Metagenomic DNA discoveries after sequencing everything (bacteria, parasites, food, gut) inside a bee

Cong Liu

2021.8

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

Submitted for the MRes/MSc in Computational Methods in Ecology and Evolution

Declaration:

Abstract

1 Introduction

The health of bees is a crucial issue that has received increasing concern (Amiri et al., 2020). As pollinators of numerous plants, bees play an important role in the stability of natural and agricultural systems (Hristov et al., 2020a, Földesi et al., 2021, Bänsch et al., 2021, Khalifa et al., 2021). However, diverse pathogens and environmental factors impose threaten on bee health. The trend of bee decline has been widely observed at a global scale and attributed to multiple factors including pesticide exposure, invasive species, habitat loss, disease prevalence and climate change (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). Bee-associated microbiome plays a role in food digestion and carbohydrate metabolism by mediating processes including breaking down plant-produced pectin and hemicellulose (Zheng et al., 2019); hydrolysis of sucrose (Engel et al., 2012, Lee et al., 2015), a predominant component of nectar (Nicolson and Thornburg, 2007) and metabolism of mannose (Engel et al., 2012, Lee et al., 2015) which presents in nectar (Adler, 2000) and is poisonous to honey bees (de la Fuente et al., 1986). Besides, it is also involved in immunity against pathogens, providing 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* (Cariveau et al., 2014, Maes et al., 2016).

Host-associated microbiome plays an important role in host homeostasis and metagenomics boosted by high throughput sequencing provides powerful tools for understanding its composition, function and dynamics. For example, amplicon sequencing is a powerful tool and has revealed incredible understanding in microbiome (Eckburg et al., 2005, Galloway-Peña and Hanson, 2020, New and Brito, 2020). However, it only focuses on a small part of the whole microbiome. Shotgun sequencing of microbiome, or metagenomics, unselectively captures DNA in a sample and is capable of providing comprehensive inventories of taxa and functional gene clusters (FGCs) with high resolution (Quince et al., 2017, New and Brito, 2020). However, utilization of shotgun metagenomics is hindered by challenges in bioinformatics. Typical goal of metagenomics is to provide taxonomic and functional profile of the community, and there is not a golden standard for bioinformatics of metagenome. Generaly, one of the first steps in metagenomic analysis is assembling short reads into long contigs, which can help improve accuracy of metagenomic annotation (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, chimeras

(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, Huson et al., 2007, Abubucker et al., 2012, Vermote et al., 2018, Bovo et al., 2018), although the accuracy may 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 method might be helpful for overcoming the complexity of metagenomic assembly and inaccuracy of short read annotation (Becker et al., 2020).

Another challenge of metagenomics is the determination of sequencing depth. It is recommended to retrieve as many reads as possible (Quince et al., 2017), since insufficient sequencing causes imprecise estimation of taxon/FGC diversity (Cattonaro et al., 2018, Zaheer et al., 2018, Pereira-Marques et al., 2019, Gweon et al., 2019). However, deep sequencing is expensive, which hinders its utilization, especially in large-scale projects. Currently, there is hardly any published guideline 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.

The determination of sequencing depth can be conducted by rarefaction analysis, a method originated from traditional ecology (Sanders, 1968, Hurlbert, 1971, Heck Jr et al., 1975, Moreno and Halffter, 2000, Hortal and Lobo, 2005, Gómez-Anaya et al., 2014, Hughes et al., 2021). In field based surveys, a random sample of individuals is drawn from a community and assigned to species. Then a rarefaction curve illustrating expected biodiversity represented by given sampling effort (often measured by number of captured individuals) is generated by random subsampling the original sample without replacement, and quantified by model fitting (Hughes and Hellmann, 2005, Gotelli and Colwell, 2011). The slope of the rarefaction curve represents the expected rise of the curve if one more individual is captured. The sample is nearly complete if and only if the final slope is small, and the point at which the slope of rarefaction curve falls to a cut-off value represents the minimal sampling effort required for assessment of biodiversity (Heck Jr et al., 1975, Moreno and Halffter, 2000, Hortal et al., 2004, Hortal and Lobo, 2005, Chao and Jost, 2012, Gómez-Anaya et al., 2014, Roswell et al., 2021). This framework for determination of sampling effort can be used in shotgun metagenomics, 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 concept corresponding to sampling effort is sequencing depth in metagenomics, which is measured by read number.

In macroecology, comprehensive analytical frameworks are available for computing expected biodiversity returned by given sample sizes using a reference sample (Heck Jr et al., 1975, Chao and Jost, 2012, Chao et al., 2014b). However, they are not suitable for metagenomics because of the assumptions that every

captured individual can be assuredly assigned to a species, and the present of rare species is as reliable as that of abundant species. In shotgun metagenomics, reads may be unannotated or annotated incorrectly due to low sequencing quality, host contamination, similar structures in genomes of different species, non-coding regions in genomes and limited sensitivity of profiling pipeline. As a result, taxa or FGCs of low relative abundance are more likely to be false positive than that of high abundance.

In this project, I aimed to (1) construct an integrated pipeline combining assembly-dependent and -independent methods for metagenomic profiling and (2) estimate minimal sequencing depth sufficient for covering species and FGC diversity of metagenome from three environmental types: honey bees, bumble bees and surface of flowers. 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 from honey bees, bumble bees and environmental DNA (eDNA) washed from a flower (*Erigeron annuus*). It was shown that the construction of reference database and assembly-free search helped identified species not represented by the assembly. Then I simulated different sequencing depth by rarefaction and profiled subsampled datasets using the integrated pipeline, generating inventories of species and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologies (KOs) (Kanehisa and Goto, 2000). Each KO represents one FGC, *i.e.* an aggregate of genes with same function. Diversity of species and KOs was measured by Hill numbers of order 0, 1 and 2. Sequencing depth was measured by number of clean read pairs (150 bp read length). The relationship between Hill numbers and sequencing depth (number of 150 bp pair-end reads) was quantified by fitting and averaging asymptotic species accumulation models, and estimation of minimal sequencing depth was given by the point where the slope of rarefaction curve drops to cut-off values. Sample would be excluded from estimation of minimal sequencing depth if the original dataset was not sufficient for providing an almost complete inventory of species or KOs, which was characterized by small final slope of rarefaction curve for species or KO richness (Hill number of order 0). As a result, when measuring species diversity by Hill number of order 0, 1 and 2, the average minimal sequencing depth for honey bee samples are 40.33 million, 18.57 million and 17.45 million; for bumble bee samples, the averages are 42.49 million, 40.33 million and 24.77 million; and for flower surface, 0.88 million sequencing depth is not sufficient for providing an almost complete inventory of species, indicating higher sequencing depth is needed. As for functional diversity, no datasets involved in this project is big enough for providing an almost complete inventory of KOs, indicating the sequencing depth need to be higher than the maximum of each sample type, *i.e.* 45.06 million for honey bees, 53.61 million for bumble bees and 0.88 million for

flower surface.

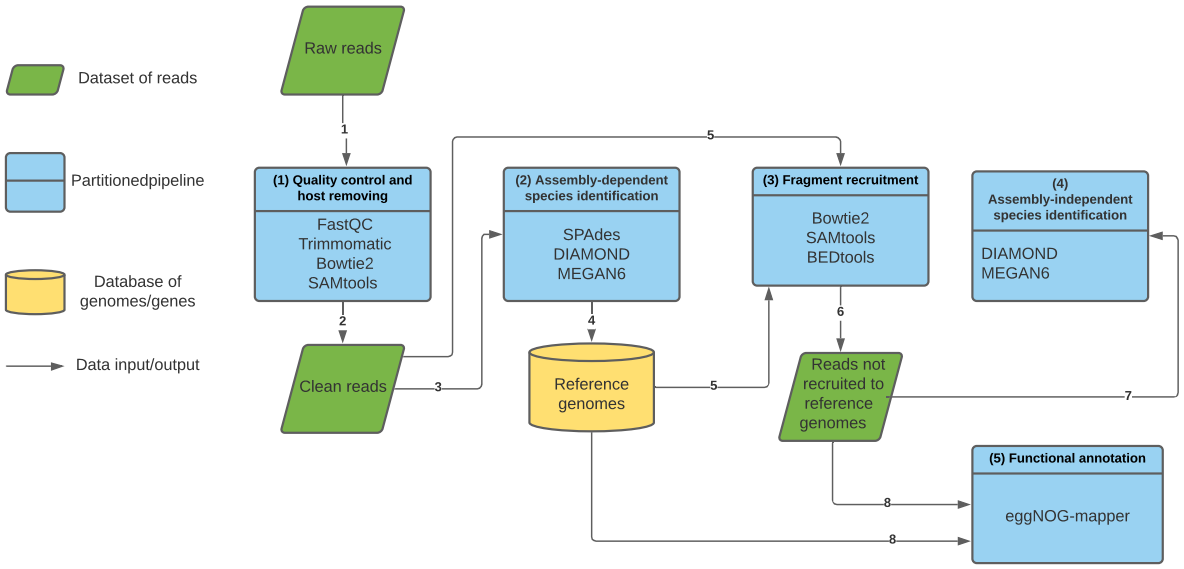

Figure 1: An overview of the integrated pipeline. The pipeline is separated into 5 modules, indicated by blue boxes. Green boxes indicate dataset of reads, while yellow boxes indicate database of genomes. Black arrows indicate the input and output of each step and the numbers on them indicate the order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

2 Materials and Methods

2.1 Samples, DNA extraction and sequencing

2.2 Metagenomic profiling using integrated pipeline

2.2.1 Integrated pipeline

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

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

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

taxa by MEGAN6 (Huson et al., 2007).

