The virtualome: a computational framework to evaluate microbiome analyses

Belén Serrano-Antón[1,§], Francisco Rodríguez-Ventura[1], Pere Colomer-Vidal[1], Riccardo Aiese Cigliano[2,*], Clemente F. Arias[1,3,*], and Federica Bertocchini[1,*]

[1]CIB, Centro de Investigaciones Biológicas Margarita Salas (CSIC), 28040 Madrid, Spain
[2]Sequentia Biotech SL, Barcelona, Spain
[3]Grupo Interdisciplinar de Sistemas Complejos de Madrid (GISC), Spain

[§]Current address:
- FlowReserve Labs SL, 15782 Santiago de Compostela, Spain
- Group of Nonlinear Physics. University of Santiago de Compostela, 15782 Santiago de Compostela, Spain

[*]Corresponding authors:
RAC: raiesecigliano@sequentiabiotech.com
CFA: tifar@ucm.es
FB: federica.bertocchini@csic.es

ABSTRACT

Microbiomes have been the focus of a substantial research effort in the last decades. The composition of microbial populations is normally determined by comparing DNA sequences sampled from those populations with the sequences stored in genomic databases. Therefore, the amount of information available in databanks should be expected to constrain the accuracy of microbiome analyses. Albeit normally ignored in microbiome studies, this constraint could severely compromise the reliability of microbiome data. To test this hypothesis, we generated virtualomes, virtual bacterial populations that exhibit the ecological structure of real-world microbiomes. Confronting the analyses of virtualomes with their original composition revealed critical issues in the current approach to characterizing microbiomes, issues that were empirically confirmed by analyzing the microbiome of *Galleria mellonella* larvae. To reduce the uncertainty of microbiome data, the effort in the field must be channeled towards significantly increasing the amount of available genomic information.

Keywords: Microbiome; Amplicon;Whole-genome sequencing; Metagenomics; Next-generation sequencing; Bacteria; Insects; *Galleria mellonella*

1 Introduction

The characterization of the microorganisms colonizing a particular ambient is becoming a gateway to the analysis of the physiological niche that the environment represents, revealing its potential functions or eventual pathological conditions. Examples in this direction are represented by the deep interest in the human gut microbiome (the compendium of microorganisms colonizing the human gut), due to the growing concern in the relationships between the microbiome and the immune system, and henceforth in the potential development of disease [1–4]; or the increasing focus on the genomic analyses of water (ocean or river) and soil samples, in search for potentially useful cataloging of the environmental niches we live in [5].

Microbiome studies are receiving increasing attention for their possible implications in the field of bioremediation, embracing issues such as degradation of organic chemicals, conversion of toxic compounds (e.g. pesticides), production of biofuels, or else from various substrates [6–9]. The strong interest in communities of microorganisms crossed fields and extended to the studies of insect guts [10, 11]. A growing body of research focuses on insect microbiomes in the quest for solutions within the bioremediation field [12]. For example, termites and beetles have raised interest due to their capacity to degrade lignin, an ability potentially dependent in some cases on

microorganisms colonizing certain portions of their intestine [13–17]. Recently, some coleopteran and lepidopteran species revealed the astonishing capability to degrade fossil fuel-derived plastics, like the sturdy polyethylene and polystyrene [18–21], opening up a new niche within the field of bioremediation by insects. Even if the molecular mechanisms responsible for this extraordinary capacity are still unknown, they are normally ascribed to the microorganisms colonizing the digestive tracts of those insects. This line of research has resulted in an ever-increasing list of microorganisms with the potential to biodegrade plastics, although with still un-concluding outcomes [20–26].

Studies in this field typically monitor the changes induced by alternative treatments in the relative abundance of the species present in the microbiome. For instance, feeding insects a diet of plastic is a standard procedure to identify the microorganisms that thrive in their gut as potential candidates for plastic metabolization [18, 20, 23, 26–28]. Analogously, the microbiome of patients suffering a given clinical condition is often compared to that of healthy control individuals, expecting the shifts in the relative abundance of microbial species to account for the observed effects or to provide effective diagnostic tools [29]. This functional approach to the study of microbiomes relies on several implicit assumptions, whose general validity is far from evident. First, a causal link is presumed between changing conditions and the differential selection (either positive or negative) of particular microbial species. In the case of plastic fed-insects, for instance, possible metabolic changes induced by a diet of plastic in the host insect are normally neglected, implying that the microorganisms colonizing the digestive tracts are considered responsible for any metabolic activity the animal embarks on.

Leaving aside its biological plausibility, this perspective takes for granted other assumptions that have to do with the methods used to study microbiomes. These methodological assumptions are the focus of this work. It is normally presupposed that currently available techniques are accurate enough to provide reliable quantitative measurements of changes in microbiome composition. Consequently, observed variations in the abundance of species (or other relevant taxa) are regarded as valid indicators of the activity of interest. A close inspection of the currently available empirical tools reveals unexpected and critical drawbacks in this approach.

The most used methodologies to study the microbiome of a chosen animal species are amplicon and whole-genome sequencing (WGS). The former is based on the identification of taxonomical markers such as the ribosomal gene 16S for bacteria. Although this relatively short gene($\sim$ 1500 bp) is universal among bacteria and archaea [30], some studies have revealed that 16S rRNA gene does not show precise phylogenetic relationships within particular taxa [31, 32]. For this reason, in recent years there has been an increase in the number of analyses carried out with WGS, which is a more expensive and computationally intensive technique than amplicon.

Amplicon and WGS analyses rely on the comparison of DNA sequences obtained from a sample of the microbial population of interest with the sequences contained in genomic databanks. Therefore, they can only identify those sequences that are already present in databases. In addition, whole genome sequencing data can also be used to generate genome assemblies of the microbial populations, the so-called MAGs (Metagenome Assembled Genomes) which can be used to reconstruct the genome of new species. However, also in this case, databases of known species are used to try to classify the taxonomy of the MAGs. As a consequence, the accuracy of these methods should be expected to depend on the number of sequences contained in the databases used for the analysis.

Remarkably, comparative studies have shown that the results of the characterization of bacterial populations by amplicon and WGS do not necessarily overlap even if they use the same reference databanks. For example, applying both techniques on the same DNA from a human fecal sample resulted in a mere 29% overlapping at the phylum level [33], and revealed differences in the metagenomic results for a human salivary sample, both in the relative abundances of species and in the identified genera [34]. Even the same method (WGS) analyzed with different pipelines yielded results that differ by up to three orders of magnitude on the same data set [35].

In addition to the issues related to the methodology used to characterize microbiomes, the ecological structure of bacterial populations (e.g. diversity or richness of species) and the sampling strategy can also bias their characterization, adding further concerns about the accuracy of microbiome analyses [36, 37]. This scenario raises the following issues: To what extent and under what circumstances can we count on the current experimental approaches to understanding the complexity of symbiotic microorganisms living within the gut (or any other tissue) of an animal? Do the available techniques have intrinsic limitations that might influence the analysis of any

microbiome, whatever the colonized environment? Or are these limitations restricted to particular habitats?

To get a deep insight into this puzzling landscape, we took an *in-silico* approach, creating what we call virtualomes: virtual microbiome communities that simulate the bacterial populations that can be found in humans, insects, etc., or soil, water, and any other medium. These virtual communities can be examined by standard genomic techniques, giving the possibility to confront the results of the analysis with the original microbiome composition.

