Large-scale Quantification of Vertebrate Biodiversity in Ailaoshan Nature Reserve from Leech iDNA

Supplementary Methods

1 Laboratory processing

DNA extraction. We extracted DNA from each replicate sample following the protocol in [31]. Leeches were transferred to a new tube to remove the preservative, soaked in a volume of digestion buffer (10 mM Tris-HCl, 10 mM NaCl, 2% SDS, 5 mM $CaCl_2$, 2.5 mM EDTA, 40 mM dithiothreitol, and 0.2 mg/ml proteinase K) equal to 5 times the volume of each sample's leeches, and incubated at 55 °C (rotating) until all the leeches were dissolved. Following this incubation, we aliquoted 0.6 ml of digestion buffer from each sample for purification with the QIAquick PCR purification kit (Qiagen, Hilden, Germany). To detect any DNA cross-contamination, negative controls were created in both steps, digestion and purification.

PCR amplification. We PCR-amplified two mitochondrial markers, one from the 16S rRNA (MT-RNR2) gene using the primers *16Smam1* forward 5'-CGGTTGGGGTGACCTCGGA-3' and *16Smam2* reverse 5'-GCTGTTATCCCTAGGGTAACT-3' [38], and the other from the 12S rRNA (MT-RNR1) gene with the primers (forward: 5'-ACTGGGATTAGATACCCC-3' and reverse: 5'-YRGAACAGGCTCCTCTAG-3') modified from [26]. Target fragments were 81 to 117 bp and 82 to 150 bp respectively, excluding primers. We hereafter refer to these two markers as LSU (16S) and SSU (12S), respectively, referring to the ribosomal large subunit and small subunit that these genes code for. The LSU primers are designed to target mammals, and the SSU primers to amplify all vertebrates. A third primer pair targeting the standard cytochrome *c* oxidase I marker [14] was tested but not adopted in this study as it co-amplified leech DNA and consequently returned few vertebrate reads.

Primers were ordered with sample-identifying tag sequences. To be able to identify (and remove) 'tag jumping' errors [30], we used a 'twin-tagging strategy,' meaning that both forward and reverse primers used the same tag sequence for a sample (e.g. F1/R1, F2/R2, F3/R3). Thus, if a library contained tag combinations F1/R1, F2/R2, and F3/R3, an F1 tag-jump would produce F1/R2 or F1/R3, which could be detected and removed, since these combinations were not used in this library. We used the DAMe protocol [40] to remove these tag-jumped Illumina reads and to identify and remove reads containing PCR and/or sequencing errors. The DAMe protocol PCR-amplifies each sample three times per marker, each time with a different twin-tag pair, which allows the PCRs to be individually identified after sequencing. Reads containing errors are more likely to show up in only one PCR and at low copy numbers, which allows them to be filtered out bioinformatically (see below).

PCR negative controls were carried out for each PCR set, and the PCR sets that revealed contamination in the negative controls were redone, or ultimately, abandoned. For each library, a sample of negative and positive controls were sent for sequencing, in order to identify contaminants and to determine a minimum read number per OTU. The 20 μL PCR reactions consisted of 2 μL of 10X buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.2 μM per primer, 5% DMSO, 0.6 U ExTaq HotStart DNA polymerase (TAKARA Biosystems, Dalian, China), and 1 μL of template DNA, with a thermal cycling profile of 95 °C for 5 min, then 40 cycles of 95 °C for 30 s, 59 °C for 30 s, and 72 °C for 45 s, with a final extension time of 7 min at 72 °C.

2 Bioinformatic pipeline and taxonomic assignment

Preprocessing. We used AdapterRemoval v2.1.7 [32] to remove adapter sequences from reads and Sickle v1.33 [11] to trim reads of low quality nucleotides. We then used BFC v181 (parameters: `-s 3g -k 25`) [15] to de-noise the reads, and we merged the read pairs with Pandaseq v2.11 [20]. Except for BFC, we used default parameters.

Demultiplexing and DAMe quality filtering. To filter out tag-jumping events and to remove artifactual reads arising from PCR or sequencing errors, we used the DAMe pipeline [40]. DAMe's `sort.py` function was used to remove reads with unused tag combinations, and the `filter.py` function was used to keep only the haplotypes that appeared in $\geq$2 PCRs, with $\geq$9 (LSU) or $\geq$20 (SSU) copies per PCR, using the logic that sequences which appear in multiple, independent PCRs and in multiple copies per PCR are more likely to be true sequences (`filter.py` parameters for 12S: `-x 3 -y 2 -p 14 -t 20 -l 81`; for 16S: `-x 3 -y 2 -p 13 -t 9 -l 82`). Filtering parameters were chosen after inspection of the control samples.