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

Reads not recruited by the reference database are subjected to assembly-independent species identification. They are aligned to NCBI nr database through DIAMOND and assigned to taxa by MEGAN6.

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

The integrated pipeline was used for analyzing metagenomic datasets involved in this study and details in parameter settings of each module are described in Supplementary 6.1.

2.2.2 Taxon/function quantification and metabolic pathway reconstruction

After profiling, identified species and KOs were quantified by calculating relative sequence abundance, *i.e.* proportion of reads assigned to a species/KO in all reads annotated. For species without available reference genomes, their abundances were calculated using reads assigned to them in assembly-independent search. As for taxa with available reference genomes, they may be identified in both assembly-dependent and -independent search due to strain-specific genomic structures that are not present in reference genomes. Their abundances were calculated by summating 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 summating 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 minimal sequencing depth required for metagenomic profiling

The impact of sequencing depth on taxon/function diversity of metagenome was simulated by rarefaction. Diversity was measured by Hill numbers, which are calculated from inventory of relative abundances. For taxon diversity, the inventory was generated by assigning reads to species; while for function diversity, it was obtained by assigning reads to KOs. Then the rarefaction curve plotting Hill number against sequencing depth was computed using a multimodel inference method, and estimation of minimal sequencing depth required for covering diversity were given by the point where the slope of rarefaction curve drops to a cut-off value.

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, *i.e.* randomly taking a proportion of the original dataset without replacement. 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).

$$D^{(q)} = (\sum_{i}(p_i)^q)^{\frac{1}{1-q}} \tag{1}$$

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

species/KOs with general abundances.

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

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

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

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

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

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

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

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

$$L = -0.5nlog(\frac{Rss}{n}) \tag{5}$$

Rss represents residual sum of squares.

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

$$\Delta_i = AICc_i - AICc_{min} \tag{6}$$

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

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

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

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

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

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

The asymptote of rarefaction curve as sequencing depth tends to infinity is given by Equation 10. It provides an estimation of the total diversity and is comparable among metagenomic DNA samples with different sequencing depth (Lamas et al., 1991, SoberónM and LlorenteB, 1993, Hortal et al., 2004, Jiménez-Valverde and Lobo, 2005, Hortal et al., 2006). However, the accuracy of asymptotic estimators is controversial in macroecology (SoberónM and LlorenteB, 1993, Colwell and Coddington, 1994, Chazdon et al., 1998, Jimenez-Valverde et al., 2006, Hortal et al., 2006), and their performance in shotgun metagenomics is seldom evaluated.

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

2.3.3 Estimating minimal sequencing depth using rarefaction curves

Minimal sequencing depth is defined as the point at which diversity starts to level off as sequencing depth increases, and its precise estimation via rarefaction is based on the assumption that the original dataset 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, Moreno and Halffter, 2000, Hortal et al., 2004, Hortal and Lobo, 2005, Gómez-Anaya et al., 2014). Then an estimation of minimal 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).

3 Results

3.1 Sequence reads and pipeline evaluation

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 dropped for further analysis since its raw read pair number (1.10 million) is significantly lower than three other samples (about 59 million). 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.

In integrated pipeline (Figure 1), clean non-host reads are first assembled to into contigs and assigned to taxa. To address false negative results in assembly-dependent taxon search (Sharon et al., 2015, Vollmers et al., 2017), a reference database composed of genomes of species represented by contigs is constructed. Reads not aligned to the reference database are subjected to assembly-independent taxon search. The reference database and assembly-independent search helped improve species identification in all three sample types, especially in simulations of low sequencing depth (Figure 2).

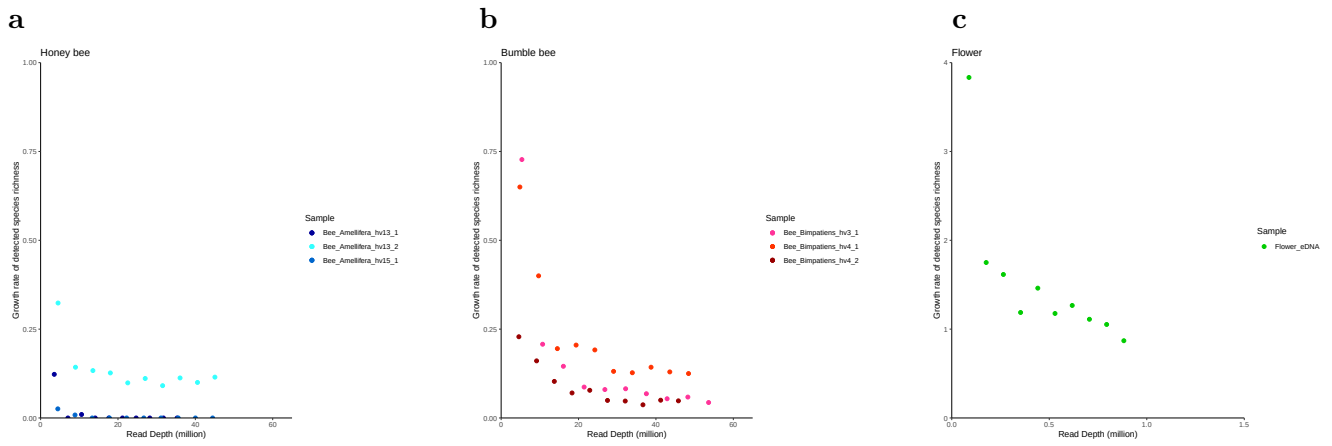

Figure 2: Integrated pipeline improves the detection of species richness. The horizonal axis represents sequencing depth, and the vertical axis represents the ratio between number of species only detected by assembly-independent search and number of species detected by assembly-dependent search. Sample type is shown in the top left of each subfigure.

3.2 Characterization of species composition

All seven samples were dominated by bacterial and/or arthropod species (Figure 3b). Honey bee samples *Bee_Amellifers_hv13_2* and *Bee_Amellifera_wild_1* were dominated by bacteria (approximate 90% relative abundance), while in *Bee_Amellifera_hv15_1*, arthropods, bacteria and viruses accounted for most annotated reads. As for bumble bee samples, they were all dominated by bacteria and arthropods. For the flower eDNA sample, arthropods were the most dominant and bacteria were the second.

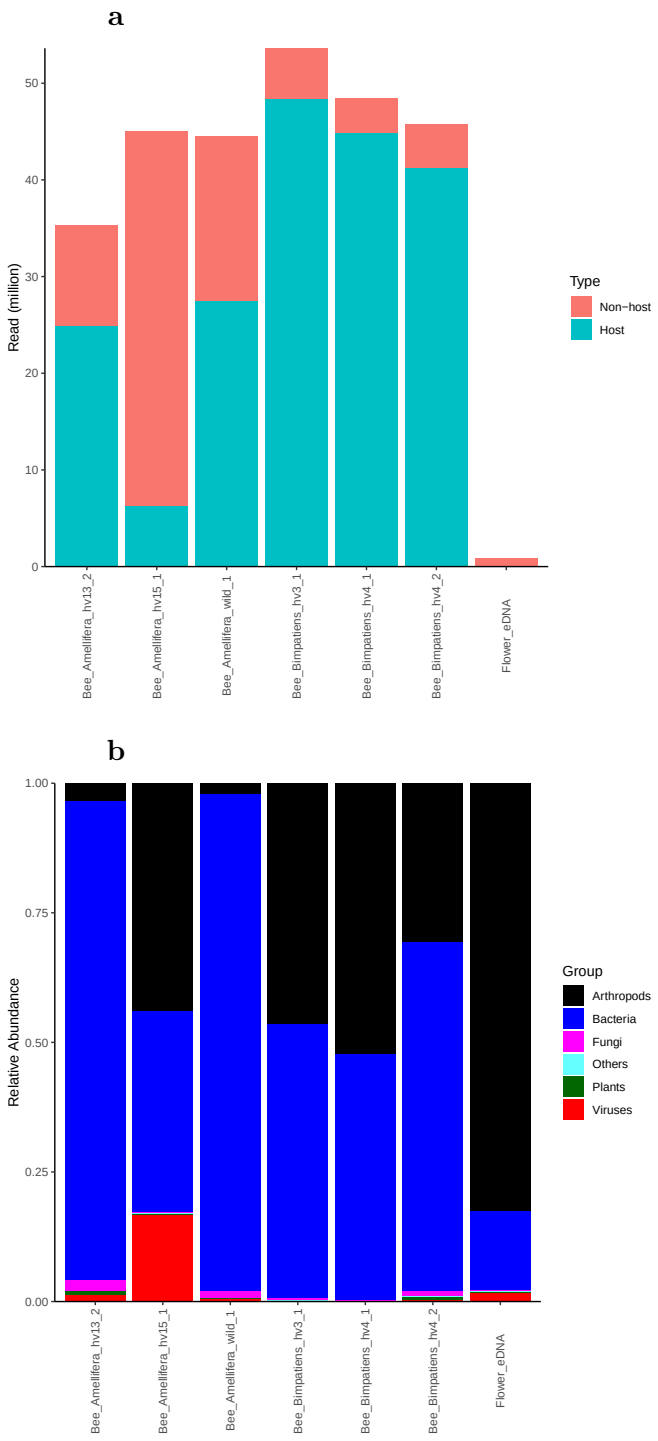

Figure 3: a. The number of host and non-host reads in each sample. b. The relative abundance of species under six taxonomic groups: superkingdom Viruses, superkingdom Bacteria, kingdom Viridiplantae (plants), kingdom Fungi, phylum Arthropoda and others (species that are not in the other five groups).

