Final report on **GENETIC SCREENING OF HAEMATOPHAGOUS LEECHES IN ORDER TO SUPPORT EFFECTIVE CONSERVATION OF THREATENED UNGULATES AND OTHER MAMMALS WITHIN THE CENTRAL ANNAMITE LANDSCAPE OF VIETNAM AND LAOS,** AGREEMENT NUMBERS: HZ-26 with WWF and 5009-0078 with GWC

Prof. Doug W. Yu, University of East Anglia and Kunming Institute of Zoology
& Dr Yinqiu Ji, Kunming Institute of Zoology

Date: 25 June 2018

Summary: The Ecology, Conservation And Environment Center (ECEC) at the Kunming Institute of Zoology (KIZ) has carried out metabarcoding analysis of 682 leech samples collected in six nature reserves in Vietnam and Laos: Quang Nam Saola Nature Reserve (QNSL), Thua Thien Hue Saola Nature Reserve (HSL), Bach Ma National Park (BM), Xe Sap National Protected Area (XS), Laving Lavern National Protected Area (LL), and Phou Si Thone Endangered Species Conservation Area (PST). Samples were collected over the period 2012-2015.

542 samples produced a PCR product. After taxonomic assignment using the new Protax pipeline, which we and our colleagues have spent the last two years implementing, we conclude that we have detected a total of 78 vertebrate species, of which 20 are birds, 4 are frogs, 1 is a bat, and 54 are non-volant mammals. Within the 54 non-volant mammal species, we detect humans (*Homo sapiens*), domestic dog (*Canis lupus familiaris*), and domestic cow (*Bos taurus*). The remaining mammal species are a diverse lot of wild species, spanning rodents, squirrels, gymnures, and shrews on one end of the size spectrum, through to macaques (*Macaca* spp.), several medium-sized carnivores (Felidae, Herpestidae, Mustelidae, Viverridae), serow (*Capricornis milneedwardsii*), sambar (*Rusa unicolor*), muntjacs (*Muntiacus vuquangensis*, *M. truongsonensis*, and *M. vaginalis / muntjak*), pig (*Sus scrofa*, probably wild), and Asiatic black bear (*Ursus thibetanus*) on the other end.

When we compare mammal communities of the four contiguous reserves (XS, HSL, QNSL, BM), we observe that Bach Ma (and to a lesser extent, Thua Thien Hue) are characterised by a relatively higher prevalence (detections) of humans, dogs, and cows, and a lower prevalence of most other mammal species. There is also a secondary gradient with Thua Thien Hue and Quang Nam on one end, and Bach Ma and Xe Sap on the other. Thua Thien Hue and Quang Nam appear to have a relatively higher prevalence of most wild mammal

species. This gradient could reflect snare removal effort but, as there was variation in the way samples were collected, further work will be needed to confirm or reject this hypothesis. Surprisingly, the presence of a few large-bodied species that are believed to be vulnerable to hunting pressure (*Rusa unicolor, Ursus thibetanus*, *Muntiacus vuquangensis*, and *Macaca arctoides*) are not obviously negatively correlated with human presence. The mammal community at Phou Si Thone is compositionally similar to Bach Ma (including high human prevalence), while Laving Lavern is compositionally most distinct from the other five reserves. These comparisons of mammal communities between reserves must be taken as preliminary, given that sample collection did not follow a single consistent protocol across reserves. Further analysis should be conducted to separate out samples collected using different protocols.

In conclusion, now that the taxonomic assignment problem appears to be largely solved, leech surveys show good potential as a method for assessing the performance of nature reserves.

Introduction

Under the terms of Agreements HZ-26 with WWF and 5009-0078 with GWC, the ECEC-KIZ were contracted to:

i. Use the methodologies based on Schnell *et al*., 2012 to extract and amplify mammalian DNA from all samples sent by WWF, WCS, and GWC, suitably modified to take advantage of high-throughput sequencing via the Illumina HiSeqs 2000/2500.

ii. Compare amplified DNA sequences from each sample to GenBank and other appropriate sources to identify all mammalian species present within each sample.

iii. Write a formal report on the methods used and analytical results.

Methods outline

Laboratory. - 1172 samples (tubes) of leeches preserved in RNALater were received from WWF, WCS, and GWC and to reduce costs, were pooled into 628 samples based on collecting information. We extracted DNA from each sample, and used the *16smamFR* primers also used by Schnell et al. (2012) to amplify a ~90 bp portion of the mitochondrial 16S (lrRNA) gene, which is designed to amplify all mammals (although there is unavoidable bias towards certain taxa). 542 samples were successful in PCR and were paired with 547 negative-control samples for sequencing, to check for sample cross-contamination, which can happen during DNA extraction and PCR.

Bioinformatics. - After sequencing, the raw sequence reads were denoised, demultiplexed into their respective samples, checked for chimeras using *uchime* (Edgar et al. 2011), and the 151,815,573 final sequences were clustered into 2133 OTUs (Operational Taxonomic Units) using *swarm*'s d = 1 and -f options (Mahé et al. 2015). We then used BLAST to filter in only Chordata-assigned OTUs, leaving 1718 OTUs. We then used *lulu* (Frøslev et al. 2017) to collapse OTUs that are likely from the same species. *lulu* outputs a representative sequence for each OTU, which was used as the representative sequence for each OTU.

Protax taxonomic assignment. - A particularly difficult challenge with *16smam*-amplified iDNA is that the OTU representative sequences are very short: ~90 bp. Even a two-to-three-nucleotide difference results in a <98-97% sequence similarity, In longer barcoding genes, two sequences that have < 97% sequence similarity are usually taken as distinct species. Since PCR + sequencing errors can easily introduce 2 or 3 errors in some of the hundreds of millions of reads that are produced in a typical sequencing run, sequences from the same species end up clustered into multiple OTUs, assigned taxonomically to different species. On top of this, sequence reference databases are incomplete, in both senses: missing species altogether and missing sequence variants for any given species. As a result, assignments are typically biased in favour of assigning to species that happen to be present in the reference base. The Protax protocol (Somervuo et al. 2016, 2017) removes this bias by taking into account the possibility of absent references, which we do by giving Protax a complete taxonomic list of the focal group, such as the mammals. Mammals and birds are ideal for Protax, because the Linnaean taxonomy is largely complete. Reference databases also contain misnamed sequences, so it is necessary to curate a reliable sequence database. Starting in 2016, Charles Xu in my lab first applied the new Protax scripts to mammals, and in 2017 and 2018, Alex Crampton-Platt (Wilting lab and NatureMetrics) finished the job. The resulting scripts and protocol are now posted on bioRxiv (http://biorxiv.org/lookup/doi/10.1101/345082) and https://github.com/alexcrampton-platt/screenforbio-mbc.

We used Protax to assign taxonomies to each of the 1718 OTUs, using a tetrapod-only reference database and a list of mammal species from the Annamites (**Weighted species for Vietnam.txt**). Any given OTU had a 90% prior probability of belonging to one of the species on this list and a 10% prior probability of coming from some other source. The 1718 OTU representative sequences are in the file: **all_2015WWFWCS_otu_table_swarm_lulu_vert16S_20180404.fas**. Of the 1718 OTUs, Protax was able to assign 1282 to a tetrapod genus with probability > 0. This number includes assignments to Protax-designated ‘unknown’ genera, which are genera that potentially are missing from the reference database. The other 436 OTUs are likely to be a combination of error-ridden sequences, non-target amplifications from nuclear genomes (known as Numts or nuclear mitochondrial DNA segments), and non-tetrapod taxa, which are not in the reference database.

I then visually inspected the Protax taxonomies and the distribution of OTU reads over the 542 samples and made a final determination, collapsing the 1282 OTUS to 78 vertebrate species, of which 55 species are mammals (1 bat and 54 non-volant mammals), our target taxon. This process is documented in the Excel spreadsheet (**analysis/leechotu_protax.xlsx**) in the column **final.taxonomy**. OTUs that Protax could not assign to a species (including to “unknown” species) were omitted. Note that of the 1282 OTUs, 650 were made up of 10 sequences or less. Such low-read-number OTUs are of course likely to have been split off from high-read-number OTUs because of sequencing and PCR errors, and it is gratifying that most of the low-read number OTUs were given the same taxonomic assignments as one of the high-read-number OTUs. Also, in previous analyses, many of the OTUs had been assigned (by other assignment methods) to seals and sea lions, an obvious error. This no longer occurred after we switched to Protax.