Within the virtualome framework, it is possible to simulate the analysis of microbial populations with partially underrepresented or limited databanks, and hence to explore the impact of the lack of genomic information on the characterization of microbiomes. This approach allows identifying the incompleteness of databases as a key constraint to the accuracy of microbiome analyses. The poor overlap between amplicon and WGS already encountered with human samples does not seem to be an exception. Furthermore, the simultaneous detection of a species by both techniques does not ensure its positive identification, since the overlap between amplicon and WGS can be dominated by false positives, i.e. by species that are not present in the original samples. This counterintuitive situation worsens notably when limited databases are used in the analyses to simulate the lack of genomic information. In this case, the detection of changes in the relative abundance of bacterial groups can be profoundly misleading. The variations in abundance detected by genomic analyses can be mostly spurious and unrelated to the actual changes taking place in the analyzed virtualomes.

To empirically test the predictions derived from our model, we proceeded with the analysis of the bacteria colonizing different tissues of the larvae of *Galleria mellonella*, a lepidopteran recognized as capable of degrading polyethylene and polystyrene [19, 23–25, 38–41]. Despite their abundance, the microbiota of lepidopteran species has been little studied [42, 43], so the bacteria associated with this group have only a marginal representation in databanks to date. This makes *G. mellonella* larvae an ideal subject to examine the limitations of genomic analyses pinpointed by the virtualome model.

To do that, we applied amplicon sequencing (16S) and WGS (using the same DNA samples) to the anterior part, the gut, and the silk glands of plastic-treated and control larvae. The results reflected the incongruences revealed by our model, with an astonishing lack of overlap between the outcomes of the two techniques at the genus level. Both techniques also showed striking discrepancies in the detection of changes in the insect microbiome.

For those animal species and environments that harbor microbial populations [44], microbiome analysis stands as a fundamental tool in the comprehension of their physiology and ecology. However, the shortcomings inherent to the used tools and the paucity in the available database hinder and mislead the outcome, and, therefore, the interpretation of microbiome analyses. This work highlights the need for an increased effort in the collection of genomic information and its eventual availability in public databases. In the meantime, the results of experimental studies about the composition or the dynamics of microbiome populations should be interpreted with caution, especially in the case of insects, where the available genomic information is still scarce.

2 Results

2.1 Virtualome models

Virtualomes are models of bacterial populations conceived to evaluate the performance of genomic analyses. Each virtualome consists of a list of species of bacteria and their respective abundances. The ecological structure of virtualomes (i.e. the taxonomic relationships among species and their relative abundance) was designed to comply with the macro-ecological rules described for real-world microbial communities (see [36] and Supplementary Material SM1).

To study how the incompleteness of genomic databanks affects microbiome analyses, we created restricted (incomplete) databases in which some families, genera, or species were removed from the complete database (i.e. the one containing all the prokaryotes genetic sequences available to date). This allowed us to simulate different scenarios in which some of the species present in a microbiome have not been previously cataloged. In particular, we generated virtualomes for which 25%, 50% and 100% of the species are represented in the incomplete databases.

Simulated amplicon sequencing and WGS reads were then generated starting from the virtualomes using complete and incomplete databases and different numbers of reads (see Methods). Thus, for each virtualome, we

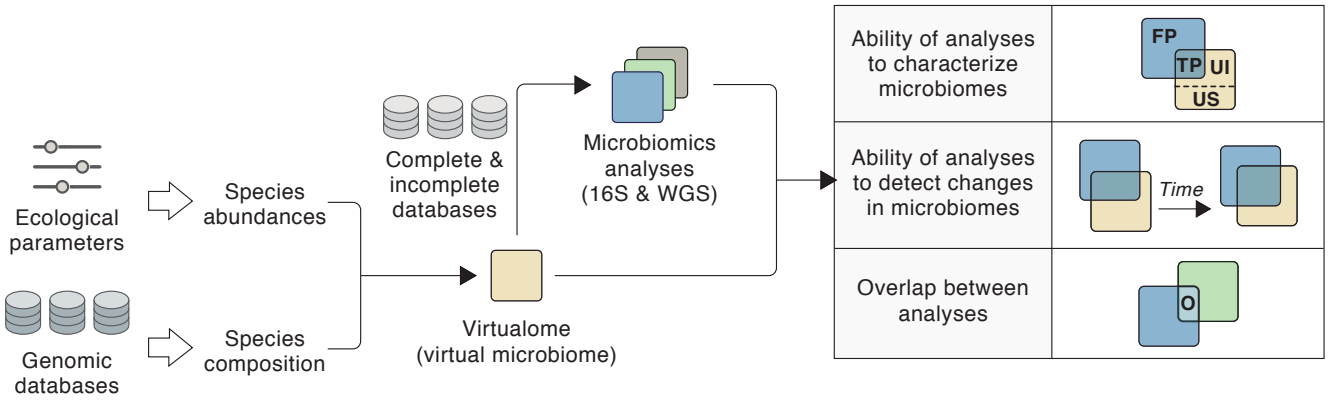

Figure 1. Rationale of the virtualome framework. See explanations in the text. (FP: false positives; TP: true positives; UI: unidentified; US: unsampled; O: overlap.)

obtained several characterizations that could be compared with the original virtualome composition and with one another (see Fig. 1). First, we evaluated how the incompleteness of genomic databanks affects the accuracy of microbiome analyses using the following variables:

1. Unsampled species and genera: percentage of species and genera in the virtualome whose abundance is below the threshold of detection imposed by the number of reads used in the analysis.
2. Unidentified species and genera: percentage of the virtualome species and genera that were sampled but did not match any of the sequences contained in the database used in the analysis.
3. False positives: percentage of the total of species and genera detected by the analysis that were not present in the original virtualome.
4. True positives: percentage of the virtualome species and genera that were detected by the analysis.
5. Abundance of species and genera predicted by the metagenomic analysis.

Second, we studied the ability of microbiome analyses to detect changes in the abundance of species and genera in the virtualomes. To that end, we compared virtualomes containing the same set of species but differing in their relative abundance, a scenario that simulates the temporal changes in a microbiome caused, for instance, by alternative treatments on the host. Finally, we measured the overlap between 16S and WGS analyses (i.e. the species and genera simultaneously detected by both techniques) when performed on the same virtualomes.

2.2 Effect of the incompleteness of databases in the characterization of virtualomes by microbiome analyses

The first step of microbiome analyses consists in sampling the microbial populations of interest. The limited size of samples imposes obvious constraints to the subsequent genomic analyses of microbiomes: these analyses are performed only on the fraction of the microorganisms present in the microbiome that are actually sampled, which may hinder the identification of species with lower abundances [36]. To test the effect of sampling on the performance of microbiome analysis, we generated virtualomes with different numbers of species. We found that in small, highly diverse communities, increasing the number of reads ensures that most of the species are sampled (Fig. 2A left). In contrast, in low-diversity communities, a greater number of reads only led to moderate growth in the percentage of species detected (light grey lines, corresponding to this type of communities, do not fall below 80% in 2A).