Figure 4 illustrates composition of bacteria, most of which are common bee-associated species within *Bifidobacterium*, *Frischella*, *Gilliamella*, *Snodgrassella* and *Lactobacillus* (Koch and Schmid-Hempel, 2011a, Moran, 2015, Kwong et al., 2017). *Apilactobacillus* and *Bombilactobacillus* are genera separated

from *Lactobacillus* recently (Zheng et al., 2020). *Fructobacillus* is 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).

Figure S2 shows composition of arthropods, and most of them are pollinators within *Apis* and *Bombus*. 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.

Figure S3 illustrates plant species, indicating foraging area of bees. Most samples were dominanted by one or two plant species, except *Bee_Amellifera_hv13_2*. Some crops were identified, including *Brassica napus* (rape), *Brassica oleracea*, *Cicer arietinum* (chickpea), *Glycine max* (soybean), *Helianthus annuus* (sunflower), *Nicotiana sylvestris* (flowering tabacco) and *Raphanus sativus* (radish).

Figure S4 shows compositions of fungal species. *Nosema ceranae*, a widespread bee pathogen, was the dominant fungal species in most samples including the flower eDNA. Three yeast species (*Clavispora lusitaniae*, *Saprochaete ingens* and *Wickerhamiella sorbophila*) were found in bees.

Figure S5 illustrates viruses identified. Several phages were identified, including *Bifidobacterium phage BitterVaud1* that infects bee-commensal bacterium *Bifidobacterium asteroides* (Bonilla-Rosso et al., 2020), *Bacteriophage sp.* that infects *Pseudomonas aeruginosa* (Essoh et al., 2015), an opportunistic pathogen that might contaminate bees (Bailey, 1968, Papadopoulou-Karabela et al., 1992, 1993), and species in Myoviridae and Siphoviridae. Most of the others are arthropod-associated viruses. *Apis mellifera filamentous virus* and *Bombus cryptarum densovirus* are bee-infecting viruses (Clark, 1978, Bailey et al., 1981, Schoonvaere et al., 2018). Several arthropod-infecting parvoviruses were found, including *Blattodean pefuambidensovirus 1*, *Hemipteran scindoambidensovirus 1*, *Hymenopteran scindoambidensovirus 1* and *Orthopteran scindoambidensovirus 1* (Pénzes et al., 2020).

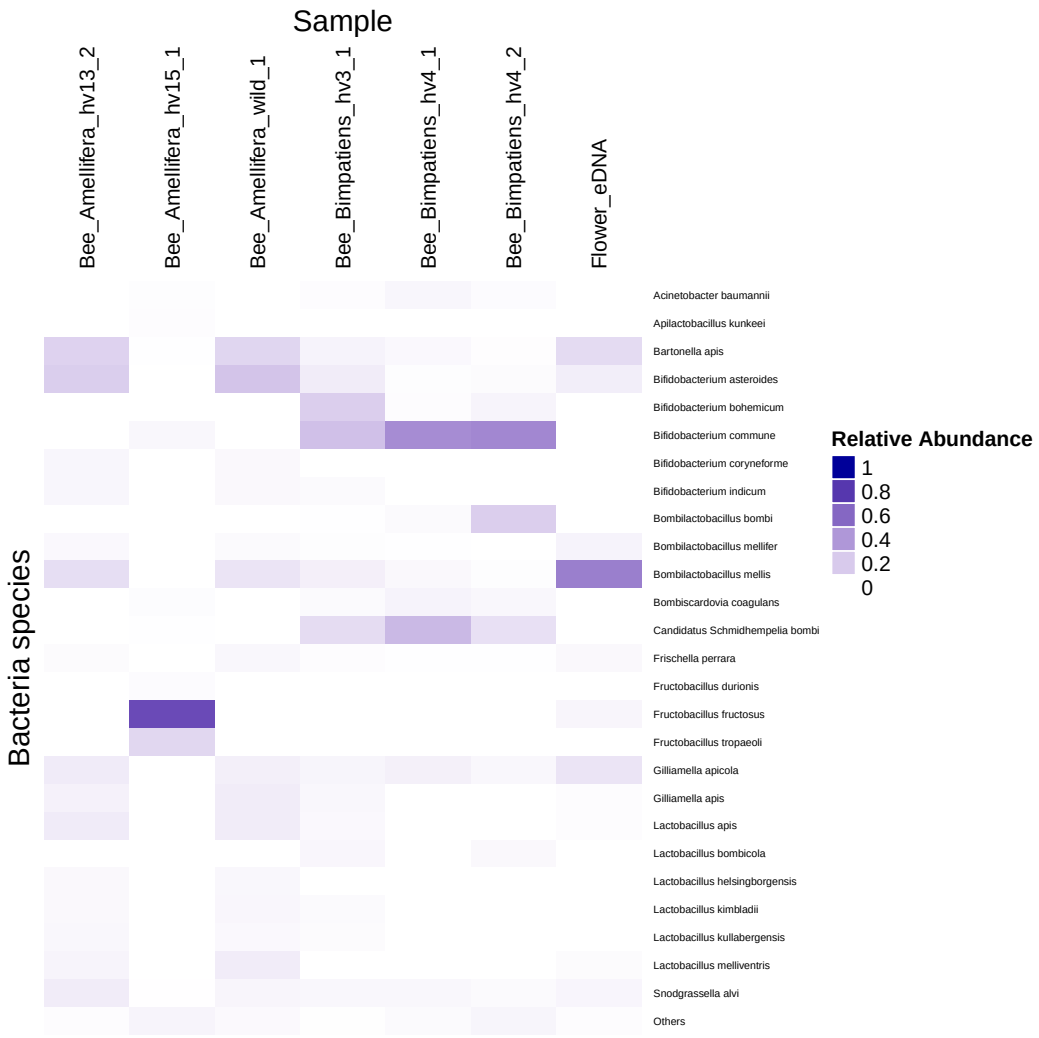

Figure 4: Heatmaps for bacterial species abundance distribution in all samples. The relative abundance takes reads assigned to bacterial species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

3.3 Characterization of functional profiling

In order to illustrate metabolic potentially of metagenomic samples, KEGG pathways were inferred using a parsimony approach. Reads assigned to plant and arthropod species were not involved in pathway inference. The coverage of a pathway was calculated by the ratio between number of annotated KOs and total number of KOs involved in that pathway.

Here concern is given to metabolism pathways of carbonhydrates and amino acids, which are crucial for bee health. Inferred pathways indicate potential capability of metabolism of sugars including fructose, sucrose, mannose and galactose (Figure S6), and all essential amino acids for honey bees (Figure 5)(Groot, 1953).

