

Optimize sequencing depth required for reliable profiling of microbiome
by rarefaction and multimodel inference, using shotgun metagenomic
data from pollination system and a modular profiling pipeline

Cong Liu

2021.8

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Research at
Imperial College London

Submitted for the MRes in Computational Methods in Ecology and Evolution

Declaration:

All datasets used in this project are provided by Dr. Peter Graystock and the project was conducted under his supervision. I am responsible for the development and conduction of the analyses presented.

Abstract

Bee-associated microbiome plays a crucial role in host health and relevant investigations have been boosted by shotgun metagenomics, which provides unique advantages by capturing diversity of multiple taxonomic clades and providing information on function potentiality. However, utilization of shotgun metagenomics is hindered by complexity of data analysis and high cost of sequencing. Here, an integrated pipeline combining assembly-dependent and -independent methods was introduced for taxonomic and functional profiling of shotgun metagenomic data, and a framework of rarefaction and multimodel inference was constructed for optimizing sequencing depth. Both the pipeline and the framework were used for analysis of real metagenomic datasets ($2 \times 150bp$ read pairs) from honey bees, bumble bees and flower washes. The integrated pipeline illustrated taxon composition and metabolic potentiality of metagenome, and provided improvement in species identification compared with an assembly-dependent procedure. Sequencing depth can be optimized for species diversity estimation especially when little emphasis is given to rare species, while functional profiling requires deeper sequencing than taxon profiling. These results are helpful for cost-effective utilization of shotgun metagenomics in investigations of pollination system.

1 Introduction

Pollinators such as bees are crucial to maintain global food security and provide stability of natural systems (Hristov et al., 2020a, Bänsch et al., 2021, Khalifa et al., 2021) but their health is under increasing concern (Amiri et al., 2020). Diverse pathogens and environmental factors threaten bee health and are contributing to their declines. A combination of stressors including parasites, pesticide exposure, invasive species, habitat loss and climate change are all contributing to declines of bee populations. (Brown and Paxton, 2009, Hristov et al., 2020b, Cheng and Ashton, 2021, Zattara and Aizen, 2021).

Bees are associated with a community of microorganisms influencing host health (Engel et al., 2016, Raymann and Moran, 2018). Within the gut of bees, stable microbial communities (microbiomes) play crucial roles in food digestion, parasite defence and chemical detoxification. They mediate food digestion processes including polysaccharides breakdown (Zheng et al., 2019), sucrose hydrolysis (Engel et al., 2012, Lee et al., 2015), and mannose metabolism (Engel et al., 2012, Lee et al., 2015). They also provide protection against pathogens including *Crithidia*, (Koch and Schmid-Hempel, 2011b, Cariveau et al., 2014), *Paenibacillus larvae*, (Ebeling et al., 2016, Forsgren et al., 2010) and *Nosema sp.* (Cariveau et al., 2014, Maes et al., 2016). Besides, bee-associated symbionts can aid in detoxification. They are involved in resistance to both metal and metalloid toxins including cadmium (Rothman et al., 2019b), copper (Rothman et al., 2020) and selenate (Rothman et al., 2019a).

Amplicon sequencing, in which a species-specific barcode region is amplified and sequenced, is a powerful

method for investigations of microbiome (Abdelfattah et al., 2018). It is vastly used in bee microbiome investigations to illustrate the taxonomic diversity of bee-associated bacteria and fungi (*e.g.* Geldert et al. (2021), Wang et al. (2021), Powell et al., Kapheim et al. (2021)). However, bees visit numerous niches during environment exploration and foraging activities, and can get contact with diverse eDNA signatures, which provide insight into bee ecology by reflecting interactions between bees and other organisms including bacteria, fungi, plants, arthropods and viruses (Bovo et al., 2018, Ribani et al., 2020, Bovo et al., 2020, Matsuzawa et al., 2020). It is difficult to explore this diversity via amplicon sequencing because it only captures a fraction of the whole community since analysis of different taxonomic groups is based on different barcode regions, *e.g.* 16S ribosomal RNA (rRNA) for bacteria (Hayashi et al., 2002, Eckburg et al., 2005), internal transcribed spacer (ITS) for fungi (Nilsson et al., 2008), cytochrome c oxidase subunit I (COI) for Animalia (Hebert et al., 2003) and plastid genes for plants (Group et al., 2009). As a result, amplicon sequencing only captures taxon diversity within a certain clade. Besides, it is difficult to illustrate function potentiality of bee-associated microbiome using amplicon sequencing since it does not provide information on content of functional gene clusters (FGCs), *i.e.* aggregates of genes with same function. As a result, functional capacity needs to be inferred based on reference genomes (Aßhauer et al., 2015, Douglas et al., 2018). However, bee bacterial symbionts are diversified at strain level (Engel et al., 2012, Powell et al., 2016, Ellegaard et al., 2020) and bacterial strains are often highly variable in gene content (Cordero and Polz, 2014, Brockhurst et al., 2019). As a result, amplicon sequencing-based inference of functional capacity of bee microbiome may not be reliable.

Shotgun metagenomics provides an alternative for microbiome investigations to overcome drawbacks of amplicon sequencing. By capturing and sequencing DNA fragments unselectively, shotgun metagenomics is capable of providing comprehensive inventories of taxa and FGCs (Quince et al., 2017, New and Brito, 2020, Galloway-Peña and Hanson, 2020). However, utilization of shotgun metagenomics is hindered by challenges in bioinformatics. Typical goal of metagenomics is to provide taxonomic and functional profile of microbiome, and there is not a golden standard for bioinformatics of metagenome. Generally, one of the first steps in metagenomic analysis is assembling short reads into long contigs, which can help improve accuracy of metagenomic annotation (Wommack et al., 2008, Carr and Borenstein, 2014, Tran and Phan, 2020) and is necessary for discovery of novel taxa and genes (Culligan et al., 2014, Youngblut et al., 2020). However, metagenome assembly is complex, compromised by fragmental assembly, chimaeras (Mikheenko et al., 2016) and loss of taxon/function diversity due to unassembled reads (Vollmers et al., 2017, Ayling et al., 2020). Probably because of these shortcomings, assembly is skipped in some researches and short reads are directly proceeded for annotation (Tringe et al., 2005, Abubucker et al., 2012, Vermote et al., 2018, Bovo et al., 2018), although the accuracy can be compromised due to low information load of short

reads (Wommack et al., 2008, Carr and Borenstein, 2014, Tran and Phan, 2020). A combination of both assembly-dependent and -free methods could overcome the complexity and improving accuracy of metagenomic profiling (Becker et al., 2020).

A remaining challenge of metagenomics is the determination of a sequencing depth that provides reliable estimation of taxon/FGC diversity without overspending. It is recommended to retrieve as many reads as possible (Quince et al., 2017) since insufficient sequencing causes compromise in metagenome profiling (Cattonaro et al., 2018, Zaheer et al., 2018, Pereira-Marques et al., 2019, Gweon et al., 2019). However, deep metagenomic sequencing is expensive, which hinders its utilization, especially in large-scale projects. Currently, there are few published guidelines for the sufficient sequencing depth of a given environment or study type in order to reach a trade-off between sequencing effort and reliable output.

In order to balance sequencing cost and reliable estimation of taxon/FGC diversity, expected diversity represented by given sequencing depth need to be computed. Since a metagenomic dataset can be viewed as a random sample of an assemblage of genomic sequences, and profiling is the process by which reads are assigned to taxa or FGCs, the relationship between sequencing depth and diversity can be illustrated by rarefaction (randomly subsampling the original dataset without replacement) and quantified by model fitting, if the original dataset can provide an almost complete inventory of taxa/FGCs (Heck Jr et al., 1975, Hortal and Lobo, 2005, Gómez-Anaya et al., 2014, Hughes et al., 2021). The sequencing depth is sufficient for reliable estimation of taxon/FGC diversity if and only if the slope of the model is small (Hortal and Lobo, 2005, Chao and Jost, 2012, Roswell et al., 2021).

In this project, I utilised metagenomes from three environmental types: the gut of a honey bee (*Apis mellifera*), the gut of a common North American bumble bee (*Bombus impatiens*) and the surface of a flower (*Erigeron annuus*). I aimed to (1) develop an integrated pipeline combining assembly-dependent and -independent methods to deliver improved taxon annotation of sequencing data and (2) optimize sequencing depth to balance sequencing cost and requirement for reliable analysis of microbial species diversity, or the description of their functional diversity. In the integrated pipeline (Figure 1), assembly-dependent taxon profiling is conducted after quality filtering and removing host contamination. To address high negative rate of assembly-dependent search, a reference database of genomes from species present in the assembly is constructed and used for filtering non-host reads. Reads not recruited by the reference database are subjected to assembly-free taxonomic search. This integrated pipeline was used to profile metagenomic datasets involved in this project. Then in order to estimate optimal sequencing depth for reliable representing taxonomic/functional diversity of metagenomes, I simulated different sequencing depth by rarefaction and profiled subsampled datasets using the integrated pipeline, generating inventories of species and FGCs represented by Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologies (KOs)

100 (Kanehisa and Goto, 2000). Diversity of species and KOs was measured by Hill numbers of order 0, 1 and
 101 2. Sequencing depth was measured by number of clean read pairs (150 bp read length). The relationship
 102 between Hill numbers and sequencing depth was quantified by fitting and averaging asymptotic species
 103 accumulation models, and estimation of optimal sequencing depth was given by the point where the slope
 104 of rarefaction curve drops to cut-off values.

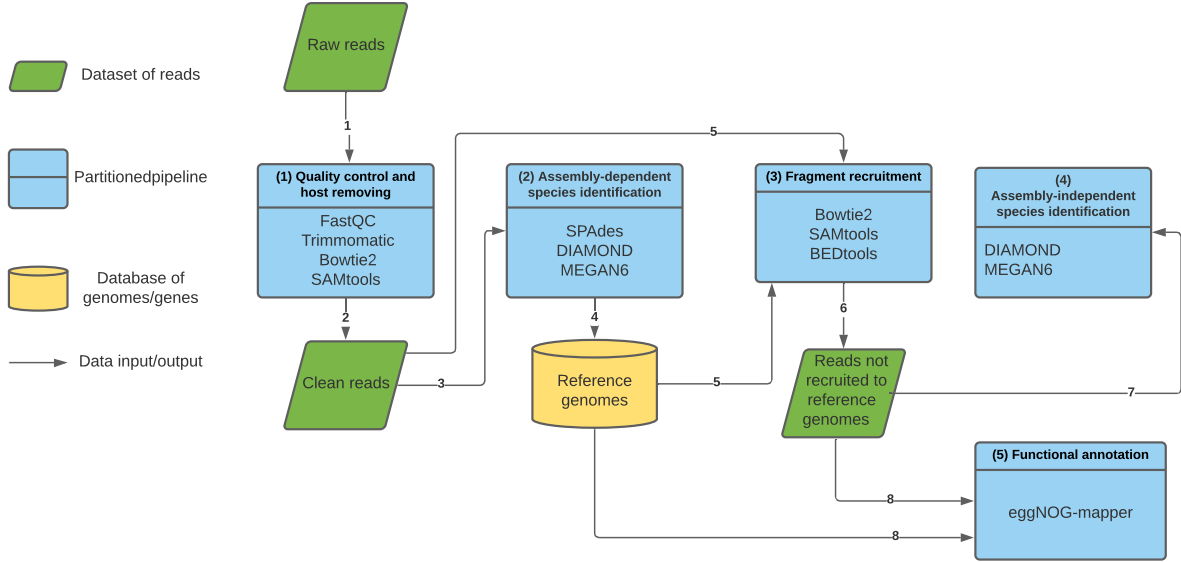


Figure 1: An overview of the integrated pipeline. The pipeline is separated into 5 modules, indicated by blue boxes. Green boxes indicate dataset of reads, while yellow boxes indicate database of genomes. Black arrows indicate the input and output of each step and the numbers on them indicate the order that each step is utilized.

105 2 Materials and Methods

106 2.1 Samples, DNA extraction and sequencing

107 Samples include four honey bees (two from the same hive (hive 13), one from a hive in the same apiary
 108 (hive 15) and one caught foraging), three bumble bees from commercially supplied (Biobest) bumblebee
 109 colonies (two from the same colony), and one buffer wash of a wild flower (*Erigeron annuus*).

110 DNA extraction was performed as in Graystock et al. (2020), followed by library preparation using a
 111 template protocol with Illumina Nextera Library Prep kits. Briefly, this involved tagmentation into
 112 fragments of 300 bases before eight samples of 10ng were pooled together and sent to Beijing Genomics
 113 Institute for further quality control and sequencing using a full lane in the X-ten platform.

114 2.2 Metagenomic profiling using integrated pipeline

115 2.2.1 Integrated pipeline

116 The integrated pipeline was designed for shotgun metagenomic profiling, *i.e.* assigning reads to taxa or
117 FGCs. It is separated into five modules (Figure 1).

118 In quality control and host removing, raw sequencing data quality is checked using FastQC v.0.11.5
119 (Andrews et al., 2010) and filtered by Trimmomatic v.0.39 (Bolger et al., 2014). Then clean reads are
120 mapped to host genome using Bowtie2 v.2.4.2 (Langmead and Salzberg, 2012) and non-host reads are
121 extracted by SAMtools v.1.11 (Li et al., 2009).

122 The non-host reads are subject to the module of assembly-dependent species identification. *De novo*
123 assembly is conducted by SPAdes v.3.15.2 (Prjibelski et al., 2020). Assembled contigs are aligned to
124 NCBI non-redundant (nr) database by DIAMOND v.2.0.7.145 (Buchfink et al., 2015), and assigned to
125 taxa by MEGAN6 (Huson et al., 2007).

126 Then fragment recruitment is conducted. A reference database is constructed and it comprises reference
127 genome dataset, *i.e.* genomic sequences in FASTA format and corresponding genome annotation in
128 general feature format (gff). For each species represented by assembly, its reference genome dataset, if
129 available, is downloaded from NCBI using its *datasets* command-line tool and added to the reference
130 database. Then non-host reads are mapped to the reference database by Bowtie2, and unmapped reads
131 are extracted by SAMtools.

132 Reads not recruited by the reference database are subjected to assembly-independent species identification.

133 They are aligned to NCBI nr database through DIAMOND and assigned to taxa by MEGAN6.

134 Finally, functional annotation is conducted by EggNOG-mapper v.2.1.2 (Huerta-Cepas et al., 2017). It
135 takes coding sequences (CDSs) of genomes in the reference database and reads subject to assembly-
136 independent species identification as input and assigns them to KOs.

137 The integrated pipeline was used for analyzing metagenomic datasets involved in this study and details
138 in parameter settings of each module are described in Supplementary 7.1.

