Direct shotgun metagenomics captures species abundance of honey samples

Anastasios Galanis, Philippos Vardakas, Martin Reczko, Efthimios M. C. Skoulakis, Georgos Pavlopoulos, Solenn Patalano

B.S.R.C «Alexander Fleming», 34 Fleming Street, 16672 Vari, Greece

# Abstract

(250 words)

# Keywords

(six)

# INTRODUCTION

Honeybees (*Apis mellifera*) are essential pollinators in agricultural landscapes, providing their services to over 70% of modern crops (UNEP, 2010), while also producing honey, beeswax, and beebread, that are often used by humans as rich food sources or raw materials. These properties have rendered honeybees as a vital part of our global economy and placed them among the few insects considered as livestock (ref).

Even though there are some 91 million beehives worldwide, honeybees succumb to anthropogenic and environmental pressures, which synergistically cause colony collapse (Vanengelsdorp et al., 2009; VanEngelsdorp et al., 2010). Due to their economic importance, honeybees are facing intense beekeeping, highlighted by their high densities. However, such high densities are linked to increased pest invasion, such as *Varroa destructor* infestation, and lower wintering survival (Dynes et al., 2019). Hives often surround fields, providing pollination services and thus increasing crop yield (Gaines-Day & Gratton, 2016), where they are exposed to various degrees of herbicide and insecticide use and lack of plant biodiversity due to monoculture cultivation. Importantly, herbicides and insecticides are known to affect honeybee survival and cognition (Chmiel et al., 2020; Farina et al., 2019; Williamson et al., 2014), while monocultures induce nutritional stress that presents an additional challenge to colony survival(Naug, 2009). Lastly, the current climate crisis is predicted to result in higher pesticide use (Delcour & Spanoghe, 2015; Matzrafi, 2019) and higher pathogen spread (such as *Nosema ceranae*)(Gisder et al., 2010), which will intensify the pressures on honeybee health and colony survival through direct effects as well as indirect due to ecosystem disturbances.

Honeybee health has recently emerged as a focal research point and measurement indices are being developed (EFSA Panel on Animal Health and Welfare, 2016; European Food Safety Authority, 2014; Gilioli et al., 2019; López-Uribe et al., 2020). To this end, honey-derived DNA is often used in order to evaluate the health status of honeybee colonies. Honey is a preferred sample for biomonitoring molecular methods, because it is produced through the regurgitation (inversion) of flower nectar, which is subsequently placed in the comb until enough water evaporates. Throughout this procedure the nectar (and honey as an extension) comes in contact with a variety of organisms and, therefore, contains DNA, termed environmental DNA (eDNA), from the flower origin, the honeybee gut microbiome, and hive organisms such as honeybees and hive pests.

Currently, biomonitoring methods rely on the molecular identification of a few specific, known pathogens and the evaluation of foraging behaviour through visual inspection of pollen grains contained in honey (melissopalynology) or honey-derived DNA amplicon sequencing of marker genes, such as *rbcL*, *matK,* ITS1 and their combinations. Conventional methods have already provided a wealth of information regarding seasonal trends of honeybee pathogens (Cornman et al., 2012; D’Alvise et al., 2019) and have helped to describe the foraging sources exploited by honeybees (Hawkins et al., 2015). However, current molecular methods are expensive, time consuming, and targeted, which presents a lost opportunity for novel discoveries. These in conjunction with the multifactorial nature of colony collapse, necessitate the development of methods that can capture the complexity of the ecological and physiological interactions from which honeybee health emerges.

Consequently, current apicultural practices are usually not evaluated on the basis of long-term honeybee health, but primarily emphasise on productivity and short-term colony stability during the productive period, usually during late Spring and Summer. Therefore, a methodology that can holistically describe the ecological niche that honeybees occupy can provide a unique biomonitoring tool for honeybee health as well as an avenue into species interactions that modulate it.

Shotgun metagenomics of honey samples is an attractive methodology to tackle this challenge, due to the increase in sequenced organisms and its unbiased nature. Honey shotgun metagenomics has only recently been applied(Utzeri et al., 2018), but has been demonstrated that it can describe the complexity of organisms that honeybees come in contact with as well as the floral origin of honey (Bovo et al., 2020). However, these previous studies have used 40-50 g of honey and a DNA extraction procedure requiring more than 4 hours to complete. These two aspects place a limit in the throughput of this methodology. Moreover, the use of materials and equipment that cannot easily be transported renders the methodology unable to be applied in the field and inaccessible to laboratories without the required equipment.

Here, we report the development of an alternative shotgun metagenomics methodology, termed direct shotgun metagenomics, coupled to a computationally validated pipeline for data analysis. Shortly, direct shotgun metagenomics makes use of 5 g of honey and DNA extraction can be completed in 30-45 minutes, with materials and equipment commonly found in laboratories. We show that direct shotgun metagenomics can describe honey samples similarly to the previously reported protocols. Through field visits and databases, we validated the presence of plants identified through our method and we provide evidence that this method can be used for monitoring of *Varroa destructor* infestation*,* a major threat to colony health. Our data shows that direct shotgun metagenomics can be used as a holistic biomonitoring method to gain insights into the ecological niche of honeybees.

* Honeybee are under strong environmental pressure (intense beekeeping, monoculture, pesticide..)
* Beekeeping sector require more monitoring of honeybee behaviour (pollination pattern and diversity) and how it related with bee health (microbiote, varroa…). However conventional methods are expensive, time consuming and most of the time targeted (Rbcl…)
* eDNA metabarcoding (Taxonomic identification of multiple species extracted from a mixed sample) complement or improved conventional methods. But current eDNA technics show also limitation due to the bias bring from DNA extraction, library preparation and sequencing coverage.
* Honey has been used as biomonitoring tools. Explain how honey is produced by bee (inversion). Contain plant but also bacteria and probably everything in a hive (ref of all stuff done to analysis honey). Too many methods, we want to develop a unique one = holistic.

AIM: develop a simple methodology and bioinformatic pipeline that can assess both foraging pattern and bee health through DNA honey content.

# METHODS

## Apiary setup and monitoring

An apiary was installed on the property surrounding "Alexander Fleming" research Centre in Vari (Greece, Attika region) in December 2018 (GPS coordinates: 37◦49’28.2" N 23◦47’25.7" E, Figure 1A). The colonies contained sister queens of the species *Apis meliferra macedonica,* all colonies were started 9 months earlier. Hives population and their degree of Varroa infestation was monitored at least once a month throughout the entire year 2019 (Sup. Figure 1). Honeybee population was estimated by measuring their coverage on each side of the frame and scoring them from 1 to 10 (10 corresponding to maximum bee coverage with no space between the bees). The absolute population was calculated based on the estimation of a complete frame coverage corresponds to 2000 honeybees (Delaplane et al., 2013). Varroa monitoring was initially performed using a technique of coating approximately 300 bees with icing sugar followed by a visual counting of the Varroa detached from the back of the bees following this treatment. Because of the approximation and invasiveness of this method, a wooden drawer was installed at the bottom of each hive from April 2019 in order to monitor the natural fall of Varroa mites without having to open the hives. To prevent fallen Varroa from escaping or returning to the hive, this monitoring was optimized by covering the drawers with olive oil on baking paper. The degree of Varroa infestation of a hive was then normalized per day. To compare the degrees of infestation throughout the year, the data obtained from the icing method were normalized against the bee population. Weather conditions around the apiary were monitored using data from a weather station at PALLADION School in Vari, Attiki located 2.2 km away from the apiary (coordinates: 37◦50’22" N, 23◦48’23"E). The temperature, wind speed, and rain precipitation across 2019 are display in (Sup. Figure 1).