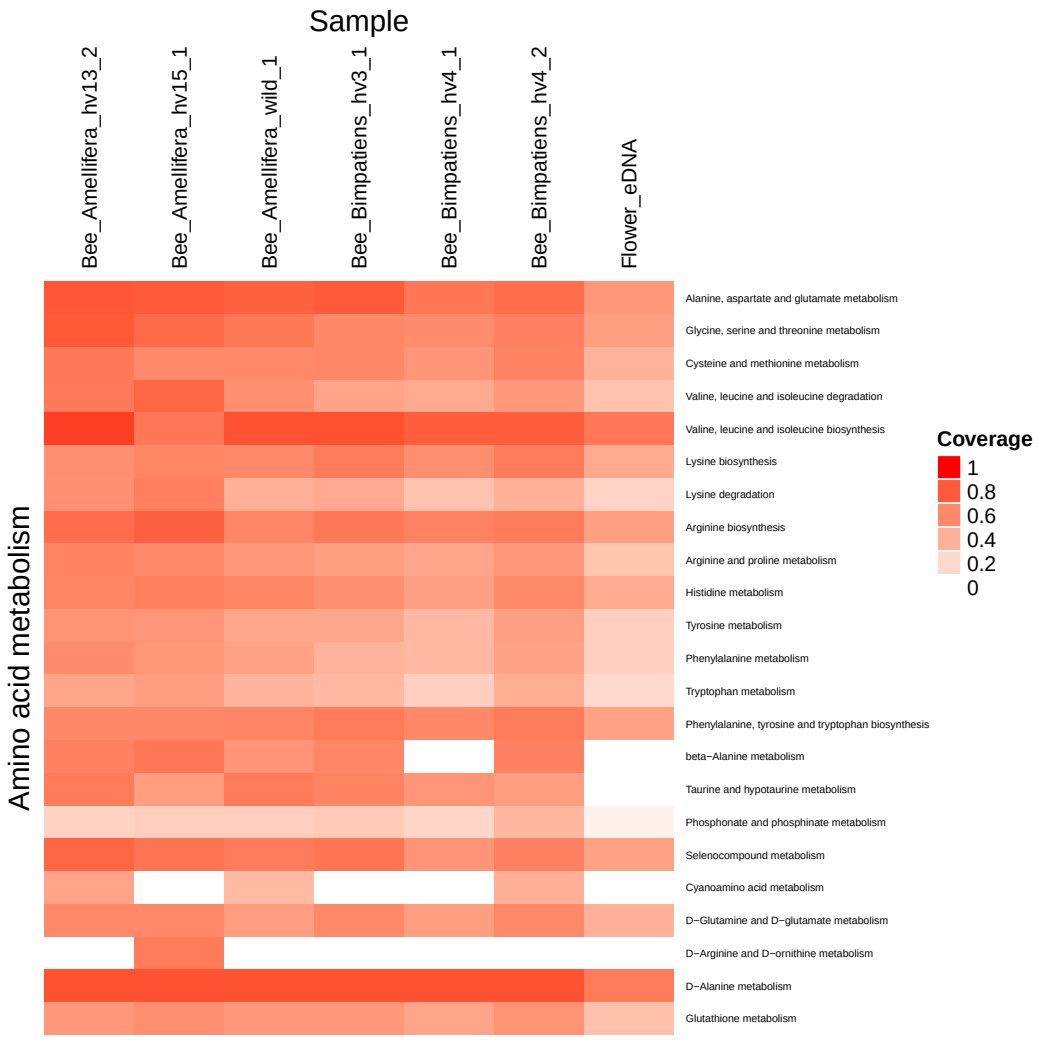

Figure 5: Heatmaps for pathways of amino acid metabolism.

3.4 Minimal sequencing depth required for detection of species and function diversity

In order to estimate minimal sequencing depth for combinations of sample type (honey bee, bumble bee and flower eDNA) and study type (species profiling and function profiling), shallow sequencing was simulated by rarefaction. The relationship between sequencing depth (clean read pairs) and species/KO diversity (Hill numbers of order 0, 1, 2) was quantified by fitting and averaging asymptotic species accumulation models. The asymptote of the model provides an estimation of total diversity and its slope reflects the increase rate of diversity. Minimal sequencing depth was estimated by the point at wich the slope of the rarefaction curve drops to a cut-off value.

Rarefaction assumes that the original dataset provides an almost complete inventory, which can be verified by final slope of rarefaction curve for Hill number of order 0 (richness). Figure 6 shows rarefaction curves for species/KO richness and Table 1 summarizes their final slopes. For species diversity rarefaction,

all bumble bee samples are sufficient, with final slopes < 0.1 and completeness (ratio between final richness and asymptote) > 0.98. As for honey bees, Bee_Amellifera_hv15_1 and *Bee_Amellifera_wild_1* are sufficient, while *Bee_Amellifera_hv13_2* is insufficient, with final slope > 1 and completeness < 0.8. For the flower eDNA sample, the final slope of species richness rarefaction curve is 10.8380 and its completeness is 1.32%, indicating more sequencing effort is needed for species profiling. As for function diversity rarefaction, the final slopes of all KO richness rarefaction curves are higher than 15, indicating no dataset can providing an almost complete inventory of KOs. Thus, estimation of minimal sequencing depth was conducted for the combinations of two sample types (honey bee and bumble bee) and one study type (species profiling), based on five datasets (two honey bees and three bumble bees).

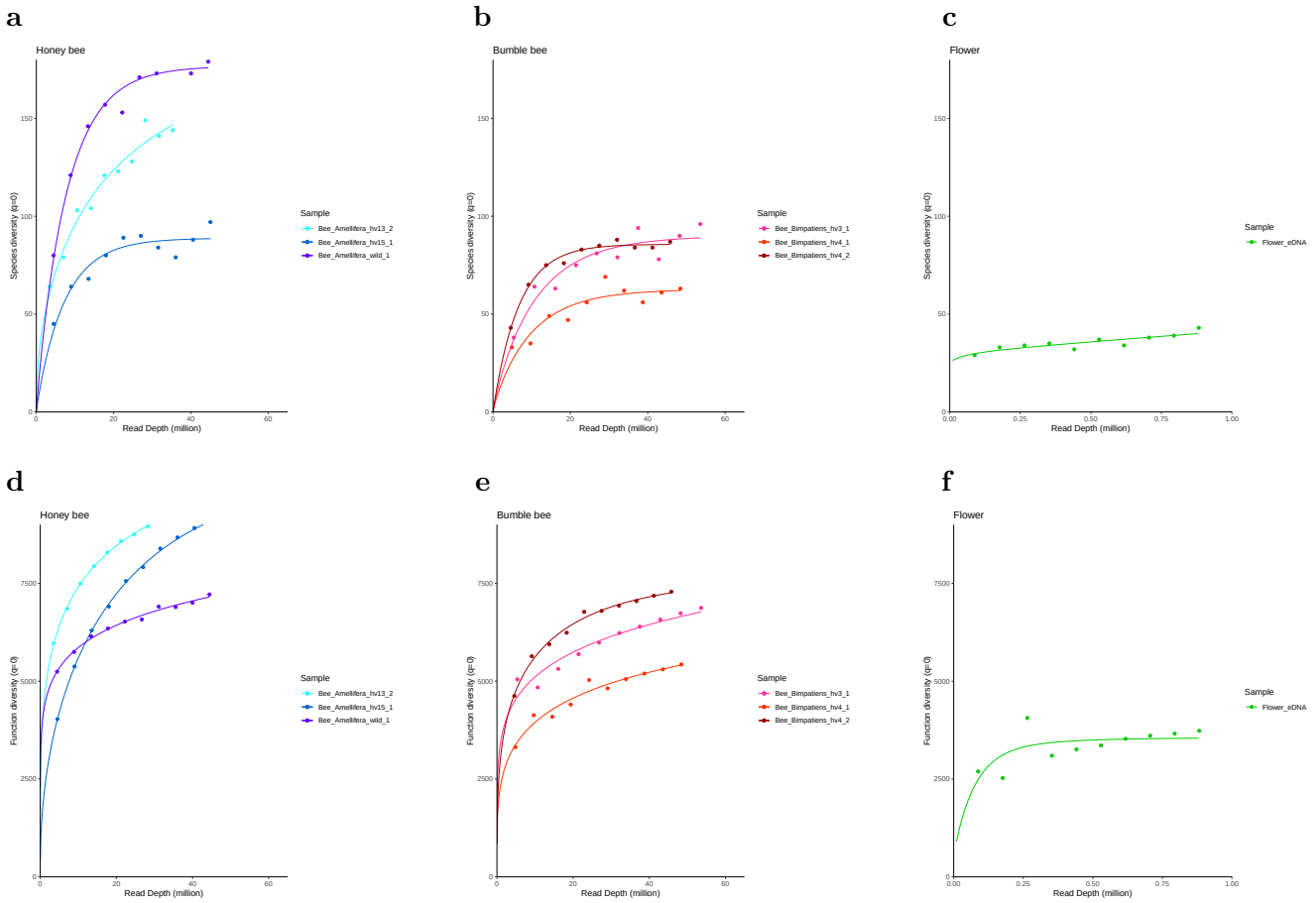

Figure 6: Rarefaction curves for species (a, b, c) or KO (d, e, f) richness (Hill number of order 0). The horizonal axis represents sequencing depth, and the vertical axis represents richness. Sample type is shown in the top left of each subfigure. Note that the scale of horizonal axis in subfigure c 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
Bee_Amellifera_hv13_2	species	35.28	144	146.69	1.0958	184.94	0.7931
Bee_Amellifera_hv15_1	species	45.05	97	88.53	0.0344	88.84	0.9966
Bee_Amellifera_wild_1	species	44.46	179	175.89	0.0991	176.84	0.9946
Bee_Bimpatiens_hv3_1	species	53.61	96	89.00	0.0897	90.79	0.9803
Bee_Bimpatiens_hv4_1	species	48.42	63	62.06	0.0626	62.92	0.9862
Bee_Bimpatiens_hv4_2	species	45.80	87	85.59	0.0146	85.69	0.9988
Flower_eDNA	species	0.88	43	40.04	10.8380	3034.88	0.0132
Bee_Amellifera_hv13_2	KO	35.28	9289	9284.45	38.6844	10641.27	0.8725
Bee_Amellifera_hv15_1	KO	45.05	9046	9102.78	43.8717	10453.98	0.8707
Bee_Amellifera_wild_1	KO	44.46	7216	7151.72	21.4368	14779.89	0.4839
Bee_Bimpatiens_hv3_1	KO	53.61	6872	6769.80	21.1588	28111.79	0.2408
Bee_Bimpatiens_hv4_1	KO	48.42	5427	5410.71	19.6932	7665.58	0.7058
Bee_Bimpatiens_hv4_2	KO	45.8	7288	7259.56	15.6309	7873.64	0.9220
Flower_eDNA	KO	0.88	3732	3544.54	60.4367	3571.11	0.9926

Minimal 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 minimal 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 minimal 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 minimal sequencing depth are 18.57 million ($q = 1$) and 17.45 million ($q = 2$). For bumble bee samples, the average minimal 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
Bee_Amellifera_hv15_1	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
Bee_Amellifera_wild_1	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
Bee_Amellifera_hv15_1	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
Bee_Amellifera_wild_1	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
Bee_Amellifera_hv15_1	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
Bee_Amellifera_wild_1	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
Bee_Bimpatiens_hv3_1	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
Bee_Bimpatiens_hv4_1	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
Bee_Bimpatiens_hv4_2	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
Bee_Bimpatiens_hv3_1	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
Bee_Bimpatiens_hv4_1	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
Bee_Bimpatiens_hv4_2	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
Bee_Bimpatiens_hv3_1	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
Bee_Bimpatiens_hv4_1	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
Bee_Bimpatiens_hv4_2	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