This effect was more pronounced in larger populations, in which greater sampling efforts did not necessarily result in a substantial reduction in the number of unsampled species, even when their ecological diversity was high

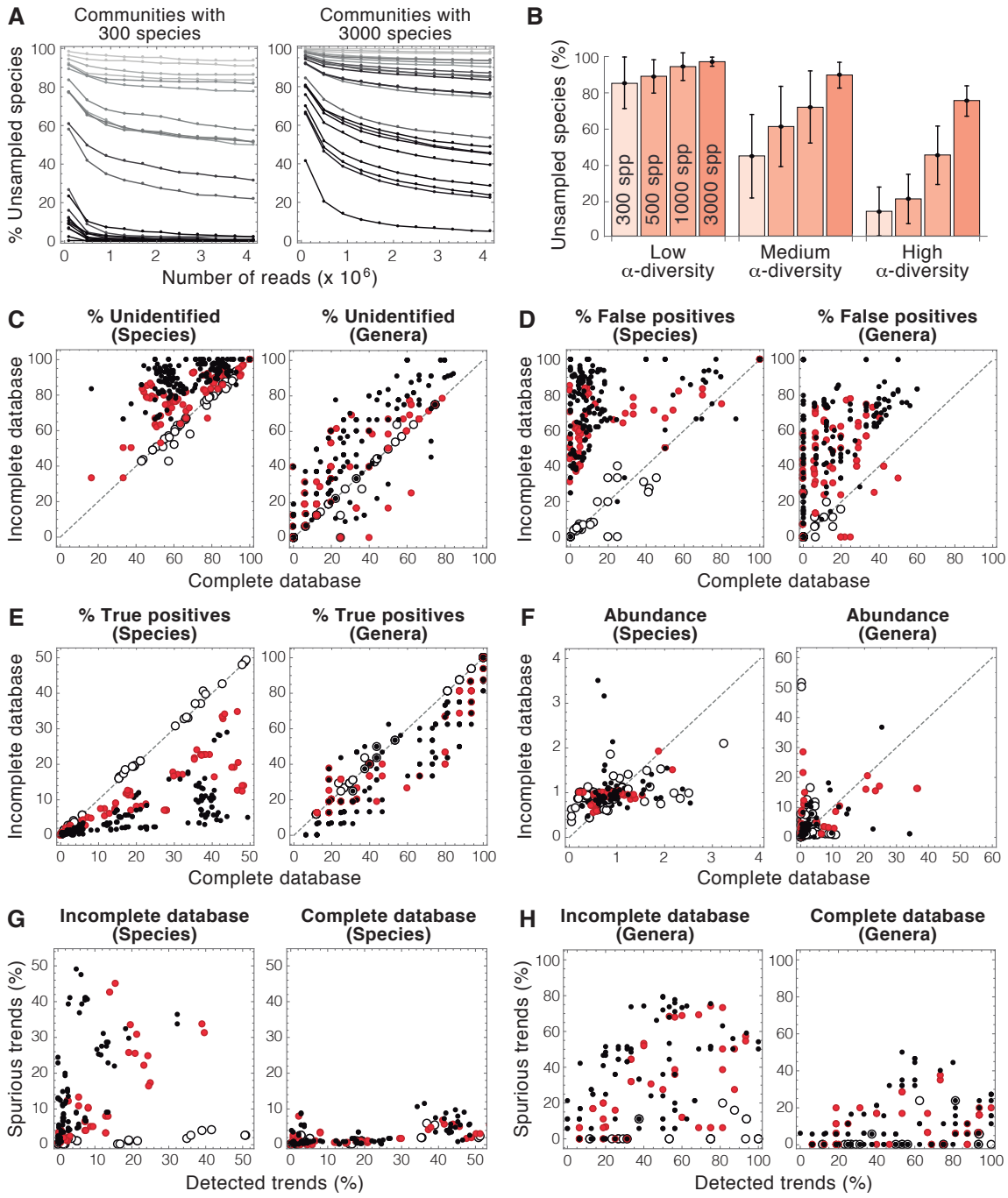

Figure 2. Evaluation of virtualome characterization by microbiome analyses. A-B) Effect of ecological parameters (richness of species and α-diversity) in the percentage of unsampled species in microbiome analyses. Darker shades of gray in A indicate greater values of α-diversity. α-diversity was calculated using the Shannon index. C-E) Effect of the incompleteness of microbiome databases on the fraction of unidentified species (C), false positives (D), and true positives (E) at the species (left) and genus (right) level. F) Effect of the incompleteness of databases on the difference between detected and actual abundances of species (left) and genera (right). This difference is defined as $D = 1/N \sum_{i=1}^{N} (\tilde{a}_i - a_i)/a_i$, where N is the number of detected species or genera, and a_i and $\tilde{a}_i$ are the actual and estimated abundances of species or genus i respectively. G-H) Ability of microbiome analysis to detect changes in the abundance of species (G) and genera (H) when incomplete (left) or complete (right) databases are used in the analysis. On the abscissa axis, the percentage of actual changes in abundance in the virtualome detected by the analysis. On the ordinate, the fraction of the changes in abundance found by the analysis that does not correspond to actual changes in the virtualome. (Open dots: 100% of the species or genera of the virtualome are in the databases; Red dots: 50% of species or genera in DBs; Black dots: 25% of species or genera in DBs. Dashed lines: $x = y$.)

(Fig. 2A right). The ecological structure of microbial communities (in terms of richness of species and diversity) imposes a major intrinsic limitation to the capacity of microbiome analyses, as shown by the loss of over 80% of the species in virtualomes with low diversity (Fig.2B).

In the subsequent evaluation of microbiome analyses, we used small virtualomes (containing only 300 species) to minimize the loss of species due to sampling. We began by examining the ability of microbiome analyses to identify sampled species and genera. The lack of identification occurs when the DNA sequences generated from a given sample do not match any of the sequences stored in the databanks. In the case of amplicon analysis, species identification is hindered by an additional factor. The region targeted by the 16S primer may be missing or unrecognizable, which prevents the amplification of the target region. For virtualomes with high diversity, between 8% and 14% of the species were lost for this reason before performing the amplicon analysis (data not shown).

The fraction of the sampled species that are unidentified by genomic analyses was above 20% in all the studied virtualomes and it was normally greater in analyses using incomplete databases (Fig. 2C left). As should be expected, the database used in the analysis does not affect the number of unidentified species for those virtualomes with all their species represented in the restricted databases (as shown by the disposition of open circles along the diagonal in Fig. 2C). The scenarios with more species absent from the incomplete databases lie above the diagonal, meaning that the lack of genomic information increases the fraction of unidentified species. In extreme cases (i.e. some of the virtualomes with 75% of the species absent from databanks), none of the sampled species were identified in the analysis. The percentage of unidentified genera was more heterogeneous than that of species, varying between 0% and 80% in the studied virtualomes (Fig. 2C right).

Unexpectedly, the incompleteness of genomic databanks also led to a sharp increase in the detection of false positives (percentage of detected species or genera detected that are not present in the original virtualome). Fig. 2D shows that most of the analyses yielded high fractions of false positives (both at the species and genus levels) even when they were performed using the complete databank. This situation dramatically worsened for virtualomes containing new species (corresponding to analyses with incomplete databases), in which false positives may represent more than half of the identified species and genera (Fig. 2D). Remarkably, the incompleteness of the databanks affected the analyses of virtualomes even when the 100% of their species were represented in the incomplete databases (the open dots do not lie in the diagonal in Fig. 2D).

