Coupling ecological network analysis with high-throughput sequencing-based surveys: lessons from the Next-Generation Biomonitoring project

Maxime Dubart[1], Pascal Alonso[2], Didac Barroso-Bergada[3], Nathalie Becker[4], Kevin Bethune[5], David A. Bohan[3], Christophe Boury[6], Marine Cambon[6,7], Elsa Canard[8], Emilie Chancerel[6], Julien Chiquet[9], Patrice David[5], Natasha de Manincor[1,10], Sophie Donnet[9], Anne Duputié[1], Benoît Facon[11,12], Erwan Guichoux[6], Tâm Le Minh[9], Sebastián Ortiz-Martínez[8], Lucie Piouceau[6], Ambre Sacco--Martret de Préville[8], Manuel Plantegenest[8], Céline Poux[1], Virginie Ravigné[4,13], Stéphane Robin[9,14], Marine Trillat[6], Corinne Vacher[6], Christian Vernière[2], François Massol[15]

1 Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France

2 PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

3 Agroécologie, AgroSup Dijon, INRA, Université Bourgogne Franche-Comté, 21000 Dijon, France

4 Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, SU, EPHE, UA. CP 50, 57 rue Cuvier, 75005 Paris, France

5 CEFE, CNRS, Montpellier, France

6 INRAE, Univ. Bordeaux, BIOGECO, Pessac, France

7 School of Natural Sciences, Bangor University, Deiniol Road, Bangor, Gwynedd, UK

8 UMR 1349 IGEPP - Institut de Génétique, Environnement et Protection des Plantes, INRAE - Agrocampus Ouest - Université Rennes 1

9 UMR MIA-Paris, Université Paris-Saclay, AgroParisTech, INRAE, Paris, France

10 Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA

11 INRAE, UMR PVBMT, Saint Pierre, France

12 CBGP, INRAE, IRD, CIRAD, Institut Agro, Univ Montpellier, Montpellier, France

13 Cirad, UMR PVBMT, Saint Pierre, France

14 LPSM (UMR 8001) - Laboratoire de Probabilités, Statistiques et Modélisations, Paris, France

15 Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France

Table of contents

Abstract 3
1. Introduction 3
2. Sampling 6
2.1. When to sample? 6
2.2. What to sample? 7
2.3. How to sample? 8
2.4. Sampling biases and potential issues 9
2.5. Preparation of samples for molecular analyses 10
2.5.1. Comparing whole-body vs. regurgitate extraction for the assessment of carabid diets (system 2) 10
2.5.2. Testing the effect of dissection and pooling samples for the assessment of fruit fly gut microbiota (system 3) 11
3. Biomolecular and bioinformatic treatments in NGB 13
3. 1. Primer choice 13
3.1.1. What is the ideal primer? 13
3.1.2. Evaluation of primer's efficiency with in silico PCR 3
3.1.3. Food web example of in silico PCR 3
3.2. Sequencing and bioinformatic treatment of sequences 4
3.2.1. Sequencing technique: the question of long vs. short read 4
3.2.2. Bioinformatic pipelines used in NGB 4
3.2.3. Database used to assign sequence: the question of public vs. locally constructed databases 4
3.3. Validation of primer efficiency across various systems studied in NGB through mock experiments 5
3.3.1 Mock communities in the plant-pollinator system (system 1) 5
3.3.2. Mock communities in the predator-prey agricultural network (system 2) 9
3.3.3. Mock communities in the microbiota network in fruit flies (system 3) 10
Short-read mock communities 10
Long-read mock communities 11
3.5. Conclusions of the mock community experiments 12
4. Network reconstruction using inference 12
4.1. Inferring networks using Poisson log-Normal models 12
4.1.1. Gaussian Graphical Models 12
4.1.2. Existing statistical methods for network from counts data 13
4.1.3. Sparse Multivariate Poisson Lognormal Model 13
4.1.4. Caveats and limitations 14
4.2. Inferring networks using Abductive/Inductive Logic Programming 14
5. Network comparison 17
5.1. Background and motivations 17
5.2. A novel unlabelled network comparison method 19
6. Discussion 20
Acknowledgements 23
References 23

Abstract

Biomonitoring ecosystems is necessary in order to evaluate risks and to efficiently manage ecosystems and their associated services. Agrosystems are the target of multiple stressors that can affect many species through effects cascading along food webs. However, classic biomonitoring, focused on species diversity or indicator species, might be a poor predictor of the risk of such whole-ecosystem perturbations. Thanks to high-throughput sequencing methods, however, it might be possible to obtain sufficient information about entire ecological communities to infer the functioning of their associated interaction networks, and thus monitor more closely the risk of the collapse of entire food webs due to external stressors.