De novo chimera removal. DAMe filtering also removes the chimeric sequences that can result from incomplete PCR extension, but we also used the *de novo* chimera detection function `uchime_denovo` in VSEARCH v2.9.0 [28] to remove any remaining chimeras after dereplicating with the `derep_fulllength` function.

Clustering into preliminary operational taxonomic units. We used SWARM v2.0 [19] to cluster the filtered sequences into preliminary OTUs ('pre-OTUs') and then used the R package `lulu` v0.1.0 [7] to merge SWARM pre-OTUs that shared high similarity and distribution across samples (i.e. over-split OTUs) and output a representative sequence for each pre-OTU. For both, we used default values.

Assigning taxonomy to preliminary operational taxonomic units. One of the more crucial steps in the iDNA bioinformatic pipeline is taxonomic assignment. With vertebrates, exact species identity can have important management consequences because some species, but not their close relatives, are given high conservation value [2]. Existing taxonomic assignment programs are typically biased toward assigning sequences to species that happen to be in a reference database, even though we know that some of our leech-derived sequences are likely from known species that have never been sequenced, or more rarely, that are undescribed. We thus used PROTAX for taxonomic assignment of the pre-OTU sequences [33, 34]. PROTAX provides an unbiased, estimated probability of assignment at each rank, where unbiased means, for example, that 70% of all assignments given a 70% probability of accuracy are indeed correct. Thus, a PROTAX assignment of a pre-OTU to Carnivora(probability=0.999)/Canidae(0.996)/*Nyctereutes*(0.821)/*Nyctereutes procyonoides*(0.557) means that this pre-OTU is very likely to be in the genus *Nyctereutes*, but there is a $(1 - 0.577)$ =44% probability that the species is not *N. procyonoides*. PROTAX can also estimate the probability that a pre-OTU sequence is 'unknown,' i.e. not in the reference database. Thus, PROTAX helps prevent mistaken assignments of sequences to species, potentially avoiding wasted management effort directed towards species that are not actually present.

We refer the reader to Somervuo *et al.* [33, 34] for in-depth discussions of PROTAX and to Axtner *et al.* [2] for details of the bioinformatic pipeline used to create the LSU and SSU reference databases and to train and assess the PROTAX models. We built the reference databases starting from the Midori Unique_20180221_lrRNA and Unique_20180221_srRNA databases [17], supplemented with mitogenomes from [22]. We used the R package `taxize` [5] to build a taxonomy database of all Tetrapoda and to harmonize species names between the Tetrapoda taxonomies and the sequences in the MidoriSalleh reference database, and we used SATIVA [13] to identify reference sequences mislabelled at family level and above, which we removed. With the curated reference database, we then trained PROTAX models for both LSU and SSU, setting a 90% prior probability for the set of Tetrapoda species known from Ailaoshan, thereby reducing false-positive assignments [27]. Raw similarities between each query and all reference sequences were calculated with LAST v.982 [12], after which the trained PROTAX models were used to assign probabilities of assignment for pre-OTUs at class, order, family, genus, and species ranks. The bioinformatic scripts, reference datasets, trained models, and bias-accuracy plots are available for download from GitHub [39].

Using pairwise correlations between SSU and LSU OTUs to reconcile taxonomies. Different marker genes have different levels of taxonomic coverage and discrimination power [33, 34], and as a result, the same species can be assigned to different taxonomies by SSU and LSU. For instance, as described above, the SSU dataset confidently detected *Nyctereutes procyonoides*, but the LSU dataset did not, although it did assign one OTU to Carnivora(probability=0.999)/Canidae(0.999)/*Canis*(0.475)/*Vulpes*, unknown species(0.231). Given the confident assignment to Canidae, this LSU OTU might also have derived from *Nyctereutes.* To combine taxonomic information across the two markers, we therefore calculated pairwise correlations of SSU and LSU pre-OTUs across the 619 replicates for which both markers had amplified and visualized the correlations as a network (Figure **??**). If an SSU and an LSU pre-OTU occur in the same subset of replicates and are assigned the same higher-level taxonomies, the two pre-OTUs are likely to have been amplified from the same set of leeches feeding on the same species. We manually inspected the network diagram and assigned such correlated pre-OTU pairs the same taxonomy.

