech-pool-specific tag combinations using a modified version of DAMe (Zepeda-Mendoza *et al.* 2016) (modified version: https://github.com/shyamsg/DAMe), retaining only sequences that contained a perfect match to both the forward and reverse primer and the tag sequences. All singletons (reads only present in one copy within each PCR replicate) were discarded, and for the Illumina sequencing output, only sequences present in two PCR replicates ((Alberdi *et al.* 2018) - restrictive approach) were kept for further analysis.

Sequences were clustered into OTUs using Sumaclust (Mercier *et al.* 2013). To choose the most appropriate values for clustering, similarity scores of all sequences within each dataset were calculated then investigated to determine whether a “barcode gap” was present (i.e. a gap between the intra- and interspecific similarity values across species (Meyer & Paulay 2005), with intraspecific similarity also comprising amplification – and sequencing errors). This investigation was done using the modified version of DAMe (https://github.com/shyamsg/DAMe), and the number of identified OTUs at variable levels of maximum ratios (-R) between the counts of two sequences was then inspected. Based on the clustering values generated, all sequences were then clustered with the exact option (-e), a similarity score of 0.96 and a maximum ratio of 0.90.

To reduce the number of erroneous OTUs (e.g. PCR and sequencing errors), the OTUs from each dataset were run through the LULU (Frøslev *et al.* 2017), a post clustering algorithm, using default settings.

The sequence with the highest copy number within each curated OTU was assessed taxonomically using blastn (default settings) (Blast+) against the NCBI non-redundant nucleotide database (Nt) (Jan 2018).

The BLAST results were imported into MEtaGenome ANalyser (MEGAN version 5.11.3, (Huson *et al.* 2007)) where they were assigned to order, family, or genus depending on the similarity with sequences in the reference database using the LCA-assignment algorithm (parameters included: Default settings except Minimum bit score = 150.0, Top percentage = 2%, Min. support = 1 sequence). Reads that could not be taxonomically assigned within these parameters were manually assessed, and subsequently assigned to appropriate taxonomic levels, based on all retained matches in Genbank (unless automatic assignment was impossible (e.g. errors in Genbank)).

The number of sequences assigned to the different OTUs, and thus which OTUs were considered as true diversity, were evaluated and to account for low-level cross contamination between samples and to reduce the effects of tag jumps between samples within datasets (Schnell *et al.* 2015a), only samples that had an OTU copy number of at least 1% of what was present in the most abundant sample were kept as true diversity. For Australian samples, where the PCR replicates had been combined ((Alberdi *et al.* 2018) – additive approach), a minimum copy number of 10 in at least one sample was also required before the OTU was considered as true diversity.

Although it is plausible that the leeches have fed on humans, human DNA is also a common laboratory/handling derived contaminant. Whether the presence of human DNA in this study originates from the blood meal of leeches, or is contamination, cannot be easily determined. Thus, as a conservative measure, all sequences assigned to Hominidae were not considered a true blood meal.

Statistical analyses

To evaluate the similarity of the two PCR replicates for all Asian and Malagasy samples the Renkonen Similarity Indeces (RSI) were computed using DAMe (Zepeda-Mendoza *et al.* 2016), and the RSI measures from the six different datasets were compared using one-way ANOVA with a post-hoc Tukey HSC test.

To better understand the underlying factors affecting the discovery of non-human OTUs from Malagasy and Asian samples, the number of non-human OTUs discovered in each leech pool was modeled in a generalised linear regression framework, using a Poisson model and a log link function. Bidirectional variable selection was performed using the Akaike Information Criteria (AIC). The effects of number of non-human reads, the number of leeches within the leech pools, the total copy number, the minimum copy number from the two PCR replicates and the RSI measure were allowed to vary between the different collection sites using an interaction term.

PCR replicates – additive or restrictive approach

The Asian and Malagasy samples were assessed to determine the relationship between bioinformatics treatment and OTU recovery. Specifically, the biodiversity identified when using a restrictive approach (i.e. retention of only OTUs found in multiple PCR replicates) was compared with that revealed when using an additive approach (the combined OTU diversity of both replicates). Sequencing analyses were identical to the one described for Australian samples (recovery of at least 10 copies in any sample in order for an OTU to be retained), but in addition, an OTU was only perceived as being present in a sample, if its copy number were higher than 3% of what was present in the most abundant sample. Lower copy numbers were perceived as possible artefacts originating from, for example, tag jumps (See (Schnell *et al.* 2015a) for identified levels of tag jumps in Illumina workflows) and cross-contamination between samples.

Identification of analysed leeches

Southeast Asian and Madagascan leech species were determined from a subset of leech pools using COI barcode primers (LCO1490/HCO2198) (Folmer *et al.* 1994), that amplified a ca. 660 bp fragment (excl. primers). Amplification followed the same protocol as described for the mammalian generic primer set, although with PCR conditions as described in Folmer *et al.* (1994). PCR products were purified and sequenced in one direction by Macrogen Europe's commercial Sanger Sequencing service.

The Australian and Tasmanian leech species composition was determined using primer set MZArtF/MZArtR (ZBJ-ArtF1c/ZBJ-ArtR2c) (Zeale *et al.* 2010), amplifying a 157 bp COI minibarcode fragment (excl. primers). Amplification, post amplification purification, as well as sequencing preparation and Ion Torrent PGM sequencing was conducted as described for the mammal and vertebrate amplicons, although with PCR conditions were as described in Zeale *et al.* (2010).

Sanger-sequenced leech DNA sequences were quality trimmed in Geneious (version 6.1.8), and those sequenced on the Ion Torrent sequencing platform were sorted following the same procedure as for the Ion Torrent-sequenced vertebrate sequences described above, although with a similarity score of 0.97 instead of 0.96 during OTU clustering.

Cluster centers from leech OTUs were globally aligned with other available COI data for haemadipsid leeches (Borda *et al.* 2010; Schnell *et al.* 2012; Tessler *et al.* 2016), and a tree was constructed using a general time reversible model with an estimated gamma distribution rate parameter and an estimated proportion invariate (GTR+I+gamma) on RaxML using http://phylogeny.lirmm.fr/ (Dereeper *et al.* 2008). On the tree, the branch lengths are proportional to the amount of change, with the scale being the estimated number of substitutions per aligned site. Both *Hirudo medicinalis* and *Macrobdella decora* were used as outgroups. In total, 166 DNA sequences were included in the alignment including nine cluster centers from the Australian samples, and the remaining 60 sequences obtained from leeches from Madagascar (20 leech-pools), Peninsular Malaysia (20 leech-pools), Laos/Vietnam (10 leech-pools) and Malaysian Borneo (10 leech-pools).

Results

Sequencing

After the initial sorting where sequences were assigned to samples (that included removal of singletons, only retaining sequences with a perfect match to both primer and tag sequence, and presence in both PCR replicates), an average of 4137 sequences per leech were generated on the Illumina sequencing platform. On average 1026 sequences per leech were retained after the initial filtering from the Ion Torrent sequencing platform (singletons removed and perfect a match to both primer - and tag sequences) (for details see Supporting Information Table S1). When combining data from all collection sites across the five regions, 443 pools (excluding 27 negative controls) of leeches were analysed during this study, with an average of 7.6 leeches per pool.

All negative controls contained no OTUs or only OTUs assigned to humans.

Identified vertebrate biodiversity

Approximately 50 % of the original OTUs were retained after post clustering curation in LULU (Supporting Information Table S2). These OTUs were assigned to 388 leech pools. After discarding 84 leech pools that only had OTUs matching human sequences (Supporting Information Table S3), the remaining 304 leech pools containing non-human vertebrate OTUs were used for further analyses. Within these, a total of 193 curated OTUs were identified (after discarding OTUs assigned to humans). These could be assigned to 75 vertebrate taxa within 42 taxonomic families: 13 OTUs assigned to amphibian taxa, 16 OTUs assigned to bird taxa, 4 scaled reptile OTUs, one OTU assigned to a turtle taxa and 135 mammalian OTUs. The remaining 24 OTUs could not be assigned to any taxa, and are therefore excluded from the final results. The OTUs that could not be assigned to any taxa were primarily present in the Australian samples (21 OTUs of the 24 unassigned OTUs) (Fig. 3 & Supporting Information Table S6). The majority of OTUs were assigned to mammals – nevertheless, from all regions except Malaysian Borneo, at least one additional class of vertebrates were identified (see Supporting Information Fig. S1A-G for charts of OTU and sample counts from each dataset). Henceforth we use "OTUs" to refer to post clustering curated, non-human OTUs only, unless stated otherwise.