139 2.2.2 Taxon/function quantification and metabolic pathway reconstruction

140 After profiling, identified species and KOs were quantified by calculating relative sequence abundance, *i.e.*
141 proportion of reads assigned to a species/KO in all reads annotated. For species without available reference
142 genomes, their abundances were calculated using reads assigned to them in assembly-independent search.
143 As for taxa with available reference genomes, they may be identified in both assembly-dependent and
144 -independent search due to strain-specific genomic structures that are not present in reference genomes.
145 Their abundances were calculated by summing number of reads that (1) mapped to coding sequences

(CDSs) of reference genomes and (2) assigned to them in assembly-independent search. Reads mapped to non-coding regions were not taken into consideration in order to avoid overestimation since the assembly-independent search was based on aligning reads to nr database, which is composed of proteins. As for KO quantification, CDSs with zero-coverage were excluded. Abundances of KOs were calculated by summing number of reads that (1) mapped to CDSs assigned to KOs and (2) assigned to KOs directly. Extraction of CDSs and calculation of their coverage were conducted by BEDtools v.2.30.0 (Quinlan and Hall, 2010). Metabolic pathways were inferred based on KOs. Reads assigned to plants and arthropods were not included since they were unlikely to represent living organisms. MinPath v.1.6 was used for pathway inference (Ye and Doak, 2009). It finds a minimal set of KEGG pathways that can explain all KOs provided as input.

2.3 Estimation of optimal sequencing depth required for metagenomic profiling

2.3.1 Simulating different sequencing depth by subsampling and measuring diversity by Hill numbers

Species/KO inventories obtained from different sequencing depth was simulated by rarefaction. Since the ratio between numbers of raw and clean reads is dependent on sequencing process and is not influenced by sample type, sequencing depth here refers to number of clean read pairs to exclude variance caused by different proportion of low quality reads in samples of same type. Besides, the expected ratio between host and non-host reads in a metagenomic dataset is dependent on the DNA sample and not impacted significantly by sequencing depth. Thus the proportion of non-host reads in each simulation is expected to be the same with that in the original dataset.

Based on these considerations, I randomly subsampled non-host dataset of each sample, taking 10%-100% of read pairs at interval of 10% by *reformat.sh* script of BBmap v.38.90 (Bushnell, 2014), and profiled subsampled datasets by the integrated pipeline (Figure 1). The sequencing depth of each subsampled dataset equals number of subsampled non-host read pairs divided by ratio between non-host and clean read pairs. Thus, each subsampled dataset of non-host reads is corresponded to an imaginary dataset of clean reads, whose proportion of non-host reads is the same with that of the original metagenomic dataset.

After profiling subsampled datasets, species/KO diversity was measured by Hill numbers of order q , defined as Equation 1 (Hill, 1973).

$$D^{(q)} = \left(\sum_i (p_i)^q \right)^{\frac{1}{1-q}} \quad (1)$$

176 p_i represents the relative abundance of i th species/KO, and q determines sensitivity to relative abundances.
 177 When $q = 0$, abundances are not taken into consideration and $D^{(0)}$ equals species/KO richness. When
 178 $q = 1$, Hill number is defined as the limit of Equation 1 as q tends to 1 (2) and emphasis is given to
 179 species/KOs with general abundances.

$$D^{(1)} = e^{-\sum_i p_i \log p_i} \quad (2)$$

180 When $q = 2$, high leverage is provided to abundant species/KOs and Hill number equals the inverse of
 181 Simpson index (Equation 3).

$$D^{(2)} = \frac{1}{\sum_i (p_i)^2} \quad (3)$$

182 **2.3.2 Quantification of relationship between sequencing depth and Hill numbers by fitting** 183 **rarefaction curves**

184 Hill number of order q (Equation 1) measures diversity of an inventory as the number of equally abundant
 185 categories in an imaginary inventory with the same diversity (Chao et al., 2014, Roswell et al., 2021).
 186 Order q determines leverage given to abundant categories. All Hill numbers behave in the following way:
 187 if a proportion of categories in an inventory was removed randomly, all Hill numbers would decrease by
 188 that proportion (Roswell et al., 2021). Thus, it can be hypothesized that as sequencing depth (number
 189 of clean read pairs) increases, the detection of novel species/KOs leads to increase of Hill numbers, and
 190 when sequencing depth is so big that all species/KOs present in the metagenomic DNA sample have been
 191 detected, Hill numbers level off. Such a relationship can be fitted by asymptotic species accumulation
 192 models.

193 Let Hill number of order q (Equation 1) be a function of sequencing depth x , which takes million read pairs
 194 as the unit. This function was fitted using a multimodel inference method. First, a total of five candidate
 195 models (Table S1) were fitted to rarefaction curves which plots Hill numbers against sequencing depth. R
 196 package *minpack.lm* v.1.2.1, which employs Levenberg-Marquardt nonlinear least-square algorithm, was
 197 used for model fitting. Then small sample unbiased Akaike information criterion (AICc) (Anderson, 2007)
 198 of each candidate model was calculated (Equation 4):

$$AICc = -2L + 2k + \frac{2k(k+1)}{(n-k-1)} \quad (4)$$

199 where n is number of observed data points ($n = 10$ in this study), k is the number of fitted coefficients,
 200 and L is maximized log-likelihood, given by Equation 5.

$$L = -0.5n \log\left(\frac{Rss}{n}\right) \quad (5)$$

201 Rss represents residual sum of squares.

202 Then model averaging was conducted. First, differences of AICc scores between i th candidate models and
 203 the model with lowest AICc value were calculated using Equation 6.

$$\Delta_i = AICc_i - AICc_{min} \quad (6)$$

204 $AICc_i$ is the AICc score of i th plausible model and $AICc_{min}$ is the lowest AICc score among all candidate
 205 models. The Akaike weight of i th model is given by Equation 7 (Anderson, 2007).

$$w_i = \frac{e^{(-0.5\Delta_i)}}{\sum_i e^{(-0.5\Delta_i)}} \quad (7)$$

206 Denote i th candidate model by $D_i^{(q)} = D_i^{(q)}(x)$, the averaged model is given by Equation 8.

$$D^{(q)}(x) = \sum_i w_i D_i^{(q)}(x) \quad (8)$$

207 The slope of rarefaction curve was calculated by first derivative of averaged model (Equation 9). It reflects
 208 the increase rate of the curve.

$$\frac{dD^q}{dx} = \sum_i w_i \frac{dD_i^{(q)}}{dx} \quad (9)$$

209 The asymptote of rarefaction curve as sequencing depth tends to infinity is given by Equation 10. It
 210 provides an estimation of the total diversity and is comparable among metagenomic DNA samples with
 211 different sequencing depth. However, the accuracy of asymptotic estimators is controversial (Colwell and
 212 Coddington, 1994, Chazdon et al., 1998, Jimenez-Valverde et al., 2006, Hortal et al., 2006).

$$\lim_{x \rightarrow +\infty} D^{(q)}(x) = \sum_i w_i \lim_{x \rightarrow +\infty} D_i^{(q)}(x) \quad (10)$$

213 2.3.3 Estimating optimal sequencing depth using rarefaction curves

214 Optimal sequencing depth is defined as the point at which diversity starts to level off as sequencing depth
 215 increases, and its precise estimation via rarefaction is based on the assumption that the original dataset
 216 is sufficient for detection of almost all species/KOs present. This assumption can be verified by looking at

the rarefaction curve that plots species/KO richness (Hill number of order 0) against sequencing depth. The original dataset is sufficient for providing a reliable inventory if and only if the rarefaction curve of richness is characterized by a small final slope (Heck Jr et al., 1975, Hortal and Lobo, 2005, Chao and Jost, 2012). Then an estimation of optimal sequencing depth is provided by the point at which the slope of rarefaction curve decreases to a given cut-off value (Hortal and Lobo, 2005, Gómez-Anaya et al., 2014).

3 Results

3.1 Sequence reads

Eight samples (four honey bees, three bumble bees and one flower eDNA) were sequenced. The quality reports of raw reads showed low-quality 3'-end (Figure S1a), uneven base content in 5'-end (Figure S1b) and the present of adaptors (Figure S1c). The quality control procedure covered these aspects and improved data quality (Figure S1d, S1e and S1f).

After quality control, read pairs aligned to host genome were removed. Table S2 reports numbers of raw, clean and non-host read pairs. Honey bee sample *Bee_Amellifera_13_1* was filtered from further analysis since its low raw read pair number (1.10 million compared to the other samples at 59 million) is suggestive of a poor quality sample. After quality control, 62.08%-76.52% of raw read pairs were retained for these three honey bee samples. Then a different proportion of non-host read pairs (29.52%-86.19%) were retained. As for bumble bee samples, about 58 million raw read pairs were obtained for each sample and 82.1%-83.8% were retained after quality control. After host removing, 7.35%-10.08% of clean reads were retained. For flower eDNA sample, 1.44 million raw read pairs were obtained and 61.15% of them were retained.

3.2 Application of integrated pipeline

The integrated pipeline was used to profile metagenomic datasets from pollination system, illustrating its capacity in presenting taxon composition and functional potentiality of microbiome. Although samples are different in proportion of host contamination (Table S2, Figure 2a), diverse communities composed of multiple taxonomic clades were identified (Figure 2b), and most species identified are common in pollination system.

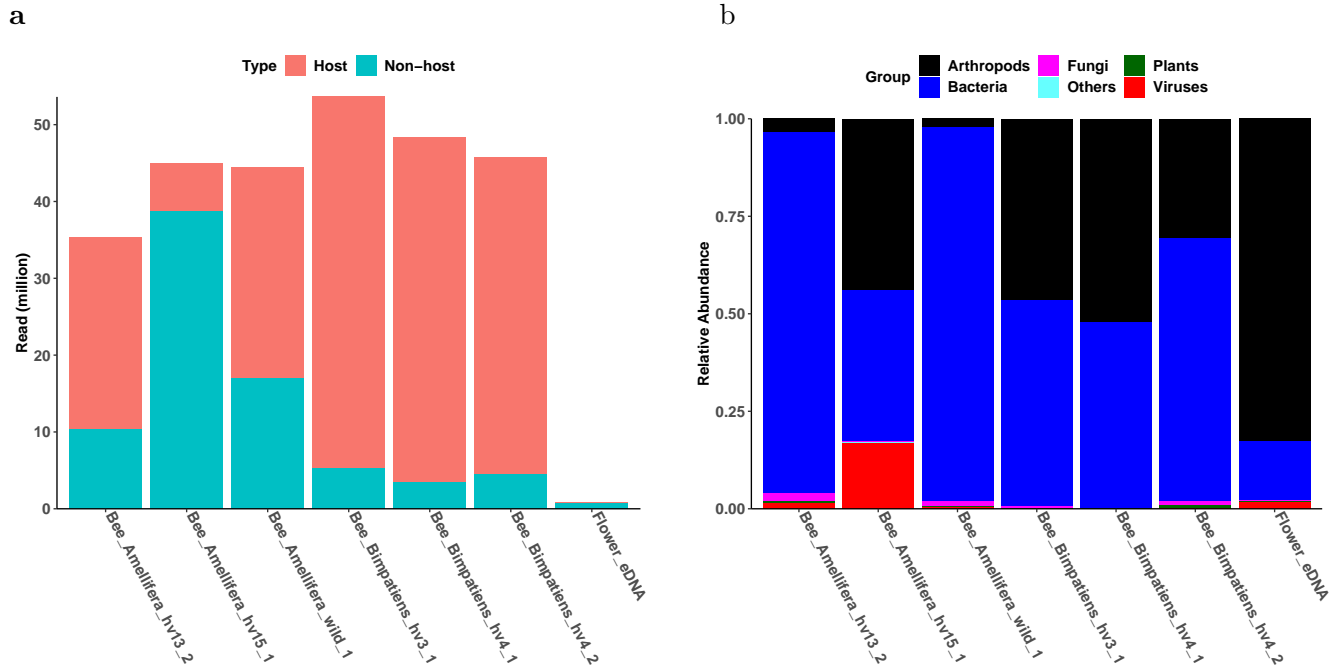


Figure 2: a. The number of host and non-host reads in each sample. b. The relative abundance of species under six taxonomic groups: superkingdom Viruses, superkingdom Bacteria, kingdom Viridiplantae (plants), kingdom Fungi, phylum Arthropoda and others (species that are not in the other five groups).

The present of core members of bee-associated bacterial community including species within *Bifidobacterium*, *Frischella*, *Gilliamella*, *Snodgrassella*, *Lactobacillus*, *Apilactobacillus* and *Bombilactobacillus* (Figure 3) (Koch and Schmid-Hempel, 2011a, Moran, 2015, Kwong et al., 2017, Zheng et al., 2020) suggests good sample quality. As for other abundant bacteria, *Fructobacillus* sp. are often found in fructose-rich environments like flowers (Endo and Dicks, 2014); *Bartonella apis* is related to animal pathogens (Kešnerová et al., 2016) and is widespread in honey bee workers (Raymann and Moran, 2018); *Candidatus Schmidhempelia bombi* is a known uncultured symbiont of *Bombus impatiens* (Martinson et al., 2014). It should be noted that some typical bee-associated bacteria were also found on the flower, including *Bartonella apis*, *Bifidobacterium asteroides*, *Bombilactobacillus mellis* and *Gilliamella apicola*. Composition of arthropods and plants indicating interactions within pollination networks. Most arthropods identified are pollinators within *Apis* and *Bombus* (Figure S2). However, some of them might be considered as false positive. For example, *Apis cerana*, *Apis dorsata* and *Apis florea* are mainly found in Asia and unlikely to present in the area where samples were collected. These might derive by similarity between genomes of *Apis mellifera* and other *Apis* species. As for plant species (Figure S3), they indicate foraging areas of bees. Several crop species were identified, including *Brassica napus* (rape), *Brassica oleracea*, *Cicer arietinum* (chickpea), *Glycine max* (soybean), *Helianthus annuus* (sunflower), *Nicotiana sylvestris* (flowering tobacco) and *Raphanus sativus* (radish).

260 Fungal and virus species were identified. In fungal communities of most samples, *Nosema ceranae*, a
 261 widespread bee pathogen, was the dominant species. (Figure S4). Besides, three yeast species (*Clavispora*
 262 *lusitaniae*, *Saprochaete ingens* and *Wickerhamiella sorbophila*) were also found in bees. As for viruses,
 263 most of them are phages or arthropod-infecting species (Figure S5). Phage species include *Bifidobac-*
 264 *terium phage BitterVaud1* infecting bee-commensal bacterium *Bifidobacterium asteroides* (Bonilla-Rosso
 265 et al., 2020); *Bacteriophage sp.* infecting *Pseudomonas aeruginosa* (Essoh et al., 2015), an opportunistic
 266 pathogen that might contaminate bees (Bailey, 1968, Papadopoulou-Karabela et al., 1992, 1993); and un-
 267 classified species within Myoviridae and Siphoviridae. Listed arthropod-infecting viruses including *Apis*
 268 *mellifera filamentous virus* and *Bombus cryptarum densovirus* that infect bees; and several parvoviruses
 269 (*Blattodean pefuambidensovirus 1*, *Hemipteran scindoambidensovirus 1*, *Hymenopteran scindoambidenso-*
 270 *virus 1* and *Orthopteran scindoambidensovirus 1*) (Pénzes et al., 2020).

