

Towards a better understanding of *Apis mellifera* and *Varroa destructor* microbiomes: introducing 'PHYLOH' as a novel phylogenetic diversity analysis tool

A. SANDIONIGI,*¹ S. VICARIO,†¹ E. M. PROSDOCIMI,‡ A. GALIMBERTI,* E. FERRI,*
A. BRUNO,* B. BALECH,§ V. MEZZASALMA* and M. CASIRAGHI*

*ZooPlantLab, Department of Biotechnology and Biosciences, University of Milan-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy, †Institute of Biomedical and Technologies (ITB), National Research Council (CNR), Via Giovanni Amendola, 122/D, 70126 Bari, Italy, ‡DEFENS, University of Milan, Via Mangiagalli, 25, 20133 Milan, Italy, §Institute of Biomembrane and Bioenergetics (IBBE), National Research Council (CNR), Via Giovanni Amendola, 165/A, 70126 Bari, Italy

Abstract

The study of diversity in biological communities is an intriguing field. Huge amount of data are nowadays available (provided by the innovative DNA sequencing techniques), and management, analysis and display of results are not trivial. Here, we propose for the first time the use of phylogenetic entropy as a measure of bacterial diversity in studies of microbial community structure. We then compared our new method (i.e. the web tool [PHYLOH](#)) for partitioning phylogenetic diversity with the traditional approach in diversity analyses of bacteria communities. We tested [PHYLOH](#) to characterize microbiome in the honeybee (*Apis mellifera*, Insecta: Hymenoptera) and its parasitic mite varroa (*Varroa destructor*, Arachnida: Parasitiformes). The rationale is that the comparative analysis of honeybee and varroa microbiomes could open new perspectives concerning the role of the parasites on honeybee colonies health. Our results showed a dramatic change of the honeybee microbiome when varroa occurs, suggesting that this parasite is able to influence host microbiome. Among the different approaches used, only the entropy method, in conjunction with phylogenetic constraint as implemented in [PHYLOH](#), was able to discriminate varroa microbiome from that of parasitized honeybees. In conclusion, we foresee that the use of phylogenetic entropy could become a new standard in the analyses of community structure, in particular to prove the contribution of each biological entity to the overall diversity.

Keywords: bioinformatics, high-throughput DNA sequencing, microbial community structure, phylogenetic entropy, symbioses

Introduction

In ecology, the analysis and interpretation of community diversity is a hot topic. In particular, with the advent of high-throughput DNA sequencing (HTS), the attention of the researchers on this topic dramatically increased. Among communities, symbioses, and particularly host-parasite interactions, are intriguing themes. In a symbiotic relationship, the partners reciprocally influence their physiology and, in general, their evolution. Nowadays, the characterization of the microbiome (intended as the sum of microscopic living beings found in a symbiotic relationship in different host body compounds, ranging

from the gut to the skin) is considered pivotal to understand physiological changes occurring in a symbiosis (Mazmanian et al. 2005). In recent years, the scientific researchers focused on microbiome composition and variation in different hosts or physiological/environmental conditions using HTS (Sanchez et al. 2012; Dimitriu et al. 2013; Meriweather et al. 2013). The link between microbiome diversity and host health condition was discernible since the first published manuscripts. For instance, there is evidence that humans and mice subjected to different kind of stresses (such as diseases, parasites or ecological factors) show intense modifications in their own microbiomes in terms of initial colonization, final composition and overall stabilization (Candela et al. 2012; Lozupone et al. 2012).

However, the comprehension of mechanisms and dynamics influencing microbial diversity in hosts and

Correspondence: Maurizio Casiraghi, Fax: +39-02-64483568; E-mail: maurizio.casiraghi@unimib.it

¹These authors contributed equally to this work.

symbionts is much more complicated due to (i) the occurrence of several interacting variables (both abiotic and biotic) and (ii) the neglected contribution of the evolutionary history of single biological entities on the overall diversity when conventional analytical methods are applied (Fig. S1, Supporting information). Currently, researchers have just started incorporating historical constraints (represented as phylogenies) into their analyses. This innovation is motivated by the aim of filling the gap between evolutionary and ecological analyses of microbial communities (Lozupone et al. 2007, 2011).

Here we tested two classes of approaches for microbial community analysis: distance method approaches (DMAs) and partitioning phylogenetic diversity (PPD). DMAs became a standard in microbiome analyses, whereas PPD, which uses the phylogenetic entropy as a measure of microbiome diversity (Jost 2007; Chao et al. 2010), is here applied for the first time. The phylogenetic entropy is a generalization of Shannon entropy based on the fact that different observed categories are not all equally different from each others, having a similar structure that could be modelled using a phylogenetic tree. We implemented PPD in the user-friendly web application, PHYLOH.

We applied DMAs and PPD to characterize microbiomes in the model honeybee (*Apis mellifera*) and its parasitic mite varroa (*Varroa destructor*, Arachnida: Varroidae). The rationale is that the comparative analysis of both honeybee and the parasitic varroa microbiomes could open new perspectives about the role of the parasite on health of honeybee colonies. Indeed, *Varroa destructor* is considered responsible of the increasing incidence of deformed wing virus (Mückel et al. 2011), and it was reported as a vector of bacterial pathogens causing for example the European foulbrood (e.g. *Melissococcus plutonius*) (Forsgren 2010; Evans & Schwarz 2011).

Varroa destructor parasites honeybee larvae in their brood cells, where female mites feed on honeybee haemolymph and lay eggs. Mites have a large dispersal capability and, in absence of reiterate chemical and/or antibiotic treatments, infested honeybee colonies typically collapse in few years. For these reasons, the occurrence of varroa has serious consequences on ecological, social and economic contexts (Rinderer et al. 2010; Rosenkranz et al. 2010; Annoscia et al. 2012; Guzman-Novoa et al. 2012).

In spite of the interest, the ecological dynamics of the honeybee–varroa parasitic symbiosis are still largely unknown. Studies conducted on adult honeybees showed a characteristic microbiome (Jeyaprasath et al. 2003; Dillon & Dillon 2004; Mohr & Tebbe 2006; Martinson et al. 2011; Sabree et al. 2012). However, the microbiome of the larval stages, as well as that of the parasitic mite, remains largely unexplored, excluding few researches on the transmission of specific pathogens

(Mouches et al. 1984; Cornman et al. 2010; Forsgren 2010; Martinson et al. 2012). At the light of present knowledge, alterations of honeybee microbiome due to the symbiosis with varroa are expected, even if several aspects influence the final outcome. The honeybee microbiome undergoes a peculiar dynamic over the life cycle of the insect: the pupa is almost sterile, as a consequence of the physiologic characteristics of the gut tract and the diet of mature larvae during the 6 days before capping (i.e. the closure of the brood cell) (Martinson et al. 2012). It is reasonable to assume that the bacterial load within the brood cells partially reflects the total bacterial count of the hives and that microbial communities characterizing the hives are partially present in the cells even after capping (Martinson et al. 2012). But, what happens when varroa alters this equilibrium? The perturbation caused by varroa in the developmental phase of honeybee larvae, and the consequent formation of the nutrition hole caused by the parasite, could lead to the intrusion of external bacteria (both from a potential vector or from the environment) into the larva, with a substantial modification of the original microbial community. The hypothesis is that varroa mites play a fundamental role in the alteration of bacterial composition of honeybee larvae, acting not only as a vector, but also as a sort of ‘swing door’ through which exogenous bacteria can enter into the larva and alter the mechanisms of primary succession of honeybee microbiome.

To validate our hypothesis, we studied varroa and honeybee bacterial communities through DNA-barcoded amplicon pyrosequencing, taking advantage of the HTS methods (Blow 2008; Metzker 2009), which also allow the detection of uncultured bacteria. We compared the results of the phylogenetic entropy-based approach (implemented in PHYLOH) with a classical method based on pairwise distances. We critically evaluated the strength and weakness of both approaches and the importance of phylogenetic constraint. At the same time, we aimed to reach a more complete vision of the relationships between honeybee and varroa in the microbiome interchange.