Shotgun metagenomics provides powerful tools for microbiome investigations, yet its utilization is hindered by bioinformatical challenges and high cost of deep sequencing. Here, I constructed an integrated pipeline, which combines assembly-dependent and -independent profiling methods to help detect species richness and profiled real metagenomic datasets from honey bees, bumble bees and flower surface. Then I simulated the relationship between species/KO diversity measured by Hill numbers (order $q = 0, 1, 2$) and sequencing depth measured by clean read pair (2×150 bp) number. It is showed that to detect species diversity, the average minimal sequencing depth for honey bee samples are 40.33 million ($q = 0$), 18.57 million ($q = 1$) and 17.45 million ($q = 2$); for bumble bee samples, the averages are 42.49 million ($q = 0$), 40.33 million ($q = 1$) and 24.77 million ($q = 2$); and for flower surface, more than 0.88 million sequencing depth is needed. As for function diversity, no datasets in this project is sufficient for providing an almost complete inventory of KOs, indicating deeper sequencing is needed, *i.e.* sequencing depth need to be higher than 45.06 million for honey bees, 53.61 million for bumble bees and 0.88 million for flower surface. Additionally, bee symbionts and pathogens were found on the flower eDNA sample, and pathway inference indicates that bee-associated microbiome may have capability of metabolizing amino acids that are essential for host.

Shotgun metagenomics provides unique advantages over amplicon sequencing, which is vastly used in bee microbiome investigations (*e.g.* Geldert et al. (2021), Wang et al. (2021), Powell et al., Kapheim et al. (2021)). In amplicon sequencing, a species-specific barcode region is amplified and sequenced (Abdelfattah et al., 2018). Since it captures a small region of the whole genome, amplicon sequencing dataset is relatively small in size, which makes it low-cost both economically and computationally. Besides, comprehensive and streamlined softwares for amplicon sequencing analysis are available, including QIIME (Bolyen et al., 2019), Mothur (Schloss, 2020) and VSEARCH (Rognes et al., 2016). However, a single amplicon sequencing dataset does not represent all taxa in the community since different barcode region is used for identification of different taxonomic group, *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). Besides, it does not provide information on FGCs. Metabolic capacity of microbiome need to be inferred based on reference genomes (Aßhauer et al., 2015, Douglas et al., 2018). Shotgun metagenomics provides an alternative to overcome these drawbacks by capturing and sequencing DNA in a sample unselectively. It is capable of representing all taxonomic groups and providing information on FGC composition.

The advantages of shotgun metagenomics over amplicon sequencing are valuable for investigations in bee-associated microbiome for two reasons. First, bees vist 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 plants, arthropods, bacteria and viruses (Bovo et al., 2018, Ribani et al., 2020, Bovo et al., 2020, Matsuzawa et al., 2020). These materials represent a diverse assemblage which is difficult to be profiled comprehensively via amplicon sequencing. Besides, bee bacterial symbionts are diversified at strain level (Engel et al., 2012, Powell et al., 2016, Ellegaard et al., 2020). Bacterial strains are often highly variable in gene content (Cordero and Polz, 2014, Brockhurst et al., 2019), and thus different in metabolic capacity. However, resolving strain-level diversity by species-specific region is difficult (Rodriguez-R et al., 2018, Ciufo et al., 2018, Olm et al., 2020), and no information on gene content is provided by amplicon sequencing. Based on these considerations, shotgun metagenomics provides valuable tools for bee microbiome investigations. However, utilization of shotgun metagenomics is hindered by unique challenges. First, there is not a golden standard for bioinformatics of metagenomic annotation pipeline to provide abundances of taxa and FGCs. Second, shotgun metagenomics is often associated with high cost since it captures a big proportion of microbial genomes.

The integrated pipeline for taxonomic and functional profiling of metagenomic dataset 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 different sequencing depth by rarefaction (Figure 2). 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 is important for troubleshooting.

Determination 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 surface. It was shown that increasing sequencing depth boosted identification of species/KO (Figure 6, 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, optimization of sequencing depth can be made when the budget is limited (Table 2). For honey bees, approximate 40 million sequencing depth can be sufficient for representing species richness. When assessing biodiversity by Hill numbers of order 1 (Equation 2) or 2 (Equation 3), 17-19 million sequencing depth can provide robust estimation. As for bumble bees, about 40-43 million sequencing depth can provide reliable estimation for Hill numbers of order 0 (richness) or 1, and about 25 million sequencing depth can be sufficient for representing biodiversity measured by Hill number of order 2.

Additionally, the results of metagenomic profiling indicate overlap between microbiome of bees and flowers, and provide evidence on potentiality of bee microbiome in metabolism of carbonhydrates and amino acids. In the flower eDNA sample, bee-associated microorganisms were identified, including both pathogens (*e.g.* *Nosema ceranae* and *Apis mellifera filamentous virus*) and typical bee symbionts (*e.g.* *Bifidobacterium asteroides*, *Bombilactobacillus mellis* and *Gilliamella apicola*) (Figure 4, S4 and S5). Shared microbiome between bees and flowers is revealed by growing evidence, although its role in pollination system 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 (arginine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan and valine) (Groot, 1953) were inferred (Figure 5). 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.

5 Data and Code Availability

6 Supplementary

6.1 Parameter settings of integrated pipeline

6.1.1 Quality control and host removing

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

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

in a pair of reads is mapped).

6.1.2 Assembly-dependent species identification

In order to identify taxa of metagenome, *de novo* assembly was conducted. Non-host read pairs were assembled using SPAdes v.3.15.2 (Prjibelski et al., 2020). Values of k-mer ranged from 21 to 101 at interval of 10. Flags *–only-assembler* and *–meta* were used. Through flag *–meta*, SPAdes runs metaSPAdes which is developed for metagenomic assembly (Nurk et al., 2017). The *–only-assembler* flag skips read error correction and runs assembly only. Its utilization is justified by the following facts. First, when *–only-assembler* is not used, SPAdes conducts error correction before assembly. It is conducted by BayesHamme, which is optimized for single cell sequencing instead of shotgun metagenomics (Nikolenko et al., 2013). Besides, reads used for *de novo* assembly had been filtered to ensure their quality.

After assembly, taxon identification was conducted using DIAMOND v.2.0.7.145 (Buchfink et al., 2015) and MEGAN6 (Huson et al., 2007). Assembled contigs with a length above 500 bp were aligned to nr database using DIAMOND v.2.0.7.145 with *–long-reads* flag. This flag triggers frame-shift aware alignment mode, which is optimized for long sequence alignment. Therefore, short contigs (length < 500 bp) were not retained. Besides, alignments with an E-value $< 1e-5$ or identity $< 50\%$ were removed, and for each contig, only alignments above 10% of the best local bit score were retained. The output of DIAMOND was analysized by the *blast2rma* tool of MEGAN6 with *-lg* flag, which runs lowest-common-ancestor (LCA)-based algorithm developed for long contigs and assigns each contig to a taxon (Huson et al., 2018). The parameter *-supp* was 0, which means the present of a taxon would be identified as long as at least one contig was assigned. This value was used because a contig is assembled from multiple short reads and represents a strong signal for the present of a taxon.

6.1.3 Fragment recruitment

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

Then fragment recruitment was conducted. The non-host read pairs were mapped to genomic sequences in the reference database using Bowtie2. Read pairs that were not recruited were extracted using SAMtools. Settings for Bowtie2 and SAMtools were the same as that described in 6.1.1. Read pairs recruited by the reference database were assigned to corresponding species, while the others were subjected to

assembly-independent search.

6.1.4 Assembly-independent species identification

In order to detect species not represented by assembly (Sharon et al., 2015, Vollmers et al., 2017), assembly-independent search was conducted, taking read pairs not recruited by the reference database as input. These reads were aligned to nr database through DIAMOND without using *–long-reads* flag, which triggers computing alignments for short metagenomic reads. Other settings were the same as described in 6.1.2. Then the output of DIAMOND was analysed by MEGAN6 (*blast2rma* tool), which assigns read pairs to taxa through LCA algorithm. Here the parameter *-supp* was 0.1, which means a taxon is reported after being represented by at least 0.1% of all assigned read pairs. It was used in order to avoid false positive results.