We next analyzed the ability of genomic analysis to detect true positives (percentage of the species and genera of the virtualome successfully identified in genomic analyses). This percentage was below 50% in all the virtualomes at the species level (Fig. 2E). As should be expected, the detection of true positives increased when larger fractions of the virtualome species were included in the genomic databases used in the analysis. At the genus level, the detection of true positives was highly variable, ranging from around 10% to 100% in the studied virtualomes.

Genomic analyses of virtualomes provided very poor estimates of the abundance of species and genera. The differences between the actual abundance in the virtualome and that observed by the analyses spanned several orders of magnitude (Fig. 2F). Remarkably, the results of this analysis were worse at the genus than at the species level, contrary to the characterization of unidentified, true positives, and false positives.

In view of their inability to accurately estimate the abundance of species and genera, it was not surprising that genomic analyses normally failed to detect changes in that abundance. We measured the success of the analyses to identify trends (i.e. increases or decreases in the abundance of species and genera), regardless of their magnitude. When incomplete databases were used in the analysis, most of the observed trends were spurious, i.e. they did not correspond to actual changes in species abundance in the virtualomes (Fig. 2G). Using complete databases greatly reduced the detection of false trends. In any case, less than half of the actual changes in abundance taking place in the virtualomes were detected by genomic analyses. The detection of both actual and spurious trends was greater at the genus level (Fig. 2G).

2.3 Overlap between 16S and WGS analysis of virtualomes

To study the differences and similitudes between 16S and WGS analyses, we began by comparing their results when applied to the characterization of the same virtualomes. As a general rule, WGS outperformed 16S in the identification of true positives but 16S analyses detected fewer false positives (Supp. Fig. S1). Owing to the

problems derived from the use of 16S primers discussed above, the fractions of unidentified species and genera were greater for amplicon than for WGS analyses (Supp. Fig. S1).

The overlap between 16S and WGS at the species level was below 50% in all the studied virtualomes. In the analyses with less species present in the incomplete databases, this overlap did not reach 10% (Fig. 3A). The coincidence between the results of 16S and WGS greatly increased with the amount of information available in the genomic databases. In contrast, using more reads in the analysis did not lead to better 16S-WGS fits (Fig. 3B). Unexpectedly, most of the species simultaneously detected by 16S and WGS using incomplete databases were false positives (Fig. 3C). Similar behaviors were observed in the analyses at the genus level (Figs. 3D-F).

These results show that the simultaneous detection of a species or genus by 16S and WGS does not provide additional information about its presence in the virtualome. The list of species and genera found in the intersection between both analyses does not characterize virtualomes more accurately than those found separately by each analysis.

Interestingly, each technique proved capable of identifying species and genera that could not be found by the other (Supp. Fig. S2.A). In this regard, the number of true positives detected exclusively by WGS was normally greater than those obtained only by 16S (Supp. Fig. S2.B). Still, some species and genera identified by 16S were not detected in WGS analysis. The number of true positives detected only by WGS was greater in analyses with incomplete databases. However, WGS also found more false positives than 16S, an effect that was more pronounced when incomplete databases were used in the analysis (Supp. Fig. S2.B).

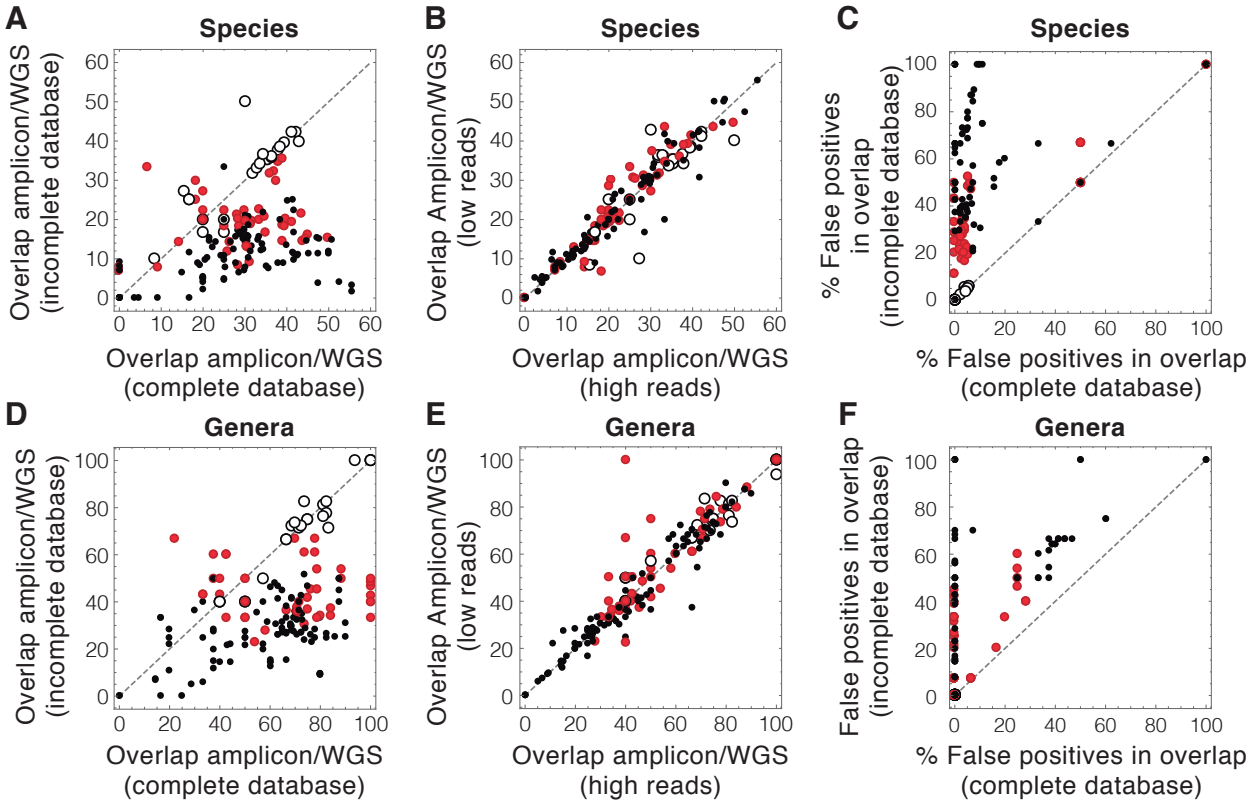

Figure 3. Overlap between 16S and WGS analyses of virtualomes. A) Percentage of the species detected by either 16S or WGS that are simultaneously detected by both techniques. B) Effect of the number of reads on the overlap between 16S and WGS. C) Percentage of the species simultaneously detected by 16S and WGS that are not originally present in the virtualome. D-F) Same as A-C for analyses at the genus level. Remark that all the open dots lie at the origin in F. (Open dots: 100% of the species or genera of the virtualome are in the databases; Red dots: 50% of species or genera in DBs; Black dots: 25% of species or genera in DBs. Dashed lines: $x = y$.)

Altogether, our results point to a very limited ability of amplicon and WGS to accurately characterize virtualomes, a conclusion that can be extrapolated to the study of real-world microbial communities. Since it is obviously impossible to know a priori if a given microbiome is underrepresented in the databanks used in its characterization, there is no way to judge the fraction of false positives found in the analysis, or how many of the taxa present in the microbiome are not identified or even sampled. This entails a high degree of uncertainty in the analysis of microbial communities, an aspect that should be explicitly taken into account in the interpretation of microbiome data. We address this issue in the next section.