Materials and methods

A schematic overview of the experimental pipeline is shown in Fig. 1. Our work is divided into three sections as follows: (i) laboratory procedures including sampling; (ii) DNA extraction, amplification and pyrosequencing; (iii) sequence analysis and microbial community differential analysis.

Laboratory procedures

Sampling—Honeybee larvae and varroa mites were sampled directly from capped brood cells in eight

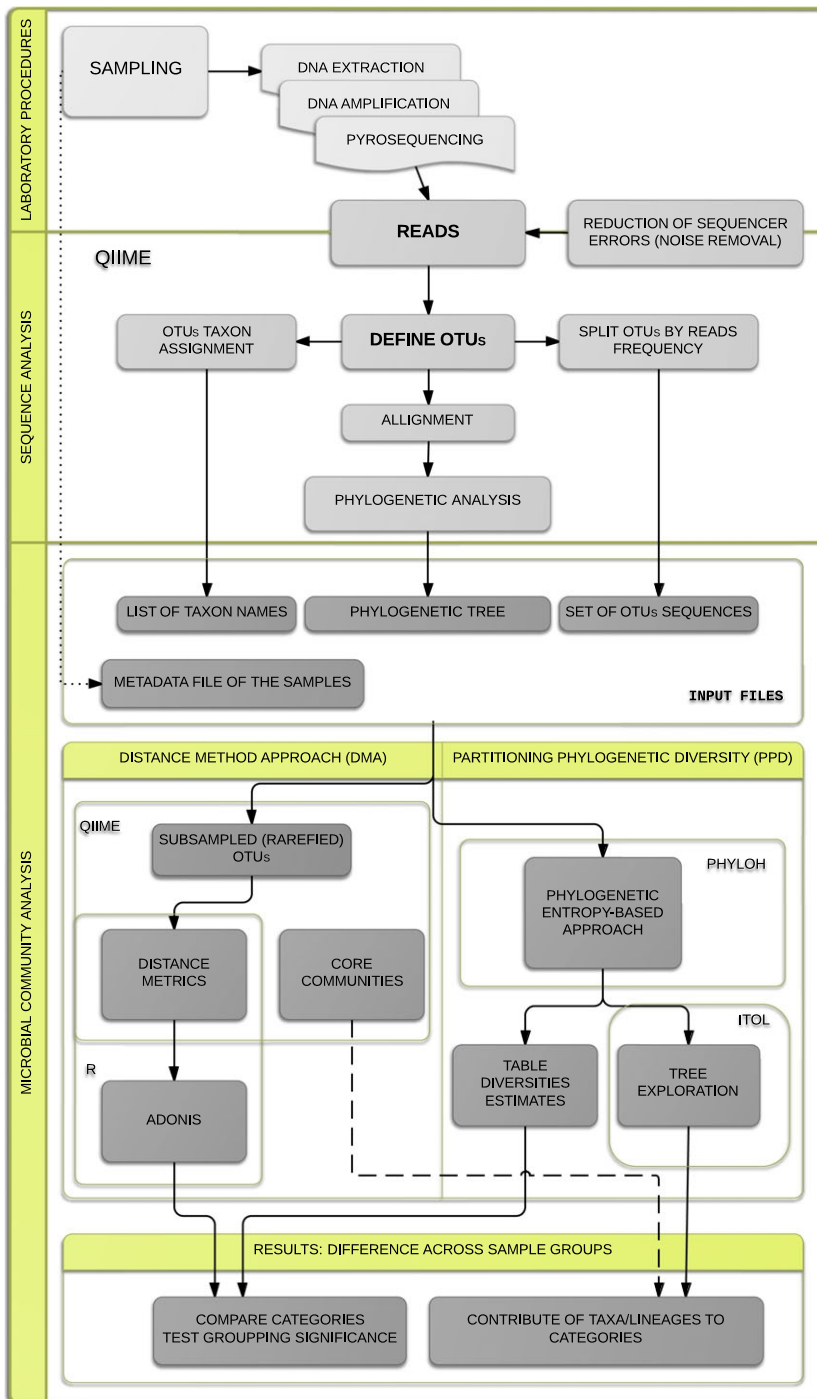


Fig. 1 Workflow of the procedures used in our project. Our work is divided into three sections (i) laboratory procedures; (ii) sequence analysis and (iii) microbial community analysis. In particular, the microbial community analysis was conducted following two approaches: microbial diversity analyses and partitioning phylogenetic distances.

apiaries in Northern Italy. We performed our analyses on a total of 21 individuals of honeybee larvae from seven different apiaries, and 21 varroa mites found in the same brood cells. As a negative control, a pool of five healthy honeybee larvae from a noninfested site was analysed.

Opercula of cells were opened with sterile instruments. Honeybee larvae and varroa were immediately

removed and put in 2-mL tubes filled with absolute ethanol. The samples were stored at 20 °C until DNA extraction. In the study area, *V. destructor* is abundant and widespread, and consequently, we found only one noninfested apiary. This apiary was determined to be healthy after a careful inspection of all the hives by expert bee-keepers. The same experts determined that the presence of varroa in the other seven apiaries was

high. In absence of preventive acaricide treatments, these honeybee colonies would have certainly collapsed before winter.

DNA extraction—All the extraction steps were performed in a sterile laminar flow cabinet. After the removal of the head, only the first segments of *A. mellifera* specimens (after cuticular removal) were used for the DNA extraction, while for *V. destructor*, DNA was extracted from the whole organism. The dissections were made in sterile conditions with a scalpel in a Petri dish. Each sample was then rehydrated for 4 h in sterile water at room temperature, and mechanically grinded with the scalpel. Total DNA was then extracted using a commercial kit (DNeasy Blood and Tissue Kit; Qiagen, Milan, Italy) and eluted in 50 µL sterile water. A pretreatment of Qiagen columns was performed to wash away any trace of contaminating bacterial DNA (Evans et al. 2003; Mohammadi et al. 2005). DNA extracts of the five larvae from the noninfested apiary were pooled.

16S rRNA amplification and pyrosequencing—The 16S rDNA gene fragment corresponding to the V3 hypervariable region was PCR-amplified with Roche 454 FLX (Titanium reagents) using the primer pair 341F (5'-CC TACGGGAGGCAGCAG-3') and 518R (5'-ATTACCGC GGCTGCTGG-3') (Watanabe et al. 2001). The reaction was performed in a 20 µL volume with the following reagents: 1X Taq-buffer with MgCl₂ 1.5 mM, dNTPs 2 mM, forward and reverse primers 1 µM each, Taq polymerase 0.5 U, DNA 50 ng, MilliQ H₂O to the volume. The thermal cycle was: 94 °C for 90 s, 29 cycles at 94 °C for 20 s, 58 °C for 30 s, 72 °C for 20 s followed by a final extension at 72 °C for 10 min and 60 °C for 5 min.

A subsequent nested PCR step using the products of the first one as template was performed with standard 52 bp primers, comprising pyrosequencing primers A or B, multiplex identifiers (MID) and 518R or 341F primers. PCR mix and reaction volumes were the same as described above, except for the primers (10 µM). The thermal cycle was 94 °C for 90 s, 40 cycles at 94 °C for 20 s, 58 °C for 30 s, 72 °C for 20 s and a final extension at 72 °C for 10 min and 60 °C for 5 min.

For each sample, we used a unique combination of MID's on the forward and reverse primers. PCR products were quantified using Bioanalyzer 2100 (Agilent) and normalized for quantity. Pyrosequencing was then performed on Roche 454 GS-FLX titanium by BMR Genomics Service at the Interdepartmental Biotechnology Centre of the University of Padua (CRIBI).

Sequence analysis—Sequence analysis was performed according to the following steps: (i) reads denoising; (ii) operational taxonomic units (OTUs) definition;

(iii) OTUs taxon assignment; (iv) phylogenetic analysis; (v) OTUs splitting based on reads frequencies.