## Honey collection and extraction

Honey was collected from four different hives in 2019 across 3 different seasons: spring (Hive 5), summer (Hives 6 and 7) and autumn (Hive 4). Collected frames from the hives were cut in small pieces (approximately 10-15 cm in length and 5 cm in width) and placed on top of a fine sieve placed on a glass bowl. This apparatus was placed in an incubator at 37◦C for approximately 16 hours. This process allowed fresh honey (uncapped) to flow naturally. After 16 hours the honey was collected and poured into clean glass jars, which were labelled and placed inside a drawer and kept at room temperature (RT: 18-24◦C) until further processing.

## DNA extraction

### Shotgun Metagenomics (SM) extraction

DNA was isolated from honey of the 4 hives similarly to (de Vere et al., 2017) with some modifications. For each hive, forty grams of honey were divided between two 50mL Falcon tube and filled with sterile distilled water up to 30mL. Tubes were incubated in a water bath at 65◦C for 30 mins, briefly shaken every 5 mins to ensure homogenisation, and ultra-centrifuged for 30 mins at 15,000 RPM using the SW50.2Ti rotor (Beckman Optima L-90K Ultracentrifuge). The supernatant was discarded and the pellets were pooled in 400μL of Buffer AP1 from the Qiagen DNeasy Plant Mini Kit (Qiagen). The mixture was homogenised progressively using the CAT X210 homogeniser for 40 seconds avoiding the formation of foam. 80μL of Proteinase K (1 mg/mL, Sigma) were added to the mixture and incubated for 50 mins at 65◦C. During the incubation the tube was further inverted a few times every 15 mins. 4μL of RNase A stock solution (100 mg/mL, from Qiagen DNeasy Plant Mini Kit) were added, the tube was briefly vortexed and incubated for 10 mins at 65◦C. Steps 3 through 12 were then followed according to the DNeasy Plant Mini Kit (Qiagen), except for the following changes: At step 1, the mixture was spun for 1 min. At step 11 and 12, elution was done with 25μL of Buffer AE and DNA concentration measured by a Nanodrop spectrophotometer. 400 ng of DNA were sonicated in a total of 50μL of Buffer AE. The solution was transferred to a covaris tube and sonicated (temperature 7◦C, 120 seconds treatment, [max] intensity, [max] factor, and [max] cycle per burst).

### Direct Shotgun Metagenomics (Direct SM)

For each hive, five grams of honey were placed into a 15 mL Falcon tube. The tube was filled with sterile distilled water up to 10 mL and incubated in a hot water bath for 10 min, briefly shaken every 3 mins to ensure homogenisation. Several successive centrifugations at 14,000 x g for 3 min in a microcentrifuge (Eppendorf 5417C with rotor FA453011) allowed to collect a single pellet in a 1.5 mL Eppendorf tube. The pellet was then dissolved in 200μL of 0.1 M NaOH, 5%Tween-20. The tube was vortexed for 30 seconds and incubated at RT for 15 min. The denaturation of the DNA was then quenched with 200μL 0.5 M Tris-HCl, 5 mM EDTA. This is further referred as the extraction mixture. Subsequently, the DNA was purified using Agencourt AMPure XP beads (Beckman Coulter). Briefly, 200μL of beads were mixed with 100μL (at a ratio of 2) of the extraction mixture and incubated for 15 mins at RT. The tube was then placed on a magnetic rack and the supernatant to be discarded. The beads were washed with freshly-prepared 70% ethanol and dried under a desk lamp for approximately 5 mins. The pellet was then resuspended in 20μL of Buffer AE and incubated at RT for 15 mins. Finally, the tube was placed on the magnet and the supernatant was transferred into a new tube for storage. Both the extraction mixture and the purified DNA were snap frozen and stored at -80◦C. Note that the DNA for direct shotgun metagenomics was not sonicated.

## Libraries preparation

A total of 8 DNA libraries (4 from the SM and 4 from the Direct SM) were build using the Ion Plus Fragment Library Kit protocol (Thermo Fisher Scientific) with the following modifications: 5 to 10 ng of DNA was diluted respectively with sterile distilled water to a final volume of 39μL. DNA was End-repaired by the addition of 10μL of End-repair buffer and 1μL enzyme per sample followed by 30min incubation at RT. Samples were purified using the AMPure XP beads (at a ratio of 1.9) and after subsequent washes with 70% ethanol, they were eluted in 20μL Low TE. Adaptors were then ligated to the DNA in the presence of 5μL ligase buffer and 1μL ligase enzyme. 1μL of universal IonXpressP1 adaptor was added in all samples with 1μL of a barcoded IonXpress adaptor, 1-96. The reaction was diluted in ddH2O to a final volume of 50μL and incubated for 30 min at RT. After a further purification with Agencourt AMPure XP beads (at a ratio of 1.5) and an elution in 17.5μL of Low TE, samples were amplified using 50μL Platinum PCR Supermix High Fidelity and 2.5μL Library amplification primer mix for 17 cycles (thermal cycling protocol: 72◦C-20’/95◦C-5’/ (97◦C -15”,60◦C-15”,70◦C-1’)\*17cycles/70◦C-5’). A final 2 step purification was performed by adding 30μL ddH2O to the 70μL reaction, purified with AMPure XP beads (at a ratio of 0.8 to remove any fragments of 400bp or more) and eluted in 20μL ddH2O. The supernatant was used for a second purification using Agencourt AMPure XP beads (at a ratio of 0.5, total ratio 1.3 of initial). Each library’s quality and quantity were assessed through a bioanalyzer using the DNA High Sensitivity Kit reagents and protocol (Agilent Technologies) (Sup. Figure 2). The quantified libraries were pooled together in 12plex with other libraries, at a final concentration of 7 pM. The pools were then processed, templated and enriched on an Ion Proton One Touch system. Templating was performed, using the Ion PITMHi-QTMOT2 200 Kit (Thermo Fisher Scientific) and sequencing, with the Ion PITMHi-QTMSequencing 200 Kit and the Ion Proton PITMV2 chips (Thermo Fisher Scientific) on a IonProtonTM System from (Rothberg et al., 2011) according to commercially available protocols.

## Mock (simulated) samples

Mock samples were created to evaluate various steps and parameters of the computational pipeline. A phylogenetic tree and details regarding the genome size and GC content of each genome, mitochondrial, chloroplast or plasmid DNA sequences used to create the mocks are displayed in (Sup. Figure 3 and Appendix 1). All mocks were created using FASTQsim (add reference) and a relative abundance (depth) was attributed to each species. First, 3 mock samples were created, using the custom IonTorrent parameters in FASTQsim, to evaluate the different taxonomic classification tools (see section below). The first mock sample contained only Viridiplantae (11 plants and 1 green alga). The second mock sample contained 11 organisms: 4 non-Viridiplantae Eukaryota, 6 Bacteria, and 1 virus. The third mock sample, was created by merging the previous 2 mock samples. Then, additional mock samples were prepared to address whether the taxonomic classification is influenced by 1) the number of species in each mock; 2) the library size; 3) the fragment size of the reads.

## Taxonomic classifiers and genomic aligners

The performance of five different computational tools for sequencing read taxonomic classification was evaluated. A brief description of the classifier and used parameters are described below.