Statistical analysis. - I used the R statistical environment to collapse OTUs with the same Protax taxonomies, and to combine the OTU taxonomies with the OTU table (sample X OTU table**: 2015WWFWCS_swarm_lulu_otu_table_20180404.xlsx**) and the sample information (nature reserve names and metadata**: allWWFWCS_sample_codes_20180621.xlsx**). I removed the negative-control samples, removed the 24 OTUs assigned to birds and frogs, and removed four samples that had only 1 species (all four of which were found in multiple other samples: *Sus scrofa*, *Mustela kathiah*, *Homo sapiens*, and *Niviventer* UNK), because such samples result in outlier points in ordination graphs. Our PCR primers were not designed for birds and amphibians, and so these taxa are more likely to be a biased subset of the non-mammal taxa in the leeches. To study these vertebrates, we would need to use a different PCR primer. Finally, with the remaining 538 samples, I generated summary tables and carried out a simple community analysis.

Further details on laboratory, sequencing, bioinformatic, and statistical methods are in the **Appendix** to this document. All scripts, input data, and generated tables and figures are provided as supplementary files in this report and can also be downloaded from **https://github.com/dougwyu/VietnamLaosLeeches**.

Results

Species list and taxonomic assignments. – In Table 1A, I list the 54 mammal species, their Protax-assigned assignment probabilities at each taxonomic rank, their incidences (number of samples in which the species was detected), and their relative incidences (incidence in each reserve / all reserves). Protax outputs are more useful than other assignment methods for two reasons visible here: each taxonomic rank is given an estimated probability of being correct, and Protax can infer the presence of taxa that are missing from the reference database (either because the species is known but its sequence is missing or because the species is not known to science at all). These inferred missing taxa are marked as UNK in the tables. Protax provides an overall estimate of how well or poorly it can assign sequences to order, family, genus, and species level (Fig. 1), almost all (>95%) probabilities at the order and family level, but 78.2% and 64.3% assignments at the genus and species levels. This estimate reflects the amount of useful taxonomic information in the reference database, so if some species in a genus are very similar, but all members of that genus are distinct from all other genera, Protax will assign a lower probability at the species level but a higher probability at the genus level. We see from Fig. 1, that the 16Smam marker is reliable down to family level, but genus and species level assignments should ideally be bolstered by other information, especially in the case of certain species (e.g. *Neofelis nebulosa*, *Lutra lutra*, *Catopuma temmincki*) whose presence could trigger management intervention.

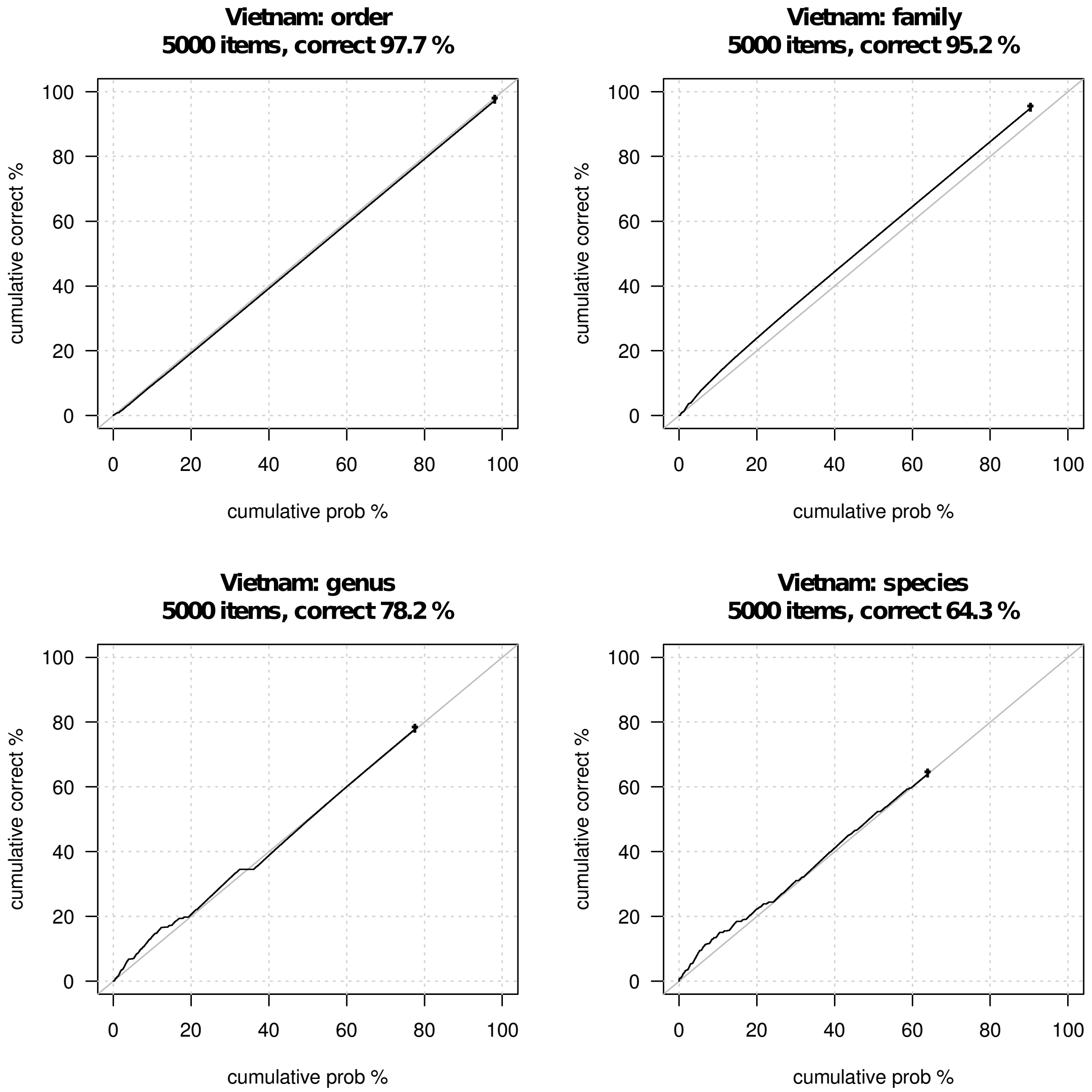


Figure 1. Protax estimates of assignment accuracy over the entire tetrapod database.

In Table 1A, the Order and Family-level assignment probabilities are mostly > 90% probability, consistent with the Protax self-assessment, except for several Carnivora taxa and the mole *Euroscaptor* UNK. We have observed that Carnivores are difficult to identify using the *16Smam* marker, and previous assignment methods have assigned some of these OTUs to seals and sea lions, which is obviously wrong. The long-term solutions to this problem will be to try amplifying other genetic ‘marker’ sequences (especially the genes *cytB* and *12S*) from the same samples, to continue populating the reference databases with more sequences from more species and more individuals per species, and to bolster these identifications with other information sources, such as camera traps (CT species lists and/or expert opinion could then be used to weight Protax assignments by reserve).

For example, we can conclude that we detected *Nomascus* in one sample (98.2% probability), but the species level assignment is mildly uncertain (76.3%). In contrast, the four macaque species are all assigned at probabilities ≥ 96%. The genus assignments of *Catopuma* and *Neofelis* are assigned at low probabilities (68.0% and 39.7%, respectively), reflecting the general difficulty of assigning felid species with this gene sequence. This information can now be used to target CT surveys for the gibbon and cats, and when particular species are known (or later identified) for a reserve, they can be added to the weighted list for that reserve for another round of assignments and future surveys.

I do not provide a probability of assignment for cow (*Bos taurus*). This is because the two OTUs making up this taxon were originally assigned by Protax to an unknown species of *Pseudoryx* UNK, but at a low probability for both genus and species (42.7% for both ranks). Given that this is a monospecific genus and that we do have the reference sequence for *Pseudoryx*, I used two other methods to assign taxonomies to these OTUs and got back *Bos taurus*, with 100% identity. The reason for the mis-assignment is that we had included *Pseudoryx nghetinhensis* in the weighted list for the Annamites, and saola and cow are closely related, so Protax ended up concluding that there was a chance that the sequences were not cow but not clearly saola either. In short, the lesson is that very rare species should probably *not* be included in the weighted list.

Finally, in Table 1B, I sort the table to show that Bach Ma has a higher detection rate for humans, dogs and cows, but a lower detection rate for many wild mammal species.

The Excel versions of these tables are in **analysis/reservecodes_by_OTU_samplecount.xlsx**.

