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

2019-2020

Keywords

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

Introduction

Beetles(Coleoptera) is the largest order of insects, it is particularly diverse and challenging to study its ecology, biodiversity and taxonomy.

Metabarcoding is a robust technique in characterising the complex community compositions and breaking through the barrier in traditional taxonomic methods. It is a method highly suitable for large-scale species richness and complex community composition surveys(Lebuhn et al., 2013).

Even though it is a widely used technique and the molecular procedures are very mature, by applying High-throughput Sequencing on arthropods(Yu et al., 2012 ; Ji et al., 2013), the sequencing data analysis stage is not standarised. Different analytical methods and procedures would potentially lead to diverged conclusions on taxon delimitation.

A previous study in UK bee surveys have built up a pipeline for the taxonomic assignment which showed a high congruence with morphological classification(Creedy et al., 2019). It provides a valid, efficient and standardised workflow in bioinformatic analysis for metabarcoding that apply on bees in UK. However, there are still challenges in selecting th High-throughput barcode sequences.

In insect barcoding analysis, cytochrome oxidase C subunit I (cox1) is commonly used as a barcode marker(Hebert et al., 2003). Amplicon sequence variants(ASVs) should be obtained from the reads after removing the sequencing error from amplication processes or variants resulting from nuclear mitochondrial DNA segments(Numts), internalsed parasits or gut contents of insects. In this project, I will construct a tropical beetles reference dataset for metabarcode analysis, including select ASVs and ignore the other non-targeted reads. IâĂŹll also test the utility of the reference dataset with data in current database and reveal the biodiversity in the regional tropical beetles community.

Methods

1)Building reference profile:

50 species of tropical beetles have been sequenced, these data would be aligned using MAFFT v1.3 (Katoh, Asimenos, Toh, 2009) and the aligned data would be used for distance-based and coalescence-based species delimitation. To separate independent coalescent groups, a phylogenetic tree would be built based on the generalised mixed yule coalescence method(Fujisawa Barraclough, 2013) with BEAST 1.8.1(Drummond Rambaut, 2007). Then a barcode identification number (BIN) for each group would be generated.

2)Handling test data and examine the utility of reference dataset:

A metabarcode library that includes about 7,000 tropical beetles sequences would be used to test the utility of the reference dataset. The sequences were conducted with Illumina MiSeq v.3 and needed to be handled before matching the reference dataset. The adaptor reads from sequencing would be removed by cutadapt(Martin, 2011) and the quality of reads would be reviewed by

FASTQC(Andrews et al.,2012). A perl script that comprises all the raw data filtering and handling functions would be developed and tested, it will eventually be able to generate a set of sequences that are feasible for reference profile testing. The sequences would be filtered by length and would recognise the number of reads in unique sequences and denoised with by using UNOISE algorithm(Edgar et al., 2011). The ASVs could be identified by selecting most frequent reads. However, it is not an ideal solution. A method that recognise the pattern of ASVs, NUMTs and other non-targeted reads would be deveoped and tested with simulation and examined validity through pylogenetic tree-based method. It would enhace the accuracy and efficiency of ASVs capture. Finally, the representitive sequences (the most abundant ones) will be subjected to BLASTn search against the NCBIn database and also subjected to the reference profile which will assess the ability of reference profile as well as concludes the community composition of regional tropical beetles.

Anticipated Outcomes

To develop an automated and valid metabarcoding analytical pipeline for large bulk tropcial beetlees biodiversity study. Also improve the secure methods of ASVs detection and non-targeted sequences removals. Finally, draw an accurate conclusion on diveristy of a regional tropical beetles community

Project Feasibility

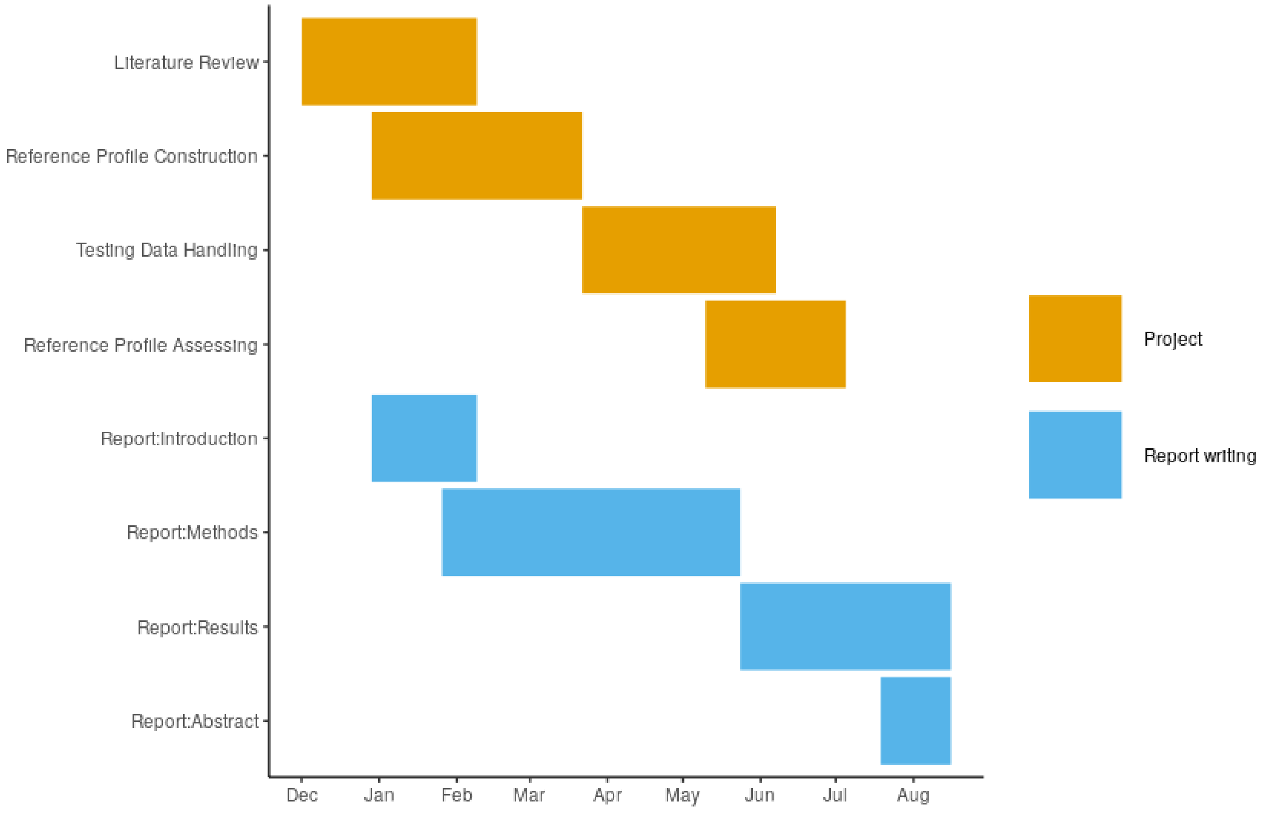

Figure 1: Project time line

Budgets

Transportation for weekly meeting: £16.55/off-peak return travel * 36 weeks Total: £595.8