2.4 Empirical test of the virtualome predictions in the microbiome of *Galleria mellonella*.

To test the predictions of the virtualome models, the microbiome of *Galleria mellonella* larvae fed with a control diet or with polyethylene was investigated using amplicon and WGS. Sixty-eight 16S samples were produced from

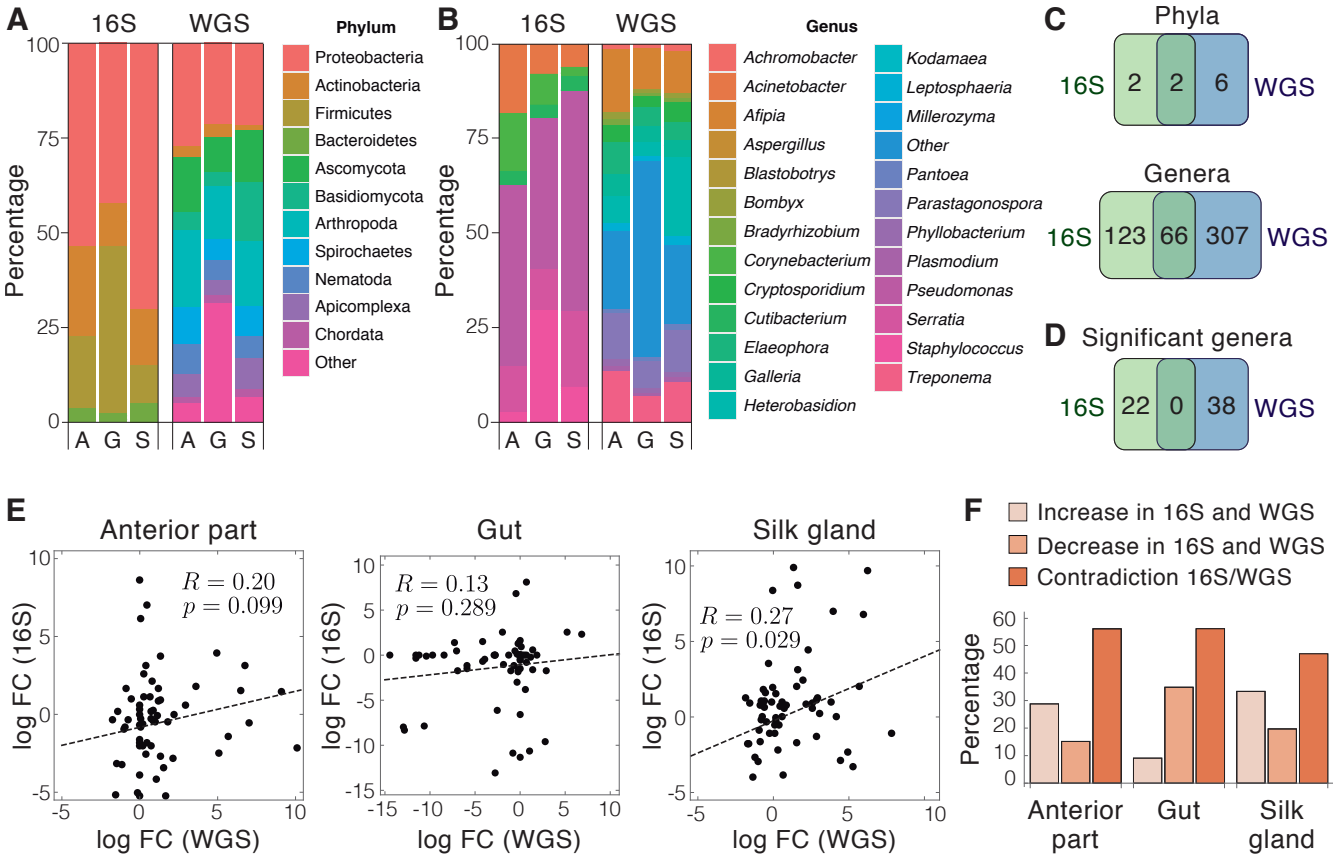

Figure 4. Overlap between 16S and WGS analyses of the microbiome of *Galleria mellonella* larvae. A) Barplot showing the percentage of reads classified in the observed Phyla with 16S and WGS respectively in three tissues of *G. mellonella* (A: anterior part; G: gut; S: silk gland). B) Barplot showing the percentage of reads classified in the observed genera with 16S and WGS respectively in three tissues of *G. mellonella* (A: anterior part; G: gut; S: silk gland). C) Venn Diagrams showing the overlap between 16S and WGS. Upper: overlap of the identified phyla; Lower: overlap of the identified genera. D) Overlap of the genera showing significant differences in abundance between treated and control samples across the gut, anterior part, and salivary glands. E) Scatterplots showing the correlation between $\log_2$ fold changes of genera simultaneously detected by WGS and 16S in the anterior part (left), gut (medium), and salivary glands (left) from treated animals with respect to controls. Pearson correlation coefficient (R) and its significance (p) is reported for each plot. F) Coincidences and discrepancies between changes in abundance detected by 16S and WGS in the anterior part, gut, and silk glands of treated animals with respect to controls.

three tissues, namely Anterior Part (AP), Gut and Silk Glands (SG), control and experimental (11 samples per tissues, plus two more samples for the AP control) producing a total of 22,548,048 MiSeq paired-end 300 bp reads (Supplementary Table ST1). About 50% of the total amount of reads could be classified at Phylum level, with Proteobacteria (52.16% of the reads), Firmicutes (22.97%), and Actinobacteria (15.82%) being the most abundant Phyla, on average (Fig. 4A). On the other hand, about 42% of the reads could be classified at genus level with *Pseudomonas* (22.92%), *Enterococcus* (13.70%), *Staphylococcus* (6.75%), *Serratia* (6.73%), and *Acinetobacter* (4.84%) being the most abundant, on average (Fig. 4B).

Thirteen samples were also processed with a WGS approach producing a total of 734,523,154 paired-end 150 bp reads (Supplementary Table ST2). After removing the reads that mapped on the *G. mellonella* reference genome, only 221,217 (0.06%) pairs could be classified at Phylum level with Proteobacteria (22.84%), Arthropoda (17.05%), Ascomycota (12.63%), Basidiomycota (9.33%), and Firmicutes (8.19%) being the most abundant, on average (Fig. 4A). The presence of Arthropoda suggests that some contamination from the host was still present. At the genus level, a total of 146,444 (0.03%) pairs could be classified, with the most abundant being *Heterobasidion* (Basidiomycota, 12.64%), *Afipia* (Proteobacteria, 12.38%), *Treponema* (Spirochaetes, 10.25%), *Parastagonospora* (Ascomycota, 10.20%), and *Enterococcus* (Firmicutes, 6.53%), on average (Fig. 4B).

As a next step, we compared the results of both sequencing strategies, taking into account that the data was produced from the same samples. For this comparison, we only considered the taxa with a median abundance higher than 1%. At the phylum level, we observed that only 50% of the phyla detected by 16S could be also identified with the WGS strategy. In addition, the WGS approach detected 6 phyla that were not found by the 16S approach. Considering the total amount of 10 detected phyla, only 2 (20%) could be detected by both methods (Fig. 4C). At the genus level, about 35% of the genera detected by 16S sequencing could also be identified by WGS. Considering the total amount of 496 detected genera, only 66 (13.3%) were simultaneously detected by both methods (Fig. 4C).