### CCMetagen

CC Metagen is a metagenomic classification pipeline which uses the KMA software for read mapping and alignment (Marcelino et al., 2019). KMA works in five steps: trimming of reads, heuristic \(k\)-mer mapping, fine alignment, ConClave scoring, and reference assembly First the reads are trimmed. Then, \(k\)-mers are mapped against the template database. The mapping \(k\)-mers are used to produce a score \([(k + 1)^{-1}]\), where mapping \(k\) mers are rewarded with a score of \(k\) and mismatches scored with -1. For a sequence to be proceed for alignment, a positive mapping score should be achieved. In this case, alignment is restricted to the template sequences the \(k\)-mers matched to. KMA uses a hash map of indexed \(k\)-mers in order to start alignment. To enable a high resolution of gaps and mismatches, KMA uses the Needleman-Wunsch algorithm. For the ConClave scoring, the alignment score for each template sequence is summed reflecting the maximum alignment score for each template. Using the ConClave score the most likely template can be chosen. Finally, ConClave alignment and scoring allows for assembly guided by a reference, resulting in a consensus sequence for a template. CCMetagen was run using the pre-indexed nt database from 2018 (downloaded from: <https://researchdata.ands.org.au/indexed-reference-databases-kma-ccmetagen/1371207>)

### DIAMOND

DIAMOND compares a file of DNA sequences against a protein database, such as the non-redundant database of NCBI (nr-NCBI) (Buchfink et al., 2015). It follows a seed and extend approach. Query sequences are mapped against a reference on which a seed (or minimiser) sequence can be found. This is then extended as far as possible. DIAMOND uses a reduced alphabet consisting of 11 letters (each bracket instance denotes 1 DIAMOND alphabet letter): [KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA]. Furthermore, DIAMOND uses double indexing of both the reference and the input queries which allows for the identification of all matches between the queries and the reference.

### kraken2

Kraken2 is an exact k-mer-based approach to the classification of metagenomic reads (Wood et al., 2019). Kraken2 uses a compact probabilistic hash table, which maps minimizers (of length l < k) to the lowest common ancestor (LCA). This results in minimizers triggering reading of the hash table instead of k-mer, which results in faster processing times (than Kraken). Kraken2 was run using an updated version of the nt database (May 2020) with the default settings.

### MG-RAST

MG-RAST is a webserver accessible through https://www.mg-rast.org. It is a pipeline consisting of approximately 22 tools which offer quality control, classification, annotation, and functional annotation of reads (Keegan et al., 2016).

### minimap2

Minimap2 uses a typical seed-chain-align procedure frequently used by other genomic aligners (Li, 2018). It creates a hash table which contains minimisers of reference sequences (e.g. the target database). For each query sequence, query minimisers are taken as seeds and minimap2 finds exact matches (anchors) to the reference, and identifies colinear anchors as chains.

## Filtering and normalisation of the libraries

### Sequence pre-processing

### (Adaptors, short reads, low abundance reads…)

### Filters based on Kraken 2 confidence score and usual laboratory cross -contaminants species

A perfect hit using Kraken2 classifier will have all the k-mers mapped. In case of one mismatch it will impact all k-mers that contain this nucleotide. The confidence score represents the number of matching k-mers divided by the total number of k-mers. To test the best confidence threshold to apply to our libraries, an extended simulated honey sample containing 64 species was evaluated (see Section 2.5 and Phylogenetic tree in Sup. Figure 3). A filtering threshold of 0.5 was selected,showing the minimal errors, high sensitivity and recall for all species of the mock samples with limited reads lost (Sup. Figure 4).

The filtered kraken2 files generated after the process of the 8 sequenced libraries were imported in R. The taxonomic attribution for each species was obtained using the *taxonomizr* package. Reads assigned to Phylum Chordata were removed as well as the family *Drosophilidae*, genus *Drosophila*, and species *Drosophila melanogaster* as likely to come from laboratory cross-contamination.

### Read normalisation across libraries

DESeq2 version v3.11 was installed from Bioconductor. Read normalisation across libraries was done using the *estimateSizeFactors* function, which applies the relative-log expression (RLE) transformation (REF DESeq2 for metagenomic normalisation).

### Compositional analysis (CoDaSeq)

Due to the nature of the data (compositional), a compositional approach was also evaluated (Gloor et al., 2017). The analysis was done by applying the centred-log transformation (clr) to perform clustering and PCA analysis.

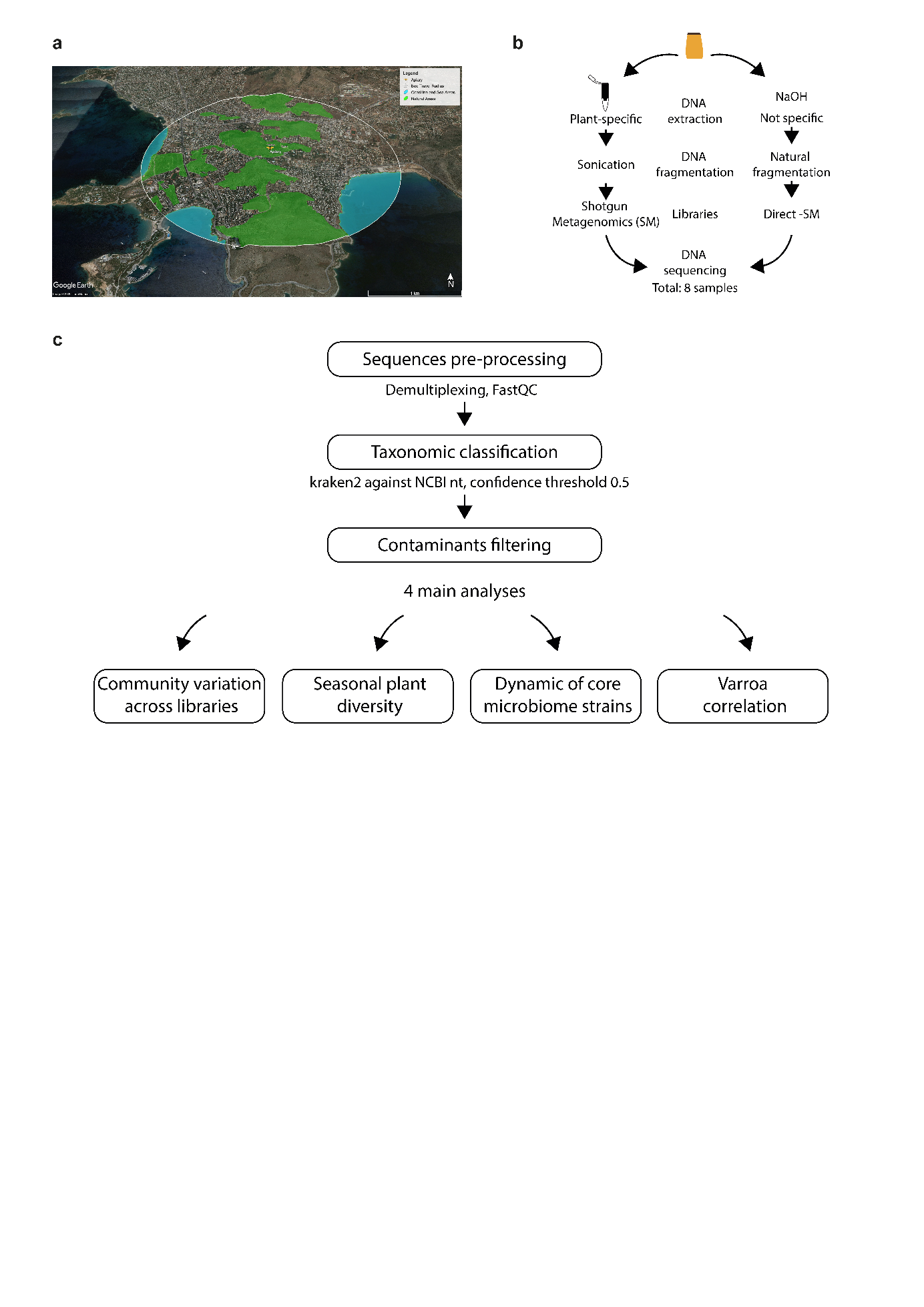
## Statistical analysis and R packages

The analysis and graphical visualisation of the results were conducted in R studio version 1.2.5033 running R version 3.6.3 on a x64 system running Microsoft Windows 10 Pro. System.