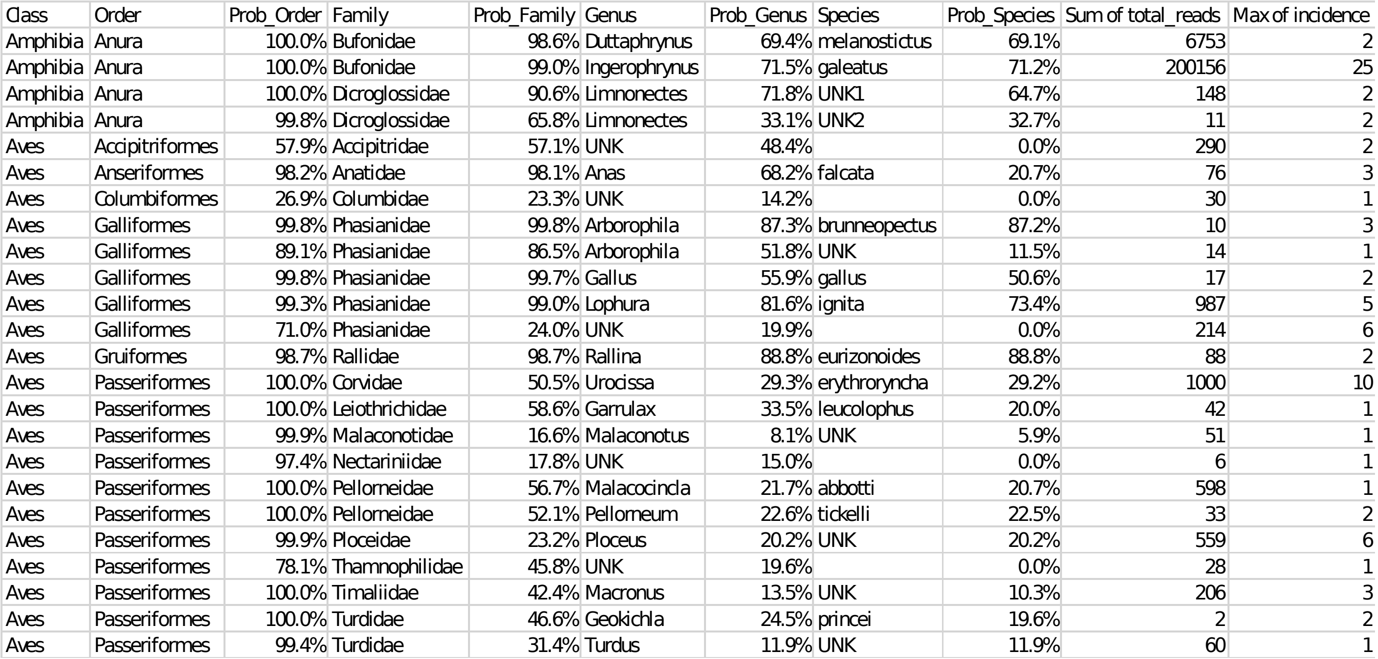
Table 1A. List of 54 mammal species detected in the six reserves.

Order	Prob_Order	Family	Prob_Family	Genus	Prob_Genus	Species	Prob_Species	Laos-LL	Laos-PST	Laos-XS	Vietnam-BM	Vietnam-HSL	Vietnam-QNSL
Artiodactyla	see legend	Bovidae	see text	*Bos*	see text	*taurus*	see text	0	4	1	3	6	9
Artiodactyla	100.0%	Bovidae	96.3%	*Capricornis*	94.0%	*milneedwardsii*	94.0%	15	14	46	11	117	122
Artiodactyla	100.0%	Cervidae	93.2%	*Muntiacus*	91.3%	*muntjak_vaginalis*	83.4%	22	20	4	1	62	87
Artiodactyla	100.0%	Cervidae	97.7%	*Muntiacus*	97.0%	*truongsonensis*	50.4%	20	15	13	10	57	118
Artiodactyla	100.0%	Cervidae	98.0%	*Muntiacus*	96.7%	UNK	93.6%	20	13	18	6	72	85
Artiodactyla	100.0%	Cervidae	96.2%	*Muntiacus*	95.7%	*vuquangensis*	95.5%	23	1	0	2	6	2
Artiodactyla	100.0%	Cervidae	99.4%	*Rusa*	93.5%	*unicolor*	93.4%	25	0	2	2	39	4
Artiodactyla	100.0%	Suidae	99.8%	*Sus*	99.5%	*scrofa*	99.1%	28	59	51	21	154	179
Artiodactyla	100.0%	Tragulidae	99.1%	*Tragulus*	98.3%	*kanchil*	95.9%	1	0	0	0	0	0
Artiodactyla	100.0%	Tragulidae	97.9%	*Tragulus*	95.4%	UNK	84.4%	0	0	0	1	3	0
Carnivora	100.0%	Canidae	63.1%	*Canis*	43.9%	*lupus_familiaris*	42.5%	2	2	7	7	19	34
Carnivora	99.9%	Felidae	73.6%	*Catopuma*	68.0%	UNK	24.1%	1	0	0	0	0	0
Carnivora	100.0%	Felidae	97.9%	*Neofelis*	39.7%	UNK	39.5%	2	5	1	2	9	35
Carnivora	99.9%	Felidae	91.9%	*Prionailurus*	43.4%	*bengalensis*	43.2%	1	0	0	1	5	17
Carnivora	99.6%	Felidae	83.2%	*UNK*	16.9%		0.0%	0	14	27	7	49	90
Carnivora	100.0%	Herpestidae	88.1%	*Herpestes*	87.1%	UNK	84.3%	6	16	15	8	67	77
Carnivora	100.0%	Mustelidae	99.9%	*Aonyx*	80.0%	*cinereus*	79.9%	0	0	3	0	0	0
Carnivora	100.0%	Mustelidae	99.9%	*Arctonyx*	96.6%	*collaris*	96.6%	2	20	2	0	0	0
Carnivora	100.0%	Mustelidae	99.0%	*Lutra*	42.6%	*lutra*	29.5%	0	4	1	1	2	20
Carnivora	100.0%	Mustelidae	99.9%	*Martes*	95.2%	*flavigula*	95.2%	2	4	3	1	1	4
Carnivora	100.0%	Mustelidae	99.9%	*Melogale*	94.6%	*moschata*	94.6%	8	38	42	13	105	161
Carnivora	100.0%	Mustelidae	99.9%	*Mustela*	80.0%	*kathiah*	80.0%	0	3	0	1	0	5
Carnivora	100.0%	Mustelidae	97.4%	*Mustela*	59.3%	UNK	59.3%	3	10	1	0	3	2
Carnivora	100.0%	Ursidae	98.9%	*Ursus*	93.6%	*thibetanus*	93.1%	0	0	0	0	9	2
Carnivora	99.3%	Viverridae	71.6%	*Arctictis*	53.8%	UNK	27.1%	0	2	0	0	3	6
Carnivora	100.0%	Viverridae	99.8%	*Paguma*	83.7%	*larvata*	83.7%	9	26	15	4	48	81
Carnivora	100.0%	Viverridae	99.9%	*Viverra*	97.2%	UNK1	69.3%	6	2	0	0	0	0
Carnivora	99.7%	Viverridae	85.4%	*Viverra*	68.6%	UNK2	6.6%	0	0	0	0	0	4
Carnivora	99.9%	Viverridae	92.7%	*Viverra*	76.3%	UNK3	10.1%	0	0	0	0	1	0
Carnivora	99.9%	Viverridae	76.8%	*Viverra*	40.0%	UNK4	12.2%	0	0	1	2	1	2
Chiroptera	98.5%	Vespertilionidae	97.1%	*Tylonycteris*	89.5%	UNK	7.0%	0	0	0	0	0	1
Erinaceomorpha	96.7%	Erinaceidae	96.6%	*Hylomys*	77.7%	*suillus*	76.2%	0	3	0	0	0	0
Lagomorpha	100.0%	Leporidae	100.0%	*Nesolagus*	99.8%	*timminsi*	99.8%	5	4	43	17	51	152
Primates	100.0%	Cercopithecidae	99.9%	*Macaca*	98.9%	*arctoides*	95.3%	5	12	5	11	56	44
Primates	100.0%	Cercopithecidae	100.0%	*Macaca*	99.4%	*assamensis*	98.8%	2	8	2	6	17	27
Primates	100.0%	Cercopithecidae	99.9%	*Macaca*	99.0%	*fascicularis*	96.3%	0	0	0	0	0	0
Primates	100.0%	Cercopithecidae	99.9%	*Macaca*	99.4%	*mulatta*	98.7%	0	4	0	0	6	0
Primates	100.0%	Cercopithecidae	99.7%	*Macaca*	97.0%	UNK	97.0%	4	7	4	8	33	34
Primates	100.0%	Hominidae	99.5%	*Homo*	99.1%	*sapiens*	99.1%	11	44	23	13	74	45
Primates	100.0%	Hylobatidae	99.3%	*Nomascus*	98.9%	*siki*	76.3%	0	1	0	0	0	0
Rodentia	100.0%	Hystricidae	99.7%	*Hystrix*	98.0%	UNK	95.9%	2	5	18	11	76	58
Rodentia	100.0%	Muridae	99.7%	*Mus*	58.2%	*musculus*	56.9%	0	0	0	0	1	0
Rodentia	100.0%	Muridae	99.4%	*Niviventer*	96.7%	UNK	96.6%	2	37	44	17	52	112
Rodentia	100.0%	Sciuridae	99.9%	*Callosciurus*	98.4%	*erythraeus*	96.2%	0	7	10	3	12	41
Rodentia	98.7%	Sciuridae	88.5%	*Callosciurus*	46.4%	UNK1	41.2%	0	6	0	1	0	0
Rodentia	100.0%	Sciuridae	99.9%	*Callosciurus*	86.1%	UNK2	8.5%	0	0	0	0	0	3
Rodentia	100.0%	Sciuridae	99.8%	*Hylopetes*	93.6%	UNK1	93.6%	0	0	1	0	0	1
Rodentia	99.9%	Sciuridae	98.7%	*Hylopetes*	86.6%	UNK2	9.9%	0	0	0	0	0	0
Rodentia	100.0%	Sciuridae	99.8%	*Petaurista*	91.5%	*philippensis*	90.6%	0	2	0	0	0	0
Rodentia	100.0%	Sciuridae	99.8%	*Ratufa*	92.1%	*bicolor*	84.7%	0	0	1	1	0	0
Rodentia	100.0%	Sciuridae	99.9%	*Tamiops*	80.2%	UNK	73.5%	1	37	3	2	7	32
Rodentia	100.0%	Spalacidae	99.9%	*Rhizomys*	99.6%	*pruinosus*	99.6%	3	9	18	4	30	109
Scandentia	100.0%	Tupaiidae	99.9%	*Tupaia*	97.8%	*belangeri*	89.5%	0	13	15	4	14	66
Soricomorpha	66.4%	Talpidae	50.2%	*Euroscaptor*	34.4%	UNK	7.4%	0	1	0	0	0	0
							Number of samples	28	63	74	24	160	189
							Species observed	28	37	32	33	36	37