Leeches from Madagascar contained DNA from 16 different taxonomic families including e.g. lemurs, tenrecs, hornbills and endemic frogs. From Southeast Asia, the leeches contained DNA from several different taxonomic families of ruminants in addition to porcupines, sunbears, treeshrews and pangolins to name but a few. From Australian leeches DNA from, among others, kangaroos, wombats, bandicoots, emus and lyrebirds were identified. All identified taxonomic families are known to be present in the geographical areas where the leeches, containing their DNA, were collected.

Use of human blocking probes

In the PCR amplifications with and without human blocking probes, 12 of the 33 leeches that were amplified individually yielded DNA in all four PCR replicates (two with and two without human blocking probes, respectively). Only these 12 leeches were included in the comparison between the two treatments. The proportion of sequences in "human" OTUs out of the total number of sequences assigned to OTUs was significantly higher in samples without human blocking probe (mean without blocker: 30.3%; with: 0%; t=2.5147, df=11, p=0.0287). No differences in detection of vertebrate OTUs could be measured.

Leech identifications

In the phylogenetic tree sequences grouped by geographical region (Fig. 4) suggesting that interregional genetic distance was larger than intraregional. The genera of leeches collected in Asia were all trignathous (three jawed) leeches whereas the rest were duognathous (two-jawed) leeches. Leeches from Madagascar grouped within two species: *Chtonobdella fallax* from Ranomafana and *Chtonobdella vagans* from Analamazaotra. The analysed leeches from Peninsular Malaysia were identified as *Haemadipsa interrupta* while those from Laos and Vietnam were all identified as

Haemadipsa sylvestris (like the Vietnamese leeches identified in Schnell *et al.*, (2012)). With the exception of two specimens of *Haemadipsa sumatrana*, samples from Sabah grouped with *Haemadipsa picta*. The highest leech diversity was found in Australia; including *Chtonobdella gloriosi*, *Chtonobdella tanae* and *Chtonobdella bilineata.* While isolate Qld2, Tas1 and Tas2 were clearly a species of *Chtonobdella*, the small COI marker precluded definitive species assignment.

Statistical analyses

There was a significant difference in average RSI measures for the different datasets (ANOVA $F_{5,396}$ = 10.2325, P<3.0886e-09). Values from Malaysian Borneo DS1 were significantly higher than the values from Madagascar, Laos/Vietnam DS1, Peninsular Malaysia and Malaysian Borneo DS2. Malagasy samples had significantly lower RSI measures than samples from Laos/Vietnam DS2 (Supporting Information Table S4). Average RSI measures for PCR replicates ± S.D. for each of the six datasets from Asia and Madagascar are listed in Table S2 (Supporting information).

The final model used for the number of identified OTUs within samples included the dataset, and a dataset specific effect for non-human reads, number of leeches in the pools, the total copy number of each pool and its RSI measure. All variables except total copy number significantly affected the final OTU count from the leech pools, with dataset being the most important predictor.

The analysis of variance for the final model is shown in supporting information Table S5A, while the effect sizes and the Z-score and the p-value are show in Supporting Information Table S5B.

For dataset Laos/Vietnam DS1, the total number of reads from each leech pool, the number of non-human reads and the RSI measure had a significant effect on the number of OTUs, whereas for Malagasy samples, the OTU count was significantly affected by the number of leeches in the leech pools.

PCR replicates – differences between the restrictive and additive approach

Using the restrictive treatment of the two PCR replicates, a total of 145 OTUs (141 assigned to vertebrate taxa) within 35 vertebrate families were identified across 268 leech pools from Madagascar and Asia. The families include six amphibian, six avian, 22 mammalian families and one family of turtles (testudines). When PCR replicates from the same samples were combined in the additive approach, 179 OTUs were identified across 301 samples; 168 OTUs that could be assigned to vertebrate taxa, and 11 that could not. The 168 OTUs were assigned to 40 families, the 35 families identified with the restrictive treatment and an additional five families including two mammalian families, Elephantidae and Tapiridae, two avian families, Cathartidae and Anatidae, and one additional family of turtles (testudines), Cheloniidae. In addition to these five taxonomic families, five additional OTUs were identified to genera within taxonomic families already present when the restrictive PCR approach was used (Supporting Information Table S6 & S7).

Elephants and tapirs are present in the areas where the leeches containing their DNA were collected (Malaysian Borneo and Peninsular Malaysia respectively). The OTU assigned to vulture (Carthartidae) was present in two samples from Laos/Vietnam DS2 in low copy numbers (10 and 11 copies respectively). Sea turtle (Cheloniidae) were present in a single leech pool from Peninsular Malaysia. We believe both vulture and sea turtle DNA derive from contamination, as discussed later. The single species of duck within the family Anatidae was found in a leech pool from Laos/Vietnam in 29 copies. The five OTUs assigned to genera within taxonomic families present with the restrictive PCR approach include the mammalian genera *Panthera*, *Hapalemur* and *Callosciurus*, in addition to the amphibian genus *Quasipaa* and the avian genus *Gallus.*

In two out of 24 sequenced PCR negative controls from Laos/Vietnam DS1, a single OTU was present. However, this OTU could not be assigned to any organism, and only 13 reads of the OTU were present in each sample.

Discussion

The aim of this study was to investigate the potential for terrestrial blood-feeding leeches across their geographical range to function as a vertebrate monitoring tool. The presence of vertebrate species across the three continents was determined using a metabarcoding approach coupled to two different second generation sequencing technologies. We also provide evidence that terrestrial blood-feeding leeches contain taxonomically informative iDNA from various vertebrate classes, and that metabarcoding coupled to high throughput sequencing is an efficient method to assess this. In addition, we provide both a wet-lab and a bioinformatic pipeline that can be used and adapted by practitioners; indeed our leech iDNA protocol is already being tested in multiple conservation projects as a wildlife monitoring method in China, Borneo, Vietnam, and Laos.

It should be highlighted that vertebrate biodiversity identified from leeches is presence data only, and the degree to which such data can be combined with state-of-the-art analytical tools to estimate distributions and abundances of the different identified vertebrate taxa is discussed in Schnell *et al.* (2015b).

The discussion in the following paragraphs is based on the use of iDNA from leeches; however, most of the topics also relate to other studies where metabarcoding is applied on more or less degraded DNA from biologically complex samples, including many eDNA studies.

Identified vertebrate taxa

Given that the primer set predominantly used in this study was designed to amplify mammalian DNA (Taylor 1996), it is unsurprising that the majority of OTUs detected in the leech blood meals were assigned to mammals (Fig. 3 & Supporting Information Fig. 1A-G). However, despite the potential primer preference towards mammalian DNA, our results show that the analysed leeches had also ingested blood from amphibians, birds and reptiles (Fig. 3). The vertebrate taxa that were identified are ecologically diverse, ranging from large-bodied animals such as bears, to small species such as rodents and amphibians. They belong to multiple trophic levels and a wide range of food preferences and strategies, including carnivores such as dogs, cats, and pythons, insectivores such as pangolins and lyrebirds, ruminant and non-ruminant herbivores, and omnivores such as pigs, lemurs, and bears. The identified vertebrates are adapted to different lifestyles, including cursorial deer, saltatorial kangaroos, arboreal possums and treeshrews, and even volant birds, which might be surprising since all leeches were collected from the ground or low-level vegetation.

It is beyond the scope of this study to delve into each OTU's taxonomic assignment and its regional significance; however, when comparing the presence of vertebrate families, and the number of samples in which these were present, the similarity between datasets was generally higher for geographically close collection sites (Fig. 5). Furthermore, endemic vertebrate families such as kangaroos and lyrebird in Australia and tenrecs in Madagascar were only identified from leeches collected within the known distribution of those vertebrate families; findings which both support our hypothesis that the identified OTUs are indeed authentic, and that leeches yield fascinating insight into the faunal composition and ecology of the study sites (Fig. 3 & Supporting Information Fig. S1A-G).