In the course of the "next-generation biomonitoring" project, we collectively sought to experiment with this idea of inferring ecological networks on the basis of metabarcoding information gathered on different systems. We here give an overview of issues and preliminary results associated with this endeavour and highlight the main difficulties that such next-generation biomonitoring is still facing. Going from sampling protocols up to methods for comparing inferred networks, through biomolecular, bioinformatic, and network inference, we review all steps of the process, with a view towards generality and transferability towards other systems.

1. Introduction

Biomonitoring comprises various methodologies which aim at observing and assessing the state and ongoing changes in ecosystems, including agricultural landscapes, especially in response to anthropogenic stressors. It also aims at providing an evaluation of risks for human health, food security, ecosystem services and environment, as well as guidelines for policy makers and agrosystem management (e.g., pest control) and preservation of ecological services. Classic biomonitoring techniques have focused on measures of biodiversity (e.g., species richness, beta diversity) and indicator species or functional groups, expected to be representative of an ecosystem state. However, indicator species are often specific to a particular ecosystem (e.g. tropical forests) or indicator of some but not all environmental changes (e.g. sensitive to eutrophication but not temperature rise), and thus not easily transferable as indicator of other similar ecosystems (e.g. temperate forests) or of different stressors.

Classic biomonitoring data may also fail at predicting consequences of human actions because measuring taxonomic diversity is not always synonymous with functional diversity, nor is it always a reliable indicator of ecosystem health (Balvanera et al., 2006, David et al., 2017). Indeed, effects of environmental changes *sensu lato* can propagate through the network of species interactions (Evans et al., 2013, Kaiser-Bunbury et al., 2017). The consideration of ecological interaction networks provides a more complete description of underlying ecological processes and consequences of human actions (Thébault and Loreau, 2006, Duffy et al., 2007, Thompson et al., 2012). Study of ecological networks have grown rapidly (see e.g. Fontaine et al., 2011, Kéfi et al., 2012 for reviews) with numerous insights regarding the effects of biodiversity loss or changes on ecosystem services (Macfadyen et al., 2009, Montoya et al., 2012, Derocles et al., 2014), functioning (Pocock et al., 2012, Astegiano et al., 2015), and stability (Evans et al., 2013, Säterberg et al., 2013), as well as effects of environmental changes on ecosystems (Blanchard, 2015, Thompson et al., 2016).

In addition to the difficulties associated with classic analyses (Blüthgen et al., 2006, Lewinsohn et al., 2006, Ings et al., 2009, Fortuna et al., 2010, Pocock et al., 2012, Poisot and Gravel, 2014), network approaches are also affected by insufficient completeness of ecological network data (Chacoff et al., 2012, Jordano, 2016). Indeed, reconstructing ecological networks using traditional approaches, i.e. observation (Ings et al., 2009, Poisot et al., 2016), is costly and labour-intensive (e.g. Derocles et al., 2012, Pocock et al., 2012). It requires, among others, a reliable identification of species as well as a precise description of interactions among organisms (Gibson et al., 2011) – this information being particularly difficult to gather in cases of poorly studied ecosystems (e.g. Hrček and Godfray, 2015).

In addition, it is practically impossible to survey the whole diversity of species and interactions with traditional tools, and networks built that way are often restricted to some macro-organisms or functional groups or a few trophic levels (e.g., microorganisms, endosymbionts, parasites, etc., are often overlooked – see Lafferty et al., 2008). Incomplete networks can be useful to answer certain questions, e.g. when assessing the community of pollinators associated with a particular plant clade in a particular region, such as orchids in Europe (Joffard et al., 2019). However, inferring networks for biomonitoring purposes, with a view toward predicting the effect of perturbations, might depend on more complete quantification of networks (Novak et al., 2011).

High-throughput sequencing (HTS) and its offshoots (e.g. metabarcoding) can be a rapid and relatively cheap method to assess biodiversity (Ji et al., 2013, Beng et al., 2016, Taberlet et al., 2018) and to build ecological networks (see e.g. Toju et al., 2014, Evans et al., 2016, Vacher et al., 2016). Coupling HTS with ecological network analysis (ENA) of reconstructed networks is therefore a promising avenue for biomonitoring (Cristescu, 2014, Vacher et al., 2016, Bohan et al., 2017, Derocles et al., 2018, Makiola et al., 2020). As DNA is common to all cellular life forms (excepting RNA-viruses), HTS is widely applicable and may allow recovering the vast majority of species present in a given ecosystem. HTS relies on the amplification and sequencing of DNA barcodes (Chakraborty et al., 2014) and is able to produce millions of reads by sequencing run. The use of tags (see also 'nested-tagging', Binladen et al., 2007, Shokralla et al., 2015, Evans et al., 2016) allows to recover data from multiple samples after sequencing, increasing the number of samples possible to sequence at once while reducing costs. Sequencing results in tables of Operational Taxonomic Units (OTU), or Amplicon Sequence Variants (ASV) representing abundances of reads in each sample. Each OTU/ASV can then be assigned to taxa using reference databases. Assuming methodological tools are available, such tables can then be used to reconstruct ecological networks (Vacher et al., 2016, Derocles et al., 2018).