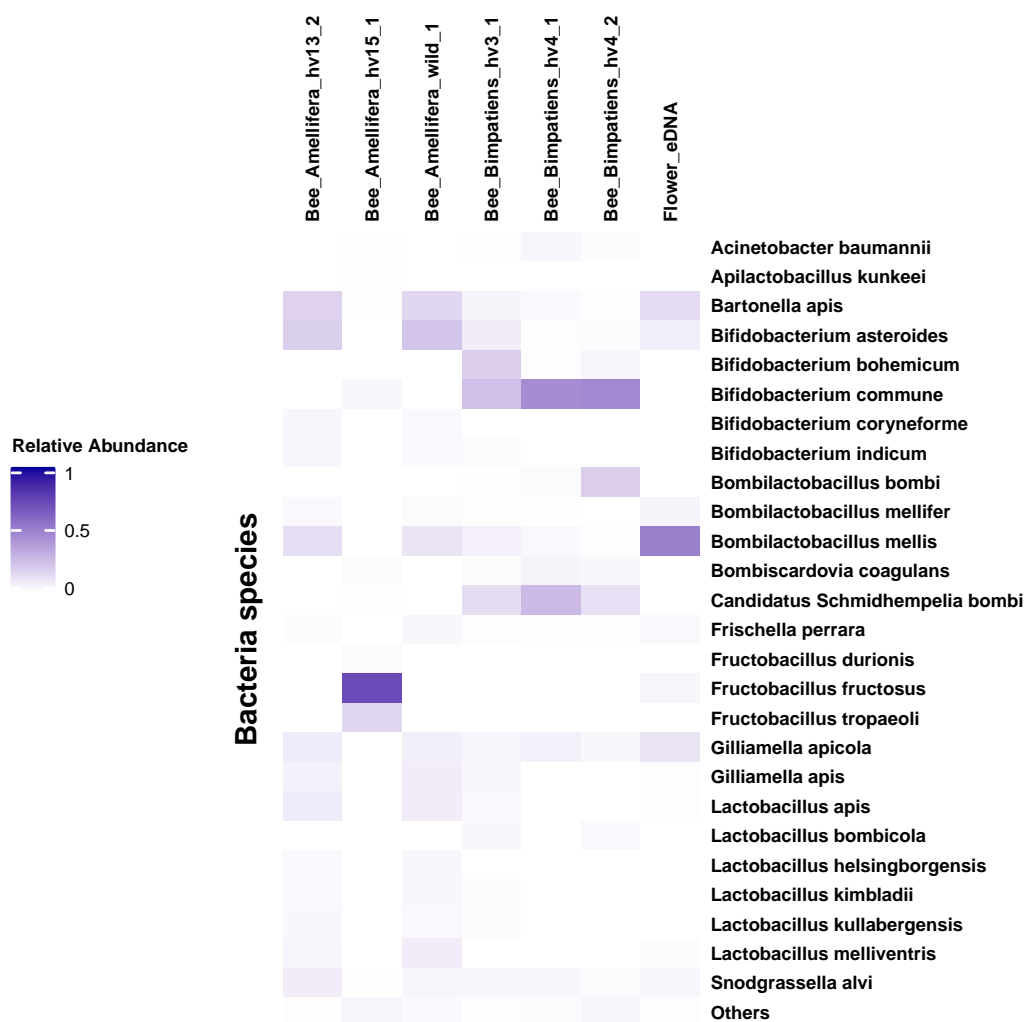


Figure 3: Heatmaps for bacterial species abundance distribution in all samples. The relative abundance takes reads assigned to bacterial species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

271 The integrated pipeline also provides information on FGC content, which shows function potentiality
 272 of metagenome. Here FGC content was represented by identified KOs, which were used for KEGG
 273 pathway inference in order to illustrate metabolic potentiality of metagenomic samples. Reads assigned to
 274 plants and arthropods were not involved in pathway inference since they are unlikely to represent living
 275 organisms. The coverage of a pathway was calculated by the ratio between number of annotated KOs
 276 and total number of KOs involved in that pathway.

277 Here concern is given to metabolism pathways of carbonhydrates and amino acids, which are crucial for
 278 bee health. Inferred pathways indicate potential capability of metabolism of sugars including fructose,
 279 sucrose, mannose and galactose (Figure S6), and all ten essential amino acids for honey bees (*i.e.* arginine,
 280 histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan and valine) (Figure
 281 S7) (Groot, 1953).

282 3.3 Evaluation of performance of integrated pipeline in species identification

283 The performance of integrated pipeline in species identification was evaluated by comparing it with
 284 assembly-dependent method, using rarefied datasets. In integrated pipeline (Figure 1), clean non-host
 285 reads are first assembled to into contigs and assigned to taxa. Then a reference database composed of
 286 genomes of assembly-represented species is constructed. Reads not aligned to the reference database
 287 are subjected to assembly-independent taxon search. The reference database and assembly-independent
 288 search helped improve species identification in all three sample types, especially in simulations of low
 289 sequencing depth (Figure 4).

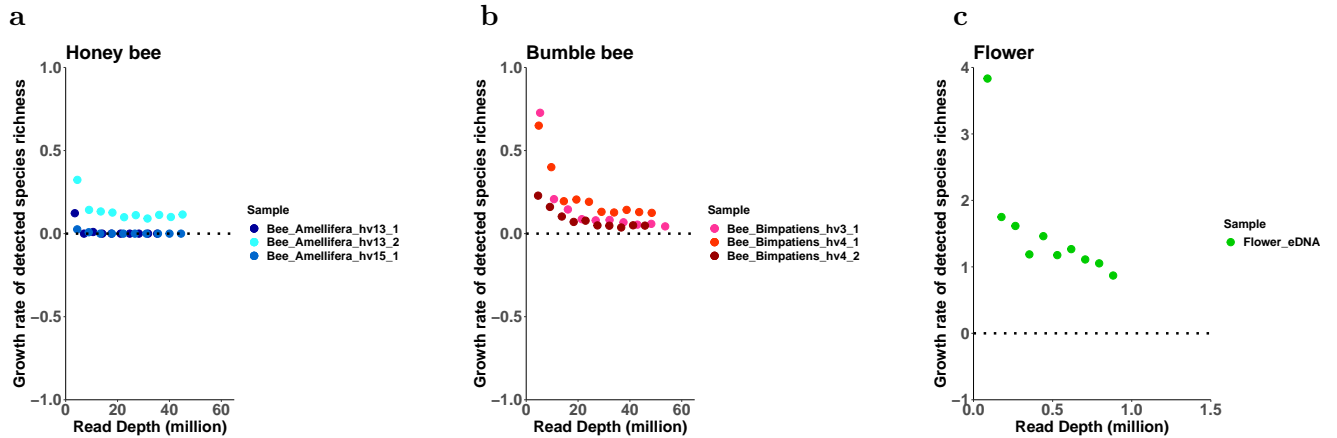


Figure 4: Integrated pipeline improves the detection of species richness. The horizontal axis represents sequencing depth, and the vertical axis represents growth rate of detected species richness comparing the integrated pipeline and assembly-dependent species identification. Different sequencing depth was simulated by rarefaction. Sample type is shown in the top left of each subfigure.

290 3.4 Optimal sequencing depth required for detection of species and function diversity

291 In order to determine the influence of sequencing depth and thus optimize sequencing depth for analysing
292 taxonomic and functional diversity, rarefaction analysis was conducted. Different sequencing depth was
293 simulated by randomly subsampling original datasets at proportions from 10% to 90% at an interval of
294 10%. The relationship between sequencing depth (clean read pair number) and species/KO diversity (Hill
295 numbers of order 0, 1, 2) was quantified by fitting and averaging asymptotic species accumulation models.
296 The slope of the model reflects the increase rate of diversity. The point at which it drops to a cut-off value
297 provides an estimation of optimal sequencing depth. Besides, the asymptote of the model as sequencing
298 depth tends to infinity provides an estimation of total diversity.

299 Rarefaction assumes that the original dataset provides an almost complete inventory, which can be verified
300 by final slope of rarefaction curve for Hill number of order 0 (richness). Figure 5 shows rarefaction curves
301 for species/KO richness and Table 1 summarizes their final slopes. For species diversity rarefaction,
302 all bumble bee samples are sufficient, with final slopes < 0.1 and completeness (ratio between final
303 richness and asymptote) > 0.98 . As for honey bees, *Bee_Amellifera_hv15_1* and *Bee_Amellifera_wild_1* are
304 sufficient, while *Bee_Amellifera_hv13_2* is insufficient, with final slope > 1 and completeness < 0.8 . For the
305 flower eDNA sample, the final slope of species richness rarefaction curve is 10.8380 and its completeness
306 is 1.32%, indicating more sequencing effort is needed for species profiling. As for function diversity
307 rarefaction, the final slopes of all KO richness rarefaction curves are higher than 15, indicating no dataset
308 can providing an almost complete inventory of KOs. Thus, estimation of optimal sequencing depth was
309 conducted for the combinations of two sample types (honey bee and bumble bee) and one study type
310 (species profiling), based on five datasets (two honey bees and three bumble bees).

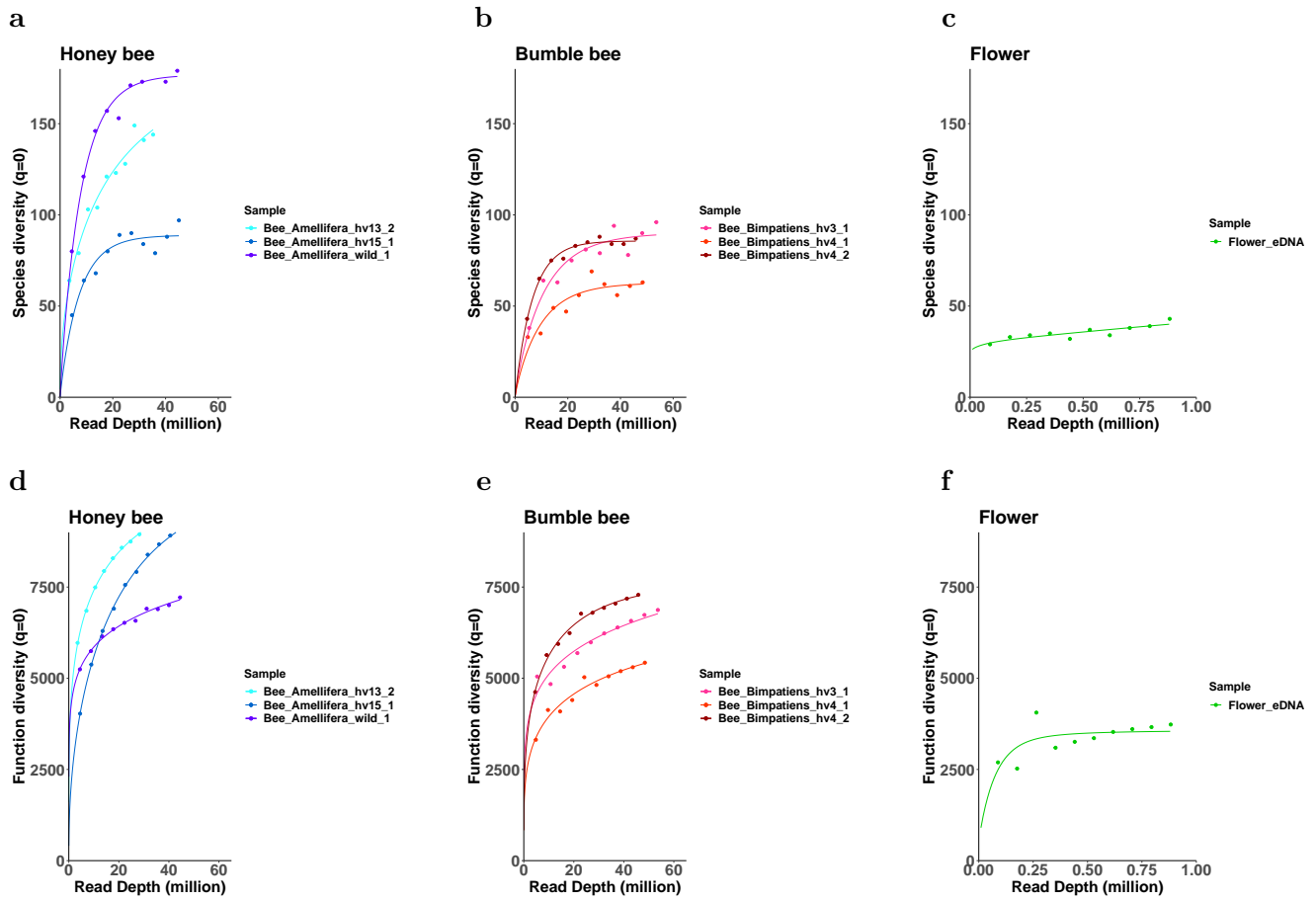


Figure 5: Rarefaction curves for species (a, b, c) or KO (d, e, f) richness (Hill number of order 0). The horizontal axis represents sequencing depth, and the vertical axis represents richness. Sample type is shown in the top left of each subfigure. Note that the scale of horizontal axis in subfigure c and f is much smaller than that in other subfigures.

Table 1: Summary of final point of rarefaction curve for species/KO richness (Hill number of order 0). Type: indicates whether this row reports rarefaction curve for species or KO richness. Depth: sequencing depth taking million read pairs as unit. OR: observed species/KO richness. ER: expected species/KO richness predicted by modeling rarefaction curve. FS: final slope of rarefaction curve. Asym: asymptote calculated by modeling rarefaction curve. Comp: completeness, represented by ratio between ER and Asym.