We then performed differential abundance analysis using 16S and WGS data to detect genera whose abundance was affected by the diet. Each tissue was analyzed separately (Supplementary Tables ST1 and ST2) and gut samples always contained the majority of differential genera. Considering all the tissues, 38 and 22 genera were found to show significant (FDR ≤ 0.05) differences in abundance with 16S and WGS respectively. Strikingly, none of these differential changes was simultaneously detected by both techniques (Fig. 4D).

Even if there was no overlap among significant results, we decided to compare the fold changes detected for common genera between WGS and 16S in order to assess whether the magnitude of change was detected in the same way by the different methods (Fig. 4E). Pearson Correlation of the log2 fold changes in the three tissues ranged from 0.13 to 0.27 between WGS and 16S. Focusing on the trend of the changes (i.e. increase or decrease) without considering their magnitude gave similar results (Fig. 4F), with both techniques detecting opposite trends in about half of the cases. The discrepancy between 16S and WGS is even more pronounced for genera showing significant changes in abundance, since the genera found to be significant are not even the same in both cases. Overall, these results point to the fact that we observed a poor agreement in terms of changes in abundance between WGS and amplicon sequencing methods.

3 Discussion

In this work, we formulate a computational framework to evaluate the performance of metagenomic analyses based on the generation of virtualomes, virtual microbiomes that simulate real bacterial communities. Using this approach, we identified critical limitations in the capability of currently available technologies to characterize microbiomes. The constrains found within the virtualome framework predicted a poor overlap between WGS and amplicon in the analysis of real-world microbiomes underrepresented in genomic databases. This prediction was confirmed by the discrepancies between both methods in the characterization of the bacteria found in *G. mellonella* larvae.

Some of the limitations of microbiome analysis result from intrinsic features of microbial populations (e.g. the loss of species due to sampling is greater in highly diverse communities). These limitations, imposed by the ecological structure of microbial populations, have been discussed elsewhere [36]. Other constraints arise from the lack of genomic information in the databanks used in metagenomic studies. It is trivial that 16S and WGS analyses

cannot detect species that are not represented in genomic databases. Unexpected, however, is that the incompleteness of the available information also leads to an increase in the detection of false positives, which may severely distort the identification of the species present in a microbiome.

It seems natural to assume that applying 16S and WGS methods to characterize the same microbial community should increase the accuracy of the analysis. Intuitively, the detection of the same species by both techniques could be interpreted as the confirmation of its presence in the microbiome. However, our results suggest that this assumption is misleading, since the observed overlap between the results of 16S and WGS can consist mostly of false positives. This result agrees with the poor overlap between 16S and WGS observed in previous studies [33, 34, 45], which, accordingly, could be explained as emerging from the combined effect of the limitations derived from the ecological structure of the microbiome and the insufficiency of the genomic information used in the analyses.

As could be expected, better results can be obtained working at the genus levels, both in terms of microbiome characterization and overlap between 16S and WGS. However, the accuracy of the analyses is still highly dependent on the amount of information present in the databases.

All these limitations critically constrain the successful identification of the bacteria present in different communities. Importantly, they also hinder the detection of the changes in microbial communities induced by external or internal factors. Our results reveal that many of the trends identified at the species level can be spurious. Again, trend detection improves when working at the genus level, but a significant gap still exists between observed and real changes in abundance. This calls into question the approach of studies claiming to find bacteria that proliferate or decline after certain treatments [18, 20, 23, 26–28]. We confirmed this point by studying the microbiome of plastic-fed and control *G. mellonella* larvae. Strikingly, the bacterial genera showing significant changes in abundance in hosts under a diet of plastic were totally different when observed by WGS or amplicon methods. This result questions the widespread experimental strategy of inducing changes in bacterial abundance to detect microorganisms with specific metabolic potentials.

Our results also confirm the utility of the virtualome as a powerful theoretical framework to scrutinize the performance of different techniques of microbiome analysis. The shortcomings of genomic analysis identified within this framework can be addressed in several ways: 1) by working at higher taxonomic levels whenever this is possible. This solution may be insufficient though if further taxonomic precision is needed; 2) continue to accumulate information in genomic databases. This is a fundamental aspect, as the lack of data leads to worse results in terms of characterization and detection of changes in abundance; 3) one must be cautious when interpreting data from metagenomic studies. As has been shown, owing to the limitations inherent to these techniques, their results should not be taken as totally reliable or conclusive.

The ever-growing interest in microbiomes and their potential applications is very much dependent on the reliability, richness, and completeness of the databanks available for their accurate description. If microbiome data are to be useful in an effective and reproducible manner, the effort in the field must be channeled towards significantly increasing the amount of available genomic information.

4 Methods

4.1 Generation of virtualomes

Virtualomes were created with a fixed number of species (300). The process of generating abundance matrices was automated by means of an algorithm that takes as input the relevant parameters of the problem (i.e. the number of populations and species and the shape parameters of the distributions, labelled as *params*) and outputs an abundance matrix (labelled as *abundances*) that complies with the macroecological laws followed by real-world microbial populations (as defined in [36]). A simplified pseudocode of the algorithm is shown below. The resulting distributions were tested using the Kolmogorov-Smirnov test for goodness of fit.

4.2 Genetic composition of virtualomes

The species of virtualomes were chosen from eight families: Clostridiaceae, Flavobacteriaceae, Enterococcaceae, Pseudomonadaceae, Acetobacteraceae, Lactobacillaceae, Enterobacteriaceae and Streptococcaceae. Virtualome

Algorithm 1 Generation of the ecological structure of the virtualomes

procedure ECOLOGYGENERATION($nCommunities, nSpecies, mu, sigma, shape, scale$)
  $MADvector \leftarrow randomLognormal(nSpecies, mu, sigma)$
  $abundances \leftarrow generateAbundances(MADvector, params)$
  $r\_value \leftarrow testTaylorsLaw(MADvector, abundances)$ ▷ Continue if correlation coefficient between mean and variance ≥ 0.95
  $testDistributions(abundances)$ ▷ Test lognormal and gamma distributions for SAD and AFD respectively
  return $MADvector, abundances, r\_value,$

species are grouped into genera and genera are grouped into families using lognormal distributions. This means that the number of genera in each family and the number of species in each genus follow lognormal distributions.

To generate the complete databank, 16S sequences from all prokaryotes and genome sequences were downloaded on February 2021 and April 2020 from NCBI Nucleotide, respectively (see Supplementary Data 1). To simulate the lack of information in genomic databases, we created incomplete databases by removing taxonomic information from this databank. We created three different databanks (see Supplementary Data 1):

1. Deleting half of the families (Incomplete database DB1).
2. Deleting half of the genera (Incomplete database DB2).
3. Deleting half of the species (Incomplete database DB3).

To simulate the lack of information in the databanks about the species present in the virtualomes, we created three scenarios (Supplementary Data 2-4):

1. 100% scenario: All the genomic information of the virtualomes are present in the incomplete database.
2. 50% scenario: Only half of the genomic information of the virtualomes are present in the incomplete databases.
3. 25% scenario: Only a quarter of the genomic information of the virtualome are present in the database.