Final operational taxonomic units and dataset filtering. After using PROTAX and then searching for network correlations, to assign taxonomies to pre-OTUs, we verified that the positive and negative control samples were free of any substantive contaminants before removing them from the dataset, along with one sample that had neither ranger nor patrol area information. We eliminated any pre-OTUs to which we were unable to assign a taxonomy; these pre-OTUs only accounted for 0.9% and 0.2% of reads in the LSU and SSU datasets respectively, and most likely represent sequencing errors rather than novel taxa. Within the LSU and SSU datasets, we merged pre-OTUs that had been assigned the same taxonomies, thus generating a final set of OTUs for each dataset. Finally, we removed the OTU identified as *Homo sapiens* from both datasets prior to analysis. As expected, since the leeches were collected with bare hands and might have in some cases been feeding on the rangers themselves, human DNA was obtained from the majority of samples in both datasets.

After excluding humans, the final LSU and SSU datasets comprised 18,502,593 and

84,951,011 reads respectively. These reads were assigned to a total of 72 OTUs across 740 replicates and 127 patrol areas in the SSU dataset, and 59 OTUs across 653 replicates and 126 patrol areas in the LSU dataset. We attached IUCN data for individual OTUs by using the R package `rredlist` v0.5.0 [4] to search for scientific names assigned by PROTAX (or synonyms where we were aware of nomenclature changes). For mammalian OTUs, we used the PanTHERIA database [10] to obtain data on adult body mass for each OTU; where species-level information was not available, we used the median adult body mass from the database for the lowest taxonomic group possible.

3 Site-occupancy modeling

Overview. We used hierarchical multispecies site-occupancy models [6] to analyze our data. The models that we used are an extension of the single-season occupancy model in [18]. For each species, the models explicitly capture (i) an 'ecological process' governing the (unobserved) presence or absence of the species in each patrol area; and (ii) an 'observation process', governing whether we detect the species' DNA in each of our replicate samples. The ecological and observation processes for individual species are linked in our model by imposing community-level priors over the parameters that describe the processes for each species.

We estimated separate models for the LSU and SSU OTU tables. For each dataset, we estimated a set of alternative models, summarized in Table S1, specifying different combinations of predictors for the ecological and observation processes. We used the deviance information criterion to compare results and select the final models as presented in the main text of this paper.

Ecological process. Each species i was assumed to be either present or absent in each patrol area j, and we used $z_{i,j}$ to denote this unobserved ecological state. We assumed the $z_{i,j}$ are constant across all replicates taken from patrol area j, consistent with the samples being taken at essentially the same point in time (sometimes referred to as the 'closure' assumption). $z_{i,j}$ was assumed to be a Bernoulli random variable governed by an occupancy parameter $\psi_{i,j}$, i.e. the probability that species i was present in patrol area j:

$$z_{i,j} \sim Bernoulli(\psi_{i,j}). \tag{S1}$$

We allowed the occupancy probability $\psi_{i,j}$ to vary among species as well as among patrol areas, to capture e.g. preferences of different species for particular habitat types, or interactions between taxa. In particular, we modelled $\psi_{i,j}$ as a function of environmental covariates that varied over the patrol areas, scaled by species-specific coefficients. Models 1a, 1b and 1c represented the full ecological model:

$$logit(\psi_{i,j}) = \beta_{0i} + \beta_{1i} elev_j + \beta_{2i} TPI_j + \beta_{3i} road_j + \beta_{4i} stream_j + \beta_{5i} reserve_j \tag{S2}$$

where $elev_j$, TPI_j, $road_j$, $stream_j$ and $reserve_j$ are, respectively, the median values of elevation, topographic position index, distance to nearest road, distance to nearest stream, and the distance from centroid to nature reserve boundary for patrol area j.