Sample	Type	Depth	OR	ER	FS	Asym	Comp
<i>Bee_Amellifera_hv13_2</i>	species	35.28	144	146.69	1.0958	184.94	0.7931
<i>Bee_Amellifera_hv15_1</i>	species	45.05	97	88.53	0.0344	88.84	0.9966
<i>Bee_Amellifera_wild_1</i>	species	44.46	179	175.89	0.0991	176.84	0.9946
<i>Bee_Bimpatiens_hv3_1</i>	species	53.61	96	89.00	0.0897	90.79	0.9803
<i>Bee_Bimpatiens_hv4_1</i>	species	48.42	63	62.06	0.0626	62.92	0.9862
<i>Bee_Bimpatiens_hv4_2</i>	species	45.80	87	85.59	0.0146	85.69	0.9988
<i>Flower_eDNA</i>	species	0.88	43	40.04	10.8380	3034.88	0.0132
<i>Bee_Amellifera_hv13_2</i>	KO	35.28	9289	9284.45	38.6844	10641.27	0.8725
<i>Bee_Amellifera_hv15_1</i>	KO	45.05	9046	9102.78	43.8717	10453.98	0.8707
<i>Bee_Amellifera_wild_1</i>	KO	44.46	7216	7151.72	21.4368	14779.89	0.4839
<i>Bee_Bimpatiens_hv3_1</i>	KO	53.61	6872	6769.80	21.1588	28111.79	0.2408
<i>Bee_Bimpatiens_hv4_1</i>	KO	48.42	5427	5410.71	19.6932	7665.58	0.7058
<i>Bee_Bimpatiens_hv4_2</i>	KO	45.8	7288	7259.56	15.6309	7873.64	0.9220
<i>Flower_eDNA</i>	KO	0.88	3732	3544.54	60.4367	3571.11	0.9926

Optimal sequencing depth is estimated by the point at which the slope of rarefaction curve drops to a cut-off value. Table 2 summarizes estimations of optimal sequencing depth for detection of species diversity, using 0.5, 0.1, 0.05 and 0.01 as cut-off values. When order q of Hill number equals 0, *i.e.* species abundances are not considered, cut-off value of 0.1 for slope of rarefaction curve is sufficient for providing completeness $> 97\%$ in all samples. The average optimal sequencing depth are 40.33 million for honey bees and 42.49 million for bumble bees. When species abundances are considered (order q equals 1 or 2), cut-off value of 0.01 can provide completeness $> 95\%$ in most pairs of sample and q value. For honey bees, the average optimal sequencing depth are 18.57 million ($q = 1$) and 17.45 million ($q = 2$). For bumble bee samples, the average optimal sequencing depth are 40.33 million ($q = 1$) and 24.77 million ($q = 2$).

Table 2: Summary of minimal sequencing depth estimated from different cut-off values of slope. Minimal sequencing depth was estimated by the point at which the slope of rarefaction curve drops to a cut-off value (0.5, 0.1, 0.05 or 0.01), which is indicated in names of columns. For example, MinD_0.5 represents estimated minimal sequencing depth taking 0.5 as cut-off value, and ED_0.5 is expected Hill number from MinD_0.5. Comp_0.5 is the ratio between OptDiv_0.5 and the asymptote (represented by Asym). Unit of sequencing depth is million read pairs. q refers to order of Hill number determining sensitivity to species abundance distribution.

Sample	q	Asym	MinD_0.5	ED_0.5	Comp_0.5	MinD_0.1	ED_0.1	Comp_0.1	MinD_0.05	ED_0.05	Comp_0.05	MinD_0.01	ED_0.01	Comp_0.01
<i>Bee_Amellifera_hv15_1</i>	0	88.84	23.54	84.85	0.9551	36.27	87.99	0.9905	41.95	88.40	0.9951	55.75	88.74	0.9989
<i>Bee_Amellifera_wild_1</i>	0	176.84	30.53	172.46	0.9752	44.39	175.88	0.9946	50.57	176.33	0.9971	65.86	176.70	0.9992
<i>Bee_Amellifera_hv15_1</i>	1	12.28	3.18	10.67	0.8699	5.83	11.33	0.9224	7.21	11.43	0.9304	15.21	11.58	0.9429
<i>Bee_Amellifera_wild_1</i>	1	14.77	0.69	13.49	0.9137	3.30	14.03	0.9498	6.20	14.23	0.9635	21.93	14.58	0.9868
<i>Bee_Amellifera_hv15_1</i>	2	8.40	3.10	6.35	0.7561	8.17	7.49	0.8909	11.97	7.75	0.9229	27.99	8.11	0.9655
<i>Bee_Amellifera_wild_1</i>	2	9.88	0.18	9.49	0.9602	0.90	9.63	0.9747	1.73	9.69	0.9805	6.91	9.80	0.9916
<i>Bee_Bimpatiensi_hv3_1</i>	0	90.79	30.80	83.69	0.9218	52.03	88.85	0.9786	62.56	89.61	0.9870	92.35	90.32	0.9948
<i>Bee_Bimpatiensi_hv4_1</i>	0	62.92	24.83	57.22	0.9093	42.76	61.61	0.9791	51.24	62.22	0.9887	73.55	62.76	0.9973
<i>Bee_Bimpatiensi_hv4_2</i>	0	85.69	21.75	82.30	0.9604	32.67	85.01	0.9920	37.39	85.35	0.9960	48.40	85.62	0.9992
<i>Bee_Bimpatiensi_hv3_1</i>	1	21.59	3.63	10.50	0.4862	9.47	11.76	0.5449	15.37	12.18	0.5639	58.09	13.04	0.6040
<i>Bee_Bimpatiensi_hv4_1</i>	1	11.85	3.80	9.45	0.7973	10.15	10.87	0.9174	14.68	11.19	0.9445	31.22	11.57	0.9766
<i>Bee_Bimpatiensi_hv4_2</i>	1	12.66	4.17	10.24	0.8087	10.47	11.67	0.9213	14.85	11.98	0.9458	31.68	12.36	0.9760
<i>Bee_Bimpatiensi_hv3_1</i>	2	8.73	2.20	6.16	0.7051	6.28	7.06	0.8088	9.45	7.28	0.8344	23.89	7.60	0.8708
<i>Bee_Bimpatiensi_hv4_1</i>	2	8.33	3.12	6.28	0.7540	8.52	7.50	0.9002	12.30	7.76	0.9323	26.68	8.09	0.9718
<i>Bee_Bimpatiensi_hv4_2</i>	2	7.12	2.90	5.34	0.7493	7.47	6.38	0.8961	10.58	6.60	0.9270	23.73	6.90	0.9683

4 Discussion

Here, I constructed an integrated pipeline combining assembly-dependent and -independent methods for taxonomic and functional profiling of shotgun metagenomics, and applied it to analysis of metagenomes from honey bees, bumble bees and flower washes. The profiling results showed that the integrated pipeline is able to capture taxonomic diversity by identifying species from multiple clades and infer metabolic potentiality of metagenome by providing information on FGC content. It can be more sensitive in species identification compared with standard assembly-dependent methods. Then I computed expected species/FGC diversity represented by given sequencing depth through rarefaction analysis, in order to optimize sequencing depth to balance cost and reliability of analysis results. Optimal sequencing depth differs by types of samples and investigations. For estimation of species diversity, sequencing depth can be optimized especially when low leverage is given to rare species. As for functional profiling, deeper sequencing depth is required.

The integrated pipeline (Figure 1) for taxonomic and functional profiling of metagenome provides several advantages. First, through combination with assembly-free taxon search, it helps solve high false negative rate associated with assembly-dependent species identification, which is caused by reads left unassembled. This was shown by analyzing real metagenomic datasets from pollination system and simulating differ-

ent sequencing depth by rarefaction (Figure 4). Second, the modularity of integrated pipeline provides flexibility for incorporation of alternative tools. For example, BWA aligner (Li and Durbin, 2009) serves as an alternative for Bowtie2, and SPAdes can be replaced by other metagenomic assemblers such as Megahit (Li et al., 2015) and IDBA-UD (Peng et al., 2012). Third, the output files generated by each step are recorded and can be inspected easily, which provides transparency for troubleshooting. However, the performance of integrated pipeline need to be evaluated more comprehensively by analysis of mock datasets and comparison with other profiling strategies. Standardized mock dataset can be used to benchmark sensitivity and accuracy of integrated pipeline since it is an artificial metagenome with predefined diversity, generated by combining known sequences from different species (Vollmers et al., 2017, ?, Becker et al., 2020). Besides, the performance of integrated pipeline need to be compared with other widely used analysis strategies, such as MG-RAST (Meyer et al., 2008), SqueezeMeta (Tamames and Puente-Sánchez, 2019) and Kraken (Wood and Salzberg, 2014).

The integrated pipeline was used to analyze metagenomes of bees and flower washes, illustrating their taxon composition and metabolic potentiality. Bee-associated microorganisms were identified in flower washes, including both pathogens (*e.g. Nosema ceranae* and *Apis mellifera filamentous virus*) and symbionts (*e.g. Bifidobacterium asteroides*, *Bombilactobacillus mellis* and *Gilliamella apicola*) (Figure 3, S4 and S5). Flowers have been indicated as hubs for transmission of bee pathogens Durrer and Schmid-Hempel (1994), Koch et al. (2017). Through shared flower use, multiple pathogens transmit between pollinators, including *Nosema ceranae* (Fürst et al., 2014) and *Crithidia bombi* (Figueroa et al., 2019). There is also growing evidence for present of bee symbionts on flower (McFREDERICK et al., 2012, McFrederick et al., 2017, Keller et al., 2020, Vannette, 2020). However, its role in bee microbiome assembly and functional importance in pollination remains an open question (Keller et al., 2020, Vannette, 2020). As for function profiling, it is indicated that bee-associated microbiome is capable of metabolizing carbohydrates such as glucose, fructose, sucrose and mannose (Figure S6). Carbohydrates are main component of bee diet and microbiome-mediated carbohydrate-processing have been vastly investigated (Engel et al., 2012, Lee et al., 2015, 2018, Taylor et al., 2019). Besides, pathways for metabolism of all ten essential amino acids for honey bees (Groot, 1953) were inferred (Figure S7). Essential amino acids are crucial for bee health (Simcock et al., 2014, Paoli et al., 2014, Stabler et al., 2015, Hendriksma et al., 2019) and influence feeding preference due to their potential deficiency in single pollen source (Cook et al., 2003, Hendriksma et al., 2014, Hendriksma and Shafir, 2016). Whether bee-associated microbiome influences host health by providing amino acids need to be further investigated.

Optimization of sequencing depth is important for shotgun metagenomics since insufficient sequencing causes underestimation of taxonomic/functional diversity (Cattonaro et al., 2018, Zaheer et al., 2018,

Gweon et al., 2019, Pereira-Marques et al., 2019), while deep sequencing is of high cost. Here, expected species/KO diversity provided by given sequencing depth was estimated by rarefaction and model fitting, using datasets from honey bees, bumble bees and flower washes. It was shown that increasing sequencing depth boosted identification of species/KO (Figure 5, Table 1), highlighting value of deep metagenomic sequencing. For function profiling, no dataset involved here is big enough to provide an almost complete inventory of KOs (Table 1) even though most samples were deeply sequenced (Table S1), indicating such task is demanding about sequencing depth. Therefore, when dealing with function potentiality of microbiome associated with bees or flowers, retrieving as many reads as possible would be recommended. As for species profiling, although deep sequencing is still valuable, sequencing depth can be optimized when the budget is limited (Table 2), especially when low leverage is given to rare species. For honey bees, approximate 40 million sequencing depth (12 Gbp) can be sufficient for representing species richness. When assessing biodiversity with reduced emphasis on rare species, 17-19 million sequencing depth (5.1-5.7 Gbp) can provide robust estimation. As for bumble bees, about 40-43 million sequencing depth (12-12.9 Gbp) can provide reliable estimation for species richness and diversity index with emphasis on species of general abundance, and about 25 (7.5 Gbp) million sequencing depth can be sufficient for estimation of biodiversity with high leverage on abundant species. It should be noted that the efficacy of optimal sequencing depth estimation is reduced by limited number of samples. Here, honey bee and bumble bee associated microbiome was represented by only three samples, and there was only one sample of flower eDNA. However, microbiomes of pollination system are highly dynamic and variable in diversity. To generate reliable guideline for sequencing depth optimization, more samples need to be evolved. Besides, there is a lack of repeat in sequencing depth subsampling. Repeated subsampling boosts precise computation of rarefaction curve and increases accuracy of estimation of optimal sequencing depth.

5 Conclusion

Shotgun metagenomics is capable of illustrating diversity of multiple taxonomic clades and gene content, and thus provides unique advantages over vastly used amplicon sequencing, particularly for investigations of highly diverse microbial communities. However, utilization of shotgun metagenomics is hindered by challenges in data analysis and high cost of sequencing. Here, I constructed an integrated pipeline for analysis of shotgun metagenomic data. It provides benefits in terms of results, flexibility and transparency. I also constructed a framework for optimizing depth of shotgun metagenomic sequencing in order to balance high cost of sequencing and reliability of analysis results. The pipeline and the framework were used for analysis of real datasets from pollination system, providing guidelines for cost-effective metagenomic investigations of bee- and flower-associated microbiomes.

401 6 Data and Code Availability

402 Scripts used for the analyses are available at [github](#). Metagenomic datasets involved in this project cannot
403 be made available publicly since they are yet to be formally published.

404 7 Supplementary

405 7.1 Parameter settings of integrated pipeline

406 7.1.1 Quality control and host removing

407 Raw data quality control was conducted to reduce compromise from low quality reads. Raw data qual-
408 ity was checked using FastQC v.0.11.5 (Andrews et al., 2010) before filtering. FastQC reports of raw
409 reads showed the following aspects need to be covered in quality control: (1) low base quality in 3'-end
410 (Figure S1a); (2) uneven base content in 5'-end (Figure S1c) and (3) the present of Nextera adaptors
411 (Figure S1e). Therefore, raw reads were filtered using Trimmomatic v.0.39 (Bolger et al., 2014), which
412 (1) trimmed adaptors; (2) cutted 15 bases from the 5'-ends of reads; (3) cutted bases off from 3'-ends of
413 reads if Phred-33 quality is below 20; (4) dropped reads shorter than 50 bp; (5) dropped reads if average
414 Phred-33 quality is below 20. Then unpaired reads were removed and quality of clean data was checked
415 using FastQC (Figure S1d, S1e, S1f).

416 After quality control, host read pairs were removed. First, clean read pairs were mapped to host
417 genome (GCA_003254395.2 for *Apis mellifera* and GCA_000188095 for *Bombus impatiens*, downloaded
418 from NCBI) using Bowtie2 v.2.4.2 (Langmead and Salzberg, 2012) with flags *-end-to-end* and *-sensitive*.
419 With flag *-end-to-end*, Bowtie2 requires the read aligned without any clipping from neither end, and
420 *-sensitive* maintains a trade-off between speed and sensitivity. SAM files generated by Bowtie2 were con-
421 verted to BAM format using SAMtools v.1.11 (Li et al., 2009). Then non-host read pairs were extracted
422 from BAM files by SAMtools according to the present of SAM flag 12 (neither forward nor reverse read
423 in a pair of reads is mapped).