(i) All reads were trimmed, filtered and assigned to the corresponding sample according to their tag. Sequences shorter than 100 bp with quality average <30 or containing unresolved nucleotides were removed from the data set. ACACIA software version 1.52 (Bragg et al. 2012) was used for pyrosequencing noise removal considering Balzer error model and a maximum k-mer distance between reads of 13 (default parameter for error correction). The detection of chimera reads was performed using a pipeline based on USEARCH (Edgar 2010) and UCHIME (Edgar et al. 2011) included in Quantitative Insights Into Microbial Ecology (QIIME) software suite (version 1.7.0) (Caporaso et al. 2010).

(ii) UCLUST wrapper was used to cluster sequences into OTUs, based on 97% sequence similarity. For each OTU, a cluster centroid (i.e. a representative sequence) was chosen. To estimate diversity and reduce noise in patterns of beta diversity, singleton OTUs (i.e. OTUs represented by a single sequence) were removed before community analysis (Zhou et al. 2011).

(iii) Using a PYTHON 2.7 script, we merged the Greengenes 16S rRNA database prefiltered at 97% identity (McDonald et al. 2011) with a bacterial OTUs data set constituted by symbionts previously described in studies conducted on *Apis mellifera* (Mohr & Tebbe 2006; Martinson et al. 2011, 2012; Mattila et al. 2012; Moran et al. 2012; Sabree et al. 2012). To create a reference database for taxonomic assignment, bacterial 16S rRNA sequences were retrieved from GenBank and clustered using UCLUST at 97% sequence similarity. The taxonomic attribution of cluster centroid sequences was carried out using RDP Bayesian classifier (Wang et al. 2007) with the new merged data set obtained using a 0.8 confidence level. OTUs were assigned by the RDP classifier, considering the fifth and sixth taxonomic levels wherever possible, which, in most cases, corresponded to family and genus ranks. If RDP assignment was uncertain (probability between 0.8 and 0.9), the QIIME-selected representative sequence was used to query with blastn algorithm on NCBI nucleotide database. Only in case of perfect match (i.e. max identity 100%, query coverage 100%), the NCBI taxonomy was used (results are shown in Table S1 in Supporting information and/or DRYAD entry doi:10.5061/dryad.j4d15).

(iv) A maximum-likelihood (ML) phylogenetic tree was built according to default parameters using FASTTREE software (Price et al. 2010) integrated in QIIME.

(v) The community abundance profile, produced by UCLUST and labelled by RDP Bayesian classifier, was split into two groups, as their global frequency was lower or higher than 1%. This threshold is the advised value in QIIME manual; it is often used in the literature and

matches a gap in the frequency distribution of this data set (see Supporting information and/or DRYAD entry doi:10.5061/dryad.j4d15).

After this step, three data sets were produced: 'All Frequency Cluster' (i.e. AFC, including all OTUs), 'Low Frequency Cluster' (i.e. LFC, including OTUs with frequency lower than 1%) and 'High Frequency Clusters' (i.e. HFC, including OTUs with frequency higher than 1%). This partitioning allowed us to explore the effect of dominant and rare taxa among the microbiomes of honeybee and varroa.

Microbial community analyses

Microbial communities were examined using two approaches: the conventional distance matrices analysis (DMA) and the partitioning phylogenetic diversity (PPD). The last method was here applied for the very first time to the analysis of microbiomes.

Distance matrices analysis, coupled with ANOVA family statistics, is a well-known statistical framework in microbial community analysis that allows comparing the effect of different explanatory variables. This method is sensitive to unbalanced sampling, and it requires data rarefaction. In a HTS framework, the term 'unbalanced sampling' refers to the amount of sequences generated for each biological sample (e.g. if you get 100 000 sequences from the organism 1 and 50 000 sequences from the organism 2, the randomization leads to a loss of information from the larger sample). The principal pitfall of this approach is that it does not take into account the influence of rare OTUs on the global diversity of a sample (see also Fig. S1, Supporting information). To overcome this limit, we here propose the use of a PPD approach (Chao et al. 2010). PPD is being framed within information theory and can deal directly with discrete values, without producing distance matrices. It can also incorporate information deriving from unbalanced sampling, therefore avoiding a preliminary step of data rarefaction. Finally, being based on the phylogenetic structure of the data, PPD takes into account the influence of rare lineages on the microbial composition. In fact, similar sequences with low counts can build up consistent contribution to beta diversity on the branch leading to their most recent common ancestor.

The distribution of variability among bacterial communities was described for three environmental variables: (i) 'Cells' to show differences between single honeybee and the corresponding parasite found in the same brood cell, highlighting possible relationships between host and parasite microbiomes; (ii) 'Localities' to draw attention to differences existing between the microbiome of the seven apiaries; (iii) 'Status' to show

microbiome differences among the pools of healthy honeybees, parasitized honeybees and mites.

Distance method approach

All the analyses were performed on the rarefied OTU tables to permit comparisons of diversity patterns within and between communities. The number of OTUs (based on the 97% sequences similarity) was determined for each sample. As depicted in Fig. 1, community analyses were performed with qualitative (jaccard and unweighted UniFrac (Lozupone et al. 2011) and quantitative distance metrics [squared chord (Cavalli Sforza & Edwards 1967; Orloci 1967) and weighted UniFrac (Lozupone et al. 2011)] using QIIME and R for statistical computing (R Development Core Team 2012).

Jaccard and squared chord were chosen as complementary metrics to unweighted and weighted UniFrac to test how the community pattern changes with or without phylogenetic information. We chose the squared-chord distance because it was identified in previous works as a metric fitting well at an exploratory analysis of communities where sampling was conducted blindly [see for example (Legendre & Gallagher 2001)].

To interpret the distance matrix, we used UPGMA hierarchical clustering method and we tested the robustness of results with jackknife analysis (1000 permutations). Further, to determine whether the grouping of samples by a given category was statistically significant, we used adonis (Oksanen et al. 2007) a permutational ANOVA allowing the use of distance matrix as dependent variable. This procedure is included within the QIIME suite and was implemented in vegan R package. The model used in adonis was the following:

$$\text{Dist}(i; j) \sim \text{Status} + \text{Localities} + \text{Cells}$$

where i and j are all possible pairs of samples without redundancy and identity.

In this model, distances were considered as the response variable, whereas each of the environmental variables was considered as predictor.

We identified the microbial core community in honeybee and in the mite using compute_core_microbiome.py script from QIIME. OTUs were grouped according to their occurrence in a specific percentage of the total samples. The grouping steps were defined as elevens threshold between 0.5 and 1, corresponding to the 50% and 100% of the samples, respectively. This allowed defining the core community of each host species and more specifically to recognize the OTUs present in the majority of the samples of a given host.

Figures S2 and S3 (Supporting information and/or DRYAD entry doi:10.5061/dryad.j4d15), showing the

taxonomic assignment and the abundance distribution, were generated with phyloseq (McMurdie & Holmes 2013) and ggplot2 (Wickham 2009) R packages.

Partitioning phylogenetic diversity

Following the framework proposed by Jost (2007), it is possible to parse the total phylogenetic entropy of a data set (c component), in intragroups entropy (a component, and intergroups entropy (b component). Jost (2007) distinguishes between entropy measures, having bit as unit (or nats, or bans, depending on the logarithm base), and diversity measure having as unit the number of equally abundant categories that would produce the same amount of entropy.

Partitioning operations are performed using entropy components, while the final result is transformed into diversity by elevating to the base of the used logarithm. Assuming that cluster label of observation is collected in vector X and that group label is collected in vector Y, this framework allows to define H_c as entropy of X, H_a as entropy of X conditional to Y, and H_b entropy as $H_c - H_a$, defined also as the mutual information between X and Y (MacKay 2003).

It is important to notice that beta diversity (D_b), the exponential of H_b , has as unit the number of equally abundant and different samples (the categories of Y), while D_c and D_a are measured in number of equally abundant clusters (the categories of X). Within microbial community analysis, the interest lays generally in estimating D_b . To assess whether this measure was significantly different from 1 (i.e. the diversity value under the hypothesis of no difference among groups), we compared the realized statistics with a null distribution obtained by a permutation of X values onto Y ones. This procedure keeps a constant number of observations per group, allowing accounting for different sampling efforts. As described here, this procedure does not consider the phylogenetic structure that links the categories of the vector X. This limitation is critical for biological data. Indeed, radically different findings from a biological perspective would produce the same Shannon-based beta diversity (see Fig. S1, Supporting information). This becomes possible using the phylogenetic entropy. The phylogenetic entropy is a generalization of Shannon entropy where the different observed categories are not all equally different from each other, but have a similarity structure that could be modelled using a phylogenetic tree. Following Chao et al. (2010), we assume that variable X is the abundance distribution of the clusters defined by UCLUST, and its similarity structure can be modelled with a phylogenetic tree t.