6.1.5 Functional annotation

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

6.2 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)

6.3 Exemplification of effect of quality control

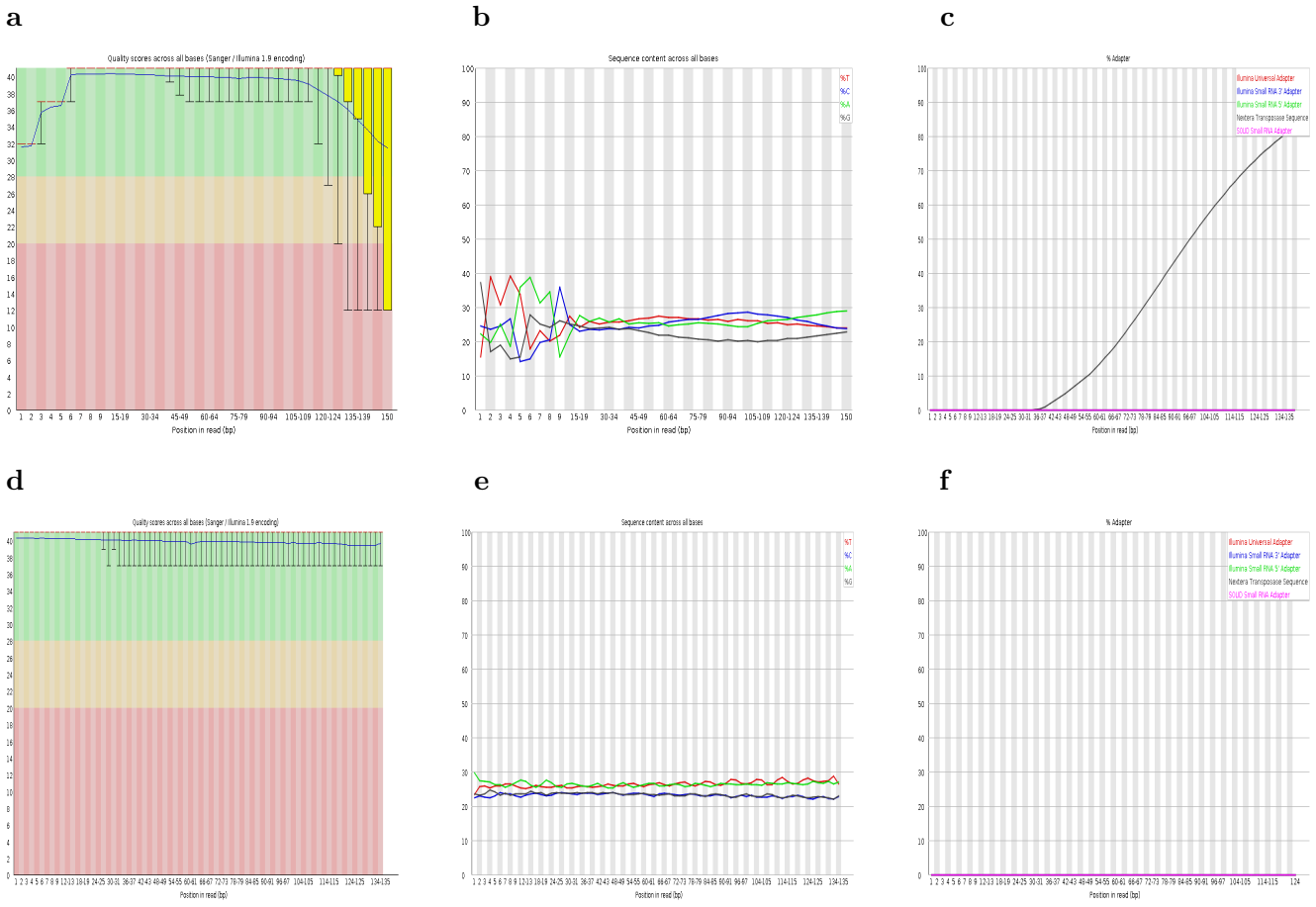

Figure S1: Data quality report of forward reads from bumble bee sample *Bee_Bimpatiens_hv4_1*. a, b and c shows low base quality in 3'-end, uneven base content in 5'-end and present of Nextera adaptors in raw data, respectively. d, e and f shows the same results from clean data.

6.4 Numbers of raw, clean and non-host read pairs

Table S2: Basic statistics of sequenced read pairs. Data in "Clean" considers as background the numbers of raw read pairs ("Raw"), while data in "Non-host" considers as background the numbers of clean read pairs ("Clean").

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

6.5 Species composition of arthropods, plants, fungi and viruses

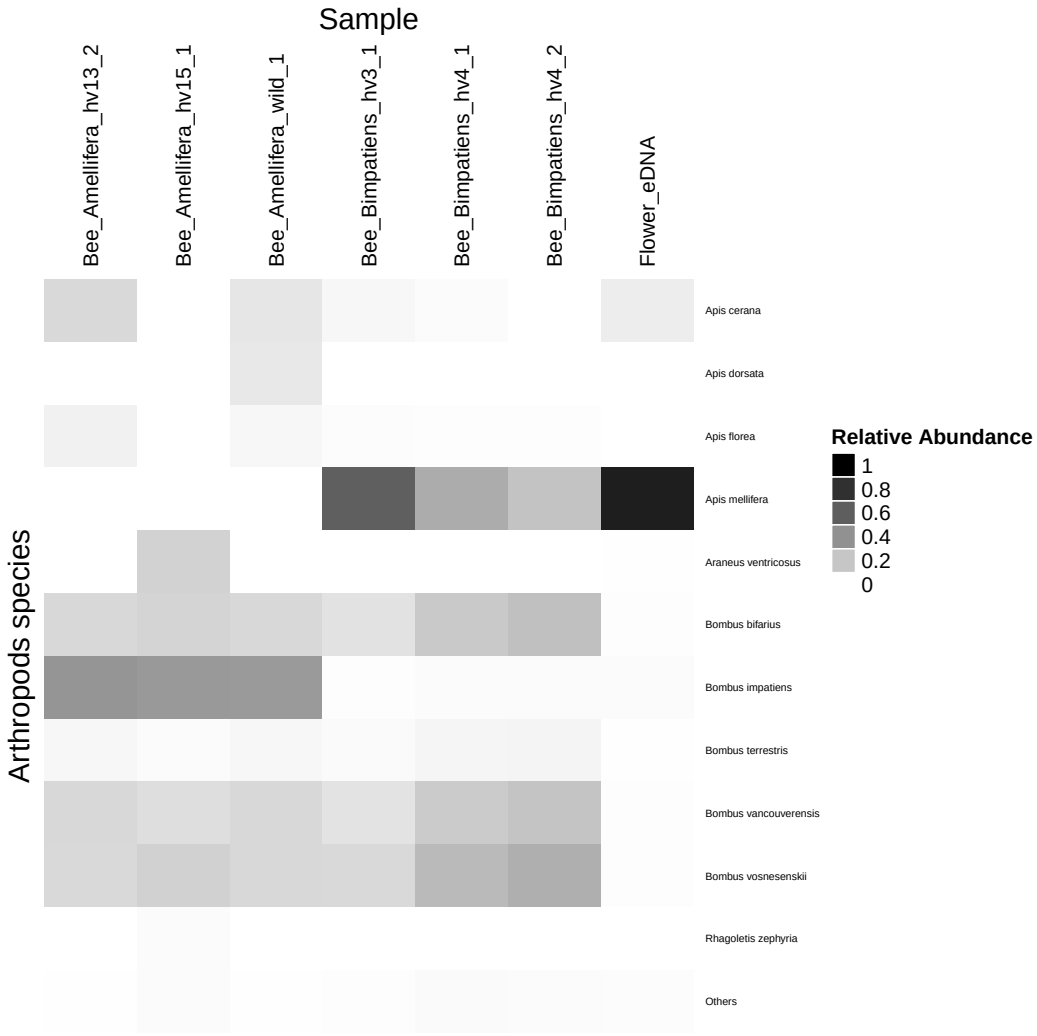

Figure S2: Heatmaps for arthropod species abundance distribution in all samples. The relative abundance takes reads assigned to arthropod species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others". It should be noted that for bee samples, host contamination 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.