Preliminary results indicated that $elev$, $reserve$ and $road$ were likely to be the most useful occupancy predictors. So, for comparison, we estimated a set of reduced models (2a, 2b and

2c) with only these occupancy covariates:

$$logit(\psi_{i,j}) = \beta_{0i} + \beta_{1i}elev_j + \beta_{2i}road_j + \beta_{3i}reserve_j. \tag{S3}$$

The covariates *elev* and *road* were positively correlated ($r = 0.6$), so we additionally estimated a set of models (3a, 3b and 3c) omitting *road*:

$$logit(\psi_{i,j}) = \beta_{0i} + \beta_{1i}elev_j + \beta_{2i}reserve_j \tag{S4}$$

and a set of models (4a, 4b and 4c) omitting *elev*:

$$logit(\psi_{i,j}) = \beta_{0i} + \beta_{1i}road_j + \beta_{2i}reserve_j. \tag{S5}$$

Observation process. Although we cannot directly observe the true ecological state $z_{i,j}$, we do know whether we detected DNA from species i in each replicate k from patrol area j. But this is an imperfect proxy for the true ecological state. For replicate k from patrol area j, we assumed that we detected DNA from species i with probability $p_{i,j,k}$ when i was truly present in patrol area j, and with probability 0 when i was absent:

$$y_{i,j,k} \sim Bernoulli(z_{i,j}.p_{i,j,k}), \tag{S6}$$

where the $y_{i,j,k}$ are the observed data (i.e. detection or non-detection of species i's DNA in each replicate). Our model therefore assumes that false positives do not occur, i.e. that

Table S1: Summary of model specifications tested. *elev* = median elevation; *TPI* = median topographic position index; *road* = median distance to nearest road; *stream* = median distance to nearest stream; *reserve* = distance from centroid to nature reserve boundary; *numleeches* = number of leeches in replicate; *othertaxa* = number of other taxa detected in replicate; *human* = fraction of reads from replicate assigned to *Homo sapiens*

Model	Occupancy covariates	Detection covariates
1a	$elev + TPI + road + stream + reserve$	$numleeches$
1b	$elev + TPI + road + stream + reserve$	$numleeches + othertaxa$
1c	$elev + TPI + road + stream + reserve$	$numleeches + human$
2a	$elev + road + reserve$	$numleeches$
2b	$elev + road + reserve$	$numleeches + othertaxa$
2c	$elev + road + reserve$	$numleeches + human$
3a	$elev + reserve$	$numleeches$
3b	$elev + reserve$	$numleeches + othertaxa$
3c	$elev + reserve$	$numleeches + human$
4a	$road + reserve$	$numleeches$
4b	$road + reserve$	$numleeches + othertaxa$
4c	$road + reserve$	$numleeches + human$

we never falsely detect species i's DNA through lab contamination or through incorrectly assigned sequence reads. On the other hand, since $p_{i,j,k}$ may be less than one, it allows for the possibility of false negatives, i.e. that we failed to detect species i's DNA when species i was actually present. Although false positives probably do occur, we focused mainly on lab procedures and the taxonomic assignment pipeline to address these, and we expect false negatives to far outstrip false positives in our final datasets.

We allowed the conditional detection probability $p_{i,j,k}$ to vary among species, to capture e.g. variation in leech feeding preferences for different taxa, as well as among replicates, to capture e.g. technical differences that might affect the probability of detecting taxa. The observed data clearly showed that the number of leeches included in a replicate was positively related to the number of taxa detected (see Figure S4b). Our baseline detection model therefore used the number of leeches in replicate k from patrol area j, denoted $numleeches_{j,k}$, as a predictor for the detection probability for each species i in that replicate:

$$logit(p_{i,j,k}) = \gamma_{0i} + \gamma_{1i} numleeches_{j,k}, \tag{S7}$$

and we used this observation model in conjunction with each of the ecological models in Equations S2 through S5 (i.e. models 1a, 2a, 3a and 4a).

We also estimated two other variants of the observation model. First, to test the idea proposed in [1] that the detection probability for species i may be lowered in the presence of DNA from other species, we calculated $othertaxa_{i,j,k}$ as the number of species other than i detected in replicate k from patrol area j. We used this along with $numleeches$ to model detection probability, and used this observation model in conjunction with each of the ecological models in Equations S2 through S5 (i.e. models 1b, 2b, 3b and 4b):

$$logit(p_{i,j,k}) = \gamma_{0i} + \gamma_{1i} numleeches_{j,k} + \gamma_{2i} othertaxa_{i,j,k}. \tag{S8}$$