424 7.1.2 Assembly-dependent species identification

425 In order to identify taxa of metagenome, *de novo* assembly was conducted. Non-host read pairs were
426 assembled using SPAdes v.3.15.2 (Prjibelski et al., 2020). Values of k-mer ranged from 21 to 101 at interval
427 of 10. Flags *-only-assembler* and *-meta* were used. Through flag *-meta*, SPAdes runs metaSPAdes which
428 is developed for metagenomic assembly (Nurk et al., 2017). The *-only-assembler* flag skips read error
429 correction and runs assembly only. Its utilization is justified by the following facts. First, when *-only-*

430 *assembler* is not used, SPAdes conducts error correction before assembly. It is conducted by BayesHamme,
 431 which is optimized for single cell sequencing instead of shotgun metagenomics (Nikolenko et al., 2013).
 432 Besides, reads used for *de novo* assembly had been filtered to ensure their quality.
 433 After assembly, taxon identification was conducted using DIAMOND v.2.0.7.145 (Buchfink et al., 2015)
 434 and MEGAN6 (Huson et al., 2007). Assembled contigs with a length above 500 bp were aligned to
 435 nr database using DIAMOND v.2.0.7.145 with *-long-reads* flag. This flag triggers frame-shift aware
 436 alignment mode, which is optimized for long sequence alignment. Therefore, short contigs (length < 500
 437 bp) were not retained. Besides, alignments with an E-value < $1e - 5$ or identity < 50% were removed,
 438 and for each contig, only alignments above 10% of the best local bit score were retained. The output of
 439 DIAMOND was analyzed by the *blast2rma* tool of MEGAN6 with *-lg* flag, which runs lowest-common-
 440 ancestor (LCA)-based algorithm developed for long contigs and assigns each contig to a taxon (Huson
 441 et al., 2018). The parameter *-supp* was 0, which means the present of a taxon would be identified as long
 442 as at least one contig was assigned. This value was used because a contig is assembled from multiple short
 443 reads and represents a strong signal for the present of a taxon.

444 **7.1.3 Fragment recruitment**

445 To integrate individual genomic data of species identified by assembly-dependent search, a reference
 446 database comprising reference genome dataset, *i.e.* genomic sequences in FASTA format and correspond-
 447 ing gff file, was constructed. For each species represented by assembly, its reference genome dataset, if
 448 available, was downloaded from NCBI using its *datasets* command-line tool and added to the reference
 449 database.

450 Then fragment recruitment was conducted. The non-host read pairs were mapped to genomic sequences
 451 in the reference database using Bowtie2. Read pairs that were not recruited were extracted using SAM-
 452 tools. Settings for Bowtie2 and SAMtools were the same as that described in 7.1.1. Read pairs recruited
 453 by the reference database were assigned to corresponding species, while the others were subjected to
 454 assembly-independent search.

455 **7.1.4 Assembly-independent species identification**

456 In order to detect species not represented by assembly (Sharon et al., 2015, Vollmers et al., 2017), assembly-
 457 independent search was conducted, taking read pairs not recruited by the reference database as input.
 458 These reads were aligned to nr database through DIAMOND without using *-long-reads* flag, which triggers
 459 computing alignments for short metagenomic reads. Other settings were the same as described in 7.1.2.
 460 Then the output of DIAMOND was analysed by MEGAN6 (*blast2rma* tool), which assigns read pairs to

461 taxa through LCA algorithm. Here the parameter *-supp* was 0.1, which means a taxon is reported after
 462 being represented by at least 0.1% of all assigned read pairs. It was used in order to avoid false positive
 463 results.

464 7.1.5 Functional annotation

465 Functional annotation was conducted by EggNOG-mapper v.2.1.2 (Huerta-Cepas et al., 2017). Sequences
 466 were searched against eggNOG database (Huerta-Cepas et al., 2019) for best seed orthologs using DIA-
 467 MOND and fine-grained orthology assignments were retrieved from pre-computed eggNOG phylogenetic
 468 trees. Then functional descriptions of retrieved orthologs including Gene Ontology (GO) terms (Consortium, 2004),
 469 KOs, Enzyme Commission (EC) numbers (Webb et al., 1992), Carbohydrate-Active Enzymes (CAZy) terms
 470 (Cantarel et al., 2009) and Clusters of Orthologous Groups (COG) functional categories (Tatusov et al., 2000)
 471 were transferred to query sequences.

472 7.2 Candidate models for fitting rarefaction curves

Table S1: Candidate species accumulation models. Dependent variable $D^{(q)}$ is Hill number of order q and independent variable x is sequencing depth. a, b, c, d are fitted coefficients.

Model	Parameter(k)	Derivative	Asymptote	Reference
$D^{(q)} = \frac{ax}{bx+1}$	2	$\frac{dD^q}{dx} = \frac{a}{(bx+1)^2}$	$\frac{a}{b}$	Clench (1979)
$D^{(q)} = a(1 - e^{-bx})$	2	$\frac{dD^q}{dx} = abe^{-bx}$	a	Miller and Wiegert (1989)
$D^{(q)} = a - bc^x$	3	$\frac{dD^q}{dx} = -bc^x \log(c)$	a	Ratkowsky (1983)
$D^{(q)} = a(1 - e^{-bx})^c$	3	$\frac{dD^q}{dx} = abce^{-bx}(1 - e^{-bx})^{c-1}$	a	Ratkowsky and Giles (1990)
$D^{(q)} = a(1 - (1 + (\frac{x}{c})^d)^{-b})$	4	$\frac{dD^q}{dx} = \frac{abd}{c}(\frac{x}{c})^{d-1}(1 + (\frac{x}{c})^d)^{-b-1}$	a	Mielke Jr and Johnson (1974)

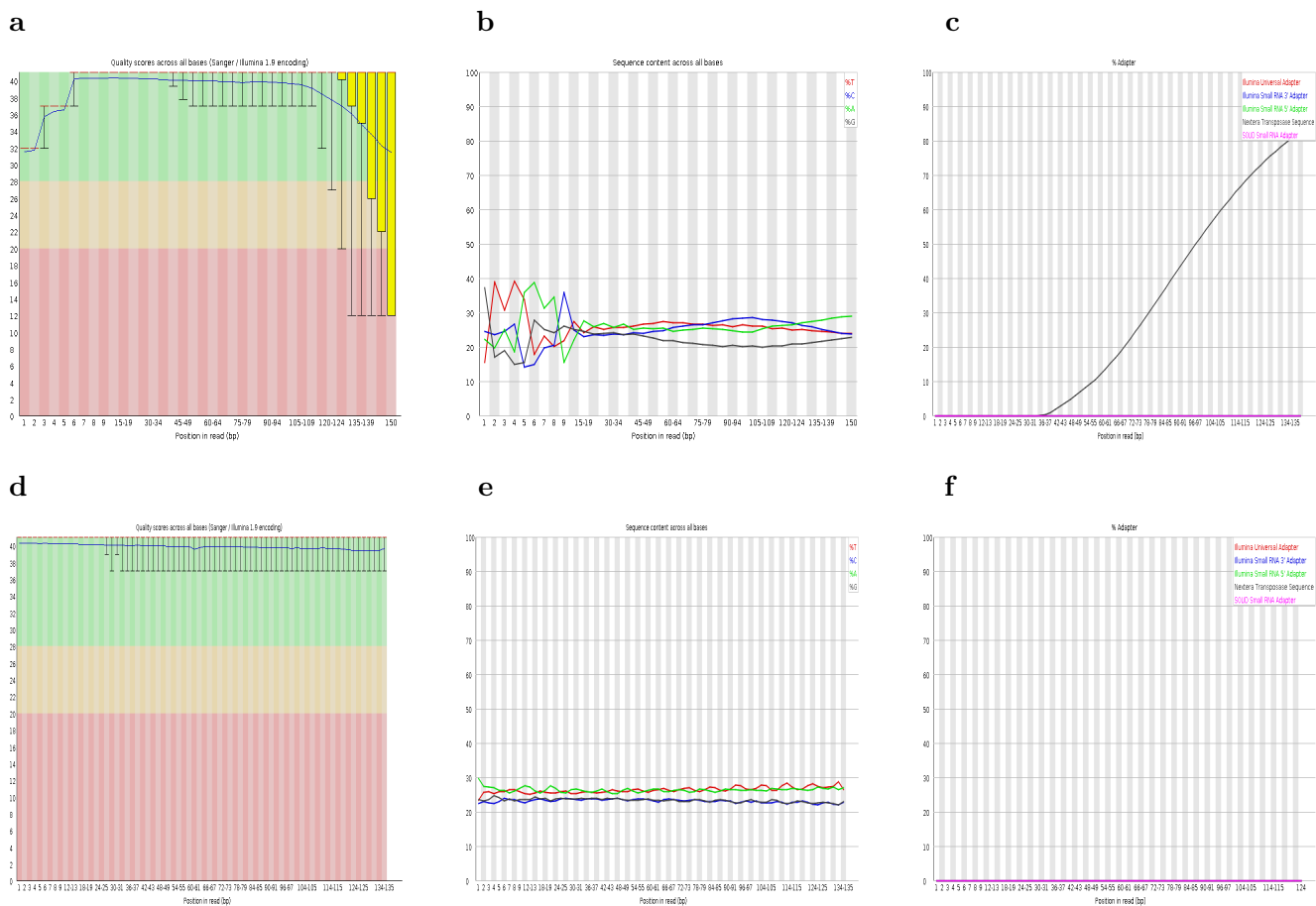


Figure S1: Data quality report of forward reads from bumble bee sample *Bee_Bimpatiens_hv4_1*. a, b and c shows low base quality in 3'-end, uneven base content in 5'-end and present of Nextera adaptors in raw data, respectively. d, e and f shows the same results from clean data.

Table S2: Statistics of sequenced read pairs.

Sample	Host	Raw read pair	Clean read pair	Non-host read pair
			[percentage of raw read pair]	[percentage of clean read pair]
<i>Bee_Amellifera_hv15_1</i>	<i>Apis mellifera</i>	63159968	45059211 [71.34%]	38837759 [86.19%]
<i>Bee_Amellifera_wild_1</i>	<i>Apis mellifera</i>	58113227	44466776 [76.52%]	17014144 [38.26%]
<i>Bee_Amellifera_hv13_2</i>	<i>Apis mellifera</i>	56836899	35282101 [62.08%]	10413704 [29.52%]
<i>Bee_Amellifera_hv13_1</i>	<i>Apis mellifera</i>	1104861	842095 [76.22%]	665507 [79.03%]
<i>Bee_Bimpatiens_hv3_1</i>	<i>Bombus impatiens</i>	63973750	53612702 [83.80%]	5300592 [9.89%]
<i>Bee_Bimpatiens_hv4_1</i>	<i>Bombus impatiens</i>	58988182	48426748 [82.10%]	3557052 [7.35%]
<i>Bee_Bimpatiens_hv4_2</i>	<i>Bombus impatiens</i>	54955553	45805759 [83.35%]	4618023 [10.08%]
<i>Flower_eDNA</i>	None	1443107	882436 [61.15%]	882436 [100%]

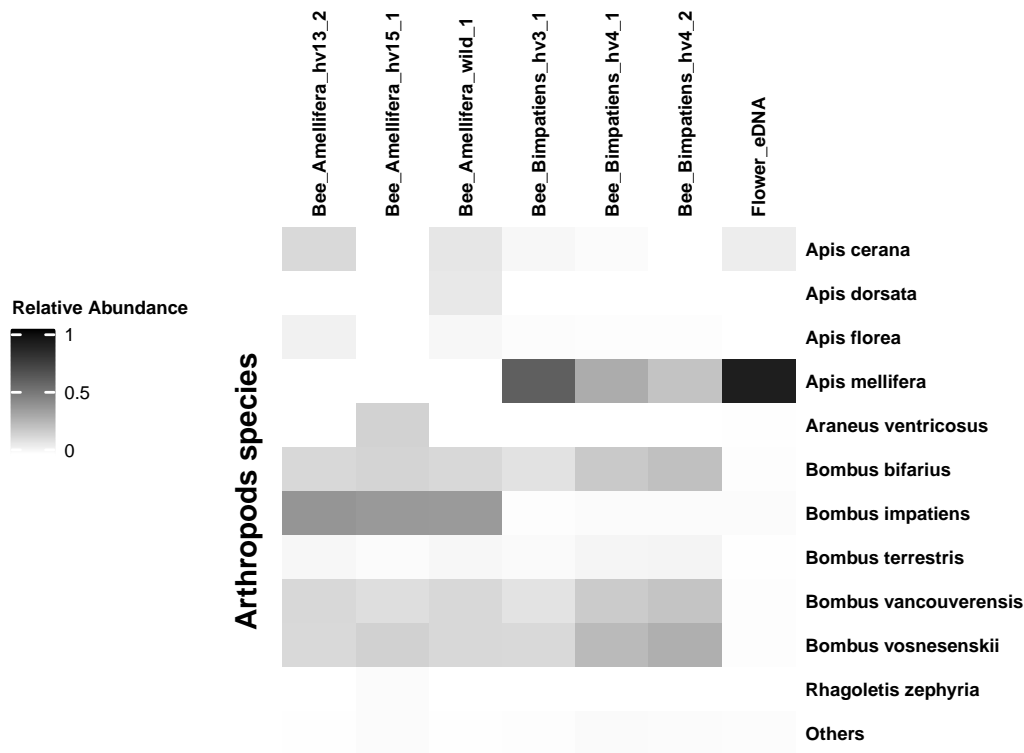


Figure S2: Heatmaps for arthropod species abundance distribution in all samples. The relative abundance takes reads assigned to arthropod species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others". It should be noted that for bee samples, host contamination was removed before taxon profiling. As a result, the relative abundances of honey bees are extremely low in three honey bee samples, and the same for bumble bees in three bumble bee samples.