Table 1B. Same list as in Table 1A but sorted so that humans, dogs, and cows are at the top of the list. I have also added new columns that calculate the percentage of samples per reserve in which each species is detected. Finally, I have formatted the four contiguous reserves (XS, BM, HSL, and QNSL) to highlight the larger numbers in each species (row). For instance, humans, dogs, and cows are detected more often in Bach Ma, but most other species are detected more often in Quang Nam. This analysis is limited to species that are detected ≥10 times.

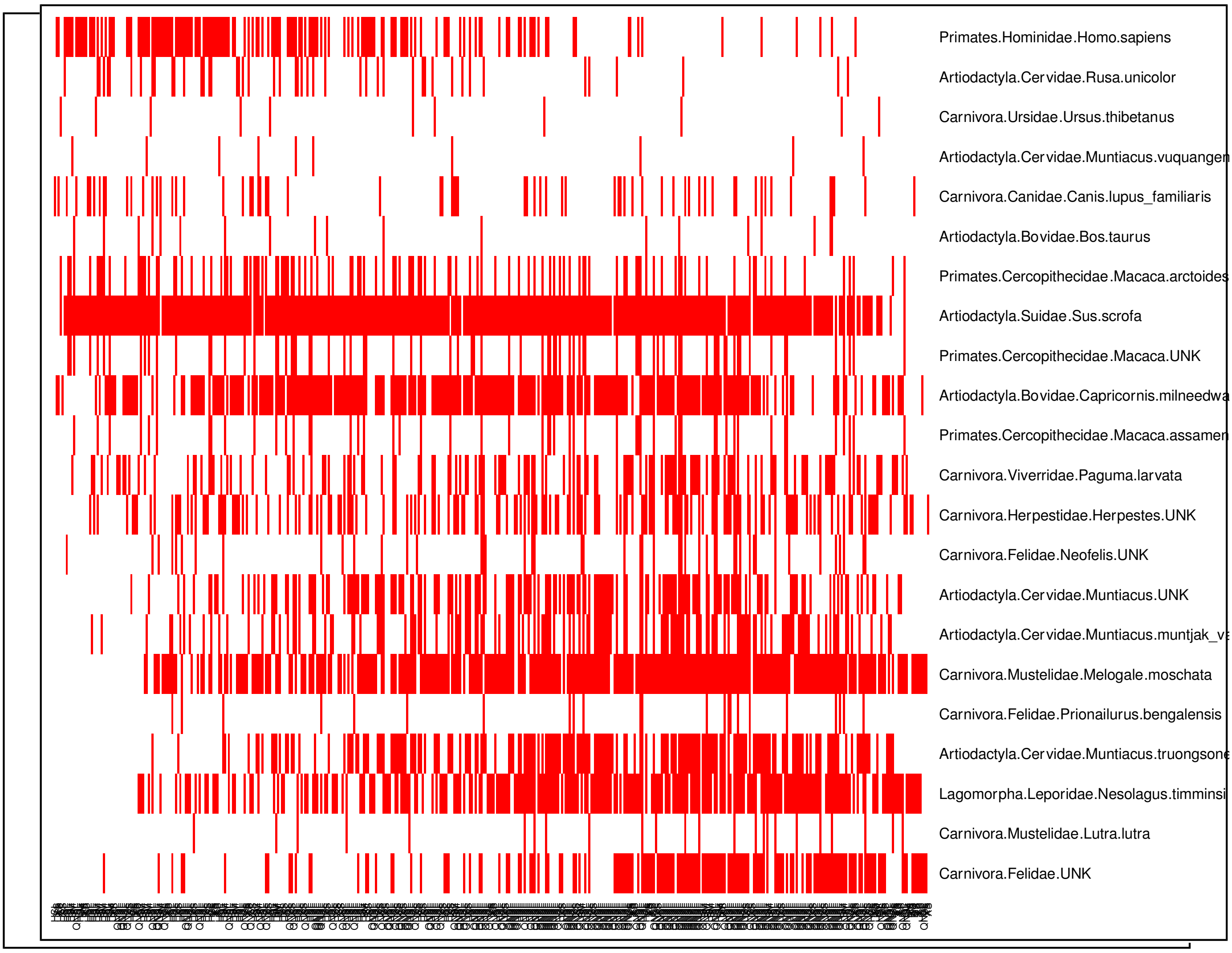
Order	Prob_Order	Family	Prob_Family	Genus	Prob_Genus	Species	Prob_Species	Laos-LL	Laos-PST	Laos-XS	Vietnam-BM	Vietnam-HSL	Vietnam-QNSL	Laos-LL	Laos-PST	Laos-XS	Vietnam-BM	Vietnam-HSL	Vietnam-QNSL
Primates	100.0%	Hominidae	99.5%	*Homo*	99.1%	*sapiens*	99.1%	11	44	23	13	74	45	39%	70%	31%	54%	46%	24%
Carnivora	100.0%	Canidae	63.1%	*Canis*	43.9%	*lupus_familiaris*	42.5%	2	2	7	7	19	34	7%	3%	9%	29%	12%	18%
Artiodactyla	see legend	Bovidae	see legend	*Bos*	see legend	*taurus*	see legend	0	4	1	3	6	9	0%	6%	1%	13%	4%	5%
Artiodactyla	100.0%	Cervidae	99.4%	*Rusa*	93.5%	*unicolor*	93.4%	25	0	2	2	39	4	89%	0%	3%	8%	24%	2%
Primates	100.0%	Cercopithecidae	100.0%	*Macaca*	99.4%	*assamensis*	98.8%	2	8	2	6	17	27	7%	13%	3%	25%	11%	14%
Rodentia	100.0%	Sciuridae	99.9%	*Tamiops*	80.2%	UNK	73.5%	1	37	3	2	7	32	4%	59%	4%	8%	4%	17%
Primates	100.0%	Cercopithecidae	99.7%	*Macaca*	97.0%	UNK	97.0%	4	7	4	8	33	34	14%	11%	5%	33%	21%	18%
Carnivora	100.0%	Felidae	97.9%	*Neofelis*	39.7%	UNK	39.5%	2	5	1	2	9	35	7%	8%	1%	8%	6%	19%
Rodentia	100.0%	Sciuridae	99.9%	*Callosciurus*	98.4%	*erythraeus*	96.2%	0	7	10	3	12	41	0%	11%	14%	13%	8%	22%
Primates	100.0%	Cercopithecidae	99.9%	*Macaca*	98.9%	*arctoides*	95.3%	5	12	5	11	56	44	18%	19%	7%	46%	35%	23%
Rodentia	100.0%	Hystricidae	99.7%	*Hystrix*	98.0%	UNK	95.9%	2	5	18	11	76	58	7%	8%	24%	46%	48%	31%
Scandentia	100.0%	Tupaiidae	99.9%	*Tupaia*	97.8%	*belangeri*	89.5%	0	13	15	4	14	66	0%	21%	20%	17%	9%	35%
Carnivora	100.0%	Herpestidae	88.1%	*Herpestes*	87.1%	UNK	84.3%	6	16	15	8	67	77	21%	25%	20%	33%	42%	41%
Carnivora	100.0%	Viverridae	99.8%	*Paguma*	83.7%	*larvata*	83.7%	9	26	15	4	48	81	32%	41%	20%	17%	30%	43%
Artiodactyla	100.0%	Cervidae	98.0%	*Muntiacus*	96.7%	UNK	93.6%	20	13	18	6	72	85	71%	21%	24%	25%	45%	45%
Artiodactyla	100.0%	Cervidae	93.2%	*Muntiacus*	91.3%	*muntjak_vaginalis*	83.4%	22	20	4	1	62	87	79%	32%	5%	4%	39%	46%
Carnivora	99.6%	Felidae	83.2%	*UNK*	16.9%		0.0%	0	14	27	7	49	90	0%	22%	36%	29%	31%	48%
Rodentia	100.0%	Spalacidae	99.9%	*Rhizomys*	99.6%	*pruinosus*	99.6%	3	9	18	4	30	109	11%	14%	24%	17%	19%	58%
Rodentia	100.0%	Muridae	99.4%	*Niviventer*	96.7%	UNK	96.6%	2	37	44	17	52	112	7%	59%	59%	71%	33%	59%
Artiodactyla	100.0%	Cervidae	97.7%	*Muntiacus*	97.0%	*truongsonensis*	50.4%	20	15	13	10	57	118	71%	24%	18%	42%	36%	62%
Artiodactyla	100.0%	Bovidae	96.3%	*Capricornis*	94.0%	*milneedwardsii*	94.0%	15	14	46	11	117	122	54%	22%	62%	46%	73%	65%
Lagomorpha	100.0%	Leporidae	100.0%	*Nesolagus*	99.8%	*timminsi*	99.8%	5	4	43	17	51	152	18%	6%	58%	71%	32%	80%
Carnivora	100.0%	Mustelidae	99.9%	*Melogale*	94.6%	*moschata*	94.6%	8	38	42	13	105	161	29%	60%	57%	54%	66%	85%
Artiodactyla	100.0%	Suidae	99.8%	*Sus*	99.5%	*scrofa*	99.1%	28	59	51	21	154	179	100%	94%	69%	88%	96%	95%
Artiodactyla	100.0%	Cervidae	96.2%	*Muntiacus*	95.7%	*vuquangensis*	95.5%	23	1	0	2	6	2	82%	2%	0%	8%	4%	1%