Second, along similar lines, we examined the possibility that the detection probability for species i may be lowered in the presence of human DNA, which in some replicates accounted for the majority of reads. We therefore calculated $human_{j,k}$ as the fraction of reads assigned to $Homosapiens$ in replicate k from patrol area j after all filtering steps in our bioinformatic pipeline. We used this along with $numleeches$ to model detection probability, and used this observation model in conjunction with each of the ecological models in Equations S2 through S5 (i.e. models 1c, 2c, 3c and 4c):

$$logit(p_{i,j,k}) = \gamma_{0i} + \gamma_{1i} numleeches_{j,k} + \gamma_{2i} human_{j,k}. \tag{S9}$$

Community model. Equations (S1) through (S9) define a set of 12 site-occupancy models for each species i with alternative specifications for modelling the ecological and observation processes (summarized in Table S1). For each of these 12 alternative model specifications, we united the species-specific models with community models for both ecological and observation processes. Specifically, we assumed that the species-level β and γ parameters are distributed according to distributions described by a set of community-level hyperparameters:

$$\beta_{mi} \sim N(\mu_{\beta_m}, \sigma_{\beta_m}) \qquad m = 1, 2, ... \tag{S10}$$

$$\gamma_{ni} \sim N(\mu_{\gamma_n}, \sigma_{\gamma_n}) \qquad n = 1, ... \tag{S11}$$

$$(\beta_{0i}, \gamma_{0i}) \sim MVN([\mu_{\beta_0}, \mu_{\gamma_0}], [\sigma_{\beta_0}, \sigma_{\gamma_0}]) \tag{S12}$$

where $N(\ )$ and $MVN(\ )$ denote normal and multivariate normal distributions, with community-level hyperparameters $\mu_\bullet$ and $\sigma_\bullet$. That is, for each model specification, m and n vary so that there is a distribution described by Equations S10 or S11 for each predictor. We used a multivariate normal prior for $(\beta_{0i}, \gamma_{0i})$ to allow non-zero covariance between species' occupancy and detection probabilities, as we might expect if, for example, variation in abundance affects both probabilities [6]. These community models allow rare species effectively to borrow information from more common ones, producing a better overall ensemble of parameter estimates [6, 16, 29].

Incompletely labelled data points (i.e. sequence data without records of which patrol areas they came from) were retained in the model by including these data points without accompanying environmental covariates. Since the identity of the collecting ranger was known and could be used to identify replicates that came from the same unknown location, this allowed these data to contribute to both detection and occupancy estimates. At the same time, we generated occupancy estimates for patrol areas without accompanying data by augmenting the data matrix with rows of missing values and including their environmental covariates.

We normalized all predictors to a mean of 0 and a standard deviation of 1 prior to modelling. We estimated all model variants in a Bayesian framework with JAGS v4.3.0 [23] in R v3.5.1 [25] via `rjags` v4.8 [24] and `R2jags` v0.5-7 [37]. We used uninformative diffuse priors for all parameters and hyperparameters. We ran each model with three chains of 40,000 generations and a burn-in of 10,000, thinning results by a factor of 20. From the retained results we calculated means for all model parameters of interest, as well as estimated species richness for each patrol area. We assessed convergence by inspecting the $\hat{R}$ statistic [8, 3], and calculated 95% credible intervals from the 2.5% and 97.5% percentiles of the posterior distribution.

Comparing model results. We used the deviance information criterion (DIC) [36] to compare the 12 model variants against one another for each dataset. This computationally straightforward approach is known to have limitations, both in general and for occupancy models in particular [35, 9], but there is a lack of consensus on superior methods, and our conclusions, in any case, are unlikely to hinge on the choice of specification.

We used `AICcmodavg:DIC` in R [21] to calculate DIC for each model, and ranked models accordingly (Table S2). In both datasets, model 3a (occupancy covariates *elev* and *reserve*; detection covariate *numleeches*) was the best ranked model. We therefore report results from this model specification in the paper. However, models 4a and 2a also performed reasonably well, and in any extension of this work it would be worth considering whether there is value in including *road* as a predictor in addition to, or instead of *elev*.

Table S2: DIC results. Models are ordered according to DIC within each dataset, with the best models first. pD = effective number of estimated parameters for each model; DIC = deviance information criterion; ΔDIC = difference in DIC compared to top-ranked model.