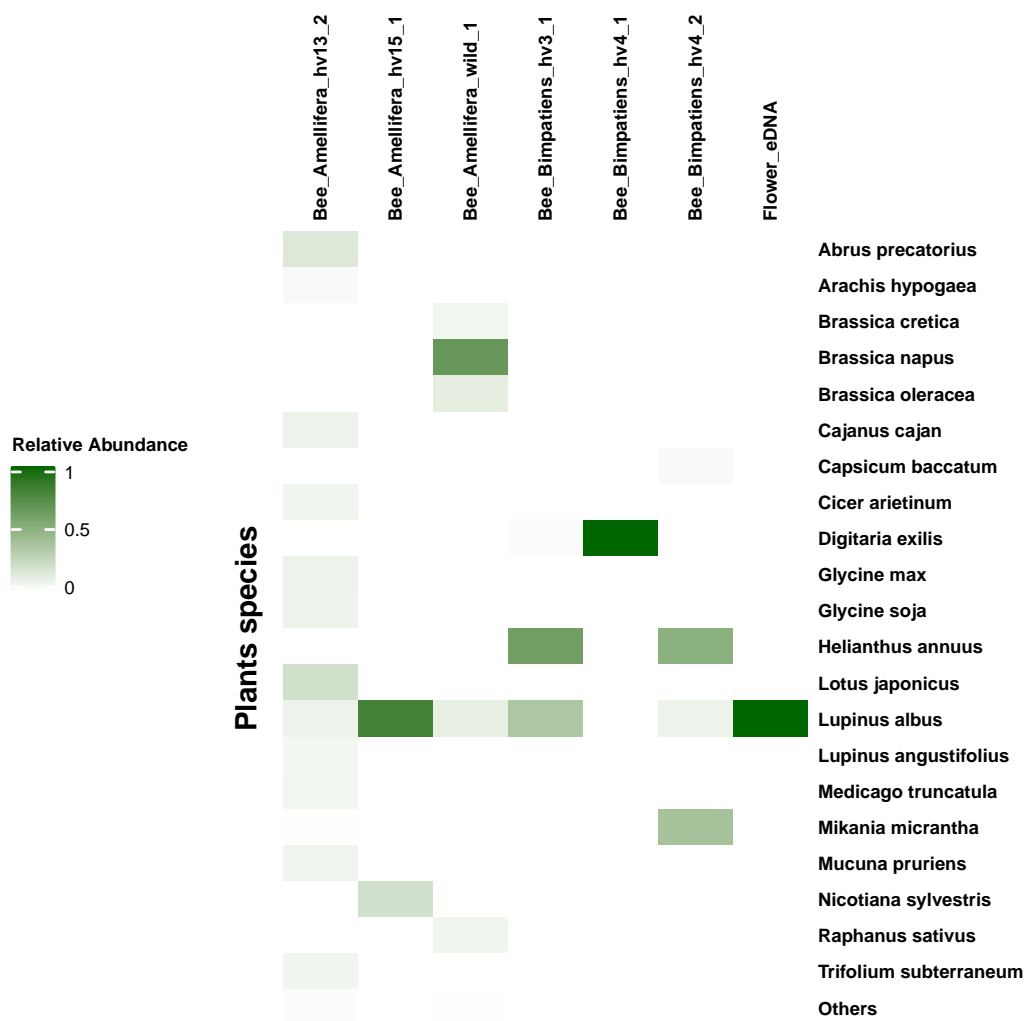


Figure S3: Heatmaps for plant species abundance distribution in all samples. The relative abundance takes reads assigned to plant species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

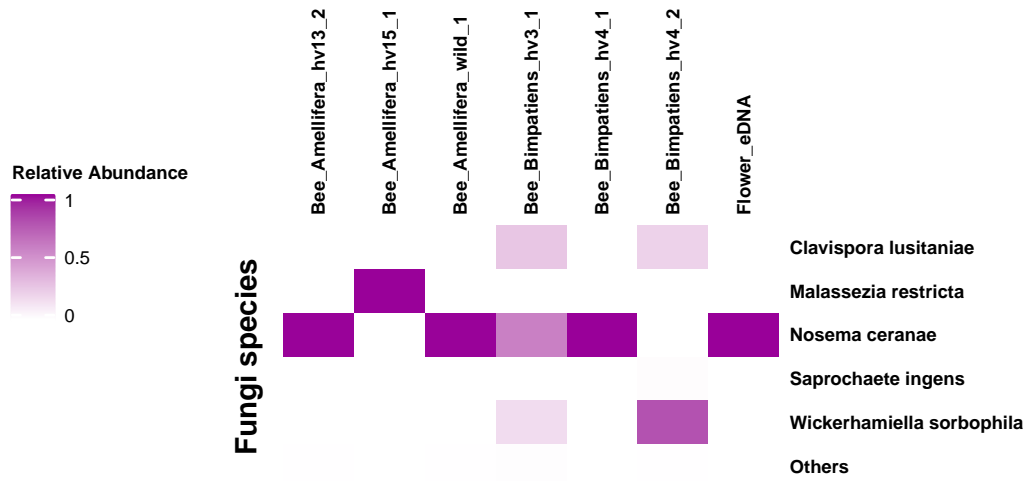


Figure S4: Heatmaps for fungal species abundance distribution in all samples. The relative abundance takes reads assigned to fungal species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

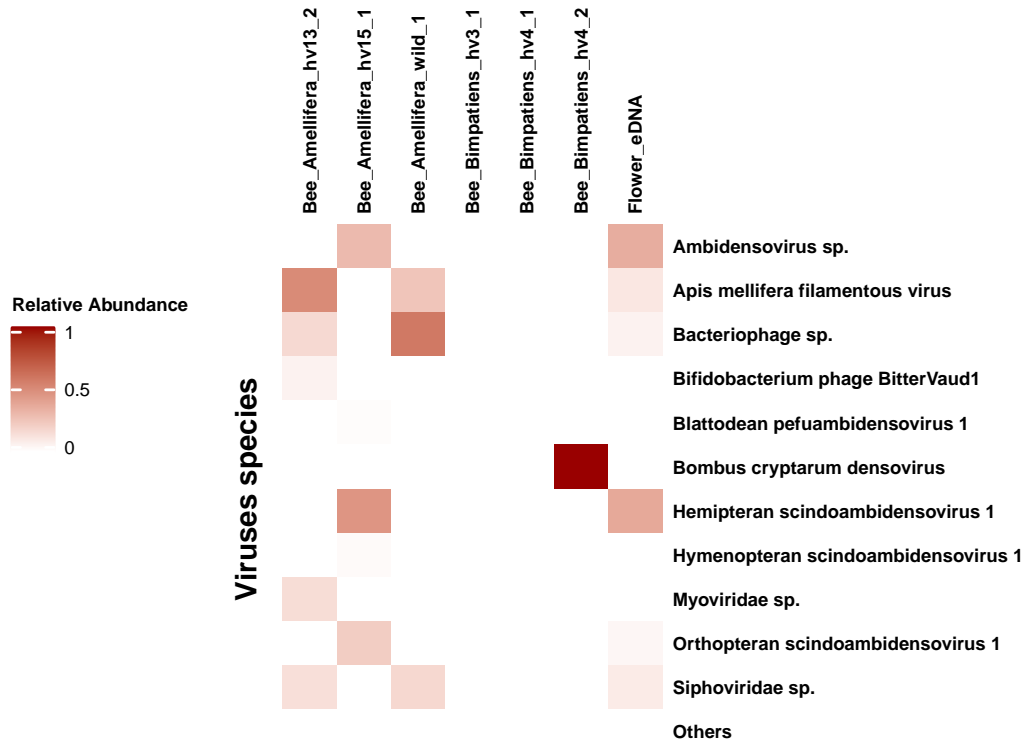


Figure S5: Heatmaps for virus species abundance distribution in all samples. The relative abundance takes reads assigned to virus species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

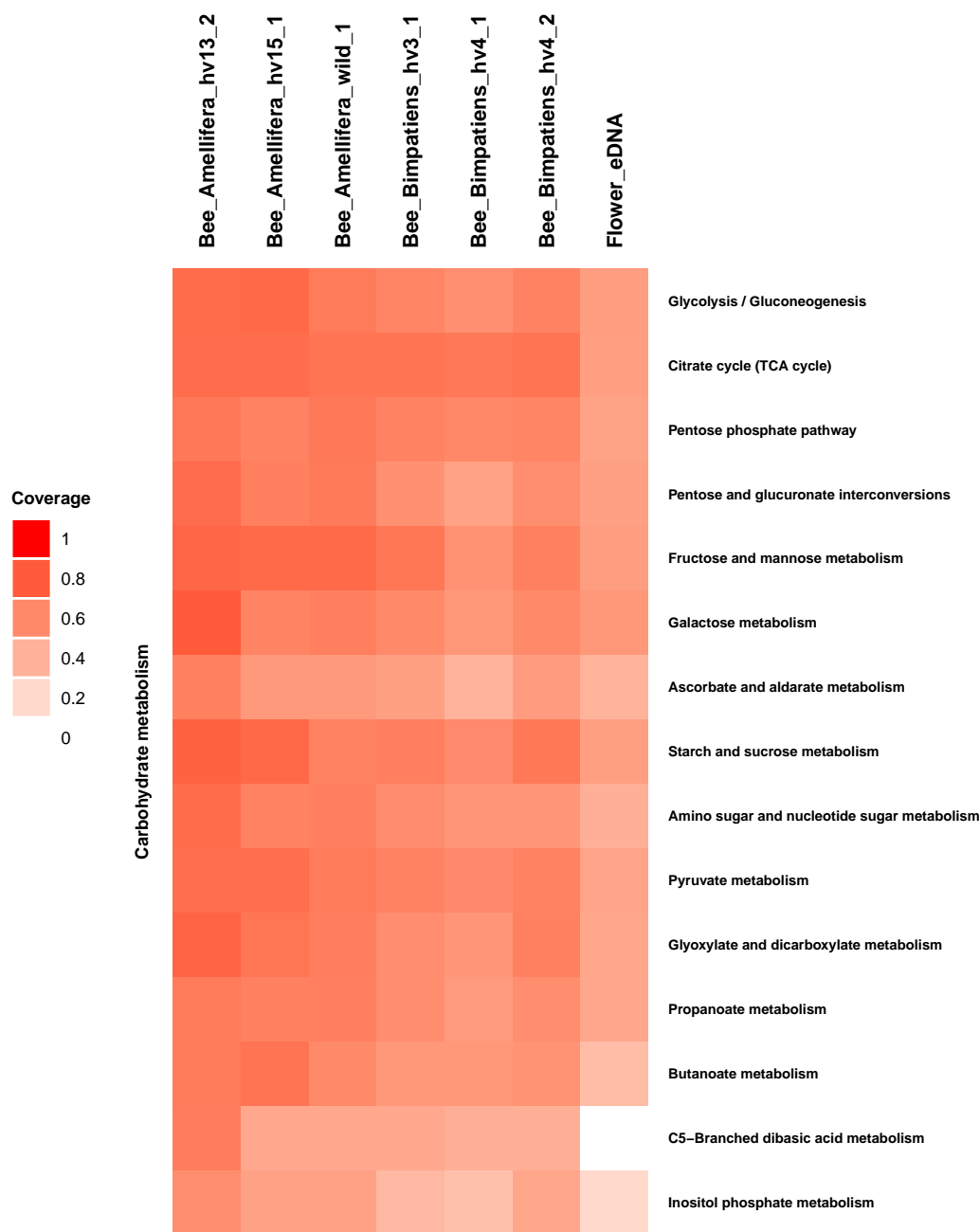


Figure S6: Heatmaps for pathways of carbohydrate metabolism.

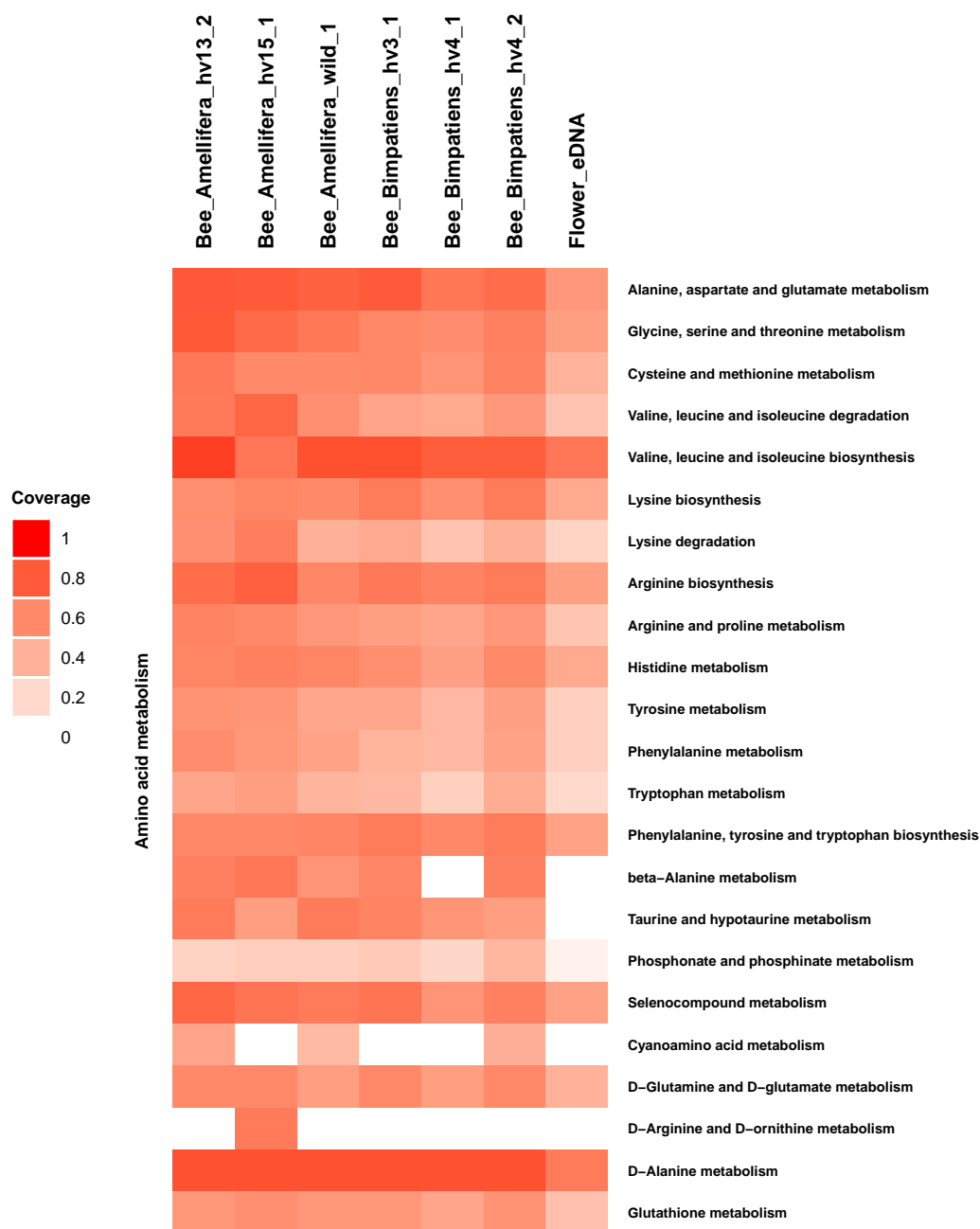


Figure S7: Heatmaps for pathways of amino acid metabolism.

References

- Ahmed Abdelfattah, Antonino Malacrino, Michael Wisniewski, Santa O Cacciola, and Leonardo Schena. Metabarcoding: A powerful tool to investigate microbial communities and shape future plant protection strategies. *Biological Control*, 120:1–10, 2018.
- Sahar Abubucker, Nicola Segata, Johannes Goll, Alyxandria M Schubert, Jacques Izard, Brandi L Cantarel, Beltran Rodriguez-Mueller, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology*, 8(6):e1002358, 2012.
- Esmail Amiri, Prashant Waiker, Olav Rueppell, and Prashanti Manda. Using manual and computer-based text-mining to uncover research trends for apis mellifera. *Veterinary sciences*, 7(2):61, 2020.
- David R Anderson. *Model based inference in the life sciences: a primer on evidence*. Springer Science & Business Media, 2007.
- Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.
- Kathrin P Aßhauer, Bernd Wemheuer, Rolf Daniel, and Peter Meinicke. Tax4fun: predicting functional profiles from metagenomic 16s rRNA data. *Bioinformatics*, 31(17):2882–2884, 2015.
- Martin Ayling, Matthew D Clark, and Richard M Leggett. New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*, 21(2):584–594, 2020.
- Leslie Bailey. Honey bee pathology. *Annual review of entomology*, 13(1):191–212, 1968.
- Svenja Bänisch, Teja Tschardt, Doreen Gabriel, and Catrin Westphal. Crop pollination services: complementary resource use by social vs solitary bees facing crops with contrasting flower supply. *Journal of Applied Ecology*, 58(3):476–485, 2021.
- Daniela Becker, Denny Popp, Hauke Harms, and Florian Centler. A modular metagenomics pipeline allowing for the inclusion of prior knowledge using the example of anaerobic digestion. *Microorganisms*, 8(5):669, 2020.
- Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
- Germán Bonilla-Rosso, Théodora Steiner, Fabienne Wichmann, Evan Bexkens, and Philipp Engel. Honey bees harbor a diverse gut virome engaging in nested strain-level interactions with the microbiota. *Proceedings of the National Academy of Sciences*, 117(13):7355–7362, 2020.