The phylogenetic entropy measure could be defined as follows:

$$H_p(X) = \sum_{i \in B_t} \frac{L_i}{T} p_i \log p_i$$

where L_i is the value of the branch length for the i th branch while T is the average distance from tip observation to root in the tree, as defined by the formula

$$T = \sum L_i p_i$$

where $i \in B_t$ is the set of branches of the tree t, and p_i is the frequency of the descendant of branch i . Once this point is set, to generalize the partitioning of diversity to include phylogenetic information, it is sufficient to apply the previous definition of c, a, b diversity using the phylogenetic entropy instead of the Shannon entropy. Phylogenetic entropy c (H_c) is equal to $H_p(X)$, while phylogenetic entropy a is equal to the weighted mean of the phylogenetic entropies per group, where weights are proportional to the number observation carried out in each group. More formally, this can be written as follows:

$$H_{pa} = \sum_{y \in Y} p_y \sum_{x \in X} H_p(X | Y = y) \quad (1)$$

where p_y is the relative frequency of observation in each group while

$$\sum_{x \in X} H_p(X | Y = y) \quad (2)$$

is the phylogenetic entropy measured in the different group y defined in vector Y. Phylogenetic entropy beta (H_{pb}) remains defined as the $H_{pc} - H_{pa}$. This phylogenetic entropy beta, or phylogenetic mutual information, behaves in analogous way as Shannon-based mutual information given the Kullback–Leiber divergence. This matches the different way to estimate the classical mutual information (Marcon et al. 2012).

Given that the H_{pb} is a difference of two summations in which each term is relative to a branch in the tree, it is possible to reorder the terms and obtain the contribution of each branch in the final H_{pb} .

The mutual information H_{pb} , or its exponential D_b , cannot be directly compared across different grouping variable Y, given that the cardinality of each variable Y (the number of possible states) defines a different upper boundary to the value of each b diversity. To normalize diversity measures across different partitioning variables, it is necessary to transform diversity in overlap or effective average proportion of shared lineages in an

individual environment. The value ranges from 1 (all lineages are shared) to 0 according to the following formula:

$$\text{Overlap} = \frac{(H_b - \max H_\beta)}{\max H_\beta}$$

where the maximum value of H_b is the logarithm of the number of groups or the entropy of Y depending on whether differential sampling needs to be included in the measure or not.

In this work, this approach was applied using the tree obtained from `FASTTREE` (Price et al. 2010) and the tree with the same topology, but internal branches with length 0 and terminal branches with length 1. The latter modified tree is used to perform the Shannon entropy analysis without taking into account the phylogenetic information, using the same software implementation. In fact, setting all internal branch lengths to zero leaves in the summation only the terms present in Shannon entropy formula. These two alternative settings allow to better evaluating the importance of phylogeny when interpreting the results.

Note that fractions of overlap are always higher in the phylogenetic entropy case, given that there are some phylogenetic similarities between different OTUs, while the classic Shannon-based approach assumes that each OTU is totally different from the others.

The mutual information (i.e. the beta entropy) deals with the covariation between two variables, so we explored the relationship between sequences and a single explanatory variable at the time. Within the framework of information theory, it could be possible to take into account the network of multiple interactions, but its application to phylogenetic entropy requires further investigations.

Partitioning phylogenetic diversity approach was implemented as a stand-alone Python script (`PHYLOH` available at <https://github.com/svicario/phyloH>) and includes a visualization routine based on `ITOL` (Letunic & Bork 2007) that allows to visualize the distribution of diversity and the contribution of the different OTUs to the partition as an html file. Input format follows `PHYLOCOM` input standard (Webb et al. 2008). To facilitate the use for the microbiologist community, we wrapped in a Web Service (https://www.biodiversitycatalogue.org/rest_methods/143) the script coupling it within a workflow (<http://www.myexperiment.org/workflows/3570.html>) with some parsing script enabling the use of output files from `QIIME` suite as input. The workflow could be run locally using a `TAVERNA` desktop engine (Wolstencroft et al. 2013) or as web application in the `BioVeL` portal. Access to the portal could be obtained from the

`BioVeL` website (www.biovel.eu) or contacting directly the authors.

Results

Results are organized according to the pipeline showed in Fig. 1.

Sequence analysis

Define OTUs—After sorting sequence reads for quality scores, sequencing errors and chimeras, we obtained 34 816 sequences. `UCLUST` returned 295 OTUs (data set AFC).

Taxonomic assignment—The complete list of OTUs, with corresponding taxa names and acronyms, is provided as Supporting information and/or `DRYAD` entry doi:10.5061/dryad.j4d15 (Table S1, Supporting information). Taxon assignment at family level of the three types of organisms involved (healthy honeybee, parasitized honeybee and varroa) is shown in Fig. 2.

Split OTUs by sequence frequency—Of 295 OTUs, 21 exceeded the threshold of 1% of minimum total observations (data set 'HFC', total sequences: 24 005), and 274 were defined as a rare OTUs not reaching the 1% threshold (data set 'LFC', total sequences: 10 811).

Phylogenetic analysis—The maximum-likelihood tree generated by `FASTTREE` is shown as internal tree in `PHYLOH` output (see Supporting information and/or `DRYAD` entry doi:10.5061/dryad.j4d15).

Microbial community analysis

Distance method approach—The `UPGMA` analysis considering jaccard distances for all the three sets of observations (AFC, HFC, LFC), shows a single cluster including all samples belonging to parasitized honeybees and mites and a separate cluster including the pool of healthy honeybees. On the contrary, with squared-chord metrics, which consider abundances information of OTUs in HFC and AFC, the analysis shows two different groups between parasitized honeybees and mites. These results are partially replicated with `UniFrac` (both weighted and unweighted), although separation is less sharp (see Fig. S4, Supporting information).

The `adonis` test was performed separately for each variable (Cells, Localities and Status) and using all variables together without interactions. Results were highly similar; for this reason only the coefficients of the model with all predictors combined are reported (Supporting

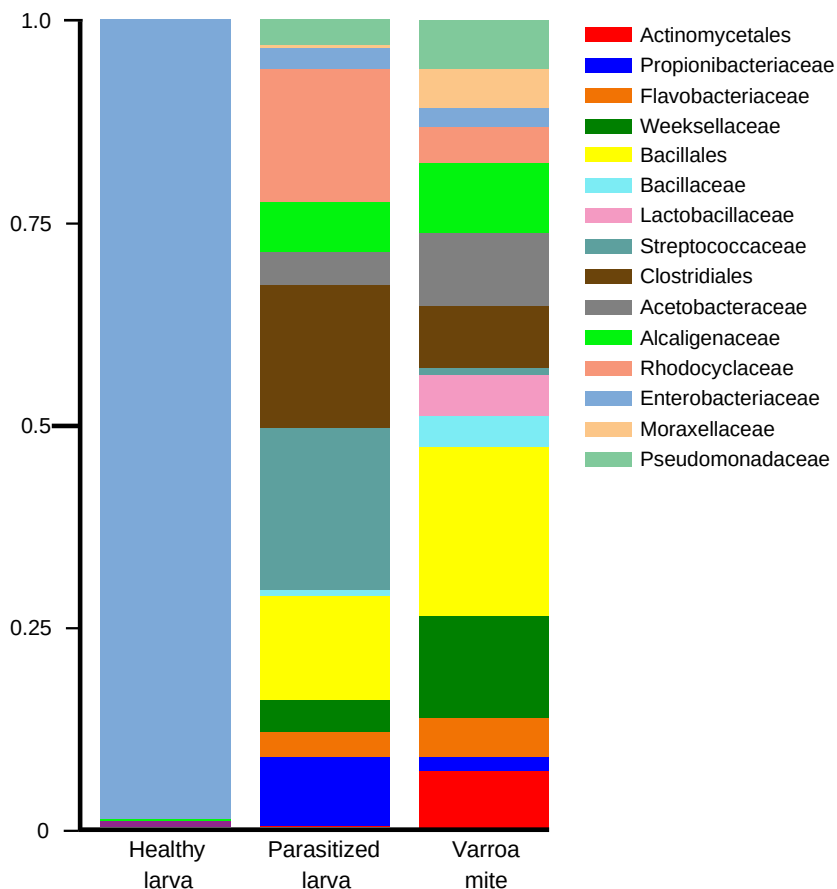


Fig. 2 Structure of microbiomes from healthy and parasitized honeybees and varroas. The histogram shows the 15 families detected and relative abundances. In two cases (Bacillales and Clostridiales), the family rank was not assigned, and consequently, the order rank is shown.

information Table S2 and/or DRYAD entry doi:10.5061/dryad.j4d15).