Carnivora	100.0%	Mustelidae	99.0%	*Lutra*	42.6%	*lutra*	29.5%	0	4	1	1	2	20	0%	6%	1%	4%	1%	11%
Carnivora	100.0%	Mustelidae	99.9%	*Arctonyx*	96.6%	*collaris*	96.6%	2	20	2	0	0	0	7%	32%	3%	0%	0%	0%
Carnivora	99.9%	Felidae	91.9%	*Prionailurus*	43.4%	*bengalensis*	43.2%	1	0	0	1	5	17	4%	0%	0%	4%	3%	9%
Carnivora	100.0%	Mustelidae	97.4%	*Mustela*	59.3%	UNK	59.3%	3	10	1	0	3	2	11%	16%	1%	0%	2%	1%
Carnivora	100.0%	Mustelidae	99.9%	*Martes*	95.2%	*flavigula*	95.2%	2	4	3	1	1	4	7%	6%	4%	4%	1%	2%
Carnivora	99.3%	Viverridae	71.6%	*Arctictis*	53.8%	UNK	27.1%	0	2	0	0	3	6	0%	3%	0%	0%	2%	3%
Carnivora	100.0%	Ursidae	98.9%	*Ursus*	93.6%	*thibetanus*	93.1%	0	0	0	0	9	2	0%	0%	0%	0%	6%	1%
Primates	100.0%	Cercopithecidae	99.9%	*Macaca*	99.4%	*mulatta*	98.7%	0	4	0	0	6	0	0%	6%	0%	0%	4%	0%
Carnivora	100.0%	Mustelidae	99.9%	*Mustela*	80.0%	*kathiah*	80.0%	0	3	0	1	0	5	0%	5%	0%	4%	0%	3%
Carnivora	100.0%	Viverridae	99.9%	*Viverra*	97.2%	UNK1	69.3%	6	2	0	0	0	0	21%	3%	0%	0%	0%	0%
Rodentia	98.7%	Sciuridae	88.5%	*Callosciurus*	46.4%	UNK1	41.2%	0	6	0	1	0	0	0%	10%	0%	4%	0%	0%
Carnivora	99.9%	Viverridae	76.8%	*Viverra*	40.0%	UNK4	12.2%	0	0	1	2	1	2	0%	0%	1%	8%	1%	1%
Artiodactyla	100.0%	Tragulidae	97.9%	*Tragulus*	95.4%	UNK	84.4%	0	0	0	1	3	0	0%	0%	0%	4%	2%	0%
Carnivora	99.7%	Viverridae	85.4%	*Viverra*	68.6%	UNK2	6.6%	0	0	0	0	0	4	0%	0%	0%	0%	0%	2%
Carnivora	100.0%	Mustelidae	99.9%	*Aonyx*	80.0%	*cinereus*	79.9%	0	0	3	0	0	0	0%	0%	4%	0%	0%	0%
Erinaceomorpha	96.7%	Erinaceidae	96.6%	*Hylomys*	77.7%	*suillus*	76.2%	0	3	0	0	0	0	0%	5%	0%	0%	0%	0%
Rodentia	100.0%	Sciuridae	99.9%	*Callosciurus*	86.1%	UNK2	8.5%	0	0	0	0	0	3	0%	0%	0%	0%	0%	2%
Rodentia	100.0%	Sciuridae	99.8%	*Hylopetes*	93.6%	UNK1	93.6%	0	0	1	0	0	1	0%	0%	1%	0%	0%	1%
Rodentia	100.0%	Sciuridae	99.8%	*Petaurista*	91.5%	*philippensis*	90.6%	0	2	0	0	0	0	0%	3%	0%	0%	0%	0%
Rodentia	100.0%	Sciuridae	99.8%	*Ratufa*	92.1%	*bicolor*	84.7%	0	0	1	1	0	0	0%	0%	1%	4%	0%	0%
Artiodactyla	100.0%	Tragulidae	99.1%	*Tragulus*	98.3%	*kanchil*	95.9%	1	0	0	0	0	0	4%	0%	0%	0%	0%	0%
Carnivora	99.9%	Felidae	73.6%	*Catopuma*	68.0%	UNK	24.1%	1	0	0	0	0	0	4%	0%	0%	0%	0%	0%
Carnivora	99.9%	Viverridae	92.7%	*Viverra*	76.3%	UNK3	10.1%	0	0	0	0	1	0	0%	0%	0%	0%	1%	0%
Chiroptera	98.5%	Vespertilionidae	97.1%	*Tylonycteris*	89.5%	UNK	7.0%	0	0	0	0	0	1	0%	0%	0%	0%	0%	1%
Primates	100.0%	Hylobatidae	99.3%	*Nomascus*	98.9%	*siki*	76.3%	0	1	0	0	0	0	0%	2%	0%	0%	0%	0%
Rodentia	100.0%	Muridae	99.7%	*Mus*	58.2%	*musculus*	56.9%	0	0	0	0	1	0	0%	0%	0%	0%	1%	0%
Soricomorpha	66.4%	Talpidae	50.2%	*Euroscaptor*	34.4%	UNK	7.4%	0	1	0	0	0	0	0%	2%	0%	0%	0%	0%
Primates	100.0%	Cercopithecidae	99.9%	*Macaca*	99.0%	*fascicularis*	96.3%	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%
Rodentia	99.9%	Sciuridae	98.7%	*Hylopetes*	86.6%	UNK2	9.9%	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%
							Number of samples	28	63	74	24	160	189	28	63	74	24	160	189
							Species observed	28	37	32	33	36	37						

Table 1C. List of frog and bird species detected using the mammal-targeted 16Smam primers. This list has not been vetted, but note that the rates of detection (incidence) are low for all species (≤ 10 detections out of 542 samples for all but one species). The purpose of this table is to document that the leech samples do appear to contain information on amphibians and birds, but more general PCR primers should be used to make estimates on distribution over sites.