506 Samuele Bovo, Anisa Ribani, Valerio Joe Utzeri, Giuseppina Schiavo, Francesca Bertolini, and Luca
507 Fontanesi. Shotgun metagenomics of honey dna: Evaluation of a methodological approach to describe
508 a multi-kingdom honey bee derived environmental dna signature. *PLoS One*, 13(10):e0205575, 2018.

509 Samuele Bovo, Valerio Joe Utzeri, Anisa Ribani, Riccardo Cabbri, and Luca Fontanesi. Shotgun se-
510 quencing of honey dna can describe honey bee derived environmental signatures and the honey bee
511 hologenome complexity. *Scientific reports*, 10(1):1–17, 2020.

512 Michael A Brockhurst, Ellie Harrison, James PJ Hall, Thomas Richards, Alan McNally, and Craig
513 MacLean. The ecology and evolution of pangenomes. *Current Biology*, 29(20):R1094–R1103, 2019.

514 Mark JF Brown and Robert J Paxton. The conservation of bees: a global perspective. *Apidologie*, 40(3):
515 410–416, 2009.

516 Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond.
517 *Nature methods*, 12(1):59–60, 2015.

518 Brian Bushnell. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley
519 National Lab.(LBNL), Berkeley, CA (United States), 2014.

520 Brandi L Cantarel, Pedro M Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and
521 Bernard Henrissat. The carbohydrate-active enzymes database (cazy): an expert resource for glycoge-
522 nomics. *Nucleic acids research*, 37(suppl_1):D233–D238, 2009.

523 Daniel P Cariveau, J Elijah Powell, Hauke Koch, Rachael Winfree, and Nancy A Moran. Variation in
524 gut microbial communities and its association with pathogen infection in wild bumble bees (*bombus*).
525 *The ISME journal*, 8(12):2369–2379, 2014.

526 Rogan Carr and Elhanan Borenstein. Comparative analysis of functional metagenomic annotation and
527 the mappability of short reads. *PloS one*, 9(8):e105776, 2014.

528 Federica Cattonaro, Alessandro Spadotto, Slobodanka Radovic, and Fabio Marroni. Do you cov me?
529 effect of coverage reduction on metagenome shotgun sequencing studies. *F1000Research*, 7, 2018.

530 Anne Chao and Lou Jost. Coverage-based rarefaction and extrapolation: standardizing samples by
531 completeness rather than size. *Ecology*, 93(12):2533–2547, 2012.

532 Anne Chao, Chun-Huo Chiu, and Lou Jost. Unifying species diversity, phylogenetic diversity, functional
533 diversity, and related similarity and differentiation measures through hill numbers. *Annual review of*
534 *ecology, evolution, and systematics*, 45:297–324, 2014.

535 Robin L Chazdon, Robert K Colwell, Julie S Denslow, and Manuel R Guariguata. Statistical methods for
536 estimating species richness of woody regeneration in primary and secondary rain forests of northeastern
537 costa rica. 1998.

538 Wenda Cheng and Louise Ashton. Ecology: What affects the distribution of global bee diversity. *Current*
539 *Biology*, 31(3):R127–R128, 2021.

540 Harry K Clench. How to make regional lists of butterflies: some thoughts. *Journal of the Lepidopterists’*
541 *Society*, 1979.

542 Robert K Colwell and Jonathan A Coddington. Estimating terrestrial biodiversity through extrapolation.
543 *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 345(1311):
544 101–118, 1994.

545 Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids*
546 *research*, 32(suppl_1):D258–D261, 2004.

547 Samantha M Cook, Caroline S Awmack, Darren A Murray, and Ingrid H Williams. Are honey bees’
548 foraging preferences affected by pollen amino acid composition? *Ecological Entomology*, 28(5):622–627,
549 2003.

550 Otto X Cordero and Martin F Polz. Explaining microbial genomic diversity in light of evolutionary
551 ecology. *Nature Reviews Microbiology*, 12(4):263–273, 2014.

552 Eamonn P Culligan, Roy D Sleator, Julian R Marchesi, and Colin Hill. Metagenomics and novel gene
553 discovery: promise and potential for novel therapeutics. *Virulence*, 5(3):399–412, 2014.

554 Gavin M Douglas, Robert G Beiko, and Morgan GI Langille. Predicting the functional potential of the
555 microbiome from marker genes using picrust. In *Microbiome Analysis*, pages 169–177. Springer, 2018.

556 Stephan Durrer and Paul Schmid-Hempel. Shared use of flowers leads to horizontal pathogen transmission.
557 *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 258(1353):299–302, 1994.

558 Julia Ebeling, Henriette Knispel, Gillian Hertlein, Anne Fünfhaus, and Elke Genersch. Biology of paeni-
559 bacillus larvae, a deadly pathogen of honey bee larvae. *Applied microbiology and biotechnology*, 100
560 (17):7387–7395, 2016.

561 Paul B Eckburg, Elisabeth M Bik, Charles N Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael
562 Sargent, Steven R Gill, Karen E Nelson, and David A Relman. Diversity of the human intestinal
563 microbial flora. *science*, 308(5728):1635–1638, 2005.

564 Kirsten M Ellegaard, Shota Suenami, Ryo Miyazaki, and Philipp Engel. Vast differences in strain-level
565 diversity in the gut microbiota of two closely related honey bee species. *Current Biology*, 30(13):
566 2520–2531, 2020.

567 Akihito Endo and Leon MT Dicks. The genus fructobacillus. *Lactic acid bacteria: biodiversity and*
568 *taxonomy*, pages 381–389, 2014.

569 Philipp Engel, Vincent G Martinson, and Nancy A Moran. Functional diversity within the simple gut
570 microbiota of the honey bee. *Proceedings of the National Academy of Sciences*, 109(27):11002–11007,
571 2012.

572 Philipp Engel, Waldan K Kwong, Quinn McFrederick, Kirk E Anderson, Seth Michael Barribeau,
573 James Angus Chandler, R Scott Cornman, Jacques Dainat, Joachim R De Miranda, Vincent Doublet,
574 et al. The bee microbiome: impact on bee health and model for evolution and ecology of host-microbe
575 interactions. *MBio*, 7(2):e02164–15, 2016.

576 Christiane Essoh, Libera Latino, Cédric Midoux, Yann Blouin, Guillaume Loukou, Simon-Pierre A
577 Nguetta, Serge Lathro, Arsher Cablanmian, Athanase K Kouassi, Gilles Vergnaud, et al. Investigation
578 of a large collection of pseudomonas aeruginosa bacteriophages collected from a single environmental
579 source in abidjan, côte d’ivoire. *PloS one*, 10(6):e0130548, 2015.

580 Laura L Figueroa, Malcolm Blinder, Cali Grincavitch, Angus Jelinek, Emilia K Mann, Liam A Merva,
581 Lucy E Metz, Amy Y Zhao, Rebecca E Irwin, Scott H McArt, et al. Bee pathogen transmission
582 dynamics: deposition, persistence and acquisition on flowers. *Proceedings of the Royal Society B*, 286
583 (1903):20190603, 2019.

584 Eva Forsgren, Tobias C Olofsson, Alejandra Váasquez, and Ingemar Fries. Novel lactic acid bacteria
585 inhibiting paenibacillus larvae in honey bee larvae. *Apidologie*, 41(1):99–108, 2010.

586 MA Fürst, Dino Peter McMahon, JL Osborne, RJ Paxton, and MJF Brown. Disease associations between
587 honeybees and bumblebees as a threat to wild pollinators. *Nature*, 506(7488):364–366, 2014.

588 Jessica Galloway-Peña and Blake Hanson. Tools for analysis of the microbiome. *Digestive diseases and*
589 *sciences*, 65(3):674–685, 2020.

590 Christina Geldert, Zaid Abdo, Jane E Stewart, and Arathi HS. Dietary supplementation with phytochem-
591 icals improves diversity and abundance of honey bee gut microbiota. *Journal of Applied Microbiology*,
592 130(5):1705–1720, 2021.

593 José Antonio Gómez-Anaya, Rodolfo Novelo-Gutiérrez, Alonso Ramírez, and Roberto Arce-Pérez. Using
594 empirical field data of aquatic insects to infer a cut-off slope value in asymptotic models to assess
595 inventories completeness. *Revista mexicana de biodiversidad*, 85(1):218–227, 2014.

596 Peter Graystock, Wee Hao Ng, Kyle Parks, Amber D Tripodi, Paige A Muñiz, Ashley A Fersch, Christo-
597 pher R Myers, Quinn S McFrederick, and Scott H McArt. Dominant bee species and floral abundance
598 drive parasite temporal dynamics in plant-pollinator communities. *Nature ecology & evolution*, 4(10):
599 1358–1367, 2020.

600 Antonius Petrus de Groot. Protein and amino acid requirements of the honeybee (*apis mellifica* l.). 1953.

601 CBOL Plant Working Group, Peter M Hollingsworth, Laura L Forrest, John L Spouge, Mehrdad Ha-
602 jibabaei, Sujeevan Ratnasingham, Michelle van der Bank, Mark W Chase, Robyn S Cowan, David L
603 Erickson, et al. A dna barcode for land plants. *Proceedings of the National Academy of Sciences*, 106
604 (31):12794–12797, 2009.

605 H Soon Gweon, Liam P Shaw, Jeremy Swann, Nicola De Maio, Manal AbuOun, Rene Niehus, Alasdair TM
606 Hubbard, Mike J Bowes, Mark J Bailey, Tim EA Peto, et al. The impact of sequencing depth on
607 the inferred taxonomic composition and amr gene content of metagenomic samples. *Environmental*
608 *Microbiome*, 14(1):1–15, 2019.

609 Hidenori Hayashi, Mitsuo Sakamoto, and Yoshimi Benno. Phylogenetic analysis of the human gut mi-
610 crobiota using 16s rdna clone libraries and strictly anaerobic culture-based methods. *Microbiology and*
611 *immunology*, 46(8):535–548, 2002.

612 Paul DN Hebert, Sujeevan Ratnasingham, and Jeremy R De Waard. Barcoding animal life: cytochrome c
613 oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London.*
614 *Series B: Biological Sciences*, 270(suppl.1):S96–S99, 2003.

615 Kenneth L Heck Jr, Gerald van Belle, and Daniel Simberloff. Explicit calculation of the rarefaction
616 diversity measurement and the determination of sufficient sample size. *Ecology*, 56(6):1459–1461, 1975.

617 Harmen P Hendriksma and Sharoni Shafir. Honey bee foragers balance colony nutritional deficiencies.
618 *Behavioral Ecology and Sociobiology*, 70(4):509–517, 2016.

619 Harmen P Hendriksma, Karim L Oxman, and Sharoni Shafir. Amino acid and carbohydrate tradeoffs
620 by honey bee nectar foragers and their implications for plant–pollinator interactions. *Journal of insect*
621 *physiology*, 69:56–64, 2014.

622 Harmen P Hendriksma, Collin D Pachow, and James C Nieh. Effects of essential amino acid supplemen-
623 tation to promote honey bee gland and muscle development in cages and colonies. *Journal of insect*
624 *physiology*, 117:103906, 2019.

625 Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432,
626 1973.

627 Joaquín Hortal and Jorge M Lobo. An ed-based protocol for optimal sampling of biodiversity. *Biodiversity*
628 *& Conservation*, 14(12):2913–2947, 2005.

629 Joaquín Hortal, Paulo AV Borges, and Clara Gaspar. Evaluating the performance of species richness
630 estimators: sensitivity to sample grain size. *Journal of animal ecology*, 75(1):274–287, 2006.

631 Peter Hristov, Boyko Neov, Rositsa Shumkova, and Nadezhda Palova. Significance of apoidea as main
632 pollinators. ecological and economic impact and implications for human nutrition. *Diversity*, 12(7):280,
633 2020a.

634 Peter Hristov, Rositsa Shumkova, Nadezhda Palova, and Boyko Neov. Factors associated with honey bee
635 colony losses: a mini-review. *Veterinary Sciences*, 7(4):166, 2020b.

636 Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen,
637 Christian Von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology
638 assignment by eggno-mapper. *Molecular biology and evolution*, 34(8):2115–2122, 2017.

639 Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen
640 Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, et al. eggno 5.0: a hierarchical,
641 functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502
642 viruses. *Nucleic acids research*, 47(D1):D309–D314, 2019.

643 Robert M Hughes, Alan T Herlihy, and David V Peck. Sampling efforts for estimating fish species richness
644 in western usa river sites. *Limnologica*, 87:125859, 2021.

645 Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. Megan analysis of metagenomic
646 data. *Genome research*, 17(3):377–386, 2007.

647 Daniel H Huson, Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Gorska, Dino Jolic, and Ro-
648 han BH Williams. Megan-lr: new algorithms allow accurate binning and easy interactive exploration
649 of metagenomic long reads and contigs. *Biology direct*, 13(1):1–17, 2018.

650 Alberto Jimenez-Valverde, Silvia Jimenez Mendoza, Jose Martin Cano, and Miguel L Munguira. Compar-
 651 ing relative model fit of several species-accumulation functions to local papilionoidea and hesperioidea
 652 butterfly inventories of mediterranean habitats. In *Arthropod diversity and conservation*, pages 163–176.
 653 Springer, 2006.

654 Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids*
 655 *research*, 28(1):27–30, 2000.

656 Karen M Kapheim, Makenna M Johnson, and Maggi Jolley. Composition and acquisition of the micro-
 657 biome in solitary, ground-nesting alkali bees. *Scientific reports*, 11(1):1–11, 2021.

658 Alexander Keller, Quinn S McFrederick, Prarthana Dharampal, Shawn Steffan, Bryan N Danforth, and
 659 Sara D Leonhardt. (more than) hitchhikers through the network: The shared microbiome of bees and
 660 flowers. *Current Opinion in Insect Science*, 2020.