The adonis test grouping the samples for brood cells (Cells) has limited significance (P-values between 0.05 and 0.01) and small effect compared to degrees of freedom used up (34–35% explained variance, 1.6% the mean value for single cell). In addition, adonis test is significant only using the jaccard distance and unweighted UniFrac in the LFC data set. We found the same variance in the AFC data set, but only using the unweighted UniFrac (35% variance explained and P-value of 0.04).

A slightly stronger signal is detected in Localities, where jaccard, unweighted UniFrac and squared chord show a significant grouping in the AFC data set (P-value <0.01 and effect 17–20% with per-Localities mean variance explained between 3% and 2%). Similar signal is also visible with HFC and LFC, but only jaccard is significant (Supporting information Table S2 and/or DRYAD entry doi:10.5061/dryad.j4d15).

The Status grouping shows a very different pattern. The mean variance explained per state of Status builds up to 4–5% in the AFC data set and 9% for the weighted UniFrac in HFC. Only the LFC subset is not significant

for the majority of distances used, and in any case the mean variance explained is quite low (1%).

Microbial core communities In healthy honeybees, we found only three OTUs, one of them (Proteo-7, a member of the genus *Serratia*) accounting for the 99% of sequences. The remaining 1% is shared between two OTUs: Proteo-2 and Firmi-7, respectively, identified as *Achromobacter* sp. and *Lactobacillus* sp.

On the curves returned by QIIME script, we defined the host and parasite core microbiomes using a 0.8 threshold (Fig. S5, Supporting information). The threshold was chosen according to the steepest point in the curve, namely the point in which the least addition of OTU in the core ensures the largest fraction of samples to be compliant with the core representation. This means that a single OTU has to be present in the 80% of samples of a certain species to be considered 'core' for that species.

The OTUs considered as microbial core for honeybee and varroa are shown in Table 1.

Partitioning phylogenetic diversity—We used PHYLOH to perform partitioning tests on three environmental variables (Cells, Localities and Status) for the three frequency data sets (AFC or LFC or HFC) considering the

Table 1 OTUs associated with healthy and parasitized honeybees and varroa by compute_core_microbiome.py script

Parasitized honeybee	Varroa mite	Shared
Propionibacterium sp. (Actino-1)	Chryseobacterium sp. (Bacte-1)	Chryseobacterium sp. (Bacte-1)
Chryseobacterium sp. (Bacte-1)	Flavobacteriaceae (Bacte-2)	Streptococcus sp. (Firmi-1)
Streptococcus sp. (Firmi-1)	Streptococcus sp. (Firmi-1)	Lactobacillus sp. (Firmi-7)
Clostridiales (Firmi-3)	Bacillus sp. (Firmi-2)	Hydrogenophilus sp. (Proteo-1)
Lactobacillus sp. (Firmi-7)	Geobacillus (Firmi-5)	Achromobacter sp. (Proteo-2)
Hydrogenophilus sp. (Proteo-1)	Anoxybacillus (Firmi-6)	
Achromobacter sp. (Proteo-2)	Lactobacillus sp. (Firmi-7)	
Escherichia sp. (Proteo-6)	Hydrogenophilus sp. (Proteo-1)	
	Achromobacter sp. (Proteo-2)	
	Pseudomonas sp. (Proteo-4)	
	Escherichia sp. (Proteo-6)	

phylogenetic information available or not. We can observe several qualitative differences among the 18 analyses transforming the mutual information into percentage of overlapping to allow a comparison among variables. As shown in Fig. 3, Localities and Cells variables are not heavily influenced by phylogenetic information. The relative position of LFC, HFC and AFC, measured in fraction of overlap counted in OTUs or lineages, is similar. In fact, LFC has, in both cases, low overlap, whereas AFC and HFC data sets show the maximum overlapping.

The variable Status describes a very different pattern from the previous two. Taking into account phylogeny, HFC differentiates more among groups (overlap 0.87), with the other two data sets showing about 0.91 overlapping. On the contrary, if phylogenetic information is not considered, the pattern is similar to the other two variables. The biggest difference is observable in LFC, while the smallest difference is observable in HFC.

The contribution of the branches to the mutual information between sequences and the Status variable. The experimental design output and the numerical partitioning of phylogenetic diversity for Status variable considering all OTUs found (AFC) are shown in Table 2. A general overview of the branches contribution to beta diversity across groups is shown in Fig. 4. The 295 AFC OTUs are well distributed on the phylogenetic tree, and consequently,

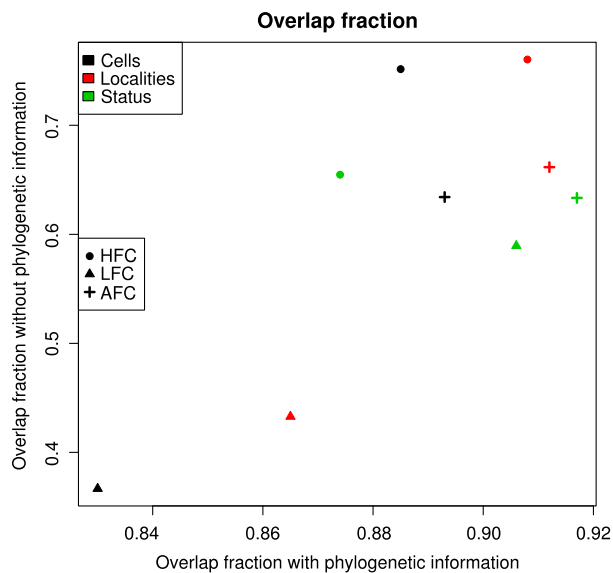


Fig. 3 Comparison of the percentage of overlap of the six data sets with or without taking into account the phylogenetic information. The percentage of overlap was calculated using estimate of beta diversity and phylogenetic beta diversity for AFC, HFC, LFC cluster frequencies and Cells, Localities, Status environmental variables.

their impact on the phylogenetic index is higher. It is observable that Bacte-1, -2, -3 are more typical of the varroa and are all grouped in the same lineage (L208). On the contrary, in Proteobacteria and Firmicutes, it is possible to find OTUs preferentially found in honeybee or varroa. Proteo-12, Proteo-14, Proteo-46, Proteo-24 and Proteo-17 OTUs belong to lineage L268 and are preferentially present in honeybee. There are three isolated OTUs preferentially found in varroa (Proteo-3, Proteo-4 and Proteo-5). Firmi-1, Firmi-9 and Firmi-4, typical of honeybee, are mixed with varroa's OTUs Firmi-6, Firmi-14, Firmi-2, Firmi-8 (both descending from lineage L387).

This lack of strong phylogenetic signal could be caused by recent specialization or random community assembly, given that Firmi-6 has a lower frequency sister taxa Firmi-14 also present mainly in varroa we prefer the first hypothesis.