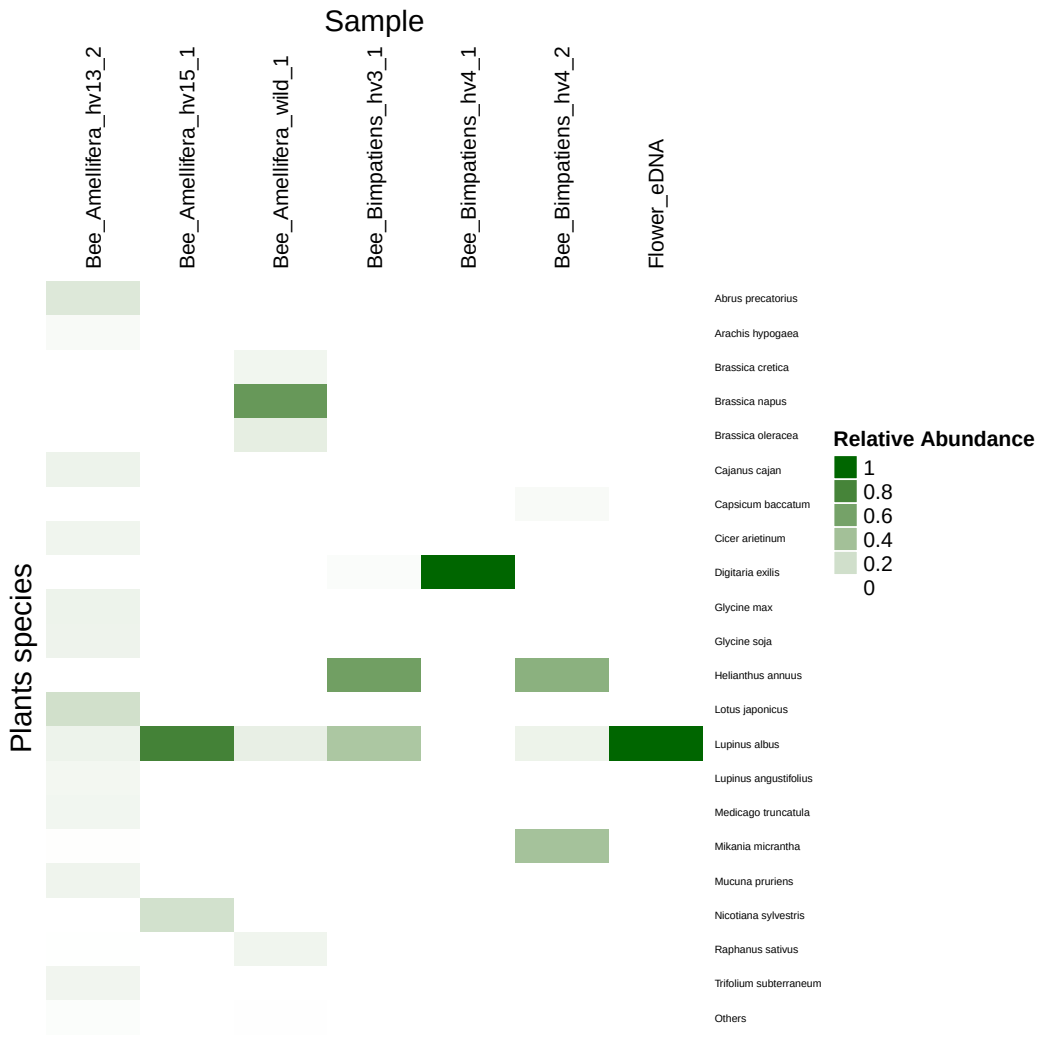

Figure S3: Heatmaps for plant species abundance distribution in all samples. The relative abundance takes reads assigned to plant species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

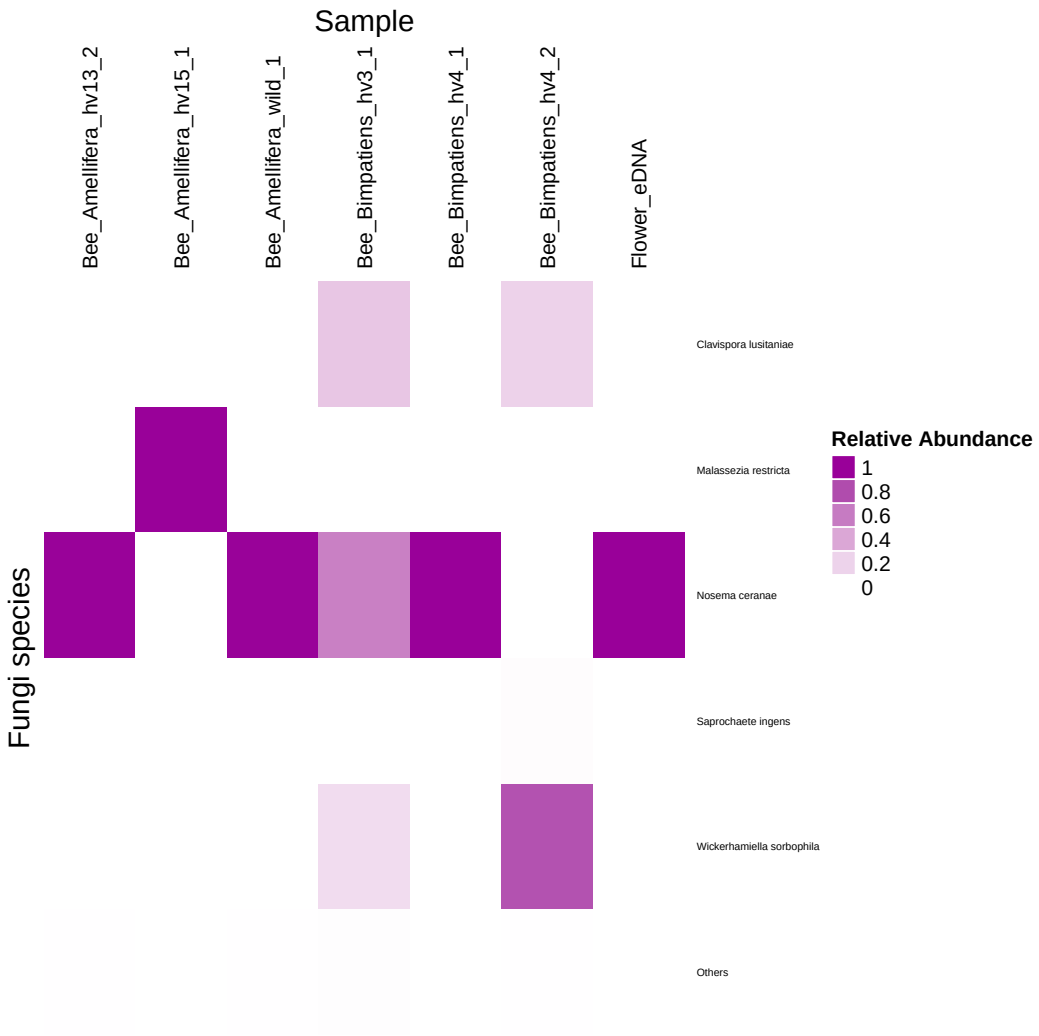

Figure S4: Heatmaps for fungal species abundance distribution in all samples. The relative abundance takes reads assigned to fungal species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

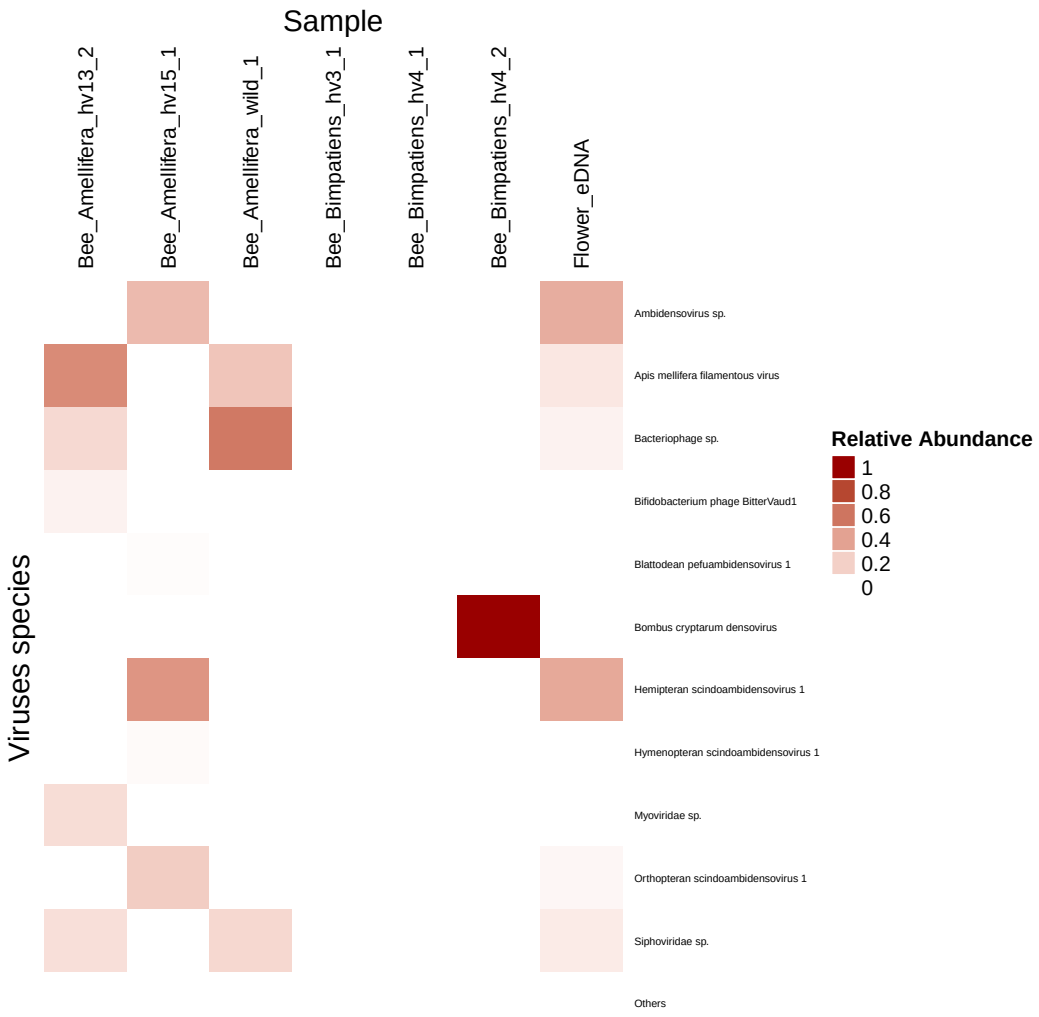

Figure S5: Heatmaps for viruse species abundance distribution in all samples. The relative abundance takes reads assigned to viruse species as background. Species with relative abundance smaller than 1% in all samples are collapsed as "others".