# RESULTS

## Experimental design, honey DNA extraction methods and overview of the bioinformatic workflow

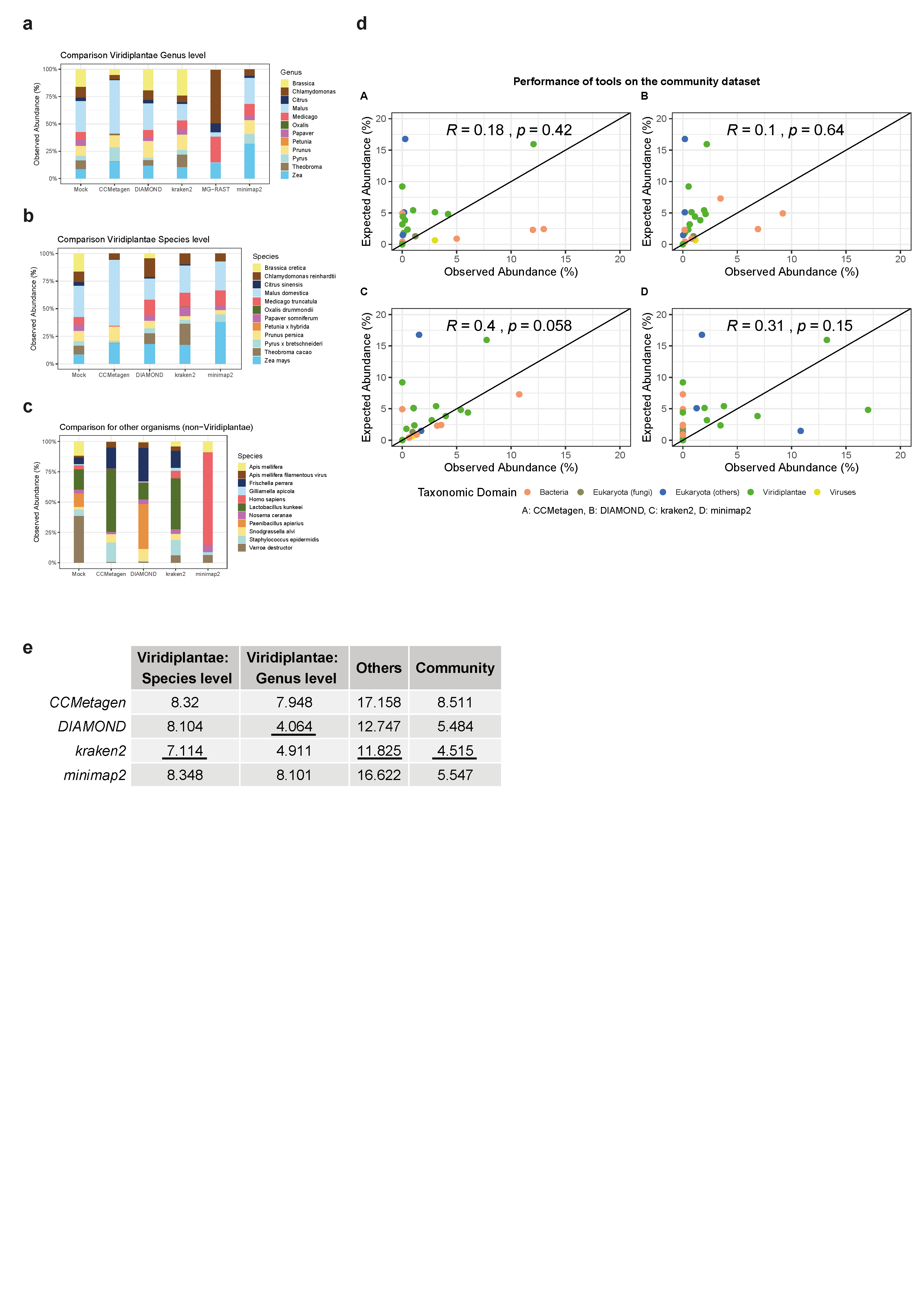
Our study took place in an apiary whose ecosystem in a 2 km radius (12.5km2) is typical of a coastal semi-arid Mediterranean climate and is mainly composed of conifers, pines, dense shrubs and olive trees (31%), as well as urban areas containing many ornamental gardens (59%), coastline and sea area (10%). It is located on the edge of the Hymettus mountain on the suburb of Athens in Greece (Figure 1a). The 4 hives of this study showed comparable evolution of both their honeybee population growth and degree of Varroa infestation during the year (Sup. Figure 1).

Three honey collections were carried out, representative of the main beekeeper harvest seasons in spring, summer and autumn. For each harvest, two DNA extraction techniques were compared (Figure 1b). The first follows a classic metagenomic shotgun (SM) protocol consisting in extracting DNA from 40g of honey and using specific extraction columns followed by a sonication step in order to fragment the DNA to a given size before the preparation of a sequencing library. While this technique has shown encouraging results in other studies (Bovo et al., 2018; Kek et al., 2017), it nevertheless brings biases that reduce the qualitative and quantitative analyses. Indeed, the DNA contained in honey is already very largely naturally fragmented (Sup. Figure 2) and the biases brought by the use of specific purification columns limit the overall analysis of the diversity of eDNAs. In order to limit these biases, a second more direct technique (Direct-SM) has been developed from 5g of honey which does not involve any specific purification column and takes advantage of the natural and experimentally induced variability of the DNA fragment extracted from honey. A total of 8 libraries were prepared and sequenced (4 technical replicates of the Direct-SM and 4 technical replicated of the SM). Analysis of the sequencing results followed the bioinformatic workflow (Figure 1c).

**Figure 1**

﻿

## Evaluation of genomic aligners and taxonomic classifiers

In order to choose the most appropriate tool for the analysis of shotgun metagenomics data from our honey samples, we evaluated several different taxonomic classifiers using mock (simulated) samples of various species communities (See Methods section). The primary reasons to apply shotgun metagenomics in honey samples were to study the foraging behaviour and health of *Apis mellifera*. Therefore, we built 3 different communities of mock samples, which included 12 known pollinated plants species (Viridiplantae community), other 11 non-plants Eukaryota, Bacteria and virus species such as *Apis mellifera*, the putative pathogens *Varroa destructor*, *Apis mellifera filamentous virus* and *Nosema ceran*a and some symbiotic bacteria (non-Viridiplantae community). Finally, we also built a simulated honey sample containing all 23 species from both Viridiplantae and non-Viridiplantae communities ****(Honey community).