(a) LSU dataset

Model	pD	DIC	ΔDIC
3a	2504	9706	0
4a	2595	9877	171
2a	2621	10317	611
1a	2594	11297	1590
2c	2377	11842	2135
3c	3035	12018	2312
4c	3109	12119	2412
1c	1922	12376	2670
3b	2637	119096	109390
2b	2552	119515	109809
4b	3729	120262	110555
1b	2705	120678	110972

(a) SSU dataset

Model	pD	DIC	ΔDIC
3a	2947	13620	0
4a	3080	13749	129
2a	3024	14204	583
1a	3734	15918	2298
4c	3339	15992	2372
3c	3426	16110	2489
2c	3470	16673	3053
1c	3186	17385	3765
2b	3131	165447	151827
3b	3700	165509	151889
4b	3977	165800	152180
1b	3083	166392	152772

References

[1] J. F. Abrams, L. Hörig, R. Brozovic, J. Axtner, A. Crampton-Platt, A. Mohamed, et al. "Shifting up a gear with iDNA: from mammal detection events to standardized surveys". *Journal of Applied Ecology* 18.3 (2019), pp. 511–512.

[2] J. Axtner, A. Crampton-Platt, L. A. Hörig, A. Mohamed, C. C. Y. Xu, D. W. Yu, et al. "An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies". *GigaScience* 8.4 (Apr. 2019), pp. 646–17.

[3] S. P. Brooks and A. Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. DOI: 10.1080/10618600.1998.10474787.

[4] S. Chamberlain. *rredlist: 'IUCN' Red List Client.* R package version 0.5.0. 2018. URL: https://CRAN.R-project.org/package=rredlist.

[5] S. Chamberlain, E. Szoecs, Z. Foster, Z. Arendsee, C. Boettiger, K. Ram, et al. *taxize: Taxonomic information from around the web.* R package version 0.9.7. 2019. URL: https://github.com/ropensci/taxize.

[6] R. M. Dorazio, J. A. Royle, B. Soderstrom, and A. Glimskar. "Estimating species richness and accumulation by modeling species occurrence and detectability". *Ecology* 87.4 (2006), pp. 842–854. DOI: 10.1890/0012-9658(2006)87[842:Esraab]2.0.Co;2.

[7] T. G. Frøslev, R. Kjøller, H. H. Bruun, R. Ejrnæs, A. K. Brunbjerg, C. Pietroni, et al. "Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates". *Nature Communications* 8 (2017). DOI: 10.1038/s41467-017-01312-x.

[8] A. Gelman and D. B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". *Statistical Science* 7.4 (1992), pp. 457–472. DOI: 10.1214/ss/1177011136.

[9] M. B. Hooten and N. T. Hobbs. "A guide to Bayesian model selection for ecologists". *Ecological Monographs* 85.1 (2015), pp. 3–28. DOI: 10.1890/14-0661.1.

[10] K. E. Jones, J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, et al. "PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals". *Ecology* 90.9 (2009), pp. 2648–2648. DOI: 10.1890/08-1494.1.

[11] J. N. Joshi and N. A. Fass. *Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33).* 2011. URL: https://github.com/najoshi/sickle.

[12] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith. "Adaptive seeds tame genomic sequence comparison". *Genome Research* 21.3 (Mar. 2011), pp. 487–493.

[13] A. M. Kozlov, J. Zhang, P. Yilmaz, F. O. Glöckner, and A. Stamatakis. "Phylogeny-aware identification and correction of taxonomically mislabeled sequences". *Nucleic Acids Research* 44.11 (June 2016), pp. 5022–5033.

[14] M. Leray, J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, V. Ranwez, et al. "A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents". *Frontiers in Zoology* 10 (June 2013), p. 34.

[15] H. Li. "BFC: correcting Illumina sequencing errors". *Bioinformatics* 31.17 (Sept. 2015), pp. 2885–2887.

[16] W. A. Link and J. R. Sauer. "Extremes in Ecology: Avoiding the Misleading Effects of Sampling Variation in Summary Analyses". *Ecology* 77.5 (1996), pp. 1633–1640. DOI: 10.2307/2265557.

[17] R. J. Machida, M. Leray, S.-L. Ho, and N. Knowlton. "Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples". *Scientific Data* 4 (2017), p. 170027. DOI: 10.1038/sdata.2017.27. Data downloaded from http://www.reference-midori.info/download.php on 9 August 2019.

[18] D. I. MacKenzie, J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. "Estimating site occupancy rates when detection probabilities are less than one". *Ecology* 83.8 (2002), pp. 2248–2255.

[19] F. Mahe, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. "Swarm v2: highly-scalable and high-resolution amplicon clustering". *PeerJ* 3 (2015). DOI: 10.7717/peerj.1420.