Two strategies may be distinguished: (i) when relationships among species are already well established, species interactions can be directly resolved through data on species presence/absence (plant-pollinator by sampling pollen on pollinators, plant-virus by sampling virus on plants – e.g., Toju et al., 2013, Derocles et al., 2014, Piñol et al., 2014, Toju et al., 2014, Wirta et al., 2014, Fayle et al., 2015, Kitson et al., 2019); (ii) when relationships among species are not known *a priori* (e.g., microbiota), HTS produces OTU tables which have to be treated using either statistical (e.g. Jakuschkin et al., 2016, Ovaskainen et al., 2017) or machine-learning- based methods (Muggleton, 1991, Tamaddoni-Nezhad et al., 2006, Bohan et al., 2011, Tamaddoni-Nezhad et al., 2013, Muggleton et al., 2015) to predict species interactions (e.g. Bohan et al., 2011, Faust and Raes, 2012, Kamenova et al., 2017, Derocles et al., 2018, Chiquet et al., 2019).

Despite examples of network inference and reconstruction from HTS data (e.g. Kitson et al., 2013, Tamaddoni-Nezhad et al., 2013, Derocles et al., 2015) and the fact that the whole field borrows heavily from proteomic and genomic network reconstruction (e.g. Shannon et al., 2003, Pržulj et al., 2006, Daudin et al., 2008), some pitfalls continue to hinder advances toward a next-generation biomonitoring framework. Indeed, each stage of the process is marked by technical or methodological difficulties. First, different strategies can be used to obtain biological samples (e.g. indirect sampling through environmental DNA or direct sampling through feces or gut extracts), which largely affect subsequent analyses and results (Dickie et al., 2018). Second, the biomolecular stage, from DNA extraction to sequencing, presents numerous traps and biases, not all restricted to HTS technologies, which are difficult to control (DNA extraction, primer choices, PCR biases and errors, different primers affinity during sequencing, etc. – Lear et al., 2018). These issues can have important impacts on the final result of the analyses, e.g. when the relationship between read counts and abundances/biomass is not straightforward, which is often the case (Takahara et al., 2012, Thomas et al., 2016, Deagle et al., 2019, Piñol et al., 2019). Third, the bioinformatic stage, from raw sequences to well-defined OTU/ASV or assigned taxa, also presents technical difficulties and results from this stage may depend on arbitrary choices (reads filtering, clustering methods, blast methods, reference database, etc. - Deiner et al., 2015, Knight et al., 2018, Porter and Hajibabaei, 2018, Bush

et al., 2019, Makiola et al., 2019, Pauvert et al., 2019, Zinger et al., 2019). Fourth, with the contingency table in hand comes the reconstruction of networks. While the reconstruction of networks when interactions (but not necessarily interactors) are known is quite straightforward, reconstruction of networks without such prior information is much more difficult. Methods generally return association matrices instead of interactions matrices (e.g. Fuhrman, 2009, Kara et al., 2013, Aires et al., 2015, Navarrete et al., 2015). For example, association between a parasite and its host, or a predator and its prey, is often positive whereas the interaction is negative.

In the context of the Next-Generation Biomonitoring project (NGB), funded by the French National Research Agency (the ANR), several teams in France are attempting to work through a full HTS-ENA protocol, from sampling to network reconstruction and comparison for six different ecosystems including temperate and tropical agricultural systems (Table 1). These systems were chosen because they are potentially economically important, and reflect the diversity and complexity of ecosystem change for biomonitoring at landscape scales. The systems come from different biomes, include antagonistic and mutualistic interactions between plants, microbes and invertebrates, and represent both microbial and macro-organism scales. Networks at the microbiome scale, representing interacting bacterial and fungal taxa, are usually identified by HTS, and networks at the macrobiome scale, representing invertebrates, can be identified using classic taxonomical approaches. The ecological network knowledge of these systems exists across a range of scientific understanding, from relatively unstudied through to well-characterised. It is this range of situations and scientific understanding that allows an exploration of the pitfalls and benefits of a biomonitoring approach, built upon the reconstruction of ecological networks from HTS data. In the following sections, we review the experience accrued by this project and illustrate through examples the difficulties found and results obtained in the different systems.