**Figure 2**

Our results show that, when evaluated at the genus level, DIAMOND followed by kraken2 classifiers produced the closest distribution of the Viridiplantae community to the one expected where respectively XX% and XX% of the classified reads aligned to the expected genus (Sup. Figure 3), while the other computational tools show strong biased toward certain genus with less that XX% of the reads correctly aligning to the expected species. However, when these classifiers are evaluated at the species level, kraken2 shows a greater performance than DIAMOND using both Viridiplantae and non-Viridiplantae communities (Figures 2a-b). Interestingly, DIAMOND is the only classifier to assign reads to the *Brassica cretica* species when none of the others tools could. MG-RAST has not been applied at the species level as it only provided genus-level resolution. Finally, when using the simulated honey community, kraken2 is the only classifier with a significant correlation between the expected abundances and the observed abundances (Figure 2c). The calculation of the mean square error (RMSE) confirmed Kraken2 reliability performance (Figure 2d). Indeed, Kraken2 having the lowest RMSE - with the exception of the evaluation of Viridiplantae at the genus level -was chosen for the rest of our analyses.

## Comparison between library preparation

Between 1.4 and 8 million reads were obtained after the sequencing of the 8 honey samples (Table 1). After their taxonomic classification with Kraken2 - with a threshold of 0.5 and additional filtering for the putative laboratory cross-contaminants (see Methods and Sup. Figure 4) - 50,000 to 600,000 reads per library were used for downstream analyses. On average, 85% of the reads per libraries were classified at the species level, 9% at the genus level and less than 4% at the family level.

The absolute abundance of species identified amongst the 8 honey samples can increase three-fold depending on the samples (from 265 to 663 species, Figure 3a). However once classified into domains, we found a remarkable conservation of species heterogeneity. Most of the species felt under the domains of Bacteria (53±6.7%), then from Viridiplantae (31 ±6%), Eukaryota (14 ±2.4%) and a minority from Virus (2.1 ±0.5 %). Overall, no significant difference of the species distribution was found when the two technics were compared (SM versus Direct-SM, Two-way RM ANOVA, p=0.5985) neither when domains were compared individually (SM versus Direct-SM, Multiple t-tests, p>0.1). These observations validate the Direct-SM sequencing technique but also show that the bias expected by the use of purification columns and sonication is not significant at this level.

To further verify the quality of our Direct-SM approach, Principal Component Analysis (PCA) and Hierarchical Clustering (HC) were performed. Initially, sequenced data were normalised across all libraries and compared at family, genus and species levels (Sup. Figure 5). PCA revealed that the two replicas collected per hive grouped with high concordance at the family level but show limitation at the species (Figure 3b-c). Instead,compositional approach has show greater concordance using species level [EXTEND]

* DESeq-2 approach: HC and Heatmap of 30 top species and heatmap of statistically variable species across technical replicate

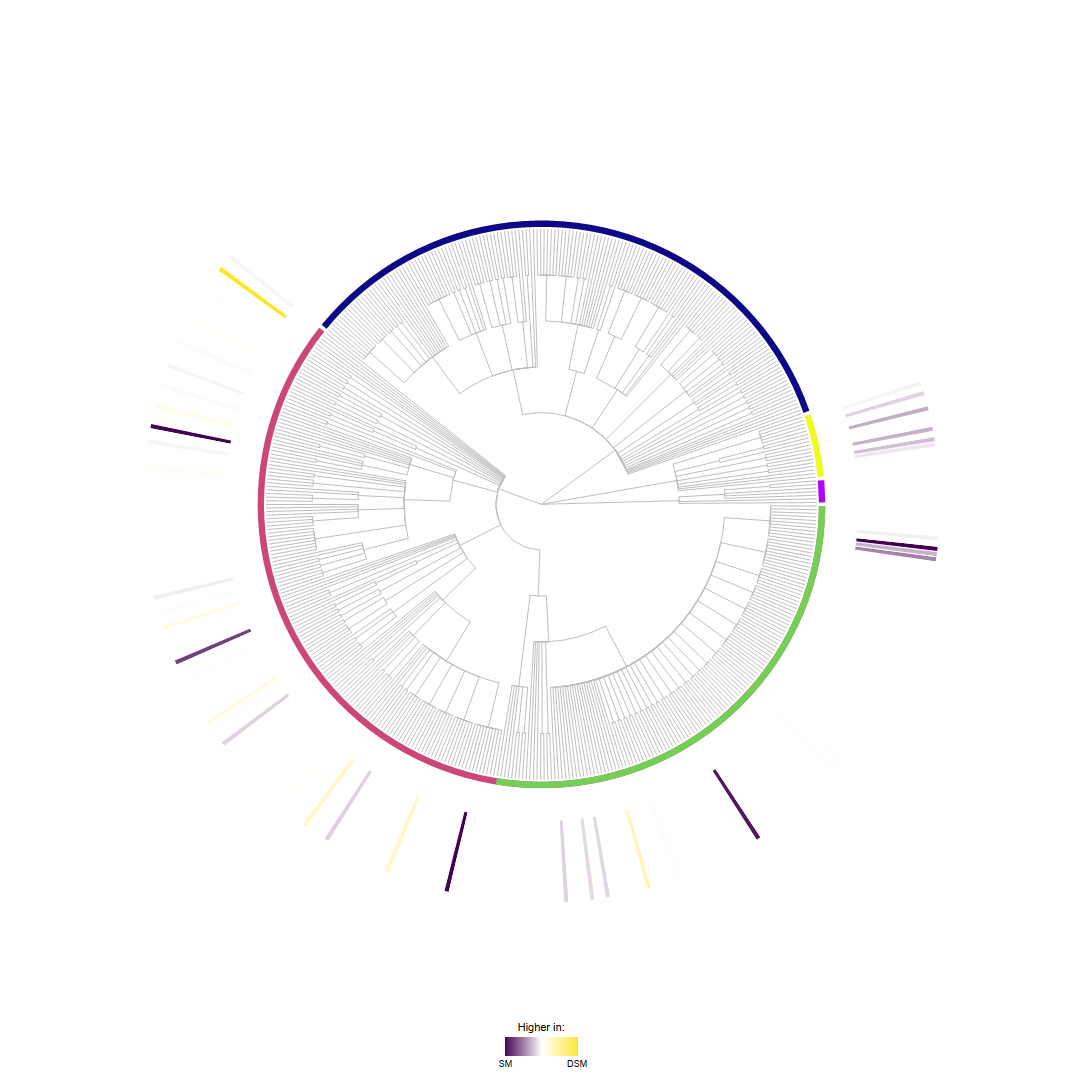
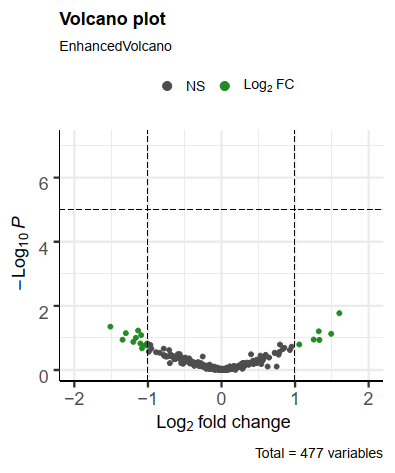
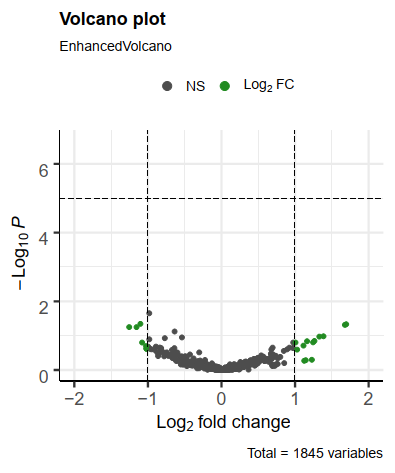
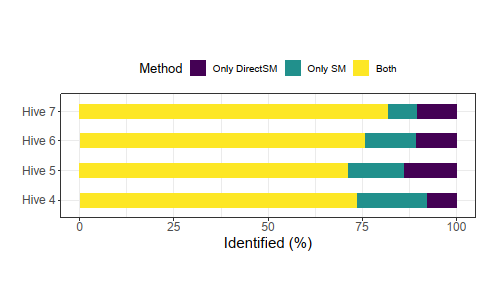
****