Class	Order	Prob_Order	Family	Prob_Family	Genus	Prob_Genus	Species	Prob_Species	Sum of total_reads	Max of incidence
Amphibia	Anura	100.0%	Bufonidae	98.6%	Duttaphrynus	69.4%	melanostictus	69.1%	6753	2
Amphibia	Anura	100.0%	Bufonidae	99.0%	Ingerophrynus	71.5%	galeatus	71.2%	200156	25
Amphibia	Anura	100.0%	Dicroglossidae	90.6%	Limnonectes	71.8%	UNK1	64.7%	148	2
Amphibia	Anura	99.8%	Dicroglossidae	65.8%	Limnonectes	33.1%	UNK2	32.7%	11	2
Aves	Accipitriformes	57.9%	Accipitridae	57.1%	UNK	48.4%		0.0%	290	2
Aves	Anseriformes	98.2%	Anatidae	98.1%	Anas	68.2%	falcata	20.7%	76	3
Aves	Columbiformes	26.9%	Columbidae	23.3%	UNK	14.2%		0.0%	30	1
Aves	Galliformes	99.8%	Phasianidae	99.8%	Arborophila	87.3%	brunneopectus	87.2%	10	3
Aves	Galliformes	89.1%	Phasianidae	86.5%	Arborophila	51.8%	UNK	11.5%	14	1
Aves	Galliformes	99.8%	Phasianidae	99.7%	Gallus	55.9%	gallus	50.6%	17	2
Aves	Galliformes	99.3%	Phasianidae	99.0%	Lophura	81.6%	ignita	73.4%	987	5
Aves	Galliformes	71.0%	Phasianidae	24.0%	UNK	19.9%		0.0%	214	6
Aves	Gruiformes	98.7%	Rallidae	98.7%	Rallina	88.8%	eurizonoides	88.8%	88	2
Aves	Passeriformes	100.0%	Corvidae	50.5%	Urocissa	29.3%	erythroryncha	29.2%	1000	10
Aves	Passeriformes	100.0%	Leiothrichidae	58.6%	Garrulax	33.5%	leucolophus	20.0%	42	1
Aves	Passeriformes	99.9%	Malaconotidae	16.6%	Malaconotus	8.1%	UNK	5.9%	51	1
Aves	Passeriformes	97.4%	Nectariniidae	17.8%	UNK	15.0%		0.0%	6	1
Aves	Passeriformes	100.0%	Pellorneidae	56.7%	Malacocincla	21.7%	abbotti	20.7%	598	1
Aves	Passeriformes	100.0%	Pellorneidae	52.1%	Pellorneum	22.6%	tickelli	22.5%	33	2
Aves	Passeriformes	99.9%	Ploceidae	23.2%	Ploceus	20.2%	UNK	20.2%	559	6
Aves	Passeriformes	78.1%	Thamnophilidae	45.8%	UNK	19.6%		0.0%	28	1
Aves	Passeriformes	100.0%	Timaliidae	42.4%	Macronus	13.5%	UNK	10.3%	206	3
Aves	Passeriformes	100.0%	Turdidae	46.6%	Geokichla	24.5%	princei	19.6%	2	2
Aves	Passeriformes	99.4%	Turdidae	31.4%	Turdus	11.9%	UNK	11.9%	60	1

Heatmap. – We can also see the anticorrelation of humans, dogs, and cows with other mammal species in a heatmap. The samples (columns) have been sorted (after running an non-metric multidimensional scaling [NMDS] ordination) so that humans are more prevalent on the left. Visually, most of the other species are more prevalent in samples that do not have humans. The order of the species (rows) is sorted so that the species least likely to appear in the same sample as a human are sorted to the bottom. Thus, species that are found with humans and/or are neutral are sorted to the top and middle, respectively. Note that the species in the next six rows include cow and dog and also four large-bodied mammals: *Rusa unicolor*, *Muntiacus vuquangensis*, *Ursus thibetanus*, and *Macaca arctoides*. This heatmap was run with only OTUs identified to Carnivora, Artiodactyla, Primates, and Lagomorpha and was limited to species that appeared in ≥10 of 542 samples (i.e. species that are more likely to have sufficient data for inferring distribution across samples).



Non-metric multidimensional scaling (NMDS) analysis. – Finally, we can visualise the same pattern, and also infer the community differences amongst reserves by running an ordination. We used the same dataset as above (Carnivora, Artiodactyla, Primates, Lagomorpha-assigned OTUs that appeared in ≥10 of 542 samples). In ordinations, the left-right axis (NMDS1) summarises the large fraction of variation amongst samples, so if we only focus on left to right, we see that humans, cows, and dogs are on the left half of the ordination, along with *Rusa unicolor*, *Muntiacus vuquangensis, Ursus thibetanus*, and *Macaca arctoides*, as was also visible in the heatmap. Most of the rest of the species are on the right hand side of the ordination, indicating anti-correlation with humans. The coloured ellipses are 95% confidence intervals of species centroids for each treatment level ('ordiellipses' [Oksanen et al. 2018]). Bach Ma and Thua Thien Hue lie more to the left, indicating a higher prevalence of human detection, whereas Xe Sap and Quang Nam lie to the right, indicating a lower prevalence of humans. As noted above leech samples do not necessarily provide a representative picture of the reserves and there may be biases due to different sampling approaches. Hence, conclusions about differences in community structure are tentative.

Now, if we switch to the up-down axis (NMDS2) summarises as much of the remaining variation as possible and shows that most species lie in the upper half, as do Thua Thien Hue, Quang Nam, and part of Bach Ma. I hypothesise that this second gradient reflects snare prevalence, with most mammal species being relatively more prevalent in Thua Thien Hue and Quang Nam, where snares have been removed. However, this is a very tentative conclusion because I do not know how well or badly the samples represent the four reserves. A better test would be a more granular sampling test, in which individual forest compartments are simultaneously surveyed for snares and leeches.

Figure 2. NMDS with only the four contiguous reserves.