We remark that all the species in the virtualomes are present in the complete database. For scenarios 2 and 3, we selected 50% and 75% of the taxonomical information from scenario 100% and replaced it with information that was not in the corresponding database, taking into account the restrictions exposed in 2.5. For the scenarios with 50% and 25% of the species present in the databases, there are several ways to remove families. In the case of removing families, the choice was made to remove those families that grouped an abundance of around 50% in the first case and 25% in the second (as close as possible within the restrictions exposed in 2.5). Therefore, the 50% scenario gives rise to two scenarios (A and B) and the 25% scenario to four (A, B, C and D). Full information on the taxonomic composition of each scenario can be found in Supplementary Data 2-4.

4.3 Bioinformatic analysis of virtualomes

The number of reads corresponding to each species was proportional to its abundance in each virtualome. We considered two scenarios regarding the total number of reads: high number of reads (100,000 for 16S and 60,000 for WGS) and low number of reads (50,000 for 16S and 20,000 for WGS). The generation of reads was done with the program Grinder [46] with the following parameters:

```
#16S reads
grinder -reference_file microb_16S_seq/$name.fasta -forward_reverse
    primer16_v2.fasta -total_reads $abundanceTotal -read_dist 300 -
    insert_dist 550 normal 55 -unidirectional 1 -length_bias 0 -
    mutation_dist poly4 -fq 1 -qual_levels 30 10 -base_name $name -output_dir
     ./reads16S
```

```
#WGS reads
grinder -reference_file microb_WGS_seq/$name.fasta -total_reads
    $abundanceTotal -read_dist 150 -insert_dist 400 normal 40 -
    unidirectional 0 -length_bias 0 -mutation_dist poly4 -fq 1 -qual_levels
    30 10 -base_name $name -output_dir ./readsWGS
```

Listing 1. Grinder commands for reads generation.

Reference files (*reference_file*) store all existing sequences for each species in the complete database. In this way, reads are generated for all the species avoiding any bias in the choice of a particular sequence. *total_reads* refers to the number of reads needed for each species. These reads were analysed by GAIA [47] using the complete and incomplete databases.

For amplicon, the length of reads (*read_dist*) was set at 300 base pairs (bp), with an overlap of 50 bp (*insert_dist* of 550 bp). For WGS the length of the reads was 150 bp and *insert_dist* of 400 bp. We used unidirectional reads (from one strand only) for amplicon and bidirectionally for WGS, as recommended by Grinder documentation (https://sourceforge.net/projects/biogrinder/).

In addition, we introduced sequencing errors in the reads (*mutation_dist*), under the form of mutations (substitutions, insertions and deletions) at positions that follow a $4th$ degree polynomial, simulating Illumina errors [48]. The quality of the reads (*qual_levels*) varies from good (30) to bad (10), for reads with insertions or substitutions. We tried to avoid length bias (i.e. greater contribution of reads from larger genomes) by setting *length_bias* to 0.

Primers used for the amplification of the V3V4 region in amplicon (parameter *forward_reverse*) were:

```
>V3V4F:34
CCTACGGGNGGCWGCAG

>V3V4R:34
GGACTACHVGGGTATCTAATCC
```

These primers have been shown to be efficient for 16S analysis [49]. However, these sequences were not able to amplify the desired region in a non-negligible percentage of the species. For this reason, a lower number of reads than expected was obtained in some *in silico* amplifications. To reach the desired number of reads, the remaining percentage of abundance was distributed according to the abundance distribution. Metagenomic analysis of the samples was done with GAIA [47], which allows to obtain a comprehensive and detailed overview at any taxonomic level of microbiomes in an accurate and easy way.

4.4 Analysis of the microbiome of *Galleria mellonella* larvae

Galleria mellonella larvae were maintained in an incubator at 28°C in the dark, and fed with beeswax from beehives.

For the experimental samples, *G. mellonella* larvae from the last larval stage (150-200mg) were kept in the presence of commercial plastic films only, during 4 to 6 days. For tissue extraction, live larvae were washed with 70% ethanol and rinsed with sterile water. Dissections were performed using sterilized microsurgery tools. DNA extraction was performed using the Quick-DNATM Tissue/Insect Miniprep Kit (Zymogen Research, Cat. No. D6016). Control larvae fed with beeswax from beehives were treated in the same way.

Amplicon sequencing was perfomed using the V3V4 primers described above to generate 16S Illumina libraries which were then sequenced with a MiSeq producing 300 bp paired-end reads. The same DNA was also used to generate TruSeq PCR-Free DNA libraries to perform whole metagenome sequencing (WGS) with an Illumina Novaseq 6000 producing 150 bp paired-end reads. Both the 16S amplicon sequencing data and the WGS data was processed with the following approach: first, the quality of the reads was assessed with the software FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), then low quality bases and adapters were removed with the software Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic); minimum quality 25 and minimum length 35 bp). Amplicon data was analyzed with the software GAIA as described above. WGS data was first mapped against the *G. mellonella* reference genome (ASM258982v1) using the software minimap2

(https://github.com/lh3/minimap2) to remove the contamination from the host. Unmapped reads were extracted with samtools (http://www.htslib.org/) and then processed with GAIA to perform the taxonomic classification. Differential abundance analyses were performed with edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html) applying the TMM normalization to the abundance matrix of the OTUs. The results of the amplicon and WGS analysis of the microbiome of *G. mellonella* can be found in Supplementary Data 5 and 6 respectively. Supplementary Data 7 and 8 show the taxa whose abundance changes significantly in larvae fed with plastic in 16S and WGS analysis respectively.

Acknowledgments

FB and CFA gratefully acknowledge support by the Roechling foundation. BS was partially supported by MINECO grant MTM2017-85020-P.

References

1. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015).

2. Vandeputte, D. *et al.* Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nat. communications* **12**, 1–13 (2021).

3. Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Sci.* **371**, eabc4552 (2021).

4. Burki, T. K. Gut microbiome and immunotherapy response. *The Lancet Oncol.* **18**, e717 (2017).

5. Wei, Z. *et al.* Initial soil microbiome composition and functioning predetermine future plant health. *Sci. advances* **5**, eaaw0759 (2019).

6. Antoniewicz, M. R. A guide to deciphering microbial interactions and metabolic fluxes in microbiome communities. *Curr. Opin. Biotechnol.* **64**, 230–237 (2020).

7. Hays, S. G., Patrick, W. G., Ziesack, M., Oxman, N. & Silver, P. A. Better together: engineering and application of microbial symbioses. *Curr. Opin. Biotechnol.* **36**, 40–49 (2015).

8. Harishankar, M. K., Sasikala, C. & Ramya, M. Efficiency of the intestinal bacteria in the degradation of the toxic pesticide, chlorpyrifos. *3 Biotech* **3**, 137–142 (2012).

9. Sgobba, E. & Wendisch, V. F. Synthetic microbial consortia for small molecule production. *Curr. Opin. Biotechnol.* **62**, 72–79 (2020).

10. Govindarajulu, S. N. *et al.* Insect gut microbiome and its applications. *Recent Adv. Microb. Divers.* 379–395 (2020).

11. Malacrinò, A. Meta-omics tools in the world of insect-microorganism interactions. *Biol.* **7**, 50 (2018).

12. Munoz-Benavent, M., Pérez-Cobas, A. E., García-Ferris, C., Moya, A. & Latorre, A. Insects' potential: understanding the functional role of their gut microbiome. *J. Pharm. Biomed. Analysis* **194**, 113787 (2021).