Figure Familes that differ between SM and DSM.

****

**Figure: Left: Volcano plot family. Right: Volcano plot species.**

****

**Fig: Number of species shared between DirectSM and SM**

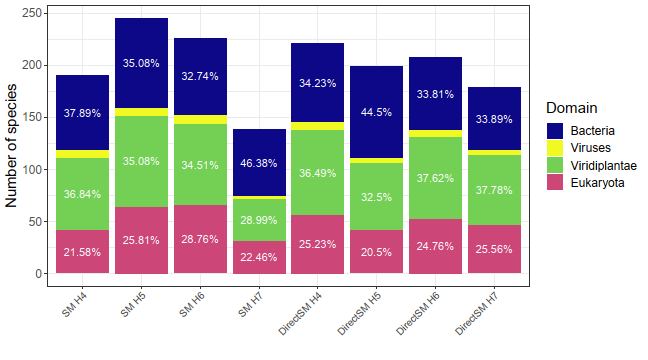
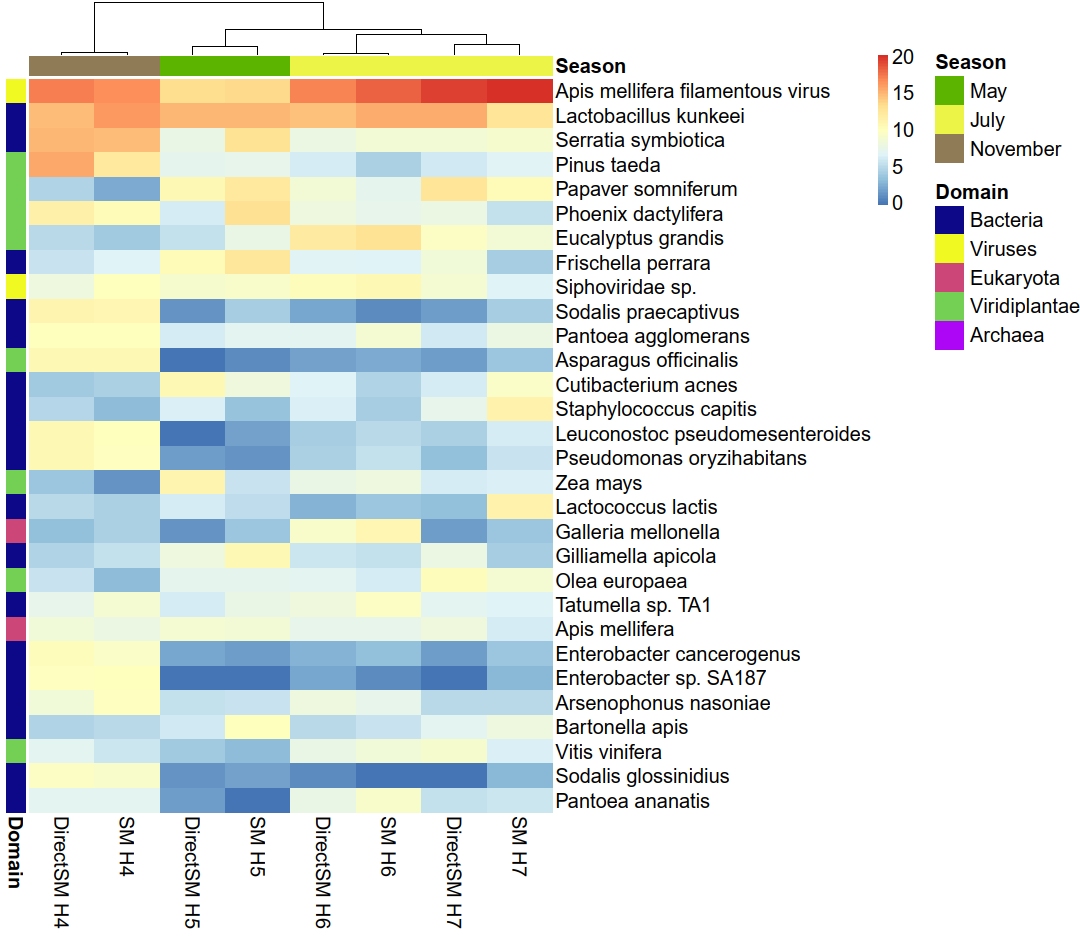
****

Figure: Barplot species (%) in each library

Fig: Top 30 species

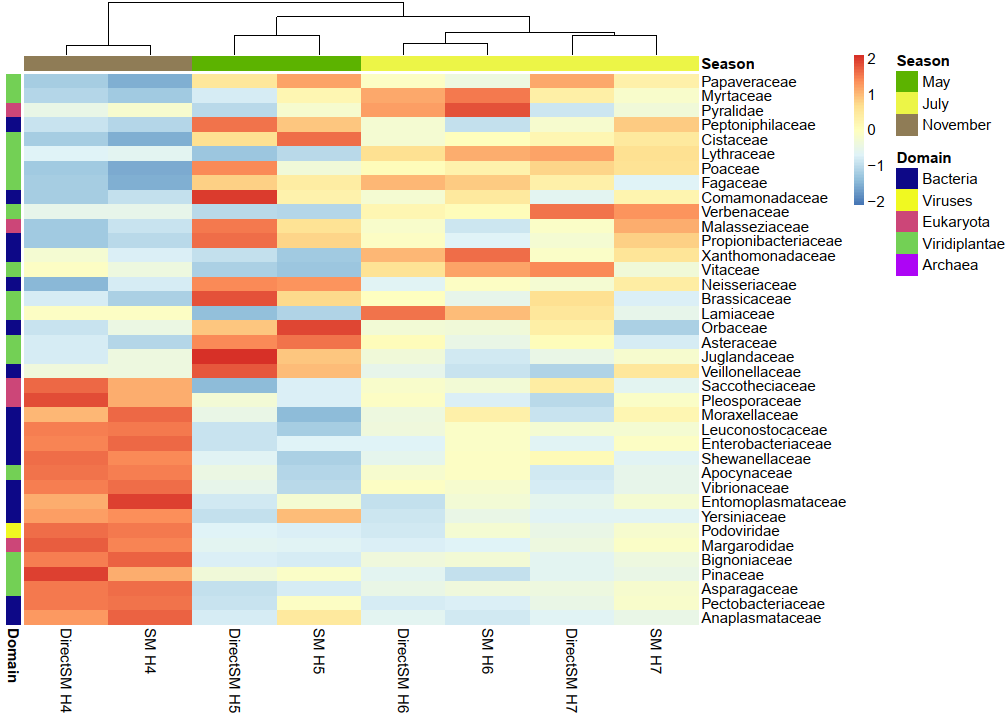


Fig: Significantly abundant families (log2 fold change) per Season.