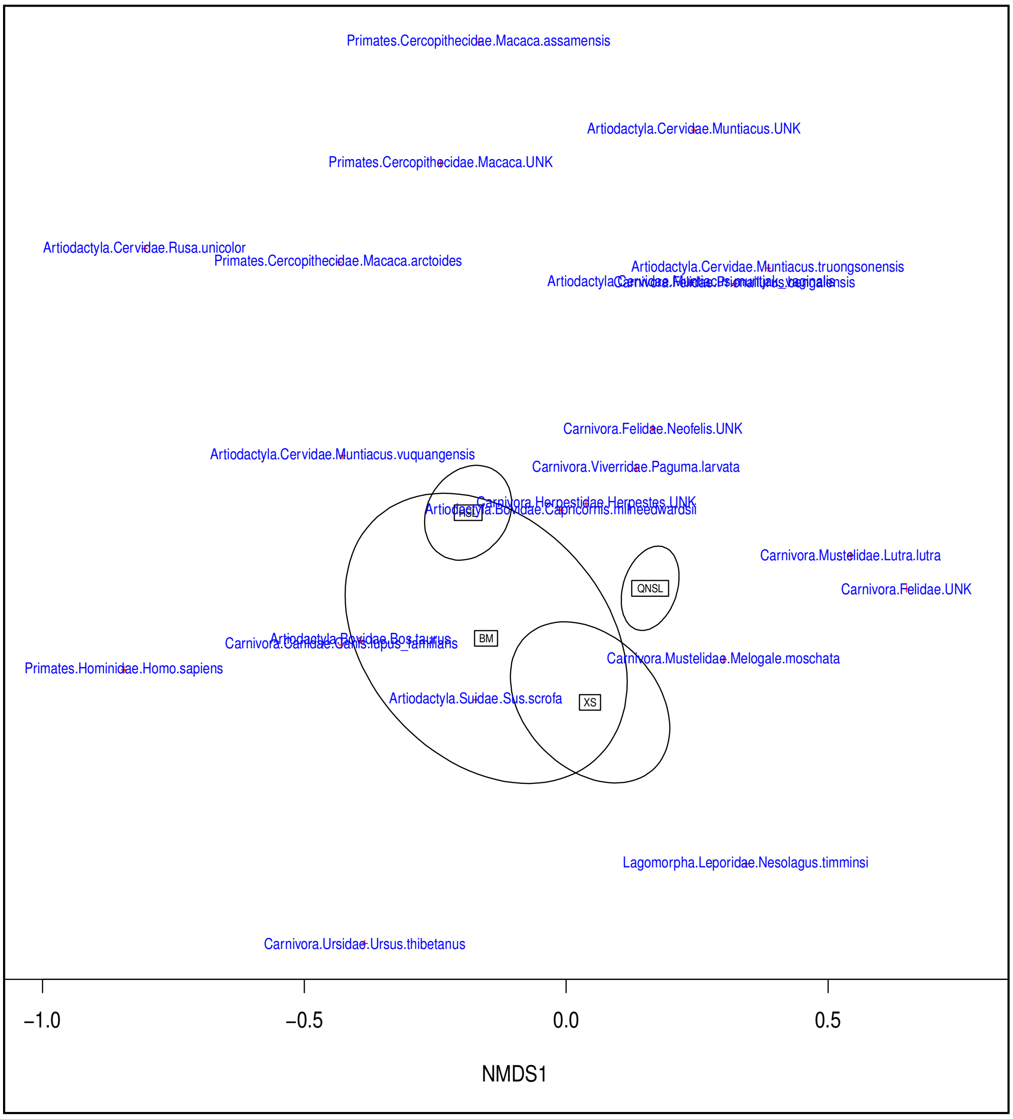
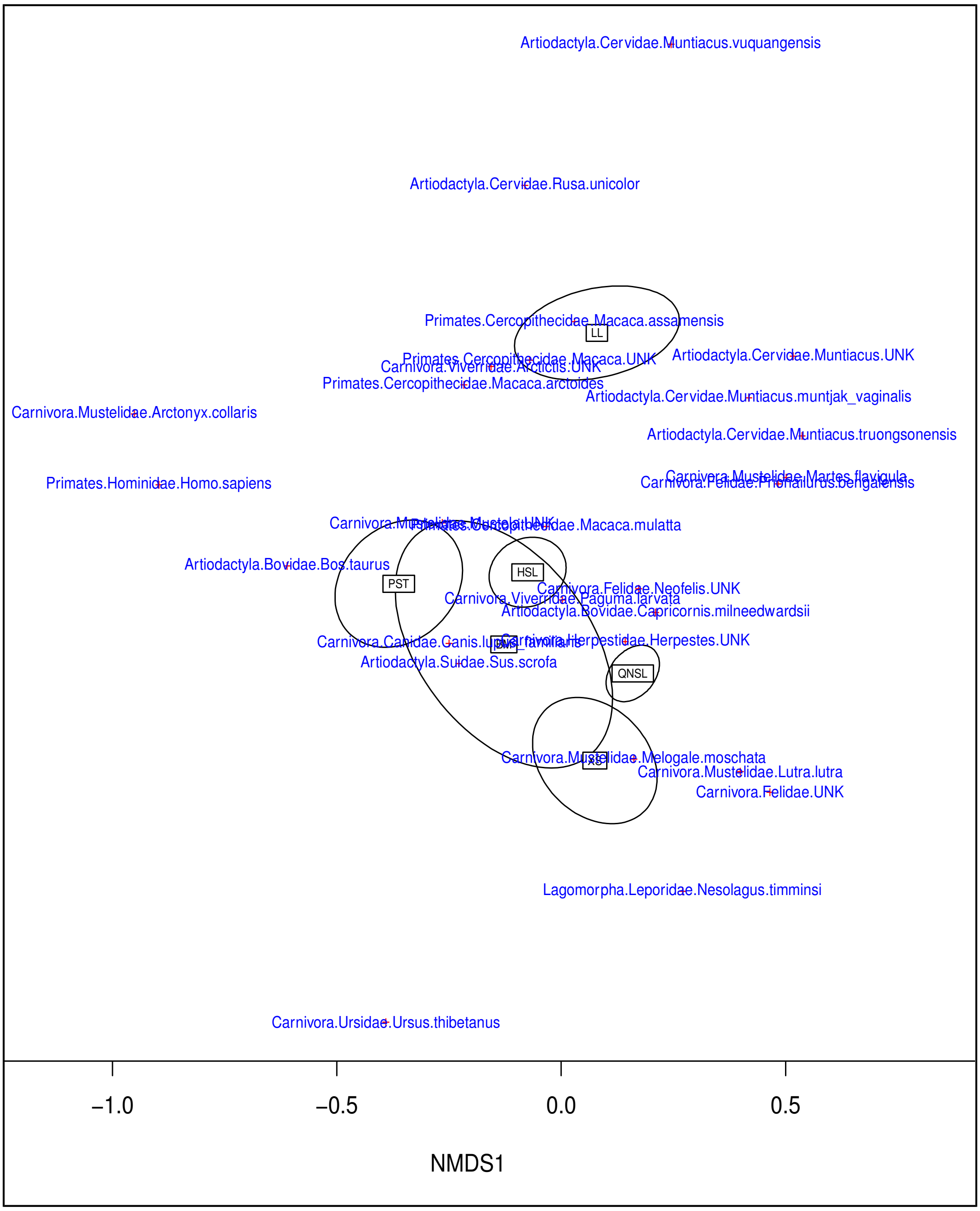


Figure 3. NMDS with all six reserves.



Discussion

I report the detection of an estimated 54 mammal species from 542 leech samples across six nature reserves in Laos and Vietnam. Community analysis suggests that the presence of most mammal species is negatively correlated with the presence of humans, dogs, and cows and that Thua Thien Hue and Quang Nam reserves have higher wild-mammal detection rates as well, which might reflect the effect of the snare removal programmes in these two reserves.

Caveats. – In addition to the normal caveats that attend all metabarcoding analyses (e.g. risk of sample cross-contamination, errors in the laboratory or bioinformatic pipeline), there are three important caveats for this study. First, human DNA can of course have been added to a sample by the collector himself, either by having touched the leeches with bare hands or having picked a feeding leech of the collector's body. We used a human blocker molecule to reduce human DNA amplification, but some human DNA is always amplified, probably in samples that have more human DNA (either because more of the leeches have human DNA on or in them or because some of the leeches have a large amount of human blood in them). On the other hand, we can think of no reason why collector human DNA would be more prevalent in Bach Ma and Phou Si Thone samples than in other samples and no reason to expect that human contamination would also add cow and dog DNA to a similar set of samples. A second caveat is that we do not know if leech sampling was equally representative of the six reserves in the choice of microhabitat, regions likely to have more or less human presence, etc. Third, I have not yet tried to control for the number of leeches per sample, which was not recorded for all leeches. This will require another analytical effort.

Future protocol improvements. – After a multi-year, multi-person effort, we finally managed to implement the Protax pipeline (Somervuo et al. 2016, 2017, Axtner et al. 2018) for the assignment of taxonomies to iDNA sequences. Protax has greatly improved the information content of the sequence outputs from the 542 leech samples. The state of the art in 2013, when we first started to receive these samples, was considerably behind what it is today, and we had to learn quite a bit to apply Illumina-based high-throughput sequencing to leech iDNA samples. In contrast, Schnell et al. (2012) processed their leeches individually and sequenced using the traditional Sanger method, which cannot be used for high-volume work.

The first of these is the use of twin-tagging of PCR primers to remove tag-jumped reads that cause species to artefactually jump from one sample to another. Fortunately, we learned of this Illumina problem early in the process, and we avoided the problem for this dataset. The second is of course is the Protax pipeline. The third, which we have been applying to current samples since 2017, is the DAMe laboratory protocol (Zepeda-Mendoza et al. 2016), in which each sample is independently PCR amplified three times, and we infer the presence of (and filter out) PCR and sequencing errors by sequences that appear in only one of the three PCRs. In contrast, highly reliable sequences appear in multiple PCRs. Fourth, reference databases continue to be improved. For instance, Salleh et al. (2017) recently published several new mitogenome sequences for Southeast Asian mammals, which have been incorporated in our reference databases. Finally, our Protax pipeline can be improved by expanding the number of taxonomic assignment methods that it uses. (Protax is a statistical wrapper method around assignment methods, not an assignment method itself. Details in Axtner et al. (2018)).

In summary, ongoing advances in laboratory and bioinformatic techniques have removed the reasons for the major delays in the reporting of these results. Given a competent molecular lab, a realistic turnaround time for 500 samples should now be around six months, most of which would be spent on DNA extraction, which is still labour intensive.

Advantages and disadvantages of leech iDNA sampling. – Leech sampling for vertebrate species has several important advantages and disadvantages.