To our knowledge, no other single census method can detect such a wide range of terrestrial vertebrate taxa. In addition, only a single researcher is required to analyse the sequencing data and provide objective information about the presence of taxa across the different vertebrate classes. This is in contrast to other monitoring methods relying on e.g. identification by visualisation, where the expertise of one person is often limited to a single class or order.

The use of blocking primers

The use of human blocking primers significantly reduced amplification of human DNA. In addition to their main purpose of decreasing competition between human and non-human DNA during PCR, blocking primers might also have reduced the costs of the analyses by reducing the number of leech pools that were identified as "DNA-positive" before library preparation, because PCR-reactions only consisting of human DNA could be excluded prior to sequencing. However, the DNA-negative pools could also have been false negatives caused by co-blocking of target DNA sequences as has previously been reported (e.g. in arthropods (Piñol *et al.* 2014)). As such, there is a need for follow-up studies to test which taxa are more likely to be lost by co-blocking. The sample size in this study was too small to detect whether the use of human blocking probes influenced the detected diversity.

Identified leech taxa

Both the duognathous (two-jawed) and trignathous (three-jawed) leeches belong to the family Haemadipsidae, but they differ in geographic distribution (Fig. 1) and morphology. The inter- and intraspecific relationships of these two clades are debated (Sawyer 1986; Borda *et al.* 2008; Borda & Siddall 2010; Tessler *et al.* 2016). The leech genera identified in Asia were trignathous (three jawed), and those identified from Madagascar and Australia were duognathous (two-jawed) leeches; this is in agreement with the currently known distribution of two- and three-jawed terrestrial leeches (Borda & Siddall 2010). Since vertebrate taxa presence was detected from all identified two- and three-jawed terrestrial leech taxa, we argue that both two- and three-jawed terrestrial haematophagous leeches can be used in terrestrial vertebrate surveys. Although the higher number of samples with OTUs identified from the Malagasy vs Asian leech pools could indicate that the duognathous leeches feed on a wider range of animals or are better at preserving the ingested DNA, we did not see this relationship when comparing Asian leeches with those from Australia (which are also duognathous). Thus we do not believe we have evidence that leech taxa is a primary driver of the observed differences. However, using the current study design, any potential bias among leech species leading to differences in vertebrate detection rate could not be determined.

PCR replicates – the restrictive or additive strategy

In this study, independently tagged PCR replicates were carried out on pools of leeches collected in Asia and Madagascar, in order to assure reproducibility and authenticity of vertebrate detections, which enabled restrictive analysis approaches to be used to reduce the error among the OTUs.

In the samples from Asia and Madagascar, five additional vertebrate families were identified using the additive PCR approach (as opposed to restrictive). These included vulture and sea turtle and their presence is without doubt a result of background contamination, likely deriving from studies previously undertaken in the Copenhagen facility (Roggenbuck *et al.* 2014, Duchene *et al.* 2012). For the five OTUs assigned to vertebrate genera that was not found with the restrictive PCR approach (*Panthera*, *Hapalemur*, *Callosciurus*, *Quasipaa* and *Gallus)* it is difficult to clearly distinguish what taxa are true diversity, and what might originating form contamination and/or sequencing error. This is the case for the OTU assigned to *Panthera* (*Panthera leo*) that is present in a single leech pool from Madagascar (41 copies), and the OTU assigned to *Hapalemur* (*Hapalemur griseus*), present in another Malagasy leech pool (459 copies). Both OTUs have a BLAST percentage identity of 100 %. The presence of *Hapalemur griseus* in Malagasy leeches is very likely since this particular vertebrate species is distributed in the area where the leeches have been collected. In contrast, the presence of DNA from lions in Malagasy leeches is indeed questionable. In this study discarding the OTU assigned to lion would not affect the final conclusions, but in other studies false positives could be conclusive especially if they cannot be easily detected.

The additive approach also yielded an increased number of OTUs that could not be assigned to any vertebrate species (data not shown). Most of these OTUs were not present in any leech pool with 10 copies or more, and therefore excluded from further analyses. However, we highlight that the decision to discard OTUs below a certain copy number threshold should be carefully considered because authentic biodiversity might be lost even when discarding only singletons (Alberdi *et al.* 2018). How to balance error removal with detection is study dependent, but one approach could be combining the two PCR replicate treatments (e.g. by having at least three PCR replicates, and only keeping sequences present in at least two replicates for further analyses) to assure both authenticity of the results and to maximise the revealed diversity, although this would not occur without increased cost and workload (see e.g. Ficetola *et al.* (2015) and Alberdi *et al.* (2018) for tests and discussions about PCR replicates).

Limits to taxonomic assignments

A major challenge in using leech iDNA to assess vertebrate diversity is incompleteness of DNA reference databases. In the present study, this resulted in error-prone or impossible assignment of OTUs to lower taxonomic levels. For instance, in leeches from Madagascar only half of the OTUs could be assigned to taxonomic order level. The number of described vertebrate species is most likely underestimated compared to true diversity, and reference DNA sequences exist only for a fraction of these (Francis *et al.* 2010). Thus, until DNA reference databases are developed further, information on vertebrate biodiversity present in leeches is not fully exploited. Geographically targeted sampling, continued growth in sequence databases, and improvements in bioinformatic methods (Somervuo et al. 2016) will all improve the quality of taxonomic assignment in future studies.

Quality vs. high throughput - a trade-off?

Metabarcoding analyses of DNA from pools of leeches using second generation sequencing platforms (as in this study) can increase the amount of PCR inherent bias, contamination, and other technical errors in comparison to the approach taken by Schnell *et al.* (2012) (single leech extractions, PCR, cloning of amplicons then Sanger sequencing). Use of generic primers on biologically complex samples might introduce PCR-inherent biases, with preferential amplification of certain templates, while others remain undetected or poorly amplified (e.g. Suzuki & Giovannoni 1996; Polz & Cavanaugh 1998; Andersen *et al.* 2003). In this study biases could be caused by differences in PCR selection (as defined by Wagner *et al.* 1994). Such PCR selection biases include differences in primer binding sites (Suzuki & Giovannoni 1996) and possibly variable levels of DNA quantity and quality (i.e. fragmentation) among leeches within a pooled sample. In addition, PCR stochasticity caused by random events in the early cycles (Wagner et al. 1994) might also have a negative impact on the identified vertebrate diversity in leech pools. Use of two or more complementary metabarcoding assays may overcome some of the PCR selection biases if they complement each other, whereas multiple PCR replicates on each sample may reduce the effect of PCR stochasticity (Wagner et al. 1994).

In this dataset, no OTUs were found in 35 % of the leech pools, and although there was a significant difference between some of the study sites, the average number of leeches required within a leech pool for providing one extra OTU was five. Thus based on this value, the number of leeches with detectable vertebrate DNA when using the presented method and data analyses was approximately 20 %; a considerably lower percentage than the findings of more than 80 % in Schnell *et al.* (2012).

In Schnell *et al.* (2015b) the components influencing the detection probability (p) of vertebrate DNA from a leech are proposed as p(*Leech feed on focal species*), p(*Fed leech collected*), p(*Target DNA extracted/amplified*) and p(*Correct identification based on the DNA*). If extraction and amplification of vertebrate DNA from leeches was perfect (p(*Target DNA extracted/amplified*) equal to one), the number of identified OTUs should be positively correlated with the number of leeches within the leech pools, with the slope reflecting p(*Leech feed on focal species*) and p(*Fed leech collected*), assuming that each leech provides unique vertebrate DNA, and thus a new OTU. Only the Malagasy samples yielded a positive correlation between OTU count and the number of leeches within the pools, but even for this study site the number of leeches required for finding an OTU was 2.6. For the remaining study sites, there was no correlation between numbers of leeches in the leech pools and the OTU counts, which indicates a limiting factor from the extraction and/or amplification on the final results. As for other metabarcoding workflows, a number of decisions were made regarding both sample processing and post sequencing data analyses in this study. Parameters such as similarity threshold during OTU clustering, values used during post-clustering curation with LULU (Frøslev *et al.* 2017), taxonomic assignment and copy number cutoffs are likely to affect the biodiversity revealed. In addition, the dissimilarity between PCR replicates (RSI measures) and the following restrictive data analyses, likely decreased the number of pools with OTUs, a hypothesis supported by the increased number of samples yielding OTUs when data was analysed using the additive PCR approach instead. Thus the importance of having at least two successful amplifications from each sample should be emphasized if restrictive PCR approaches are used, alongside an acknowledgement of the many other decisions made during a metabarcoding workflow, and how they are likely to influence the final outcome (Alberdi *et al.* 2018).