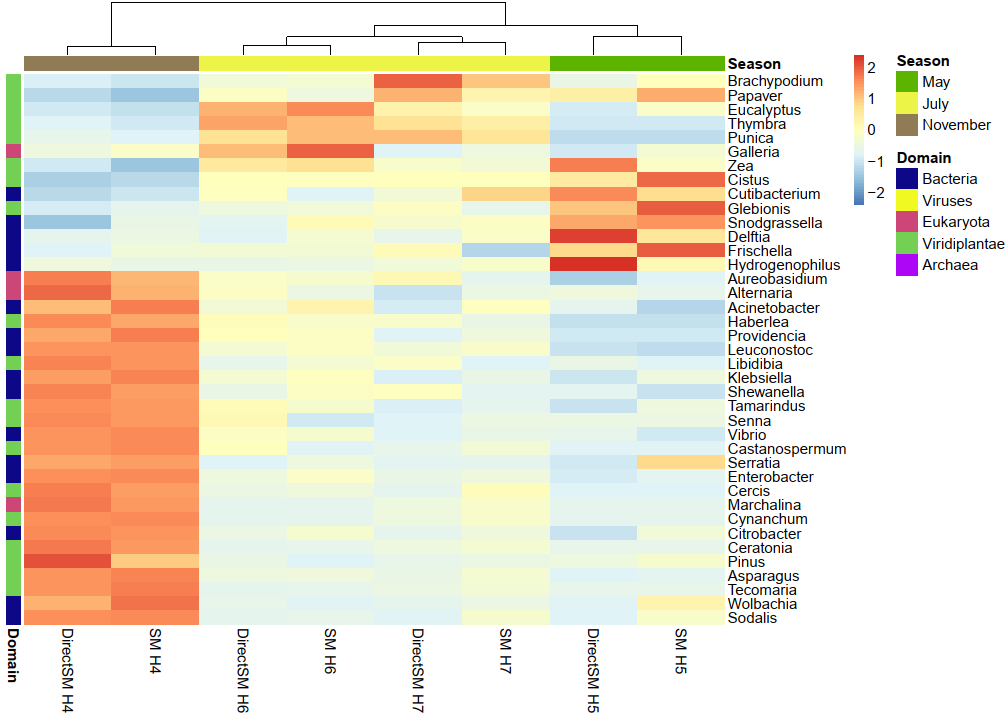


Fig: Significantly abundant genera (log2 fold change) per Season.

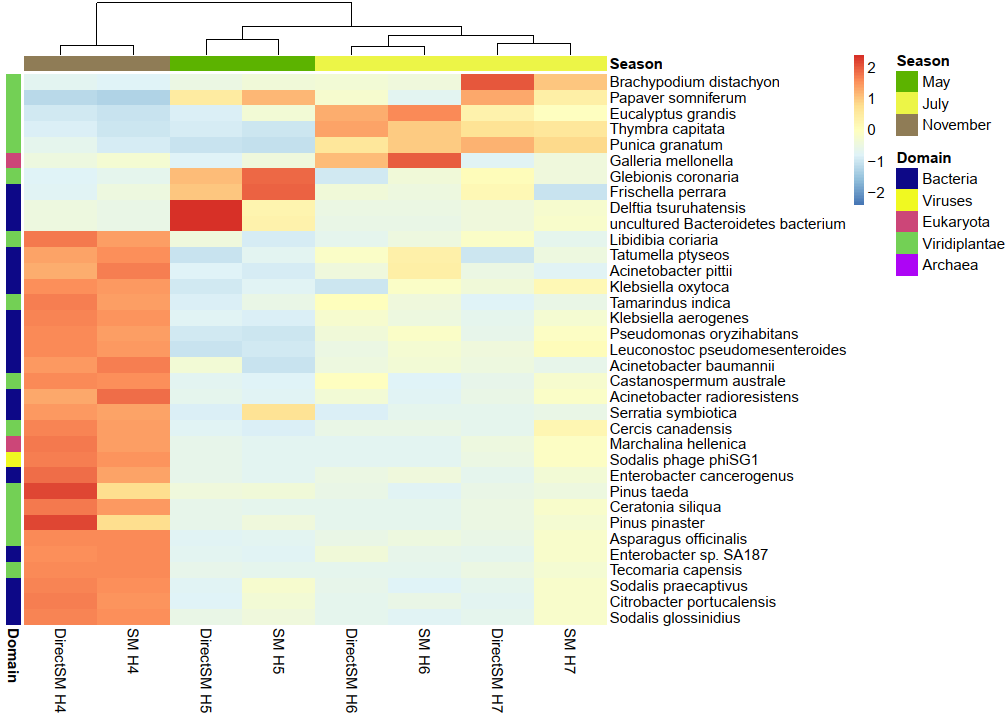


Fig: Significantly abundant species (log2 fold change) per Season.

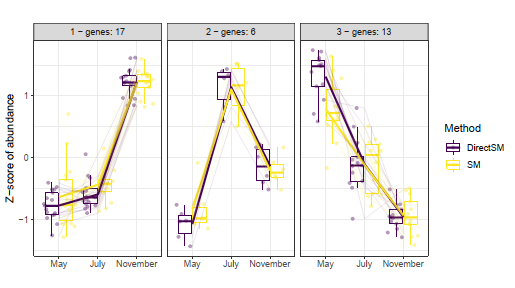


Fig: Cluster of families exhibiting same trends in abundance

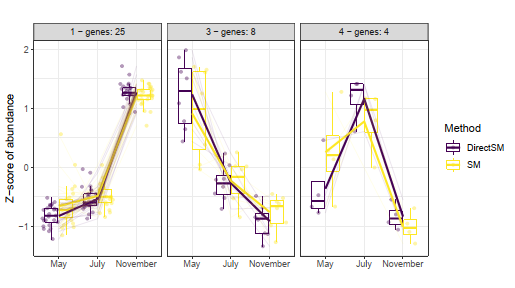


Fig: Cluster of genera exhibiting same trends in abundance

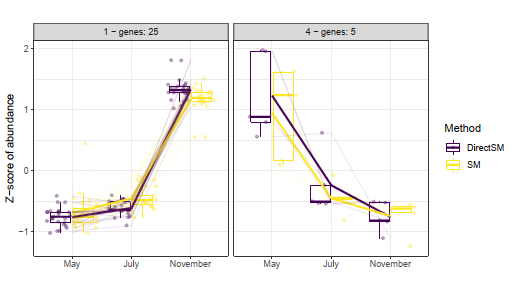


Fig: Cluster of species exhibiting same trends in abundance

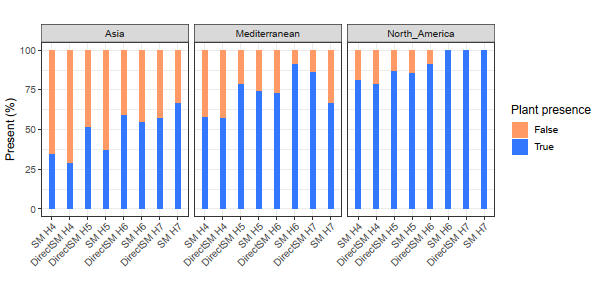


Fig: Plant presence in different locations. SAME latitude for all, DIFFERENT longitude. Note: North America has the HIGHEST number of total observations in the database and it may be why we see most of them there. Alternatively, I can also use same longitude and different latitude (e.g. Africa)

Conclusion: (validation of our direct-SM technic: strong metagenomic signature by hive with also seasonal clustering)

## Plant diversity and variation of foraging profiles through the seasons

### Plant diversity around the apiary

* 382 absolute number of plants species collected in the apiary (calculate taxa diversity / richness (see De Vere et al.,)? )
* Distribution of plant for each librairies: dotplots (Figure 4A)
* Validation using Mediterranean plant reference library (Figure 4B):
* Validation by botanic collection and identification (Greek flora database) (Appendix 2)
* Validation by gyroscopic analysis (in process)

Conclusion: (XX % of plants collected in the honey were validated at the genus level. However, some limitations were found for species identification).