661 Lucie Kešnerová, Roxane Moritz, and Philipp Engel. *Bartonella apis* sp. nov., a honey bee gut symbiont
 662 of the class alphaproteobacteria. *International journal of systematic and evolutionary microbiology*, 66
 663 (1):414–421, 2016.

664 Shaden AM Khalifa, Esraa H Elshafiey, Aya A Shetaia, Aida A Abd El-Wahed, Ahmed F Algethami,
 665 Syed G Musharraf, Mohamed F AlAjmi, Chao Zhao, Saad HD Masry, Mohamed M Abdel-Daim, et al.
 666 Overview of bee pollination and its economic value for crop production. *Insects*, 12(8):688, 2021.

667 Hauke Koch and Paul Schmid-Hempel. Bacterial communities in central european bumblebees: low
 668 diversity and high specificity. *Microbial Ecology*, 62(1):121–133, 2011a.

669 Hauke Koch and Paul Schmid-Hempel. Socially transmitted gut microbiota protect bumble bees against
 670 an intestinal parasite. *Proceedings of the National Academy of Sciences*, 108(48):19288–19292, 2011b.

671 Hauke Koch, Mark JF Brown, and Philip C Stevenson. The role of disease in bee foraging ecology. *Current*
 672 *opinion in insect science*, 21:60–67, 2017.

673 Waldan K Kwong, Luis A Medina, Hauke Koch, Kong-Wah Sing, Eunice Jia Yu Soh, John S Ascher,
 674 Rodolfo Jaffé, and Nancy A Moran. Dynamic microbiome evolution in social bees. *Science Advances*,
 675 3(3):e1600513, 2017.

676 Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):
 677 357, 2012.

678 Fredrick J Lee, Douglas B Rusch, Frank J Stewart, Heather R Mattila, and Irene LG Newton. Saccharide
679 breakdown and fermentation by the honey bee gut microbiome. *Environmental microbiology*, 17(3):
680 796–815, 2015.

681 Fredrick J Lee, Kayla I Miller, James B McKinlay, and Irene LG Newton. Differential carbohydrate
682 utilization and organic acid production by honey bee symbionts. *FEMS microbiology ecology*, 94(8):
683 fty113, 2018.

684 Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-
685 fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph.
686 *Bioinformatics*, 31(10):1674–1676, 2015.

687 Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform.
688 *bioinformatics*, 25(14):1754–1760, 2009.

689 Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo
690 Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25
691 (16):2078–2079, 2009.

692 Patrick W Maes, Pedro AP Rodrigues, Randy Oliver, Brendon M Mott, and Kirk E Anderson. Diet-
693 related gut bacterial dysbiosis correlates with impaired development, increased mortality and nosema
694 disease in the honeybee (*apis mellifera*). *Molecular Ecology*, 25(21):5439–5450, 2016.

695 Vincent G Martinson, Tanja Magoc, Hauke Koch, Steven L Salzberg, and Nancy A Moran. Genomic
696 features of a bumble bee symbiont reflect its host environment. *Applied and environmental microbiology*,
697 80(13):3793–3803, 2014.

698 Tomonori Matsuzawa, Ryo Kohsaka, and Yuta Uchiyama. Application of environmental dna: Honey
699 bee behavior and ecosystems for sustainable beekeeping. In *Modern beekeeping-bases for sustainable*
700 *production*. IntechOpen, 2020.

701 QUINN S McFREDERICK, William T Wcislo, Douglas R Taylor, Heather D Ishak, Scot E Dowd, and
702 Ulrich G Mueller. Environment or kin: whence do bees obtain acidophilic bacteria? *Molecular Ecology*,
703 21(7):1754–1768, 2012.

704 Quinn S McFrederick, Jason M Thomas, John L Neff, Hoang Q Vuong, Kaleigh A Russell, Amanda R
705 Hale, and Ulrich G Mueller. Flowers and wild megachilid bees share microbes. *Microbial ecology*, 73
706 (1):188–200, 2017.

707 Folker Meyer, Daniel Paarmann, Mark D’Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias
708 Paczian, Alex Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics rast server—a public
709 resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*,
710 9(1):1–8, 2008.

711 Paul W Mielke Jr and Earl S Johnson. Some generalized beta distributions of the second kind having
712 desirable application features in hydrology and meteorology. *Water Resources Research*, 10(2):223–226,
713 1974.

714 Alla Mikheenko, Vladislav Saveliev, and Alexey Gurevich. Metaquast: evaluation of metagenome assem-
715 blies. *Bioinformatics*, 32(7):1088–1090, 2016.

716 Ronald I Miller and Richard G Wiegert. Documenting completeness, species-area relations, and the species-
717 abundance distribution of a regional flora. *Ecology*, 70(1):16–22, 1989.

718 Nancy A Moran. Genomics of the honey bee microbiome. *Current opinion in insect science*, 10:22–28,
719 2015.

720 Felicia N New and Ilana L Brito. What is metagenomics teaching us, and what is missed? *Annual Review*
721 *of Microbiology*, 74:117–135, 2020.

722 Sergey I Nikolenko, Anton I Korobeynikov, and Max A Alekseyev. Bayeshammer: Bayesian clustering for
723 error correction in single-cell sequencing. In *BMC genomics*, volume 14, pages 1–11. Springer, 2013.

724 R Henrik Nilsson, Erik Kristiansson, Martin Ryberg, Nils Hallenberg, and Karl-Henrik Larsson. Intraspe-
725 cific its variability in the kingdom fungi as expressed in the international sequence databases and its
726 implications for molecular species identification. *Evolutionary bioinformatics*, 4:EBO–S653, 2008.

727 Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaspades: a new versatile
728 metagenomic assembler. *Genome research*, 27(5):824–834, 2017.

729 Pier P Paoli, Luisa A Wakeling, Geraldine A Wright, and Dianne Ford. The dietary proportion of essential
730 amino acids and sir2 influence lifespan in the honeybee. *Age*, 36(3):1239–1247, 2014.

731 K Papadopoulou-Karabela, N Iliadis, V Liakos, and E Bourdzy-Hatzopoulou. Experimental infection of
732 honeybees by pseudomonas aeruginosa. *Apidologie*, 23(5):393–397, 1992.

733 K Papadopoulou-Karabela, N Iliadis, and V Liakos. Haemocyte changes in honeybee (apis mellifera l)
734 artificially infected by pseudomonas aeruginosa. *Apidologie*, 24(1):81–86, 1993.

735 Yu Peng, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. Idba-ud: a de novo assembler for single-
736 cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428,
737 2012.

738 Judit J Péntzes, Maria Söderlund-Venermo, Marta Canuti, Anna Maria Eis-Hübinger, Joseph Hughes,
739 Susan F Cotmore, and Balázs Harrach. Reorganizing the family parvoviridae: a revised taxonomy
740 independent of the canonical approach based on host association. *Archives of Virology*, 165(9):2133–
741 2146, 2020.

742 Joana Pereira-Marques, Anne Hout, Rui M Ferreira, Michiel Weber, Ines Pinto-Ribeiro, Leen-Jan van
743 Doorn, Cornelis Willem Knetsch, and Ceu Figueiredo. Impact of host dna and sequencing depth
744 on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in*
745 *microbiology*, 10:1277, 2019.

746 Elijah Powell, Nalin Ratnayeke, and Nancy A Moran. Strain diversity and host specificity in a specialized
747 gut symbiont of honeybees and bumblebees. *Molecular ecology*, 25(18):4461–4471, 2016.

748 J Elijah Powell, Zac Carver, Sean P Leonard, and Nancy A Moran. Field-realistic tylosin exposure
749 impacts honey bee microbiota and pathogen susceptibility, which is ameliorated by native gut probiotics.
750 *Microbiology Spectrum*, 9(1):e00103–21.

751 Andrey Prjibelski, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. Using
752 spades de novo assembler. *Current Protocols in Bioinformatics*, 70(1):e102, 2020.

753 Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun
754 metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844, 2017.

755 Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features.
756 *Bioinformatics*, 26(6):841–842, 2010.

757 David Ratkowsky. Nonlinear regression modelling. 1983.

758 David A Ratkowsky and David EA Giles. *Handbook of nonlinear regression models*. Number 04; QA278.
759 2, R3. M. Dekker New York, 1990.

760 Kasie Raymann and Nancy A Moran. The role of the gut microbiome in health and disease of adult honey
761 bee workers. *Current opinion in insect science*, 26:97–104, 2018.

762 Anisa Ribani, Valerio Joe Utzeri, Valeria Taurisano, and Luca Fontanesi. Honey as a source of environ-
763 mental dna for the detection and monitoring of honey bee pathogens and parasites. *Veterinary sciences*,
764 7(3):113, 2020.

765 Michael Roswell, Jonathan Dushoff, and Rachael Winfree. A conceptual guide to measuring species
766 diversity. *Oikos*, 130(3):321–338, 2021.

767 Jason A Rothman, Laura Leger, Peter Graystock, Kaleigh Russell, and Quinn S McFrederick. The bumble
768 bee microbiome increases survival of bees exposed to selenate toxicity. *Environmental microbiology*, 21
769 (9):3417–3429, 2019a.

770 Jason A Rothman, Laura Leger, Jay S Kirkwood, and Quinn S McFrederick. Cadmium and selenate
771 exposure affects the honey bee microbiome and metabolome, and bee-associated bacteria show potential
772 for bioaccumulation. *Applied and environmental microbiology*, 85(21):e01411–19, 2019b.

773 Jason A Rothman, Kaleigh A Russell, Laura Leger, Quinn S McFrederick, and Peter Graystock. The
774 direct and indirect effects of environmental toxicants on the health of bumblebees and their microbiomes.
775 *Proceedings of the Royal Society B*, 287(1937):20200980, 2020.

776 Itai Sharon, Michael Kertesz, Laura A Hug, Dmitry Pushkarev, Timothy A Blauwkamp, Cindy J Castelle,
777 Mojgan Amirebrahimi, Brian C Thomas, David Burstein, Susannah G Tringe, et al. Accurate, multi-kb
778 reads resolve complex populations and detect rare microorganisms. *Genome research*, 25(4):534–543,
779 2015.

780 Nicola K Simcock, Helen E Gray, and Geraldine A Wright. Single amino acids in sucrose rewards modulate
781 feeding and associative learning in the honeybee. *Journal of insect physiology*, 69:41–48, 2014.

782 Daniel Stabler, Pier P Paoli, Susan W Nicolson, and Geraldine A Wright. Nutrient balancing of the adult
783 worker bumblebee (*bombus terrestris*) depends on the dietary source of essential amino acids. *The*
784 *Journal of experimental biology*, 218(5):793–802, 2015.

785 Javier Tamames and Fernando Puente-Sánchez. Squeezemeta, a highly portable, fully automatic metage-
786 nomic analysis pipeline. *Frontiers in microbiology*, 9:3349, 2019.

787 Roman L Tatusov, Michael Y Galperin, Darren A Natale, and Eugene V Koonin. The cog database: a
788 tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33–36,
789 2000.

790 Michelle A Taylor, Alastair W Robertson, Patrick J Biggs, Kate K Richards, Daniel F Jones, and Shan-
791 thi G Parkar. The effect of carbohydrate sources: Sucrose, invert sugar and components of mānuka
792 honey, on core bacteria in the digestive tract of adult honey bees (*apis mellifera*). *PloS one*, 14(12):
793 e0225845, 2019.

794 Quang Tran and Vinhthuy Phan. Assembling reads improves taxonomic classification of species. *Genes*,
795 11(8):946, 2020.

796 Susannah Green Tringe, Christian Von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W
797 Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, et al. Comparative metagenomics
798 of microbial communities. *Science*, 308(5721):554–557, 2005.

799 Rachel L Vannette. The floral microbiome: plant, pollinator, and microbial perspectives. *Annual Review*
800 *of Ecology, Evolution, and Systematics*, 51:363–386, 2020.

801 Louise Vermote, Marko Verce, Luc De Vuyst, and Stefan Weckx. Amplicon and shotgun metagenomic
802 sequencing indicates that microbial ecosystems present in cheese brines reflect environmental inoculation
803 during the cheese production process. *International Dairy Journal*, 87:44–53, 2018.

804 John Vollmers, Sandra Wiegand, and Anne-Kristin Kaster. Comparing and evaluating metagenome
805 assembly tools from a microbiologists perspective-not only size matters! *PloS one*, 12(1):e0169662,
806 2017.

807 Kai Wang, Jiahuan Li, Liuwei Zhao, Xiyan Mu, Chen Wang, Miao Wang, Xiaofeng Xue, Suzhen Qi, and
808 Liming Wu. Gut microbiota protects honey bees (*apis mellifera* l.) against polystyrene microplastics
809 exposure risks. *Journal of Hazardous Materials*, 402:123828, 2021.

810 Edwin C Webb et al. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of*
811 *the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification*
812 *of Enzymes*. Number Ed. 6. Academic Press, 1992.

813 K Eric Wommack, Jaysheel Bhavsar, and Jacques Ravel. Metagenomics: read length matters. *Applied*
814 *and environmental microbiology*, 74(5):1453–1463, 2008.

815 Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using
816 exact alignments. *Genome biology*, 15(3):1–12, 2014.

817 Yuzhen Ye and Thomas G Doak. A parsimony approach to biological pathway reconstruction/inference
818 for genomes and metagenomes. *PLoS computational biology*, 5(8):e1000465, 2009.

819 Nicholas D Youngblut, Jacobo De la Cuesta-Zuluaga, Georg H Reischer, Silke Dauser, Nathalie Schuster,
820 Chris Walzer, Gabrielle Stalder, Andreas H Farnleitner, and Ruth E Ley. Large-scale metagenome as-
821 sembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic
822 diversity. *Msystems*, 5(6):e01045–20, 2020.

823 Rahat Zaheer, Noelle Noyes, Rodrigo Ortega Polo, Shaun R Cook, Eric Marinier, Gary Van Domselaar,
824 Keith E Belk, Paul S Morley, and Tim A McAllister. Impact of sequencing depth on the characterization
825 of the microbiome and resistome. *Scientific reports*, 8(1):1–11, 2018.

826 Eduardo E Zattara and Marcelo A Aizen. Worldwide occurrence records suggest a global decline in bee
827 species richness. *One Earth*, 4(1):114–123, 2021.

828 Hao Zheng, Julie Perreau, J Elijah Powell, Benfeng Han, Zijing Zhang, Waldan K Kwong, Susannah G
829 Tringe, and Nancy A Moran. Division of labor in honey bee gut microbiota for plant polysaccharide
830 digestion. *Proceedings of the National Academy of Sciences*, 116(51):25909–25916, 2019.

831 Jinshui Zheng, Stijn Wittouck, Elisa Salvetti, Charles MAP Franz, Hugh Harris, Paola Mattarelli, Paul W
832 O’Toole, Bruno Pot, Peter Vandamme, Jens Walter, et al. A taxonomic note on the genus *Lactobacillus*:
833 Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union
834 of *Lactobacillaceae* and *Leuconostocaceae*. 2020.