Discussion

The analysis of complex communities (such as microbiomes) is today standard in different fields of biology and medicine. The number of published works is increasing daily, but there are some concerns on the real quality of the results showed. We are here comparing a 'traditional' approach to the community analysis [i.e. distance method approach (DMA)], vs. the innovative partitioning phylogenetic diversity (PPD, here implemented in

Table 2 Variability in reads count for each sample. Variability in reads count and sample number in groups. Overall gamma in each group. Overall alpha and contribution to alpha in each group. Beta across samples within groups and between groups. E is always within S, given that each sample belongs to only one environment type or sample group

Experimental design diversity: entropy and diversity of observation in the different groups							
H_Environment							
H(E)		MaxDiversity		Diversity			
0.691		2		2			
Experimental design diversity: entropy and diversity of observation in the sample within the groups							
H_Sample							
H(E)		MaxDiversity		Diversity			
3.68		42		39.5			
Gamma diversity: diversity using all data and in each group							
H_gamma		H_gamma_parasitized_honeybee		H_gamma_mite			
H(T)	alpha_Diversity	H(T E = host_p)	gamma_Diversity	H(T E = pars)	gamma_Diversity		
1.76	5.83	1.75	5.73	1.69	5.42		
Alpha diversity: mean within group diversity							
H_alphaByEnvironment			H_alphaBySamples				
H(T E)		alpha_Diversity	H(T S)		alpha_Diversity		
1.72		5.56	1.35		3.85		
Beta diversity or mutual information between the phylogeny and a given grouping: diversity across group and across sample within same group							
MI_treeAndEnvironment				MI_treeAndSampleGivenEnvironment			
I(T,E)	beta_Diversity	Percentage_Overlap	P value	I(T,S E)	beta_Diversity	Percentage_Overlap	P value
0.0465	1.05	0.933	0	0.369	1.45	0.466	0
Difference of each group from total: phylogenetic Kullback–Leiber divergence between each group and the overall sample							
KL_of_host_p					KL_of_pars		
Observed					Observed		
0.0162					0.0303		

$H_{\text{gamma}} = H_{\text{alphaByEnvironment}} + H_{\text{beta}} = H(T) = H(T|E) + I(T,E)$ while taking into account sample info: $H_{\text{gamma}} = H_{\text{alphaBySample}} + H_{\text{betaBySampleGivenEnvironment}} + H_{\text{beta}} = H(T) = H(T|S) + I(T,S|E) + I(T,E)$.

our newly proposed software (PHYLOH). The final aim is to propose our approach as a new standard in community diversity analysis. We tested the performance and utility of DMAs and PPD in the case of honeybee and varroa microbiomes, a biological scenario intriguing and complex at the same time. Our results showed that healthy honeybees have a simplified microbiome, constituted of few bacterial OTUs, while varroa is characterized by a more complex microbiome, qualitatively not different from that of parasitized honeybee. A simple conclusion could be that varroa microbes infected honeybees. However, the most abundant OTUs of parasitized honeybee do not derive directly from the mite, but are generalist or environmental bacteria. This is undoubtedly a peculiar result suggesting that these microbes could play a role of pioneer species, with a potential pathogenic activity (i.e.

Firmi-1, -9, L387, Streptococcus; Firmi-4, -14, L387, Clostridiales) (Lozupone et al. 2012).

Our results show that both DMAs and PPD approach give comparable results, but only PPD shows explicit support and allows, within the same statistical framework, to observe our data both from the general pattern to the contribution of single or group of OTUs.

We partitioned data according to three environmental variables, which represent three different forces shaping microbiome diversity: Localities, Cells and Status. When Localities variable is considered, in DMAs approach, the UPGMA method and the adonis test analyses show a weak, but significant signal for most of the distances and data set used. When all OTUs (AFC) combined with unweighted UniFrac are considered, the signal is stronger.

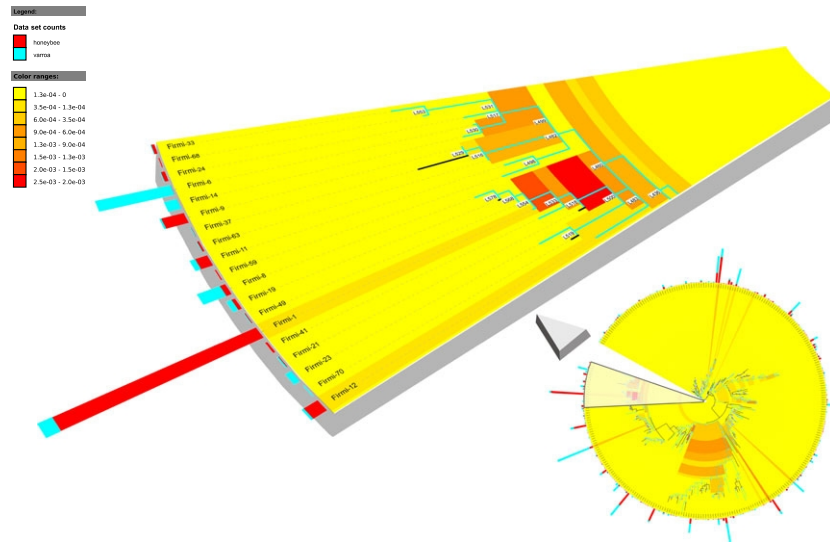


Fig. 4 In the figure are shown the main tips and relative branches of one of the most discriminating lineages across Status groups [L387, see tree.html (Appendix S1) in Supporting information and/or DRYAD entry doi:10.5061/dryad.j4d15 for further details] as example of the 'hairy pacman graphical output' derived from a *PHYLOH* analysis. Three types of data are shown on the tree: (i) the cyan colour of the branches indicates a significant contribution to I(T|E) (beta entropy or mutual information); (ii) the background of each branch is a gradient (colour range) from yellow to red for increased contribution to I(T|E); (iii) bar plot on each tip indicates the number of reads count in each group. For more details, see README file (Appendix S1) in Supporting information and/or DRYAD entry doi:10.5061/dryad.j4d15.

The results are similar using PPD, but with a stronger support. In fact, the variable Localities produces a significant effect, although this is better appreciable for rare OTUs (LFC) rather than considering the most abundant OTUs (HFC). In other words, low-frequency OTUs (better defined using PPD) represent the fraction of bacteria characterizing the single apiary. This finding is in agreement with previous results, showing that geographical distance is not crucial in shaping the core microbiome of beehives (Sabree et al. 2012).

In our hypothesis, the bacteria are transmitted from varroas to the honeybee larvae. This is corroborated by the analysis of the variable Cells, where differences between each larva and the corresponding mite were estimated. This grouping has limited effects in terms of mean explained variance when the DMA approach is used, but it is still significant for some distances in AFC and LFC data sets. In PPD, the overlap across categories is obtained by Cells with the LFC data set, showing a strong effect. This situation is difficult to explain: only the more frequent bacteria are shared among cells, whereas rare bacteria are not. However, the rare bacteria are more interesting because they act as a clear signature to discriminate different apiaries.

Considering the variable Status, both DMAs and PPD approaches can discriminate the three different categories: healthy honeybees, parasitized honeybees and varroas. Healthy honeybees have a simply and distinctive community with only one dominant OTU and two other

low-frequency OTUs. The most representative phylotype (Proteo-7) belongs to the genus *Serratia* that was isolated from the intestinal contents of healthy foraging worker honeybees (Jeyaprakash et al. 2003), a well-known symbiont (generally harmless) in many insect taxa (Dillon & Dillon 2004).

The scarcity of bacteria in the healthy larvae could be attributed to their particular gut morphology, physiology and nutrition (Martinson et al. 2012). Indeed, the larva retains its faeces from the early days of development, due to the temporary absence of a connection between the large mid-gut and the hindgut. The mature larva defaecates just before spinning a cocoon, when the capping has already happened. As the cocooned pupa does not eat, we can assume that there is no further colonization by bacteria present in the brood cell. Through these mechanisms, the early microbiome characterizing honeybee larvae is maintained constant in composition and ubiquitous in space (Jeyaprakash et al. 2003; Mohr & Tebbe 2006).

The higher level of bacterial communities diversity showed in varroa and parasitized larvae suggests a transmission from parasite to the host. These two microbiomes are identical from a qualitative point of view (taking into account the HFC), but the relative abundance of the different OTUs clearly differentiates the two bacterial communities. When phylogenetic information is included, the level of discrimination between parasitized larvae and varroa is even higher. In fact, only using

PPD and weighted UniFrac in DMA (i.e. the approaches that take into account the phylogenetic signal), the two clusters are highly discriminated. However, only in PPD, it is possible to show the contribution of the single OTU and groups of OTUs to the differences observed among samples.