With high-throughput sequencing platforms, the consequence of contaminants present in even very low frequency may be particularly problematic, because the sequencing depth is much higher than with Sanger sequencing (Porter *et al.* 2013). In addition, iDNA contains biologically complex substrates of various degradation and fragmentation levels, and this makes it challenging to distinguish true biodiversity present in low template number from amplicons created by PCR and sequencing errors, and contaminant sequences. Thus, methods to limit and detect contamination not only between samples within a project, but also cross contamination between different projects should be prioritized as discussed in Murray *et al.* (2014) and Schnell *et al.* (2015a).

In this study, we only retained sequences that were present in multiple PCR replicates (for Asian and Malagasy leeches), and during OTU clustering we used a low similarity score (0.96) in comparison to other studies using the same primer sets (e.g. Murray et al. 2013; Harris et al. 2014). Despite these efforts, it was evident that PCR and sequencing errors inflated the number of OTUs, and as a consequence, the LULU post-clustering algorithm was used to curate the clustering output. Alternative OTU clustering algorithms were not investigated, but this could of interest in future iDNA studies, especially because the number of OTUs assigned to e.g. humans within each dataset was still artificially inflated even after post-clustering curation with LULU. In addition, ways to simplify and/or optimize bioinformatics pipelines specifically for iDNA are of high priority, and the confidence that leech iDNA contains considerable ecological information on vertebrate diversity may help to justify these efforts.

Primer choice for iDNA studies

The PCR primer set predominantly used in this study (16Smam1/mam2 (Taylor 1996)) is designed to amplify mammalian DNA, however *in silico* PCR analyses have estimated low taxonomic coverage for other vertebrate classes (Ficetola *et al.* 2010). Only the Australian leech samples were amplified with the vertebrate generic 12S primers in addition to the 16S mammal primers. Here, the 12SA/12SO primer set revealed the presence of an emu (*Dromaius novaehollandiae*) that was not detected with the 16S mammal primers (Supporting Information Table S3). Within a few of the mammalian families, the number of detected OTUs from the two primer sets was also slightly different. Hence, even though non-mammalian vertebrates were identified using primer set 16Smam1/mam2, we hypothesise that use of additional primer sets targeting other vertebrate classes (e.g. bird or amphibian) could have disclosed a greater non-mammalian diversity.

Even though blood-feeding leeches are known for their long inter-meal intervals and slow digestion rate, the exact rate with which ingested DNA gets fragmented during digestion remains unknown. To accommodate this unknown fragmentation level, we assume an inverse relationship between amplification efficiency and length of amplification target similar to what is observed in studies with ancient/degraded DNA (Handt *et al.* 1994; Deagle *et al.* 2006). Here, we only targeted short vertebrate mtDNA markers (approximately 100 bp excl. primer sequences, Table 2). The 16S marker used has a low resolution capacity for identification at species level, but a high level of resolution for genus and family identification (Ficetola *et al.* 2010). Meanwhile, the 12S marker has in previous studies been used to identify both birds and mammals to species and genus levels (e.g. Poinar *et al.* 1998; Kuch *et al.* 2002; D'Costa *et al.* 2011; Porter *et al.* 2013).

In future studies, other vertebrate generic primer sets amplifying short mitochondrial DNA markers (e.g. Riaz et al. 2011) could be used with the aim of improving taxonomic resolution and coverage. It is worth noting that reference database coverage should be taken into account when selecting markers in order to optimise the taxonomic assignments. Furthermore, multiple primer sets targeting different fragment lengths and/or groups of animals could be applied in order to optimise both amplification success and taxonomic assignment, and thereby ensure that results based on leech iDNA is not limited by the PCR assays. Lastly, approaches where PCR is minimised or redundant, e.g. shotgun sequencing or third generation sequences techniques, have shown promise for overcoming some of the PCR-induced challenges that are associated with metabarcoding of biologically complex samples (Zhou et al. 2013).

Conclusions

A principal goal of many new wildlife-survey methods is to increase cost-effectiveness and/or accuracy, since money spent on monitoring is money not available for conservation actions (Lindenmayer *et al.* 2011; Possingham *et al.* 2012). Despite our relatively conservative approach of principally restricting our PCRs to one genetic target region, leech iDNA captured nearly the full range of vertebrate diversity, both taxonomically (mammals, birds, amphibians, and reptiles), and functionally (arboreal and terrestrial, small and large, predators and prey). Based on our results, we argue that iDNA holds great potential to complement, camera-trapping and visual surveys, depending on management needs. This method allows vertebrate surveys to be successful without expert knowledge in taxonomy, and leeches iDNA can be collected during single expeditions contrary to other survey techniques (e.g. camera trapping) where multiple visits are required.

We expect a continuing decrease in sequencing costs to allow for commensurate increase in leech screening ability, which make iDNA surveys even more applicable in the future.

In this study, the primers predominantly used were not optimised to assign vertebrate taxa to species level. Future studies should choose primers tailored to specific management needs to fully exploit the taxonomic information available from iDNA. In addition, as DNA reference databases are expanded, assignment of sequences to lower taxonomic levels will be further facilitated.

Future monitoring studies relying on iDNA from terrestrial haematophagous leeches, and other iDNA sources, need to thoroughly consider a range of pros and cons. Of particular importance is choosing the right laboratory and analytical methods in a given case as to optimise assignment of more OTUs to species rank and for obtaining abundance and distribution information rather than simply presence data. Indeed, abundance and distribution information at the species level is important when monitoring vertebrate population trends and thus whether the world is meeting the Aichi Biodiversity targets, as proposed by the United Nations (Convention on Biological Diversity 2010).

Acknowledgements

We thank the staff of the Danish National High-Throughput Sequencing Centre for technical assistance and acknowledge research grants by the Danish National Research Foundation (DNRF94) and University of Copenhagen for financial support, as well as the Australian Research Council Future Fellowship to MB (FT0991741). We appreciate help by Benjamin Andriamihaja from Institute for the Conservation of Tropical Environments and Tianasoa Ratolojanahary from Association Mitsinjo for permits and collection assistance. We thank MICET for facilitating permits for fieldwork and export of samples from Madagascar. We thank Madagascar National Parks and the Ministry of Water and Forests for permissions. Centre ValBio (Ranomafana) assisted with logistics and provided laboratory facilities for preparation and storage of samples.

Fieldwork in Vietnam/Laos was part of the CarBi project of WWF Greater Mekong largely funded by KfW Bankengruppe and WWF Germany.

Samples were collected in Borneo under permission numbers JKM/MBS.10000-2/2(120), DVMC 2012/07, and MBMC 2012/14. We thank members of Oxford University Tropical Forest Ecology Field Course (2012), led by Owen Lewis, and members of The Natural History Museum-funded field trip to Borneo for field collection assistance, and the Royal Society’s South East Asia Rainforest Research Programme (SEARRP) and the Borneo Rainforest Lodge for logistical support and Suzan Benedick for collaborative support. We thank Yayasan Sabah, Danum Valley Management Committee, the State Secretary, Sabah Chief Minister’s Departments, the Economic Planning Unit of the Prime Minister’s Department, and the Sabah Biodiversity Council for permission to conduct research.

Permissions to conduct research in Australia and Tasmania was provided by DPIWE, Department of Primary Industries, Parks, Water and Environment (Tasmania), Department of Environment and Heritage Protection (Queensland), and NSW Environment and Heritage.

