# Dissertation Proposal

# Metagenomic DNA Discoveries after Sequencing Everything (bacteria, parasites, food, gut) inside a Bee

Cong Liu

cong.liu20@imperial.ac.uk

Supervisor: Dr. Peter Graystock

p.graystock@imperial.ac.uk

Faculty of Natural Sciences, Department of Life Sciences (Silwood Park), Imperial College London

**Key words**

Bee; microbiome; metagenomics; taxonomy composition; functional profile; host-microbiome interactions.

## Introduction

Microbiome studies have been boosted by high throughput sequencing technology, which enables investigations of uncultured microorganisms. For example, amplicon sequencing is a powerful method and has revealed incredible understanding in microbiome (Eckburg et al. 2005, New and Brito 2020, Galloway-Peña and Hanson 2020). However, amplicon sequencing only focuses on a small part of the whole microbial community. Shotgun sequencing of microbiome, or metagenomics, captures all DNA in a sample and can be used in both analysis of taxon composition and function potentiality of microbiome (Quince et al. 2017).

Metagenomics can provide insight into the diversity and functions of microbiome. However, its utilization is hindered by relatively high cost. Shallow shotgun metagenome sequencing is a cheaper method due to its low sequencing depth and modified protocol. It can be a useful tool in metagenomic research (Cattonaro et al. 2018, Hillmann et al. 2018). However, there is also evidence for the impact of sequencing depth on taxonomic and functional profiling of microbiome. (Pereira-Marques et al. 2019, Zaheer et al. 2018, Gweon et al. 2019). Another challenge in utilization of metagenomics is assembling short-read sequencing data. Huge amount of reads in metagenomic data makes assembly via overlapping reads computationally impractical, while assembly by de Bruijn graph approach might not be able to fully cover both obtainment of accurate fragments from high-abundance genomes and detection of low-abundance genomes, for the unequal sequencing coverage of genomes in metagenomic data (AylingClark and Leggett 2020). As for reference-based assembly methods, the availability and quality of database impose limit to their effects.

Here, a hybrid strategy of reference-based and *de novo* assembly is proposed and will be used in assembly of high-sequencing-depth metagenomic data (more than 100,000,000 clean reads for most samples) from bees to investigate potential molecular mechanisms for host-microbiome interactions. Then impact of sequencing depth and availability of reference genomes on microbiome profiling and recovery of known genomic structures will be assessed.

## Proposed Methods

I propose a pipeline for analysis of metagenomic data including the following steps: (1) conduct taxonomic profiling by aligning reads to NCBI non-redundant (nr) database; (2) map reads to genomic data of species identified and (3) conduct *de novo* assembly and gene prediction using un-mapped reads. Assembled fragments are aligned to NCBI nr database to assess their taxon,

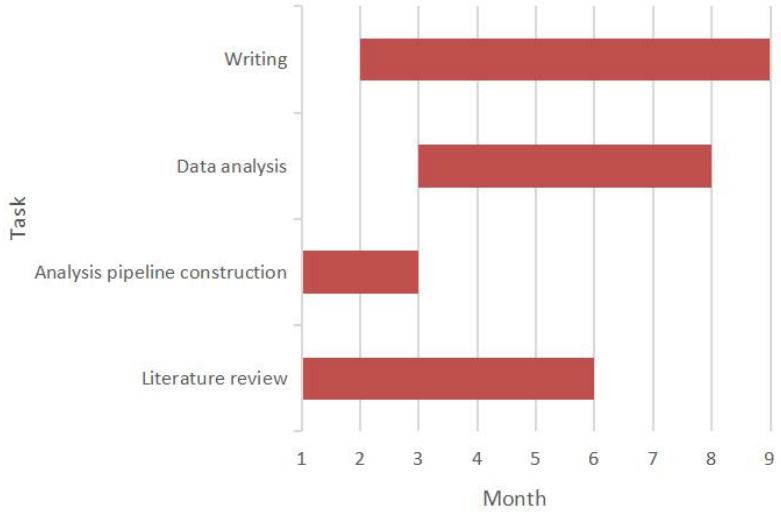
and functional annotations of genes are used to investigate potential mechanisms of host-microbiome interactions.

To simulate the impact of sequencing depth, the original dataset will be randomly subsampled at different threshold and analyzed using the hybrid strategy with a consistent set of references genomes derived from original dataset. Given sequencing depth, the average lateral coverage of reference genomes, the number of genes and taxa at different level will be calculated. All these statistics will be determined as functions of sequencing depth through non-parametric method. Minimal sequencing depth required for the following tasks will be estimated: (1) identification of 95% of genes or taxa compared with original dataset; (2) recovery of reference genomes with coverage larger than 98%. Then I will subsample the set of reference genomes randomly at different threshold and estimate the minimal sequencing depth to explicit its impact on microbiome profiling and recovery of known genomic structures.

**Anticipated outputs and outcomes (including stakeholders involved if applicable)**

Anticipated outputs including a reproducible pipeline for the analysis of metagenomic data and minimal sequencing depth required for microbiome profiling or recovery of specific genomic structures.

**Project feasibility supported by a timeline of tasks (including a Gantt chart)**




**An itemized budget**

8Tb External Harddrive: £200

**References:**


Ayling, M., M. D. Clark & R. M. Leggett (2020) New approaches for metagenome assembly with short reads. *Brief Bioinform,* 21**,** 584-594.

Cattonaro, F., A. Spadotto, S. Radovic & F. Marroni (2018) Do you cov me? Effect of coverage reduction on species identification and genome reconstruction in complex biological matrices by metagenome shotgun high-throughput sequencing. *F1000Res,* 7**,** 1767.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson & D. A. Relman (2005) Diversity of the human intestinal microbial flora. *science,* 308**,** 1635-1638.

Galloway-Peña, J. & B. Hanson (2020) Tools for Analysis of the Microbiome. *Dig Dis Sci,* 65**,** 674-685.

Gweon, H. S., L. P. Shaw, J. Swann, N. De Maio, M. AbuOun, R. Niehus, A. T. Hubbard, M. J. Bowes, M. J. Bailey & T. E. Peto (2019) The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiome,* 14**,** 1-15.

Hillmann, B., G. A. Al-Ghalith, R. R. Shields-Cutler, Q. Zhu, D. M. Gohl, K. B. Beckman, R. Knight & D. Knights (2018) Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems,* 3.

New, F. N. & I. L. Brito (2020) What Is Metagenomics Teaching Us, and What Is Missed? *Annu Rev Microbiol,* 74**,** 117-135.

Pereira-Marques, J., A. Hout, R. M. Ferreira, M. Weber, I. Pinto-Ribeiro, L. J. van Doorn, C. W. Knetsch & C. Figueiredo (2019) Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front Microbiol,* 10**,** 1277.

Quince, C., A. W. Walker, J. T. Simpson, N. J. Loman & N. Segata (2017) Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol,* 35**,** 833-844.

Zaheer, R., N. Noyes, P. R. Ortega, S. R. Cook, E. Marinier, G. Van Domselaar, K. E. Belk, P. S. Morley & T. A. McAllister (2018) Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci Rep,* 8**,** 5890.

**Supervisor Statement**

**I have seen and approved the proposal and the budget.**

Supervisor name: Peter Graystock

Supervisor signature

Date 21/12/2020