The enhanced performance of PPD is exemplified in PHYLOH output tree [Supporting information tree.html (Appendix S1) and/or DRYAD entry doi:10.5061/dryad.j4d15 and as an example Fig.4] where the lineages L208, L268 and L387 are those discriminating the microbiome of varroa and infected honeybee. In particular, L208 lineage encompasses OTUs present in varroa, while L268 and L387 contain OTUs present in both varroa and honeybees. The most discriminating OTUs of lineage L208 belong to the genus *Chryseobacterium* (Bacte-12, -3). In contrast to the majority of bacteria belonging to Flavobacteria, typically found in soil and water environments, two (Bacte-1, -2) of these three OTUs were found as pathogens of soft ticks (Bure sova et al. 2006). Given the phylogenetic closeness between mites and ticks, it would be interesting to investigate the role of this genus in mites. The OTUs belonging to the lineage L268 were assigned to genus *Haemophilus* in which bacteria recognized as pathogenic to bees but not associated with the presence of varroa were found. Indeed, there are numerous studies related to the antimicrobial properties of honey with references to *Haemophilus* (i.e. Jeffrey & Echarreta 1996; Antunez et al. 2008; Al-Waili et al. 2011).

Proteo-4 (genus *Pseudomonas*) is strongly present in varroa, probably because it is common on the mite cuticle (Tang et al. 2012), which had not been removed in our study.

Lineage L387 (Bacillaceae Firmi-6, -8) includes OTUs associated with honeybee (Mohr & Tebbe 2006; Evans & Schwarz 2011; Moran et al. 2012), but here we found associated mainly with varroa. It is possible that these bacteria may be generalist present in the hive.

In conclusion, we would like to underline that the partitioning of phylogenetic diversity is a powerful method to analyse community diversity. Using PPD, it is possible to evaluate the different analyses using the percentage of overlap across groups as a comparable statistic. The use of relative read frequencies as distance among samples, instead of the summary statistics typical of DMAs, allows to identify which lineages, or groups of lineages, generate the significant differences. Furthermore, the permutation procedure on all sequences prevents any subsampling procedure (i.e. rarefaction), a practice that has been recently criticized because it reduces the resolution power hiding the signal coming from rare OTUs (McMurdie & Holmes 2014).

Recently, Chiu et al. (2014) proposed a different formulation for alpha diversity than the one proposed in

Jost (2006, 2007). Here, the alpha entropy becomes the joint entropy of observation and environment minus the logarithm of the number of environments. The new formulation allows beta diversity to reach its theoretical maximum (number of environments) whatever unbalanced design is used, but it causes the lower bound of the beta diversity to be higher than 1 depending on sampling design. Furthermore, the alpha diversity does not match anymore the concept of mean diversity within each environment. Consequently, we are not following this new formulation because it is distant from the canonical information theory and further evaluations are needed.

A brief example could illustrate the reason of our preference. Let us assume two communities with the same six equal abundant species, but one locality has 100 observations and the other 1000. According to our definitions, alpha and beta diversities values are, respectively, 6 and 1, while according to Chiu et al. (2014) definition corresponding values are 4.07 and 1.475. We think much more logic, and close to the original information theory realm, the values of 6 and 1 that reflect the fact that in each environment we expect to find six species and that the two samples behave as 1. Chiu et al. (2014) could oppose that using a very similar example, but with the six species being different in the two environments, the value of beta diversity would be 1.35 for the formulation of this article and 2.00 for theirs. This apparent mismatch could be corrected by realizing that the real maximum beta diversity is given by the exponential of the entropy of the sampling vector that is exactly 1.35.

Partitioning phylogenetic diversity method was here implemented in our newly proposed tool PHYLOH. We tested our analytic software using large trees (i.e. constituted by several thousand tips) with no significant delay. The only problem we observed is related to the number of permutation that is proportional to the number of observations following the rule of $N \log(N)$ permutation for N observations. There are no possibilities to avoid this, and several thousands of observations require hours of computations. However, the service, given by the National Institute for Nuclear Physics (INFN), allows 24 h as maximal running time, sufficient to cope with medium-large data sets. Larger data sets could be handled downloading the software and using it in local computation facilities.

The software is not parallelizable, and replicates cannot be distributed on several CPUs. We do not think that parallelization is needed, but we will work on distributing permutations on several CPUs. A complete tutorial of PHYLOH is available on Bioveit web page. Input data used in the present article are available on DRYAD entry doi:10.5061/dryad.j4d15.

On the whole, we think that PHYLOH will be a relevant innovation to study community diversity, a field where data analysis is complex, but representation of results is even more difficult. Indeed, PHYLOH produces clear vectorial graphical outputs, in html format, that are easy to explore using functions such as text finder and zoom without reducing image quality. PHYLOH shows its performances when large amounts of inter-linked data are available. This is the case of the researches involving high-throughput DNA sequencing, such as our pilot study on microbiomes. Furthermore, it can be used whenever a depth analysis of diversity distribution, from microorganisms to macroorganisms, is required.

Acknowledgements

This work was supported by Fondazione Cariplo Grant C91H09000010003 with the project entitled 'Insetti Pronubi: mezzi di connessione e diffusione di specie vegetali rare ed endemiche del parco regionale della Grigna Settentrionale'. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are thankful to BIOVEL (FP7 project no. 283359) for funding and Stefano Pagnotta (University of Sannio) for early discussion on the use of phylogenetic entropy. A special thank is for all the bee-keepers who gave us the possibility to obtain biological samples from their hives. Finally, we would like to thank the anonymous reviewers for their helpful and stimulating comments.