We thank the Saola Working Group, especially William Robichaud, for guidance and motivation, and Tanja Normann Gade for help with sampling processing.

KB was supported by The Danish Council for Independent Research grant DFF - 5051-00140. DWY was supported by the National Natural Science Foundation of China (31400470, 41661144002), the Ministry of Science and Technology of China (2012FY110800), and the State Key Laboratory of Genetic Resources and Evolution at the Kunming Institute of Zoology (GREKF13-13, GREKF14-13, GREKF16-09).

Authors have no conflicts of interest.

References

Agca Y, Monson RL, Northey DL *et al.* (1998) Normal calves from transfer of biopsied, sexed and vitrified IVP bovine embryos. *Theriogenology*, **50**, 129–145.

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, **9**, 134–147.

Binladen J, Gilbert MTP, Bollback JP *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**, e197.

Boessenkool S, Epp LS, Haile J *et al.* (2012) Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*, **21**, 1806–1815.

Bohmann K, Evans A, Gilbert MTP *et al.* (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution* **29**, 358–367.

Bohmann K, Schnell IB, Gilbert MTP (2013) When bugs reveal biodiversity. *Molecular Ecology*, **22**, 909–911.

Borda E, Siddall ME (2004) Arhynchobdellida (Annelida: Oligochaeta: Hirudinida): phylogenetic relationships and evolution. *Molecular Phylogenetics and Evolution*, **30**, 213–225.

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

Borda E, Oceguera-Figueroa A, Siddall ME (2008) On the classification, evolution and biogeography of terrestrial haemadipsoid leeches (Hirudinida: Arhynchobdellida: Hirudiniformes). *Molecular Phylogenetics and Evolution*, **46**, 142–154.

Calvignac-Spencer S, Leendertz FH, Gilbert MTP, Schubert G (2013a) An invertebrate stomach's view on vertebrate ecology. *BioEssays*, **35**, 1004–1013.

Calvignac-Spencer S, Merkel K, Kutzner N *et al.* (2013b) Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. *Molecular Ecology*, **22**, 915–924.

CBD (Convention on Biological Diversity) (2010). COP 10 Decision X/2. Strategic Plan for Biodiversity 2011-2020. Available from: http://www.cbd.int/decision/cop/?id=12268

Ceballos G, Ehrlich PR (2006) Global mammal distributions, biodiversity hotspots, and conservation. *Proceedings of the National Academy of Sciences*, **103**, 19374–19379.

D'Costa VM, King CE, Kalan L *et al.* (2011) Antibiotic resistance is ancient. *Nature*, **477**, 457–461.

Deagle BE, Evesom JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples – a case study on DNA in faeces. *Frontiers in Zoology*, **3**, 11.

Dereeper, A., Guignon, V., Blanc, G., Audic, S. & Buffet, S. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**, W465–W469.

Duchene S, Frey A, Alfaro-Núñez A *et al.* (2012) Marine turtle mitogenome phylogenetics and evolution. *Molecular Phylogenetics and Evolution*, **65**, 241–250.

Ficetola GF, Coissac E, Zundel S *et al.* (2010) An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, **11**, 434.

Ficetola GF, Pansu J, Bonin A (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, **15.3**, 543-556.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology*, **3**, 294–299.

Francis CM, Borisenko AV, Ivanova NV *et al.* (2010) The Role of DNA Barcodes in Understanding and Conservation of Mammal Diversity in Southeast Asia (S Joly, Ed,). *PLoS ONE*, **5**, e12575.

Frøslev TG, Kjøller R, Bruun HH *et al.* (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, **8**, 1188.

Gariepy TD, Lindsay R, Ogden N, Gregory TR (2012) Identifying the last supper: utility of the DNA barcode library for bloodmeal identification in ticks. *Molecular Ecology Resources*, **12**, 646–652.

Handt O, Höss M, Krings M, Pääbo S (1994) Ancient DNA: Methodological challenges. *Experientia*, **50**, 524–529.

Harris RB, Jiake Z, Yinqiu J, Kai Z, Chunyan Y (2014) Evidence that the Tibetan fox is an obligate predator of the plateau pika: conservation implications. *Journal of Mammalogy* **95**: 1207–1221.

Herr CM, Reed KC (1991) Micronanipulation of bovine embryos for sex determination. *Theriogenology*, **35**, 45–54.

Hope PR, Bohmann K, Gilbert MTP *et al.* (2014) Second generation sequencing andmorphological faecal analysis revealunexpected foraging behaviour by Myotisnattereri (Chiroptera, Vespertilionidae) in winter. *Frontiers in Zoology*, **11**, 1–15.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Jones KE, Safi K (2011) Ecology and evolution of mammalian biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2451–2461.

Kocher A, de Thoisy B, Catzeflis F *et al.* (2017) iDNA screening: Disease vectors as vertebrate samplers. *Molecular Ecology*, **26**, 6478–6486

Kreader CA (1996) Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein. *Applied and Environmental Microbiology*, **62**, 1102–1106.

Kuch M, Rohland N, Betancourt JL *et al.* (2002) Molecular analysis of a 11 700- year- old rodent midden from the Atacama Desert, Chile. *Molecular Ecology*, **11**, 913–924.

Lassen SB, Nielsen SRA, Kristensen M (2012) Identity and diversity of al hosts of biting midges

(Diptera: Ceratopogonidae:Culicoides Latreille) in Denmark. *Parasites & Vectors*, **5**, 1–1.

Lindenmayer DB, Gibbons A, Bourke M *et al.* (2011) Improving biodiversity monitoring. *Austral Ecology*, **37**, 285–294.

Lindgreen S (2012) AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. *BMC research notes*, **5**, 337.

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

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

Meyer M, Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, **2010**, 1–10.

Murray DC, Haile J, Dortch J *et al.* (2013) Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports*, **3**, 3371.

Murray DC, Coghlan ML, Bunce M (2015) From Benchtop to Desktop: Important Considerations when Designing Amplicon Sequencing Workflows. *PLoS ONE* **10(4)**, e0124671.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca, Gustavo AB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Phillips AJ, Siddall ME (2009) Poly-paraphyly of Hirudinidae: many lineages of medicinal leeches. *BMC evolutionary Biology*, **9**, 246.

Piñol J, Mir G, Gomez-Polo P, Agustí N (2014) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, **15**, 819–830.

Poinar HN (1998) Molecular Coproscopy: Dung and Diet of the Extinct Ground Sloth Nothrotheriops shastensis. *Science*, **281**, 402–406.

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.

Porter TM, Golding GB, King C *et al.* (2013) Amplicon pyrosequencing late Pleistocene permafrost: the removal of putative contaminant sequences and small-scale reproducibility. *Molecular Ecology Resources*, **13**, 798–810.

Possingham HP, Wintle BA, Fuller RA, Joseph LN (2012) The conservation return on investment from ecological monitoring (HP Possingham, P Gibbons, Eds,). *Biodiversity Monitoring in Australia*, 49–58.

Riaz T, Shehzad W, Viari A *et al.* (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, **39**, e145.

Roggenbuck M, Schnell IB, Blom N *et al.* (2014) The microbiome of New World vultures. *Nature Communications*, **5**.

Romanowski G, Lorenz MG, Wackernagel W (1993) Use of polymerase chain reaction and electroporation of Escherichia coli to monitor the persistence of extracellular plasmid DNA introduced into natural soils. *Applied and Environmental Microbiology*, **59**, 3438–3446.

Sawyer RT (1986) *Leech Biology and Behaviour*. Clarendon Press, Oxford.

Schipper J, Chanson JS, Chiozza F *et al.* (2008) The status of the world's land and marine mammals: Diversity, threat, and knowledge. *Science*, **322**, 225–230.

Schnell IB, Bohmann K, Gilbert MTP (2015a) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, **15**, 1289-1303.

Schnell IB, Sollmann R, Calvignac-Spencer S *et al.* (2015b) iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool - prospects, pitfalls and avenues to be developed. *Frontiers in Zoology*, **12**, 24.