[20] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld. "PANDAseq: paired-end assembler for illumina sequences". *BMC Bioinformatics* 13.1 (Nov. 2012), pp. 1–7.

[21] M. J. Mazerolle. *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*. R package version 2.2-2. 2019. URL: https://cran.r-project.org/package=AICcmodavg.

[22] F. Mohd Salleh, J. Ramos-Madrigal, F. Peñaloza, S. Liu, S. S. Mikkel-Holger, P. P. Riddhi, et al. "An expanded mammal mitogenome dataset from Southeast Asia". *GigaScience* 6.8 (Aug. 2017), pp. 1–8.

[23] M. Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Computer Program. Version 4.3.0. 2017. URL: https://sourceforge.net/projects/mcmc-jags.

[24] M. Plummer. *rjags: Bayesian graphical models using MCMC*. Computer Program. R package version 4.8. 2018. URL: https://CRAN.R-project.org/package=rjags.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.

[26] T. Riaz, W. Shehzad, A. Viari, F. Pompanon, P. Taberlet, and E. Coissac. "ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis". *Nucleic Acids Research* 39.21 (Sept. 2011), e145–e145. DOI: 10.1093/nar/gkr732.

[27] T. W. Rodgers, C. C. Y. Xu, J. Giacalone, K. M. Kapheim, K. Saltonstall, M. Vargas, et al. "Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community". *Molecular Ecology Resources* 17.6 (Nov. 2017), e133–e145.

[28] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. "VSEARCH: a versatile open source tool for metagenomics". *PeerJ* 4 (2016), e2584.

[29] D. B. Rubin. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". *The Annals of Statistics* 12.4 (1984), pp. 1151–1172.

[30] I. B. Schnell, K. Bohmann, and M. T. P. Gilbert. "Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies". *Molecular Ecology Resources* 15.6 (2015), pp. 1289–1303. DOI: 10.1111/1755-0998.12402.

[31] I. B. Schnell, P. F. Thomsen, N. Wilkinson, M. Rasmussen, L. R. Jensen, E. Willerslev, et al. “Screening mammal biodiversity using DNA from leeches”. *Current Biology* 22.8 (2012), R262–R263. DOI: 10.1016/j.cub.2012.02.058.

[32] M. Schubert, S. Lindgreen, and L. Orlando. “AdapterRemoval v2: rapid adapter trimming, identification, and read merging”. *BMC Research Notes* 9 (Feb. 2016), p. 88.

[33] P. Somervuo, S. Koskela, J. Pennanen, R. H. Nilsson, and O. Ovaskainen. “Unbiased probabilistic taxonomic classification for DNA barcoding”. *Bioinformatics* 32.19 (2016), pp. 2920–2927. DOI: 10.1093/bioinformatics/btw346.

[34] P. Somervuo, D. W. Yu, C. C. Y. Xu, Y. Q. Ji, J. Hultman, H. Wirta, et al. “Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding”. *Methods in Ecology and Evolution* 8.4 (2017), pp. 398–407. DOI: 10.1111/2041-210x.12721.

[35] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. “The deviance information criterion: 12 years on”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.3 (2014), pp. 485–493. DOI: 10.1111/rssb.12062.

[36] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. “Bayesian measures of model complexity and fit”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4 (2002), pp. 583–639. DOI: 10.1111/1467-9868.00353.

[37] Y.-S. Su and M. Yajima. *R2jags: Using R to Run ‘JAGS’*. Computer Program. R package version 0.5-7. 2015. URL: https://CRAN.R-project.org/package=R2jags.

[38] P. G. Taylor. “Reproducibility of ancient DNA sequences from extinct Pleistocene fauna”. *Molecular Biology and Evolution* 13.1 (Jan. 1996), pp. 283–285.

[39] D. Yu. *Ailaoshan version with unweighted and weighted PROTAX and MIDORI 1.2*. 2020. URL: https://github.com/dougwyu/screenforbio-mbc-ailaoshan/releases/tag/1.3.

[40] M. L. Zepeda-Mendoza, K. Bohmann, A. Carmona Baez, and M. T. Gilbert. “DAMe: a toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses”. *BMC Research Notes* 9 (2016), p. 255. DOI: 10.1186/s13104-016-2064-9. Downloaded 9 August 2019 from forked version at https://github.com/shyamsg/DAMe.