**Figure 4**

### Foraging profiles through the season

Because our technics show no significant variation, we could use them as technical replicate for the foraging profiles through the season

* Most abundant plants per season and across seasons: are they correspond to flowering profil? (Figure 4A, 4C) (any colonie specificity? As show in Devere)
* Quantitative validation using specific genome alignments (Martin)

Conclusion:

## Insight into the bee health through honey

### Overall

* Increase species identification using GJI database (Pavlopoulos)
* Proportion of non-plant DNA in honey, pathogen versus symbiotic (meta-analysis from Pavlopoulos)

### Microbiote

* Description of the main 10 core and non-core phylotypes and their dynamic (see Engel publications)
* Validate consistency and variability with previous publications (Any way to make nice comparison? Pavlopoulos)Functional diversity (Pavloopulous
* (Validation using specific genome alignment (as for plants) )

Conclusion: we found all mained CORE and 2 non-CORE family

### Varroa

* Varroa natural fall counting correlation with read count (Figure 5C): better correlation with DSM

# DISCUSSION

* Summary : 1/ Improvement of library preparation bring greater quantitative accuracy, environmental impact of the method. 2/Build new standard pipeline for honey analysis, 3/Assess quantitatively the DNA honey diversity to understand foraging behaviour and key variable of bee health (no killing)
* Discuss the limitation of species identification and non-used reads: Need more reference genomes to accurately identified the species.
* Discussion about Microbiote and virus. Justify the use of genus level because others study show strong strain divergence (Ellegaard & Engel, 2019)
* Varroa underestimation with Kraken2,
* Bee health: integration of multiple factor across level of biological organisation (López-Uribe, Ricigliano, & Simone-Finstrom, 2020). (Some focus of Frishella and Sodalis in winter).
* Bring information for future applied and fundamental research on bee adaptation to new environment
* Future: monitor to adjust variables to improve the health of pollinators and to determine if there has been a change (manipulation)

# Acknowledgements

Sequencing facility (Pantelis and Vagelis), MCSA, other funds?

# References

Bovo, S., Ribani, A., Utzeri, V. J., Schiavo, G., Bertolini, F., & Fontanesi, L. (2018). 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). doi: 10.1371/journal.pone.0205575

Buchfink, B., Xie, C., & Huson, D. H. (2014, January 1). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, Vol. 12, pp. 59–60. doi: 10.1038/nmeth.3176

De Vere, N., Jones, L. E., Gilmore, T., Moscrop, J., Lowe, A., Smith, D., … Ford, C. R. (2017). Using DNA metabarcoding to investigate honey bee foraging reveals limited flower use despite high floral availability. *Scientific Reports*, *7*(42838). doi: 10.1038/srep42838

Delaplane, K. S., Steen, J. Van Der, & Guzman-novoa, E. (2013). *Standard methods for estimating strength parameters of Apis mellifera colonies Métodos estándar para estimar parámetros sobre la fortaleza de las colonias de Apis mellifera*. *I*(1), 1–12. doi: 10.3896/IBRA.1.52.1.03

Ellegaard, K. M., & Engel, P. (2019). Genomic diversity landscape of the honey bee gut microbiota. *Nature Communications*, *10*(1). doi: 10.1038/s41467-019-08303-0

Gloor, G. (2016). CoDaSeq: Analyzing HTS using compositional data analysis. *F1000Research*, *5*. doi: 10.7490/F1000RESEARCH.1112250.1

Keegan, K. P., Glass, E. M., & Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. In *Methods in Molecular Biology* (Vol. 1399, pp. 207–233). doi: 10.1007/978-1-4939-3369-3\_13

Kek, S. P., Chin, N. L., Tan, S. W., Yusof, Y. A., & Chua, L. S. (2017). Molecular identification of honey entomological origin based on bee mitochondrial 16S rRNA and COI gene sequences. *Food Control*, *78*, 150–159. doi: 10.1016/j.foodcont.2017.02.025

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. doi: 10.1093/bioinformatics/bty191

López-Uribe, M. M., Ricigliano, V. A., & Simone-Finstrom, M. (2020). Defining Pollinator Health: A Holistic Approach Based on Ecological, Genetic, and Physiological Factors. *Annual Review of Animal Biosciences*, *8*(1), 269–294. doi: 10.1146/annurev-animal-020518-115045

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., … Holmes, E. C. (2020). CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 103. doi: 10.1186/s13059-020-02014-2

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. doi: 10.1038/nature10242

Shcherbina, A. (2014). FASTQSim: Platform-independent data characterization and in silico read generation for NGS datasets. *BMC Research Notes*, *7*(1), 533. doi: 10.1186/1756-0500-7-533

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. doi: 10.1186/s13059-019-1891-0

# Data Accessibility Statement

# Author Contributions

# Tables and Figures

# Tables

**Table 1: Distribution of raw, classified and filtered sequencing reads across the 8 libraries**

# Figure Legends

**Figure 1: Overview of the experimental design and bioinformatic workflow.** (a) Apiary location and habitat structure. (b) 8 libraries were prepared from 4 honey collection across 3 seasons. Step variation between Shotgun Metagenomic (SM) and direct-SM libraries preparation. (c) Downstream analysis of the metagenomic data and main output analysis.

**Figure 2: Evaluation of the taxonomic classifiers and genomic aligners.** (a) Viridiplantae at the species level. (b) Non-Viridiplantae at the species level. (c) Correlation between expected and observed abundance. (d) Summary table of the mean square error (RMSE).

**Figure 3: Community variation across libraries**

**Figure 4: Seasonal plant diversity**

**Figure 5: Bee health**

# Supplementary Information

**Supplementary Figure 1: Apiary monitoring.** (a) Population. (b) Varroa. (c) Temperature. (d) Rain.

**Supplementary Figure 2: Quality control of the library preparation.** (a) Bioanalyser before/ (b) after

**Supplementary Figure 3: Phytogenic distribution and genomic characteristics of the species used to build the mock (simulated) samples.** (a) phylogenetic tree. (b) Length and C/G content of DNA sequences. (c) output of various classifier tool for the Viridiplantea mock sample at the genus level

**Supplementary Figure 4**: **Evaluation of the confidence thresholds filtering.** (a) ppv. (b) sensitivity. (c) confidence versus genus abundance

**Appendix 1:** NCBI reference of the DNA sequences used to build the mock (simulated) samples and their relative abundance.