Interim 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 NUMBER: HZ-26

Date: 15 October 2014

**Summary**: The Ecology, Conservation And Environment Center (ECEC) at the Kunming Institute of Zoology (KIZ) has carried out metagenetic analysis of 80 leech samples collected by WWF in Vietnam. The 80 samples reveal a diverse set of mammal species, including sambar (*Rusa unicolor*), serow (*Capricornis* spp.), Asiatic black bear (*Ursus thibetanus*), possibly leopard (*Panthera pardus*), wild pig (*Sus* cf. *scrofa*), a variety of rodents, small cervids, and carnivores, plus a few detections of birds, fish, and frogs (non-target taxa). The assignment of taxonomies to the genetic sequences required the use and development of a range of methods, some unpublished, so all assignments should taken as provisional. Furthermore, there always remains the possibility of sample cross-contamination at all stages of the field to lab pipeline. Nevertheless, these data do suggest that leech metagenetics can be used to characterise large-scale spatial distributions of mammals at near-species-level resolution for at least some important taxa. The next steps in this project will be to analyse the rest of the WWF samples (plus the samples from Laos), which we expect will be completed by the end of January 2015, to compare the results against existing data on species distributions, and to publish the results in a scientific journal. In the longer term, it will be necessary to further refine taxonomic assignment methods, which includes augmentation of mammal reference databases and development of a way to reconcile different assignment methods.

# Introduction

Under the terms of Agreement HZ-26, WWF contracted ECEC-KIZ to:

1. Use the methodologies based on Schnell *et al*., 2012 [Current Biology Vol 22(8):R262-R263] to extract and amplify mammalian DNA from all samples sent by WWF to ECEC, suitably modified to take advantage of high-throughput sequencing via the Illumina HiSeq 2000.
2. Compare amplified DNA sequences from each sample to GenBank and other appropriate sources to identify all mammalian species present within each sample.
3. Writing of a formal report on the methods used and analytical results.

# Methods outline

*Laboratory*. - 468 samples of leeches preserved in RNALater were received from WWF. We extracted DNA from each sample and tested each sample for PCR product using the 16s\_leech\_FR (=16Smam) primer set without human blocker, which is designed to amplify all mammals, and using the saola-specific primer set SaolaATP6FR, designed by Ida Schnell.

28 samples showed no 16s\_leech\_FR PCR product, and of these, 4 showed a PCR-product in the size range that is expected for saola (228 bp). These 4 were Sanger-sequenced and did not produce a saola sequence. The 28 samples were therefore set aside.

The remaining samples showed at least some 16s\_leech\_FR product, and a subset of these also showed a SaolaATP6FR product. We chose 80 samples for HiSeq 2500 sequencing, 54 of which had at least some SaolaATP6FR product, and 26 of which did not, as a control. The remaining samples are currently (October 2014) being prepared for sequencing in December 2014. Sample status details are in the supplementary file WWF-Samples-Status-2014Sep.xlsx.

*Bioinformatics*. - After sequencing, the raw read data were denoised, split into their respective samples, checked for chimeras, and clustered into OTUs (Operational Taxonomic Units) at 98% similarity. The clustering program CROP outputs the ‘seed’ sequence of each cluster (OTU), and this was used as the representative sequence for each OTU.

Further details on laboratory, sequencing, and bioinformatic methods are in the **Appendix** to this document.

*Taxonomic assignment*. - Finally, we assigned taxonomies to the representative sequences, and during this stage, we made the determination that existing methods to assign taxonomies to the amplified DNA sequences were unlikely to produce reliable results, and we therefore began a collaboration with Dr. Douglas Chesters at the Institute of Zoology in Beijing to develop additional assignment methods. The combined output from existing and new methods were then used to assign a consensus taxonomy to each OTU’s representative sequence.

The assignment methods used were: *SAP* (Munch et al. 2008), *RDP* (Wang et al. 2007), *Geneious* 8 beta’s *Sequence Classifier* plug-in (www.biomatters.com), *BAGpipe Phylo* and *BAGpipe Dist* (adapted by Chesters from Papadopoulou et al. [2014] and unpublished), *Claident* (Tanabe & Toju 2013), *BAGpipe Phylo + pplacer* (Matsen et al. 2010), and *BAGpipe Phylo + epa* (Berger et al. 2011). All these methods search for Genbank-deposited sequences that are similar to each query sequence, after which the software makes a determination regarding the taxonomic rank to which the query can be confidently placed. Thus, if a query closely matches only one species from Genbank, the query can be identified to species level, but if a query matches several species, then the query is assigned to the consensus higher-level taxon that includes all the species to which the query had a close match. How this consensus is chosen is a major way by which these methods differ. Another difference is that SAP and Claident use the entire Genbank database. The other methods use 16S databases that we downloaded from Genbank and curated. Some of the methods (*Sequence Classifier, Bagpipe Phylo, pplacer,* and *epa*) use placement of queries on a reference phylogeny of mammals that Chesters constructed, and then infer assignment from the placement of a query (near a tip or more basal) and the branch distance between the query and its nearest neighbour reference sequence. All methods have their biases, some of the methods are more conservative than others, and the reference databases and trees are themselves incomplete and likely to have some errors, such as subtle alignment errors in the trees. On the other hand, phylogenetic placement methods appear to be less prone to mis-assignment when the reference database is incomplete. In particular, we find that several queries get assigned to seals (Phocidae) quite strongly, and since seals are not found in the sampling locations, we have ignored those hits and assigned the queries to Carnivora (because some of the methods assign these seal-hitting queries to Carnivora only). Taxonomic assignment protocols continue to need a lot of work.