6.6 Inferred carbohydrate metabolism pathways

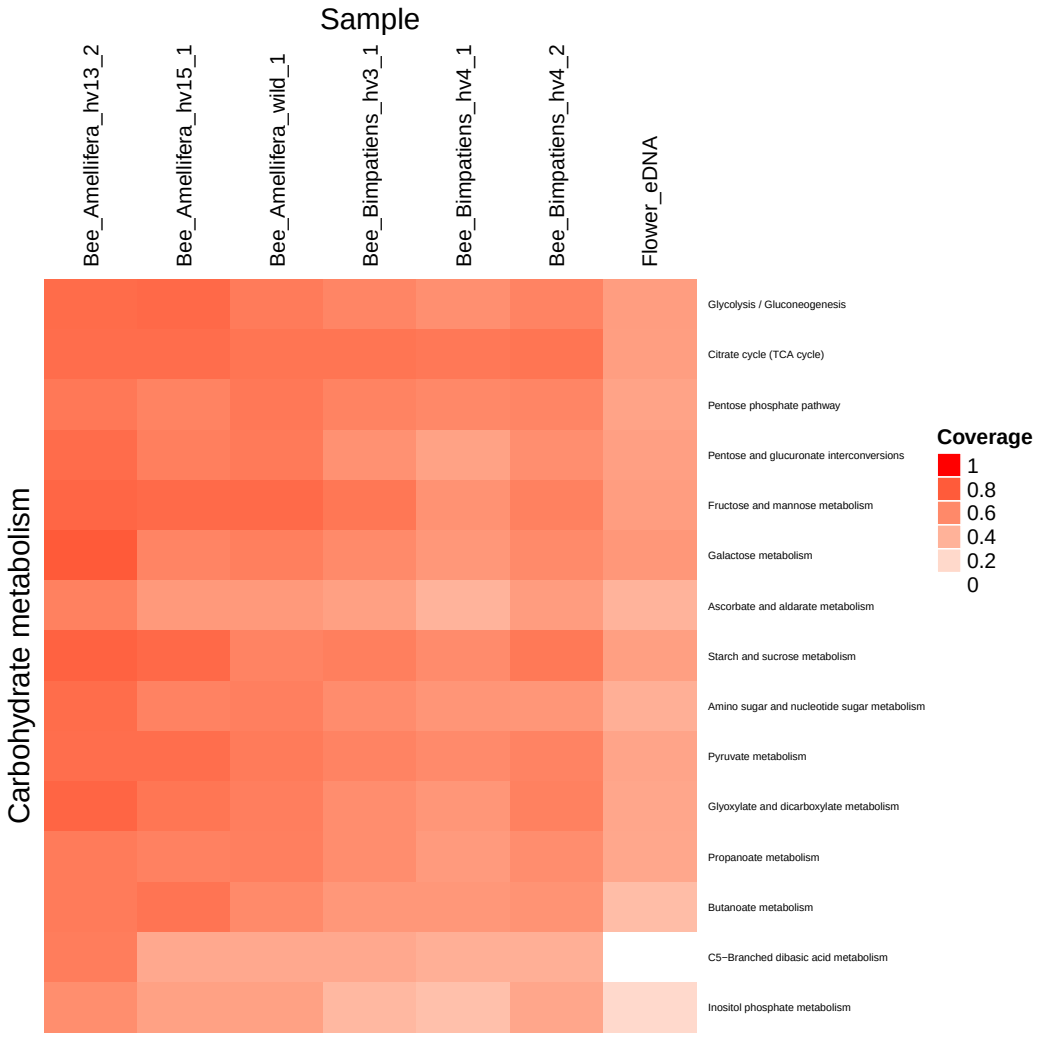

Figure S6: Heatmaps for pathways of carbohydrate metabolism.

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.

Lynn S Adler. The ecological significance of toxic nectar. *Oikos*, 91(3):409–420, 2000.

Esmaeil 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.

L Bailey, JM Carpenter, and RD Woods. Properties of a filamentous virus of the honey bee (apis mellifera). *Virology*, 114(1):1–7, 1981.

Leslie Bailey. Honey bee pathology. *Annual review of entomology*, 13(1):191–212, 1968.

Svenja Bänsch, Teja Tscharntke, 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.

Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, 37(8):852–857, 2019.

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.

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

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

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

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

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

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

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

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

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

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

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

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

Anne Chao, Nicholas J Gotelli, TC Hsieh, Elizabeth L Sander, KH Ma, Robert K Colwell, and Aaron M Ellison. Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological monographs*, 84(1):45–67, 2014b.

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

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

Stacy Ciufo, Sivakumar Kannan, Shobha Sharma, Azat Badretdin, Karen Clark, Seán Turner, Slava Brover, Conrad L Schoch, Avi Kimchi, and Michael DiCuccio. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the ncbi. *International journal of systematic and evolutionary microbiology*, 68(7):2386, 2018.

Truman B Clark. A filamentous virus of the honey bee. *Journal of Invertebrate pathology*, 32(3):332–340, 1978.

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

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

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

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

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

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

Mónica de la Fuente, Pablo F Penas, and Alberto Sols. Mechanism of mannose toxicity. *Biochemical and biophysical research communications*, 140(1):51–55, 1986.

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

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

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

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

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

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

Philipp Engel, Waldan K Kwong, Quinn McFrederick, Kirk E Anderson, Seth Michael Barribeau, James Angus Chandler, R Scott Cornman, Jacques Dainat, Joachim R De Miranda, Vincent Doublet,

et al. The bee microbiome: impact on bee health and model for evolution and ecology of host-microbe interactions. *MBio*, 7(2):e02164–15, 2016.

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

Rita Földesi, Brad G Howlett, Ingo Grass, and Péter Batáry. Larger pollinators deposit more pollen on stigmas across multiple plant speciesa meta-analysis. *Journal of Applied Ecology*, 58(4):699–707, 2021.

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

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

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

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

Nicholas J Gotelli and Robert K Colwell. Estimating species richness. *Biological diversity: frontiers in measurement and assessment*, 12(39-54):35, 2011.

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

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

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

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

Paul DN Hebert, Sujeevan Ratnasingham, and Jeremy R De Waard. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(suppl_1):S96–S99, 2003.

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

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

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

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

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

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

Joaquín Hortal, Patrícia Garcia-Pereira, and Enrique García-Barros. Butterfly species richness in mainland portugal: predictive models of geographic distribution patterns. *Ecography*, 27(1):68–82, 2004.

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

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

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

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

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

656 Jennifer B Hughes and Jessica J Hellmann. The application of rarefaction techniques to molecular inven-
657 tories of microbial diversity. *Methods in enzymology*, 397:292–308, 2005.

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

660 Stuart H Hurlbert. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*,
661 52(4):577–586, 1971.

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

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

667 Alberto Jiménez-Valverde and Jorge M Lobo. Determining a combined sampling procedure for a reliable
668 estimation of araneidae and thomisidae assemblages (arachnida, araneae). *The Journal of Arachnology*,
669 33(1):33–42, 2005.

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

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

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

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

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

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

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

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

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

Gerardo Lamas, Robert K Robbins, and Donald J Harvey. A preliminary survey of the butterfly fauna of pakitza, parque nacional del manu, peru, with an estimate of its species richness. *Publicaciones del Museo de Historia Natural Universidad Nacional Mayor de San Marcos*, 1991.

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

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

Fredrick J Lee, Kayla I Miller, James B McKinlay, and Irene LG Newton. Differential carbohydrate utilization and organic acid production by honey bee symbionts. *FEMS microbiology ecology*, 94(8): fiy113, 2018.

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

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

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

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

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

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

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

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

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

Claudia E Moreno and Gonzalo Halffter. Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied ecology*, 37(1):149–158, 2000.

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

Susan W Nicolson and Robert W Thornburg. Nectar chemistry. In *Nectaries and nectar*, pages 215–264. Springer, 2007.

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

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

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

Matthew R Olm, Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B Matheus Carnevali, and Jillian F Banfield. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *Msystems*, 5(1):e00731–19, 2020.

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

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

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

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

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

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

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

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

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

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

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

David Ratkowsky. Nonlinear regression modelling. 1983.

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

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

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

Luis M Rodriguez-R, Juan C Castro, Nikos C Kyrpides, James R Cole, James M Tiedje, and Konstantinos T Konstantinidis. How much do rrna gene surveys underestimate extant bacterial diversity? *Applied and Environmental Microbiology*, 84(6):e00014–18, 2018.

Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

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

Howard L Sanders. Marine benthic diversity: a comparative study. *The American Naturalist*, 102(925): 243–282, 1968.

Patrick D Schloss. Reintroducing mothur: 10 years later. *Applied and environmental microbiology*, 86(2): e02343–19, 2020.

Karel Schoonvaere, Guy Smagghe, Frédéric Francis, and Dirk C de Graaf. Study of the metatranscriptome of eight social and solitary wild bee species reveals novel viruses and bee parasites. *Frontiers in microbiology*, 9:177, 2018.

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

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

Jorge SoberónM and Jorge LlorenteB. The use of species accumulation functions for the prediction of species richness. *Conservation biology*, 7(3):480–488, 1993.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jinshui Zheng, Stijn Wittouck, Elisa Salvetti, Charles MAP Franz, Hugh Harris, Paola Mattarelli, Paul W O'Toole, Bruno Pot, Peter Vandamme, Jens Walter, et al. A taxonomic note on the genus lactobacillus: Description of 23 novel genera, emended description of the genus lactobacillus beijerinck 1901, and union of lactobacillaceae and leuconostocaceae. 2020.