13. Hongoh, Y. *et al.* Intra-and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. *Appl. environmental microbiology* **71**, 6590–6599 (2005).

14. Brune, A. & Dietrich, C. The gut microbiota of termites: digesting the diversity in the light of ecology and evolution. *Annu. review microbiology* **69**, 145–166 (2015).

15. Ni, J. & Tokuda, G. Lignocellulose-degrading enzymes from termites and their symbiotic microbiota. *Biotechnol. advances* **31**, 838–850 (2013).

16. Jeffries, T. W. *et al.* Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat. biotechnology* **25**, 319–326 (2007).

17. Chellappan, M. & Ranjith, M. Metagenomic approaches for insect symbionts. In *Microbial Approaches for Insect Pest Management*, 271–313 (Springer, 2021).

18. Yang, Y. *et al.* Biodegradation and mineralization of polystyrene by plastic-eating mealworms: Part 1. chemical and physical characterization and isotopic tests. *Environ. science & technology* **49**, 12080–12086 (2015).

19. Bombelli, P., Howe, C. J. & Bertocchini, F. Polyethylene bio-degradation by caterpillars of the wax moth *Galleria mellonella*. *Curr. Biol.* **27**, R292–R293 (2017).

20. Yang, J., Yang, Y., Wu, W.-M., Zhao, J. & Jiang, L. Evidence of polyethylene biodegradation by bacterial strains from the guts of plastic-eating waxworms. *Environ. science & technology* **48**, 13776–13784 (2014).

21. Yang, Y. *et al.* Biodegradation and mineralization of polystyrene by plastic-eating mealworms: part 2. role of gut microorganisms. *Environ. science & technology* **49**, 12087–12093 (2015).

22. Yang, Y., Wang, J. & Xia, M. Biodegradation and mineralization of polystyrene by plastic-eating superworms *Zophobas atratus*. *Sci. total environment* **708**, 135233 (2020).

23. Lou, Y. *et al.* Biodegradation of polyethylene and polystyrene by greater wax moth larvae (*Galleria mellonella* l.) and the effect of co-diet supplementation on the core gut microbiome. *Environ. science & technology* **54**, 2821–2831 (2020).

24. Cassone, B. J., Grove, H. C., Kurchaba, N., Geronimo, P. & LeMoine, C. M. Fat on plastic: Metabolic consequences of an ldpe diet in the fat body of the greater wax moth larvae (*Galleria mellonella*). *J. Hazard. Mater.* **425**, 127862 (2022).

25. Zhang, J. *et al.* Biodegradation of polyethylene microplastic particles by the fungus *Aspergillus flavus* from the guts of wax moth *Galleria mellonella*. *Sci. Total. Environ.* **704**, 135931 (2020).

26. Brandon, A. M. *et al.* Biodegradation of polyethylene and plastic mixtures in mealworms (larvae of *Tenebrio molitor*) and effects on the gut microbiome. *Environ. science & technology* **52**, 6526–6533 (2018).

27. Yang, S.-S. *et al.* Impacts of physical-chemical property of polyethylene (pe) on depolymerization and biodegradation in insects yellow mealworms (*Tenebrio molitor*) and dark mealworms (*Tenebrio obscurus*) with high purity microplastics. *Sci. The Total. Environ.* 154458 (2022).

28. Cassone, B. J., Grove, H. C., Elebute, O., Villanueva, S. M. & LeMoine, C. M. Role of the intestinal microbiome in low-density polyethylene degradation by caterpillar larvae of the greater wax moth, *Galleria mellonella*. *Proc. Royal Soc. B* **287**, 20200112 (2020).

29. Gupta, V. K. *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nat. communications* **11**, 1–16 (2020).

30. Khachatryan, L. *et al.* Taxonomic classification and abundance estimation using 16s and wgs—a comparison using controlled reference samples. *Forensic Sci. Int. Genet.* **46**, 102257 (2020).

31. Dewhirst, F. E. *et al.* Discordant 16s and 23s rRNA gene phylogenies for the genus *Helicobacter*: implications for phylogenetic inference and systematics. *J. bacteriology* **187**, 6106–6118 (2005).

32. Ceuppens, S., De Coninck, D., Bottledoorn, N., Van Nieuwerburgh, F. & Uyttendaele, M. Microbial community profiling of fresh basil and pitfalls in taxonomic assignment of enterobacterial pathogenic species based upon 16s rRNA amplicon sequencing. *Int. journal food microbiology* **257**, 148–156 (2017).

33. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967–977 (2016).

34. Lazarevic, V. *et al.* Analysis of the salivary microbiome using culture-independent techniques. *J. Clin. Bioinforma.* **2**, 4 (2012).

35. McIntyre, A. B. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology* **18**, 1–19 (2017).

36. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nat. Commun.* **11** (2020).

37. Zhang, N. *et al.* Contribution of sample processing to gut microbiome analysis in the model lepidoptera, silkworm *Bombyx mori*. *Comput. structural biotechnology journal* **19**, 4658–4668 (2021).

38. Kong, H. G. *et al.* The *Galleria mellonella* hologenome supports microbiota-independent metabolism of long-chain hydrocarbon beeswax. *Cell Reports* **26**, 2451–2464 (2019).

39. Ren, L. *et al.* Biodegradation of polyethylene by *Enterobacter sp. d1* from the guts of wax moth *Galleria mellonella*. *Int. journal environmental research public health* **16**, 1941 (2019).

40. Peydaei, A., Bagheri, H., Gurevich, L., de Jonge, N. & Nielsen, J. L. Impact of polyethylene on salivary glands proteome in *Galleria melonella*. *Comp. Biochem. Physiol. Part D: Genomics Proteomics* **34**, 100678 (2020).

41. LeMoine, C. M., Grove, H. C., Smith, C. M. & Cassone, B. J. A very hungry caterpillar: polyethylene metabolism and lipid homeostasis in larvae of the greater wax moth (*Galleria mellonella*). *Environ. Sci. & Technol.* **54**, 14706–14715 (2020).

42. Paniagua Voirol, L. R., Frago, E., Kaltenpoth, M., Hilker, M. & Fatouros, N. E. Bacterial symbionts in lepidoptera: their diversity, transmission, and impact on the host. *Front. microbiology* **9**, 556 (2018).

43. Mereghetti, V., Chouaia, B. & Montagna, M. New insights into the microbiota of moth pests. *Int. J. Mol. Sci.* **18**, 2450 (2017).

44. Hammer, T. J., Sanders, J. G. & Fierer, N. Not all animals need a microbiome. *FEMS microbiology letters* **366**, fnz117 (2019).

45. Hall, J. B. *et al.* Isolation and identification of the follicular microbiome: Implications for acne research. *J. Investig. Dermatol.* **138**, 2033–2040 (2018).

46. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* **40**, e94–e94 (2012).

47. Paytuví, A., Battista, E., Scippacercola, F., Cigliano, R. A. & Sanseverino, W. GAIA: an integrated metagenomics suite. *bioRxiv* 804690 (2019).

48. Korbel, J. O. *et al.* Pemer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology* **10**, 1–14 (2009).

49. Klindworth, A. *et al.* Evaluation of general 16s ribosomal RNA gene pcr primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research* **41**, e1–e1 (2013).