References

- Al-Waili NS, Salom K, Butler G, Al Ghamdi AA (2011) Honey and microbial infections: a review supporting the use of honey for microbial control. *Journal of Medicinal Food*, 14, 1079–1096.
- Annoscia D, Del Piccolo F, Nazzi F (2012) How does the mite *Varroa destructor* kill the honeybee *Apis mellifera*? Alteration of cuticular hydrocarbons and water loss in infested honeybees. *Journal of Insect Physiology*, 58, 1548–1555.
- Antunez K, Harriet J, Gende L, Maggi M, Eguaras M, Zunino P (2008) Efficacy of natural propolis extract in the control of American Foulbrood. *Veterinary Microbiology*, 131, 324–331.
- Blow N (2008) Metagenomics: exploring unseen communities. *Nature*, 453, 687–690.
- Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW (2012) Fast, accurate error correction of amplicon pyrosequences using Acacia. *Nature Methods*, 9, 425–426.
- Buresova V, Franta Z, Kopacek P (2006) A comparison of *Chryseobacterium indologenes* pathogenicity to the soft tick *Ornithodoros moubata* and hard tick *Ixodes ricinus*. *Journal of Invertebrate Pathology*, 93, 96–104.
- Candela M, Biagi E, Maccaferri S, Turroni S, Brigidi P (2012) Intestinal microbiota is a plastic factor responding to environmental changes. *Trends in Microbiology*, 20, 385–391.
- Caporaso JG, Kuczynski J, Stombaugh J et al. (2010) QIIME allows analysis of high throughput community sequencing data. *Nature Methods*, 7, 335–336.
- Cavalli Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, 19(3 Pt 1), 233.
- Chao A, Chiu CH, Jost L (2010) Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 3599–3609.
- Chiu CH, Jost L, Chao A (2014) Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, 84, 21–44.
- Cornman RS, Schatz MC, Johnston JS et al. (2010) Genomic survey of the ectoparasitic mite *Varroa destructor*, a major pest of the honey bee *Apis mellifera*. *Bmc Genomics*, 11, 602.
- Dillon RJ, Dillon VM (2004) The gut bacteria of insects: nonpathogenic interactions. *Annual Reviews in Entomology*, 49, 71–92.
- Dimitriu PA, Boyce G, Samarakoon A, Hartmann M, Johnson P, Mohn WW (2013) Temporal stability of the mouse gut microbiota in relation to innate and adaptive immunity. *Environmental Microbiology Reports*, 5, 200–210.
- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27, 2194–2200.
- Evans JD, Schwarz RS (2011) Bees brought to their knees: microbes affecting honey bee health. *Trends in Microbiology*, 19, 614–620.
- Evans GE, Murdoch DR, Anderson TP, Potter HC, George PM, Chambers ST (2003) Contamination of Qiagen DNA extraction kits with *Legionella* DNA. *Journal of Clinical Microbiology*, 41, 3452–3453.
- Forsgren E (2010) European foulbrood in honey bees. *Journal of Invertebrate Pathology*, 103, S5S9.
- Guzman-Novoa E, Emsen B, Unger P, EspinosaMontano LG, Petukhova T (2012) Genotypic variability and relationships between mite infestation levels, mite damage, grooming intensity, and removal of *Varroa destructor* mites in selected strains of worker honey bees (*Apis mellifera* L.). *Journal of Invertebrate Pathology*, 110, 314–320.
- Jeffrey AE, Echazarreta CM (1996) Medical uses of honey. *Revista Biomedica*, 7, 4349.
- Jeyaprakash A, Hoy MA, Allsopp MH (2003) Bacterial diversity in worker adults of *Apis mellifera capensis*, *Apis mellifera scutellata* (Insecta: Hymenoptera) assessed using 16S rRNA sequences. *Journal of Invertebrate Pathology*, 84, 96–103.
- Jost L (2006) Entropy and diversity. *Oikos*, 113, 363–375.
- Jost L (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, 88, 2427–2439.
- Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271–280.
- Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23, 127–128.
- Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73, 1576–1585.
- Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R (2011) UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal*, 5, 169.
- Lozupone C, Faust K, Raes J et al. (2012) Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Research*, 22, 1974–1984.
- MacKay DJ (2003) *Information Theory, Inference, and Learning Algorithms*, vol. 7. Cambridge University Press, Cambridge.
- Marcon E, Herault B, Baraloto C, Lang G (2012) The decomposition of Shannon's entropy and a confidence interval for beta diversity. *Oikos*, 121, 516–522.
- Martinson VG, Danforth BN, Minckley RL, Rueppell O, Tingey S, Moran NA (2011) A simple and distinctive microbiota associated with honey bees and bumble bees. *Molecular Ecology*, 20, 619–628.
- Martinson VG, Moy J, Moran NA (2012) Establishment of characteristic gut bacteria during development of the honeybee worker. *Applied and Environmental Microbiology*, 78, 2830–2840.
- Mattila HR, Rios D, WalkerSperling VE, Roeselers G, Newton IL (2012) Characterization of the active microbiotas associated with honey bees

- reveals healthier and broader communities when colonies are genetically diverse. *PLoS One*, 7, e32962.
- Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, 122, 107–118.
- McDonald D, Price MN, Goodrich J et al. (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6, 610–618.
- McMurdie PJ, Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8, e61217.
- McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10, e1003531.
- Meriweather M, Matthews S, Rio R, Baucom RS (2013) A 454 Survey reveals the community composition and core microbiome of the common bed bug (*Cimex lectularius*) across an urban landscape. *PLoS One*, 8, e61465.
- Metzker ML (2009) Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11, 31–46.
- Möckel N, Gisder S, Genersch E (2011) Horizontal transmission of deformed wing virus: pathological consequences in adult bees (*Apis mellifera*) depend on the transmission route. *Journal of General Virology*, 92, 370–377.
- Mohammadi T, Reesink HW, Vandenbroucke-Grauls CM, Savelkoul PH (2005) Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *Journal of Microbiological Methods*, 61, 285–288.
- Mohr KI, Tebbe CC (2006) Diversity and phylotype consistency of bacteria in the guts of three bee species (*Apoidea*) at an oilseed rape field. *Environmental Microbiology*, 8, 258–272.
- Moran NA, Hansen AK, Powell JE, Sabree ZL (2012) Distinctive gut microbiota of honey bees assessed using deep sampling from individual worker bees. *PLoS One*, 7, e36393.
- Mouches C, Bove JM, Albisetti J (1984) Pathogenicity of *Spiroplasma apis* and other spiroplasmas for honey-bees in Southwestern France. *Annales de l'Institut Pasteur/Microbiologie*, 135, 151–155.
- Oksanen J, Kindt R, Legendre P, O'Hara B (2007) *Vegan: community ecology package* R package version 1.8-5 Available from <http://cran.r-project.org/> (accessed 07 November 2014).
- Orlaci L (1967) An agglomerative method for classification of plant communities. *The Journal of Ecology*, 55, 193–206.
- Price MN, Dehal PS, Arkin AP (2010) *FASTTREE 2*—approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, e9490.
- R Development Core Team (2012) *R: A language and environment for statistical computing*. R Development Core Team, Vienna, Austria. Available from <http://www.r-project.org/> (accessed 07 November 2014).
- Rinderer TE, Harris JW, Hunt GJ, de Guzman LI (2010) Breeding for resistance to *Varroa destructor* in North America. *Apidologie*, 41, 409–424.
- Rosenkranz P, Aumeier P, Ziegelmann B (2010) Biology and control of *Varroa destructor*. *Journal of Invertebrate Pathology*, 103, S96–S119.
- Sabree ZL, Hansen AK, Moran NA (2012) Independent studies using deep sequencing resolve the same set of core bacterial species dominating gut communities of honey bees. *PLoS One*, 7, e41250.
- Sanchez LM, Wong WR, Riener RM, Schulze CJ, Lington RG (2012) Examining the fish microbiome: vertebrate-derived bacteria as an environmental niche for the discovery of unique marine natural products. *PLoS One*, 7, e35398.
- Tang X, Freitag D, Vogel H et al. (2012) Complexity and variability of gut commensal microbiota in polyphagous lepidopteran larvae. *PLoS One*, 7, e36978.
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73, 5261–5267.
- Watanabe K, Kodama Y, Harayama S (2001) Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of Microbiological Methods*, 44, 253–262.
- Webb CO, Ackerly DD, Kembel SW (2008) *PHYLOCOM*: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, 24, 2098–2100.
- Wickham H (2009) *ggplot2: Elegant graphics for data analysis*. Available from <http://cran.r-project.org/> (accessed 07 November 2014).
- Wolstencroft K, Haines R, Fellows D et al. (2013) The TAVERNA workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41, W557.
- Zhou J, Wu L, Deng Y et al. (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME Journal*, 5, 1303–1313.

A.S., S.V. and M.C. wrote the manuscript. E.M.P., A.G., E.F., A.B. and B.B. implemented the manuscript. E.M.P., E.F. and M.C. designed the project. S.V. designed and implemented the software *PHYLOH*. A.S., S.V. and B.B. evaluated the codes. A.S., A.G., E.F., A.B., V.M. and M.C. provided biological background.

Data Accessibility

DNA sequences: NCBI SRA: SRP046312.

Source code of *PHYLOH*, user manual and example data set are available on github: <https://github.com/svica rio/phyloH>.

Access to the web page of *PHYLOH*: https://www.bi odiversitycatalogue.org/rest_methods/143.

Input and output data of *PHYLOH*, Supporting information Figures and Tables: DRYAD entry doi:10.5061/ dryad.j4d15.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 The figure depicts how Shannon entropy is totally blind to some meaningful biological pattern.

Fig. S2 Distribution of OTUs of High Frequency Cluster (HFC) between parasitised honey bee and *varroa* considering relative abundance.

Fig. S3 Distribution of OTUs of High Frequency Cluster (HFC) between healthy and parasitised honey bee and *varroa*.

Fig. S4 UPMGA trees showing the different relationships among samples for all the distance metrics considered in DMA (jaccard, squared-chord, unweighted UniFrac and weighted UniFrac) considering Localities and Status variables.

Fig. S5 Output of the script `compute_core_microbiome.py`.

Table S1 Summary of all OTUs taxon assignment.

Table S2 Summary of results of adonis statistic.

Appendix S1 *PHYLOH* output in html format for 'Status' variable.

Appendix S2 Phylogenetic tree used in *PHYLOH* analysis in Newick tree format.