Validating a laboratory pipeline for accurate reconstruction of metabarcode data

Yige Sun

MRes. Computational Methods in Ecology and Evolution

Supervised by

Prof. Alfried Vogler

Faculty of Natural Sciences, Department of Life Sciences(Silwood Park), ICL
Department of Life Sciences, Natural History Museum, London
a.vogler@imperial.ac.uk

2019-2020

Keywords

community barcoding, metabarcoding analysis, metagenomics, high-throughput sequencing, tropical beetles, species identification.

Introduction

Metabarcoding is a robust technique in characterising the complex community compositions and breaking through the barrier in traditional taxonomic methods. It is a highly suitable method for the study of large-scale species richness and complex community composition surveys (Lebuhn et al., 2013). It also is a widely used technique for the study of arthropods with High-throughput Sequencing (Ji et al., 2013).

In insect barcoding analysis, cytochrome oxidase C subunit I (cox1) is commonly used as a barcode marker (Hebert et al., 2003). Most existing studies are clustering the sequence variants into a Operational Taxonomic Units (OTUs) that roughly correspond to the species in the Linnaean taxonomy, which can accommodate PCR and sequencing errors in the individual reads, but this removes genetic variation.

Newer approaches attempt to retain all true genetic variants, the so-called amplicon sequence variants (ASVs) after removing the sequencing errors. However, this leads to a further challenge from amplification of nuclear mitochondrial DNA segments (NUMTs), which are pseudogenes derived from the mitochondrial copies that persist in the nuclear genome to increase the apparent diversity of genotypes. In addition, internalised parasites or gut contents of insects are co-amplified and represent a useful source of ecological information but unhelpful for taxonomic study.

A previous study of UK bee survey has built up a pipeline for the taxonomic assignment which showed a high congruence with morphological classification (Creedy et al., 2019). It removed the pseudogenes based on reads abundance and phylogenetic relateness to references sets. Moreover, it reveals that there still have rooms for bioinformatically improvement in true ASVs retention and generate a more accurate taxonomic assignment.

According to such standarised analysis have not been applied on tropical beetles survey. In this project, I will analyse several metabarcode datasets of tropical beetles to improve the construction of clean ASV information and detection of spurious NUMTs copies and reveal the biodiversity in the regional tropical beetles community.

Methods

Various metabarcoding library that include up to 7,000 tropical beetles sequences generated with Illumina MiSeq v.3 will be subjected to data handling and filtering by quality procedures, including adaptor removal by cutadapt (Martin, 2011) and quality filtering with FASTQC (Andrews et al.,2012), followed by denoising using UNOISE (Edgar et al., 2011) under various parameter settings.

The resulting reads will be further filtered by relatedness with reference ASVs from mitochondrial whole-genome sequencing. These procedures would be carried out on individual specimens and assigned to a correct haplotype. Then, specimens would be classified by relatedness in to batches of 50 and set up reads abundance threshold values for true ASVs retention based on relatedness

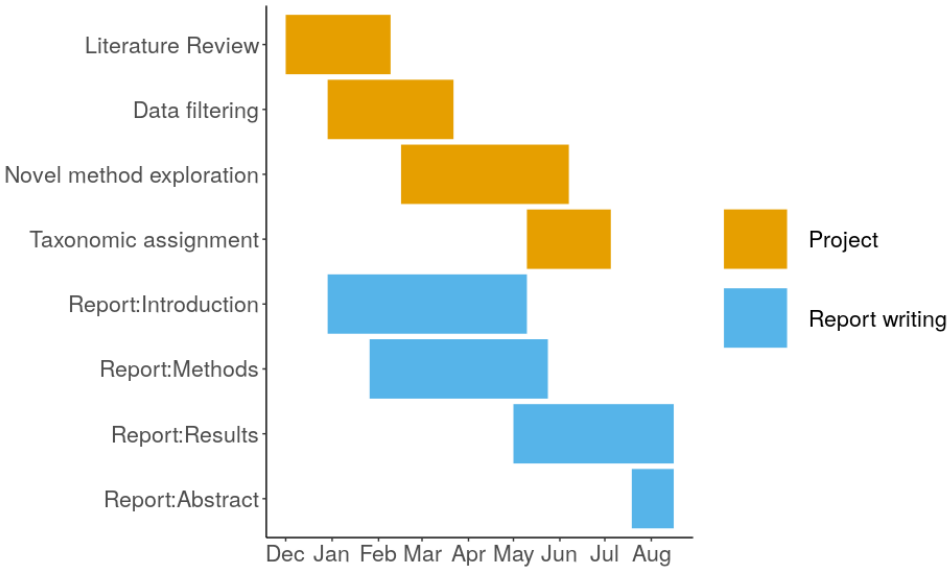
of representative reference ASVs. It is a most plausible method for ASV retention that founded in previous studies, but there is a trade-off between insufficient removal of NUMTs and poor retention of true ASVs copies, in particular if these correspond to rare species or low biomass.
Therefore, apart from applying the developed ASV retention method on tropical beetles dataset, I am also aiming to explore a improved method to recognise the abundance pattern of true ASV copies, NUMTs and other non-targeted reads(e.g. parasite or gut contents) based on relatedness to reference ASVs. The reads that are phylogenetically distant from reference ASV would be considierd as non-target reads then directly get omitted. The reads site on closely related branches would be assign abundance threshold values for true ASV copies retention and NUMTs removals. The new method would be tested with simulating a mock community or more complex community dataset.
Finally, the retained ASV copies will be subjected to BLASTn search against the NCBIn database to concludes the community composition of regional tropical beetles.

Anticipated Outcomes

To apply an automated and valid metabarcoding analytical pipeline for large bulk tropical beetles biodiversity study. Draw an conclusion on diversity of a regional tropical beetles community. Also enhance the accuracy and efficiency of ASVs retention, ultmately provides a more accurate method of taxonomy assignment on complex community.

Project Feasibility

Work timeline represented in the Gantt Chart bleow:



Budgets

Transportation to NHM	£16.55/travel	36 weeks
Total		£595.8

References

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., Wingett, S. (2012). FastQC: A quality control tool for high throughput sequence data. Babraham, UK: Babraham Institute.

Creedy T.J, Norman H, Tang C.Q, Chin K.Q, Andujar C, Arribas P, O'Connor R.S, Carvell C, Notton D.G, Volger A.P,(2019). A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. Molecular Ecology Resource,00, 1-14.

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.

Hebert, P. D. N., Cywinska, A., Ball, S. L., DeWaard, J. R. (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1512), 313-321.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A.,Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters, 16, 1245-1257.

Lebuhn, G., Droege, S., Connor, E. F., GemmillâĂŘHerren, B., Potts, S. G., Minckley, R. L., Parker, F. (2013). Detecting insect pollinator declines on regional and global scales. Conservation Biology, 27, 113-120.

Martin, M. (2011). Cutadapt removes adapter sequences from highthroughput sequencing reads. Embnet, 17, 10-12.

"I have seen and approved the proposal and the budget"

Prof. Alfried Vogler

Dec 11 2019