Schnell IB, Thomsen PF, Wilkinson N *et al.* (2012) Screening mammal biodiversity using DNA from leeches. *Current biology*, **22**, 262–263.

Schubert G, Stockhausen M, Hoffmann C *et al.* (2014) Targeted detection of mammalian species using carrion fly-derived DNA. *Molecular Ecology Resources*, **15**, 285-294.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic biology*, *51*(3), 492-508.

Somervuo P, Yu DW, Xu C *et al.* (2016) Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *bioRχiv* preprint: doi: http://dx.doi.org/10.1101/070573

Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.

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

Taylor PG (1996) Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular biology and evolution*, **13**, 283–285.

Tessler M, Barrio A, Borda E *et al.* (2016) Description of a soft-bodied invertebrate with microcomputed tomography and revision of the genus Chtonobdella (Hirudinea: Haemadipsidae). *Zoologica Scripta*, **5**, 552-565.

Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples – a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, **5**, 12.

Wagner A, Blackstone N, Cartwright P, Dick M, Misof B (1994) Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. *Systematic Biology*, **43(2)**, 250-261.

Zeale MRK, Butlin RK, Barker G, Lees DC, Jones G (2010) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, **11**, 236–244.

Zepeda-Mendoza ML, Bohmann K, Carmona Baez A, Gilbert MTP (2016) DAMe: a toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC research notes*, **9**, 255.

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

Data Accessibility:

Data and inputfiles: Dryad doi:10.5061/dryad.5mr8v1v

Table 1 Collection sites and number of leeches analysed from the five different geographical regions. * Analysed in two separate datasets

Geographic region	Collection sites	No. individual leeches analysed
Madagascar	Analamazaotra Forest Station Ranomafana National Park	815
Peninsular Malaysia	Krau Wildlife Reserve, Pahang	287
Laos/Vietnam*	Thua Thien Hue and Quang Nam Saola Reserves, Vietnam Nam Kading National Protected Area, Laos Xe Xap National Protected Area, Laos	1623
Malaysian Borneo* (Sabah)	Danum Valley Conservation Area, Sabah Maliau Basin Conservation Park, Sabah Sandakan Rainforest Discovery Centre, Sabah Kinabalu National Park, Sabah	510
Australia	Atherton Tablelands (unspecified locations), QLD Dinden NP, QLD Lake Barrine (Crater Lakes NP), QLD Lumholtz Lodge, QLD Mission Beach, QLD Mt. Hypipamee NP, QLD Paluma Range, QLD Wooroonooran NP/Mt. Bartle Frere, QLD Royal NP, NSW Washpool NP, NSW Waldheim, Cradle Mountain NP, TAS	192

Table 2 List of oligonucleotides used in this study, their main target organisms and amplicon length (excluding primer sequences). Original sources of the primers are given in the main text. (*Amplicon length varies between genera).

Primerset	**Main target**	**Forward (5’ - 3’)**	**Reverse (5’ - 3’)**	**App. Length***
12SA/12SO	Vertebrate 12S	CTGGGATTAGATACCCCACTAT	GTCGATTATAGGACAGGTTCCTCTA	80-110 bp
16Smam1/mam2	Mammal 16S	CGGTTGGGGTGACCTCGGA	GCTGTTATCCCTAGGGTAACT	85-115 bp
MZArtF/MZArtR	Arthropod CO1	AGATATTGGAACWTTATATTTTATTTTTGG	WACTAATCAATTWCCAAATCCTCC	157 bp
LCO1490/HCO2198	Invertebrate CO1	GGTCAACAAATCATAAAGATATTGG	TAAACTTCAGGGTGACCAAAAAATCA	660 bp
Human blocking probe (16Smam_blkhum2)		GCGACCTCGGAGCAGAACCC		

Figure 1 Map of collection sites in this study. Blue dots represents collection sites where duognathous leeches (two-jawed) are expected, and red dots are sites where trignathous (three-jawed) leech species are expected.

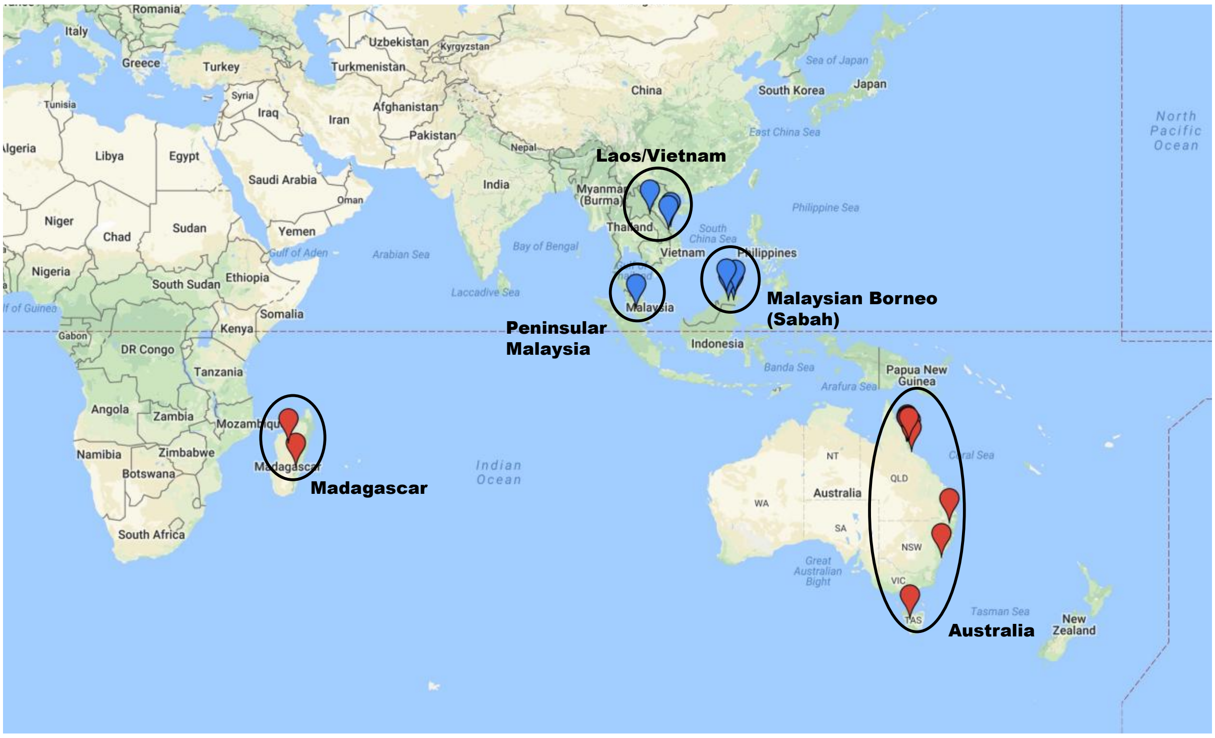


Figure 2 Workflow from collection to BLAST result including Collection (1), Digestion (2), Purification of leech-pools (3), PCR – Library Build – Sequencing - workflow for Ion Torrent (4a - Blue) and Illumina MiSeq (4b - Red), Sequence sorting (5) and BLASTn (6).