We still do not have a formal way to choose amongst the different assignments provided by the methods, so for this report, DY examined the 98 OTUs with the most sequence reads (stopping after the OTU that was assigned to *Ursus thibetanus*) and assigned a consensus taxonomy for each OTU. DY’s main rule was to trust the assignments by SAP, which is quite conservative, so when it assigns to species, it seems reasonable to take that assignment.

Low-read-number OTUs are likely to have been split off from high-read-number OTUs because of sequencing or PCR errors for some reads, which makes assignment of low-read-number OTUs difficult (as can be seen by a higher frequency of assignments of low-read-number OTUs to species that are not in Vietnam [DY’s visual impression]). However, we are probably missing a few true species, like *Manis pentadactyla*, which shows up in the low-read-number OTUs.

We also BLASTed each sequence directly to Genbank, to allow comparison with other studies, but we did not refer to the BLAST results until after the consensus assignment was made. This allows some independence from the methods that we used. N.B. Most of the methods use BLAST at some point in their pipelines and then refine the results.

The OTU X Sample data tables are in WWF-Samples-Status-2014Sep.xlsx.

The raw assignments are in: 14SepWWF\_all\_CROP98\_merge\_otu\_16Svert\_DC.xlsx

It is very important for everybody to examine carefully the raw assignments and to keep in mind that clustering raw reads into species regularly results in splitting and lumping. That is, some species will be represented by more than one OTU, and they will tend to be assigned the same or similar taxonomies (and have very similar distributions across samples). Split species that receive the same taxonomy get merged again in the final OTU X Sample data table (known shorthand as an ‘OTU table’). Other species might be lumped within the same OTU, and since we only choose one representative sequence from each OTU, we might miss those species. Thus, we also searched the raw reads before clustering for sequences that match saola but did not find any.

Finally, it is instructive to examine the placements of the OTUs on the reference trees (see the two huge PDFs and the PowerPoint slide). All OTU names are prefixed by “uniques\_” plus a number.

# Financial statement

To date, our expenditures have been ca. $17,423 USD, with PCR, sequencing, and bioinformatic analysis still to be done for the remaining samples. It appears that we will come in at budget.

|  |  |
| --- | --- |
| **Activity to date** | **Cost (RMB)** |
| DNA Extraction | ¥51,992.00 |
| PCR test for Mammal 16S | ¥3,663.10 |
| PCR and Sanger sequencing for SaolaATP6FR | ¥13,969.10 |
| Hiseq 2500 sequencing for Mammal 16S | ¥38,399.00 |
| Total | ¥108,023.20 |
| **USD @ 6.2 RMB/USD** | **$17,423** |
| Project Budget | $25,466 |
| Person-weeks to date | 19 weeks and 4 days |

# References

Berger, S.A., Krompass, D. & Stamatakis, A. (2011). Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol*, 60, 291–302.

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.

Greenfield, P., Duesing, K., Papanicolaou, A. & Bauer, D.C. (2014). Blue: correcting sequencing errors using consensus and context. *Bioinformatics*, 30, 2723–2732.

Hao, X., Jiang, R. & Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, 27, 611–618.

Matsen, F.A., Kodner, R.B. & Armbrust, E.V. (2010). *pplacer*: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11, 538.

Munch, K., Boomsma, W., Huelsenbeck, J., Willerslev, E. & Nielsen, R. (2008). Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics. *Syst Biol*, 57, 750–757.

Papadopoulou, A., Chesters, D., Coronado, I., la Cadena, De, G., Cardoso, A., Reyes, J.C., *et al.* (2014, in press). Automated DNA-based plant identification for large-scale biodiversity assessment. *Mol Ecol Res*.

Tanabe, A.S. & Toju, H. (2013). Two New Computational Methods for Universal DNA Barcoding: A Benchmark Using Barcode Sequences of Bacteria, Archaea, Animals, Fungi, and Land Plants. *PLoS ONE*, 8, e76910.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007). Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol*, 73, 5261–5267.

# 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), and homogenizing using Omni tissue homogenizer with CLEAN hybrid probes. 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, 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**

For the Saola-specific primer, ATP6, the DNA samples were amplified using the specific primers, SaolaATP6F 5’-AGCACCAAAACAAAGTATCCCT-3’ and SaolaATP6R 5’-CCCTTTGACGTTATGGATGCT- 3’. PCRs were performed in 25 μL reaction volumes containing 2.5 μL of 10x buffer, 2.5 mM MgCl2, 0.2 mM dNTPs, 0.2 μM each primer, 1.25 U AmpliTaq Gold polymerase (Applied Biosystems), and 1 μL DNA. We used a thermocycling profile of 94°C for 3 min; 45 cycles of 94°C for 45 s, 58°C for 45 s, 72°C for 45 s; a final extension of 72°C for 5 min. PCR products were visualized on 2% agarose gels and were bidirectionally sequenced using BigDye version 3.1 on an ABI 3730xl DNA Analyser (Applied Biosystems).

For mammal 16S rRNA, 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. PCR products were visualized on 2% agarose gels, gel-purified using the Qiagen QIAquick PCR purification kit, quantified using the Quant-iT PicoGreen dsDNA Assay kit, pooled to construct Illumina insert-size libraries 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 BLUE (Greenfield et al. 2014) and then split by samples 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 at 98% similarity by using CROP (Hao et al. 2011). Finally, the OTU sequences were BLASTed online in Genbank, and the OTUs that were not identified as vertebrate 16S were deleted.