Advantages: Leech sampling is more taxonomically comprehensive than camera trapping and can be more efficient under certain conditions: single-visit expeditions, areas where theft of cameras is likely, and when multiple individuals can contribute to the sampling. As an example of the latter advantage, the Ailaoshan national nature reserve in Yunnan, China is 678 km^2, approximately the size of Singapore. Ailaoshan is divided into 101 ranger patrolling areas, and my lab has successfully used these rangers to collect over 30,000 leeches from most of the reserve in just two months. Moreover, we did not need to train the rangers; we just provided a collecting bag and multiple sampling tubes with RNALater. Also, because each ranger collected his leeches into multiple tubes, we will treat these tubes as repeat samples of the same patrol area, which will allow for occupancy modelling. Finally, the collections themselves require very little money to add to an expedition, unlike a camera-trap network.

Disadvantages: As we have seen with this project, the two major problems with leech iDNA are an unavoidable (but diminishing) level of uncertainty in taxonomic assignment and long (but also diminishing) turnaround times in lab and bioinformatic processing. Both of these problems are being reduced by advances in laboratory and bioinformatic protocols, as I have described above. A third potential problem is that we also still do not know whether leech sampling is inherently biased against detecting rare species. The reason is that cost considerations force us to process leeches in bulk (i.e. multiple leeches per sampling tube, extracted and sequenced as a group). Rare species are inherently more likely to be present in one or just a few leeches, while abundant species are likely to be present in many leeches per tube. If so, then it is possible that the more abundant species (sequences) will outcompete the less abundant species during PCR amplification, in a process known as PCR runaway, even if the primer regions are identical. We can test experimentally for this possibility, and we can mitigate it by using fewer PCR cycles (assuming that the sample successfully produces product with low numbers of cycles). Another possibility is to subdivide samples that have large numbers of leeches, at additional cost and time. However, despite these limitations we were able to detect species such as *Ursus thibetanus*, *Arctonyx collaris* and *Rusa unicolor* which are believed to persist at low density in the Annamites on the basis of camera trap studies and other evidence. Furthermore the presence of some of these species in or near the areas where we detected them has since been independently confirmed by surveys using other methods.

Recommendations for PA managers and future surveys. – In sum, given the recent advances in laboratory and bioinformatic protocols, I think leech iDNA sampling can form a valuable complementary method for vertebrate surveys in the Paleotropics. The method is able to be applied over larger areas and detect a larger range of species with less capital investment (albeit with higher post-processing costs and more taxonomic uncertainty than camera trapping). This suite of advantages and disadvantages probably renders leech iDNA more suitable for assessing the conservation performance of reserves than to try to find specific rare species. Fortunately, if leech iDNA surveys can be justified for broad surveys, then the same, already paid-for datasets can also be interrogated for rare species of high conservation importance.

References

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, 7, 335–336.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27, 2194–2200.

Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A.K., Pietroni, C., & Hansen, A.J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, 8(1), 104. doi:10.1038/s41467-017-01312-x.

Li, H. (2015) BFC: correcting Illumina sequencing errors. *Bioinformatics*, 31, 2885–2887.

Mahé, F., T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3:e1420.

Mohd Salleh, F., J. Ramos-Madrigal, F. Peñaloza, S. Liu, S. S. Mikkel-Holger, P. P. Riddhi, R. Martins, D. Lenz, J. Fickel, C. Roos, M. S. Shamsir, M. S. Azman, K. L. Burton, J. R. Stephen, A. Wilting, and M. T. P. Gilbert. 2017. An expanded mammal mitogenome dataset from Southeast Asia. GigaScience 6:1–8.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, McGlinn, P.D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. and Wagner, H. (2018). vegan: Community Ecology Package. R package version 2.5-2. https://CRAN.R-project.org/package=vegan.

Schnell, I. B., K. Bohmann, and M. T. P. Gilbert. 2015. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular Ecology Resources.

Schnell, I. B., P. F. Thomsen, N. Wilkinson, M. Rasmussen, L. R. D. Jensen, E. Willerslev, M. F. Bertelsen, and M. T. P. Gilbert. 2012. Screening mammal biodiversity using DNA from leeches. Current Biology 22:R262–R263.

Somervuo, P., D. W. Yu, C. C. Y. Xu, Y. Ji, J. Hultman, H. Wirta, and O. Ovaskainen. 2017. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. Methods in Ecology and Evolution 8:398–407.

Somervuo, P., S. Koskela, J. Pennanen, R. Henrik Nilsson, and O. Ovaskainen. 2016. Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics 32:2920–2927.

Zepeda Mendoza, M. L., K. Bohmann, A. Carmona Baez, and M. T. P. Gilbert. 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.

Appendix

Detailed methods

DNA extraction

First, the leeches from each sample (tube) were transferred to a new tube to remove the RNALater.

If the volume of the leeches was no more than 2 ml, we prepared the leech soup (homogenate) by adding five volume times of the lysis buffer (10mM Tris-HCl, 10mM NaCl, 2% SDS, 5mM CaCl2, 2.5mM EDTA, 40mM Dithiothreitol and 0.2mg/ml proteinase K), incubating overnight at 55°C (rotating). Then the DNA was extracted from about 2.5% of the leech soup by using the Qiagen QIAquick PCR purification kit.

If the volume of the leeches was greater than 2 ml, in order to reduce costs, we added 3 times that volume of PCR-grade water with 0.02mg/ml proteinase K, incubating overnight at 55°C (rotating), then homogenizing using the Omni tissue homogenizer with CLEAN hybrid probes, transferring 10% of the leech soup to a new tube, adding 0.2 ml concentrated lysis buffer (25mM Tris-HCl, 25mM NaCl, 5% SDS, 12.5mM CaCl2, 6.25mM EDTA, 100mM Dithiothreitol and 0.5mg/ml proteinase K) for every 1 ml start volume of leech and incubating overnight at 55°C (rotating). Then the DNA was extracted from about 25% of the lysis mix by using the Qiagen QIAquick PCR purification kit.

The extracted DNA was stored at -20°C.

PCR amplification and sequencing

PCR amplifications were performed in 2 rounds and negative controls were set in both 2 rounds. In first PCR, the DNA samples were amplified using the mammal-universal primers, *16Smam1* 5'-CGGTTGGGGTGACCTCGGA-3' and *16Smam2* 5'-GCTGTTATCCCTAGGGTAACT- 3'. A unique 8 bp MID (Multiplex Identifier) tag for each sample were attached to the forward and backward primers. Each sample was amplified in three independent reactions and pooled. Because human DNA might be dominant in the leech samples and might obscure the detection of wildlife, human blocker, Human_block_16sF_long (3'-spacer C3) 5'-CGGTTGGGGCGACCTCGGAGCAGAACCC-3', was used to prevent human DNA from binding the primers. PCRs were performed in 20 µL reaction volumes containing 2 µL of 10x buffer, 1.5 mM MgCl2, 5% DMSO, 0.2 mM dNTPs, 0.4 µM each primer, 4 µM human blocker, 0.6 U Ex Taq polymerase (TaKaRa Biosystems), and 1 µL DNA. We used a thermocycling profile of 95°C for 5 min; 40 cycles of 95°C for 12 s, 59°C for 30 s, 72°C for 25 s; a final extension of 72°C for 7 min. In the second PCR, the PCR products from first PCR were amplified using the corresponding 8 bp MID tailed with the Illumina TruSeq adapters, 5'IlluminaLinker 5'-CAAGCAGAAGACGGCATACGAGAT(6bp Illumina index)GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT(8bp MID tag)-3', and 3'IlluminaLinker 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT(8bp MID tag)-3'. The second PCR's conditions were same as the first PCR's except without human blocker in PCR mix. Then the PCR products from second PCR were visualized on 2.5% agarose gels, quantified by Image Lab 2.0.1 (Bio-Rad), pooled to construct Illumina libraries (each library had at least one negative control), gel-purified using the Qiagen QIAquick PCR purification kit, and sequenced on Illumina HiSeq 2000. A base-calling pipeline (Sequencing Control Software, SCS; Illumina, San Diego, California, USA) was used to process the raw fluorescent images and to call sequences.

Raw reads were denoised with *bfc* (Li 2015) and then demultiplexed in QIIME (Caporaso et al. 2010) with the script split_libraries.py. Chimeras were detected and removed with UCHIME (Edgar et al. 2011). The remaining reads were clustered into OTUs by using SWARM 2.2.2 (d =1 and -f) (Mahé et al. 2015) and LULU (Frøslev et al. 2017). Then, the OTU sequences were BLASTed online in Genbank, and the OTUs that were not identified as vertebrate 16S were deleted. Finally, the vertebrate 16S OTUs were assigned taxonomies using Protax (Axtner et al. 2018).