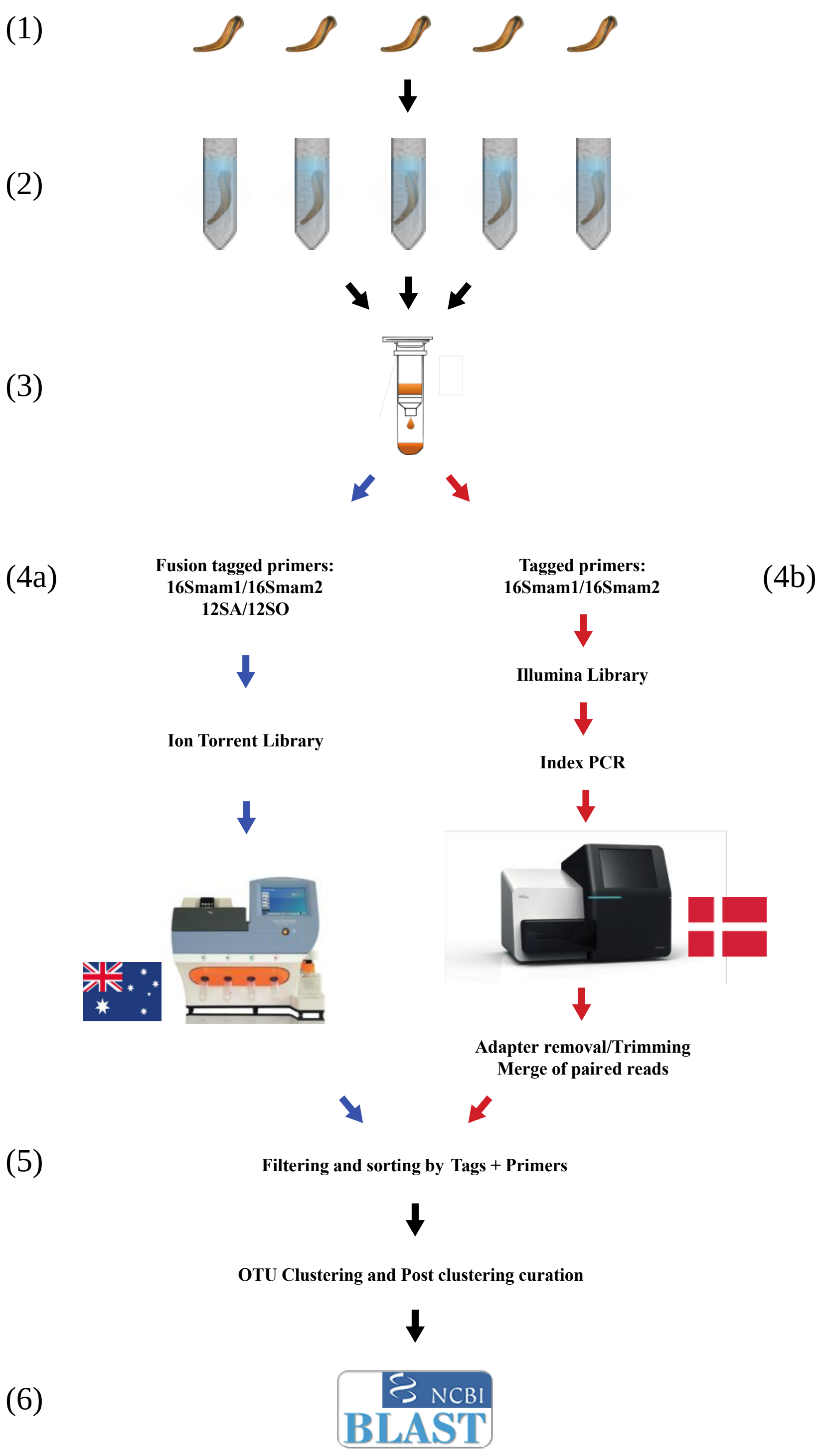

(1)
(2)
(3)
(4a)
Fusion tagged primers: 16Smam1/16Smam2 12SA/12SO
Ion Torrent Library
Tagged primers: 16Smam1/16Smam2
(4b)
Illumina Library
Index PCR
Adapter removal/Trimming Merge of paired reads
(5)
Filtering and sorting by Tags + Primers
OTU Clustering and Post clustering curation
(6)
NCBI BLAST

Figure 3. Heat map visualising numbers of vertebrate OTUs in the analysed leeches. Data are split between the five different geographical regions and presented on taxonomic family level (for detailed list of identified taxa see Supporting Information Table S3).

(*Leeches from Australia were amplified using both 'mammalian specific' and general vertebrate primers, thus observation of a higher diversity on non-mammalian classes may be expected)

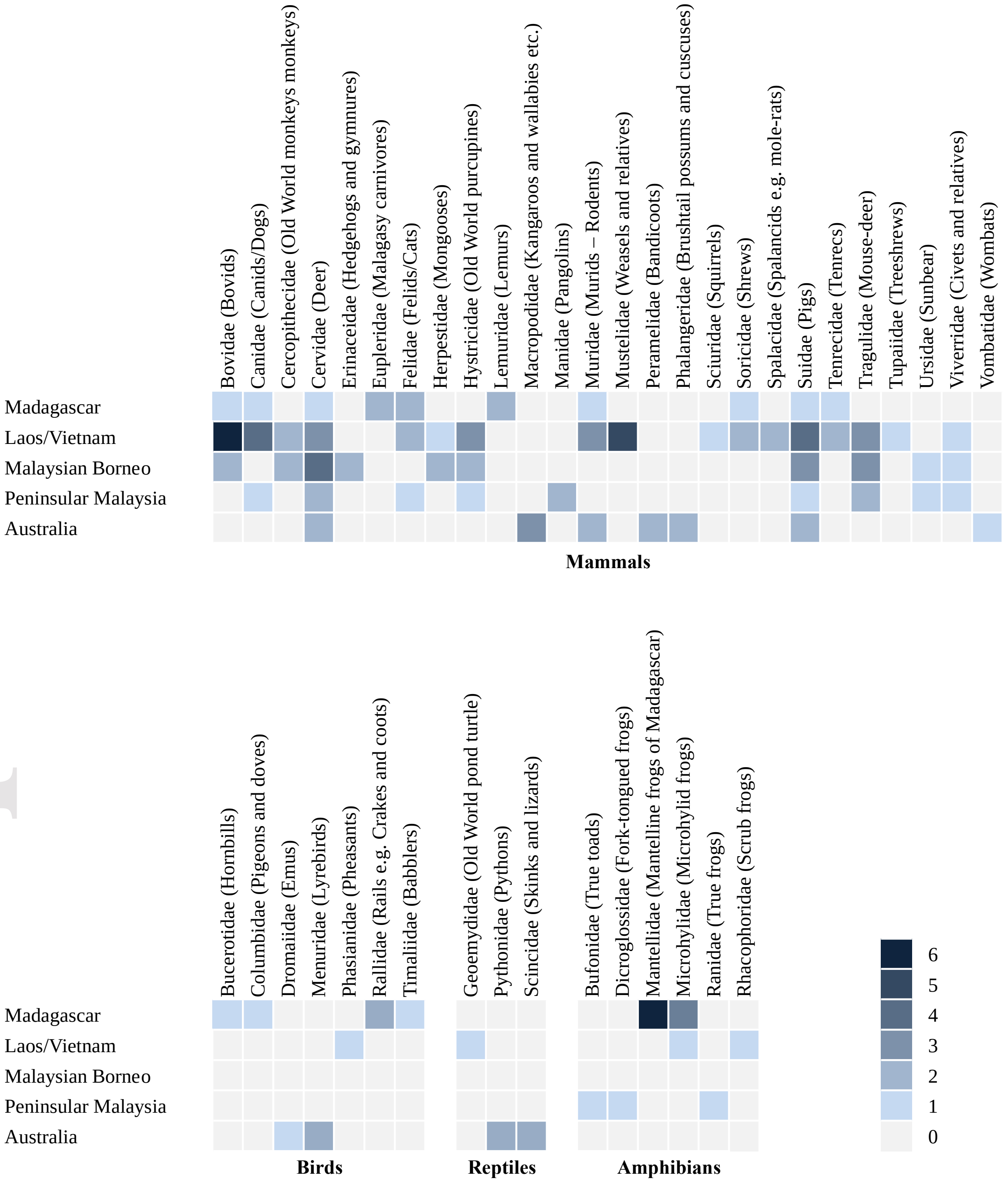


Figure 4.

Identified leech taxa presented as a phylogenetic tree using both *Hirudo medicinalis* and *Macrobdella decora* as outgroups. OTU cluster centers (Ion Torrent NGS reads, Australia) and sequences from analysed leech-pools (Sanger sequenced DNA, SE Asia and Madagascar) were globally aligned with other available CO1 data for haemadipsid leeches (Borda *et al.* 2010; Schnell *et al.* 2012; Tessler *et al.* 2016). Support values are the obtained using the Approximately Unbiased Test (Shimodaira, H. 2002).

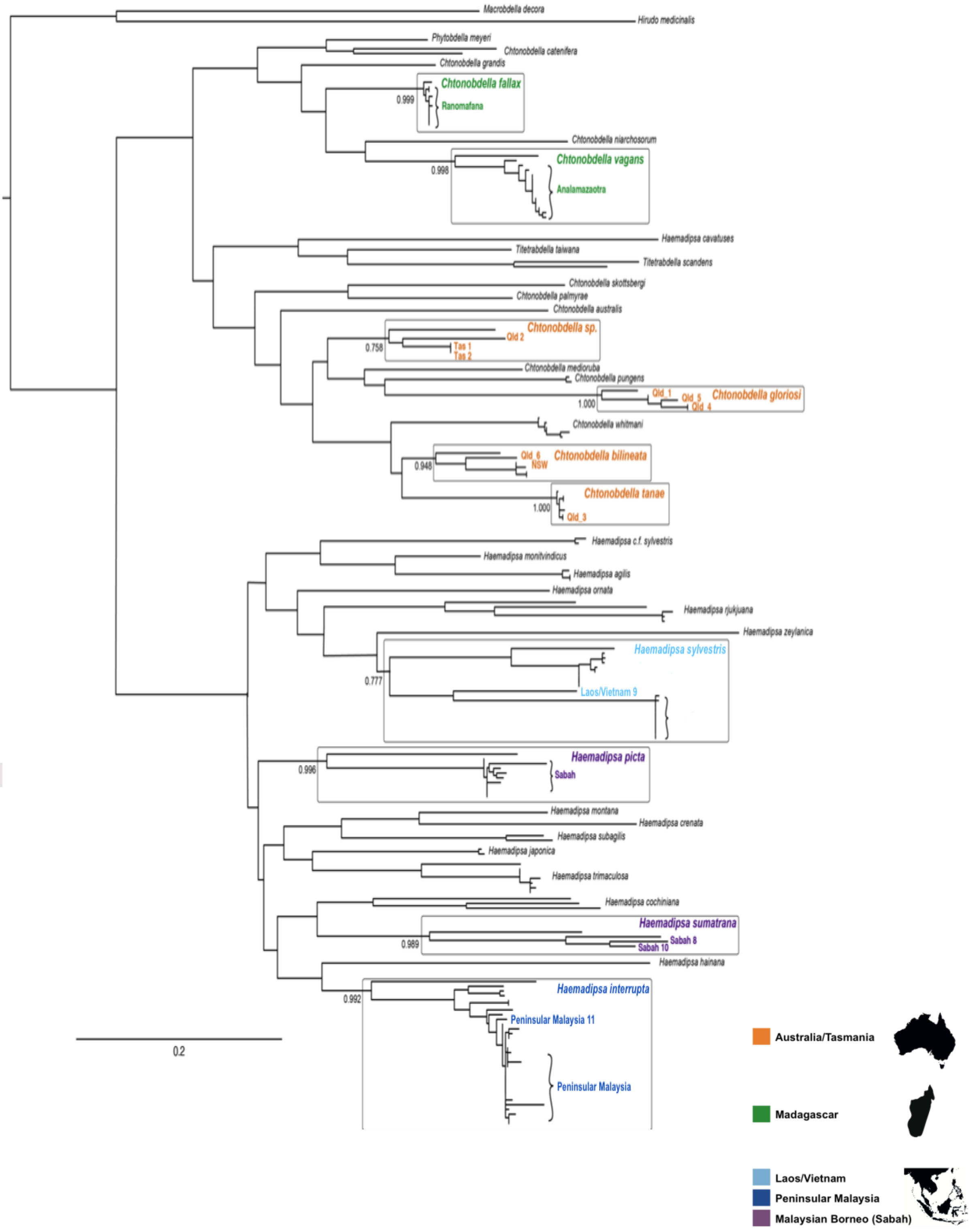
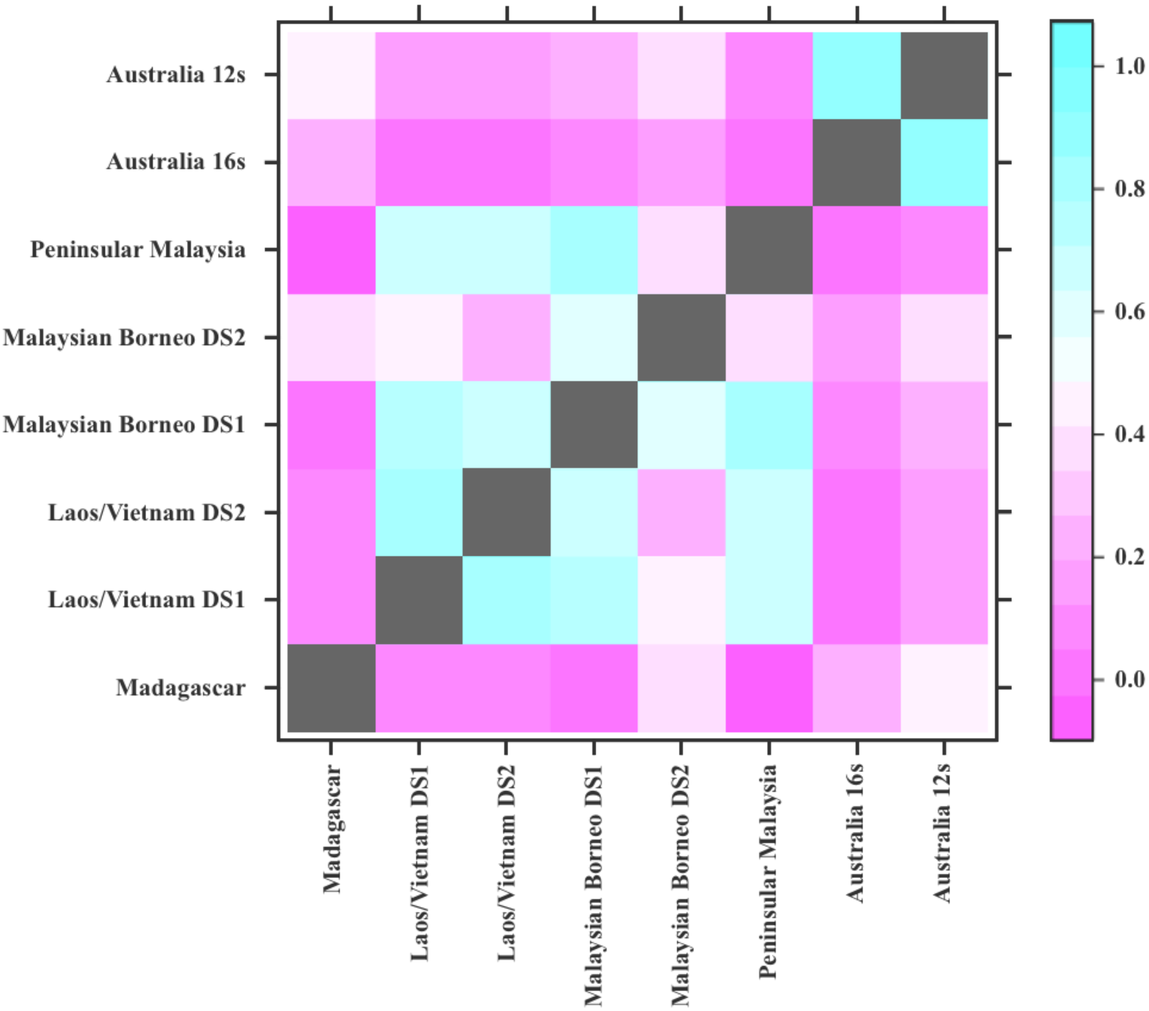


Figure 5. Similarity between identified vertebrate families and the number of leech pools in which these are found across the different datasets.


Australia 12s
Australia 16s
Peninsular Malaysia
Malaysian Borneo DS2
Malaysian Borneo DS1
Laos/Vietnam DS2
Laos/Vietnam DS1
Madagascar
Madagascar
Laos/Vietnam DS1
Laos/Vietnam DS2
Malaysian Borneo DS1
Malaysian Borneo DS2
Peninsular Malaysia
Australia 16s
Australia 12s