Table 1 – The different systems used in the NGB project

System	Nature of the network	Ecosystems	Sampling material	Network already known from other approaches?
1	Plant-pollinator network	Calcareous grasslands in France	Pollen pellets collected on pollinators	yes
2	Predator-prey network [carabid beetles and their prey]	Agricultural landscapes in Brittany	Carabid gut content (regurgitates and whole body extracts)	no
3	Host-gut microbiota network [fruit flies and their gut microbiota]	Agricultural landscapes in Réunion island	Fruit fly guts	no
4	Host-parasite network [trematodes infecting gastropods]	Ponds in Guadeloupe island (gastropod composition known)	Water	+/-
5	Microbial networks associated with tree leaves under drought stress	Experimental forest in south of France	Oak and birch leaves and pine needles	no
6	Plant microbiota of modern varieties and traditional landraces	Rice fields in China	Rice leaves and roots	no

2. Sampling

To define and compare strategies for sampling ecological systems in order to infer ecological networks of interactions between species, we need to address three critical points:

(i) *when to sample*, i.e. at what frequency samples should be obtained to cover variation in species phenologies (in terms of species presence but also differences in their abundance) and what time window should be covered by one sample;

(ii) *what to sample*, considering (1) the type of sample, i.e. whether to extract environmental DNA from tissues (e.g. foliar disks), entire organs (e.g. guts or insect legs), or even entire organisms (observed in an interaction, e.g. pollinator visiting a flower), and (2) the heterogeneity of the target species and conditions, i.e. whether to sample from only one specific habitat or many, with effort to correct for the dominance of abundant species or not (i.e. oversampling rare species to obtain rare interactions);

(iii) *how to sample*: which technique to use and for what purpose (i.e. whether to exhaustively infer the network or to focus on a particular sub-network of interest). For interaction networks which have to be inferred from gut contents (predator-prey or host-gut microbiota networks), the technique of dissection is also of paramount importance. For microbial networks, the method of sample preservation and storage (e.g. in silicagel, liquid nitrogen, buffer) is important because it determines the type of nucleic acid that can be extracted (DNA or RNA) and thus the type of microbial community that can be recovered (whole community or active community)

The methods used to infer and analyze the obtained networks will greatly depend on the answers to these three questions, e.g. whether to analyze it as a single unit composed by the sum of its interactions or to consider the network as a dynamic unit in which the timing of sampling should be taken into account. Consideration of these three points will also allow us to explore the potential biases associated with the sampling scheme, and to consider potential means to correct them.

2.1. When to sample?

Ecological interaction networks are dynamic systems, in which links (i.e. interactions) and nodes (i.e. species) can vary greatly among seasons and locations. Taking into account this variation is particularly important when studying a system that experiences important disturbances, such as agroecosystems (Fig. 1). The frequency and time-extent of sampling (e.g. several times in the year vs. only one important period) can affect the method chosen to investigate the relationship between the ultimate network structure and its functioning. Linked to this particular methodological choice, a crucial question is to determine whether to consider the network as a single unit comprising the sum of its interactions, which can happen at different times, or as a dynamic unit.

Sampling at high frequency but on short time periods (e.g. sampling plant-pollinator interactions every two weeks, for 2-4 hours a day), has the advantage of potentially covering a large phenological period and accounting for species turnover. For instance, under temperate climates, it is easier to have samples over the whole flowering period if one is willing to sample for only two hours every two weeks than if one is allocating the same amount of sampling hours to a single week. In other words, a sparse distribution of sampling time can allow covering a larger time window overall, and thus paves the way for the investigation of the effect of species phenology and synchrony on network structure. This also helps build studies aiming at uncovering potential phenological mismatches, leading to network rewiring. Conversely, a few massive samplings over a limited time-window can be geographically redistributed (i.e. comparing many different locations at once, rather than following a given site throughout the season).

When sampling occurs at a given frequency across a rather long time period, it can be convenient to consider the network as a dynamic object, rather than a constant one. Techniques exist to analyze networks that vary in time, and the addition of time as an explanatory variable can be a valuable

asset. However, one should also keep in mind that the assessment of explanatory factors of dynamic objects lends itself to other kinds of statistical issues (e.g. taking into account the effect of autocorrelation and non-independence of network realizations, which generally undermines statistical power).

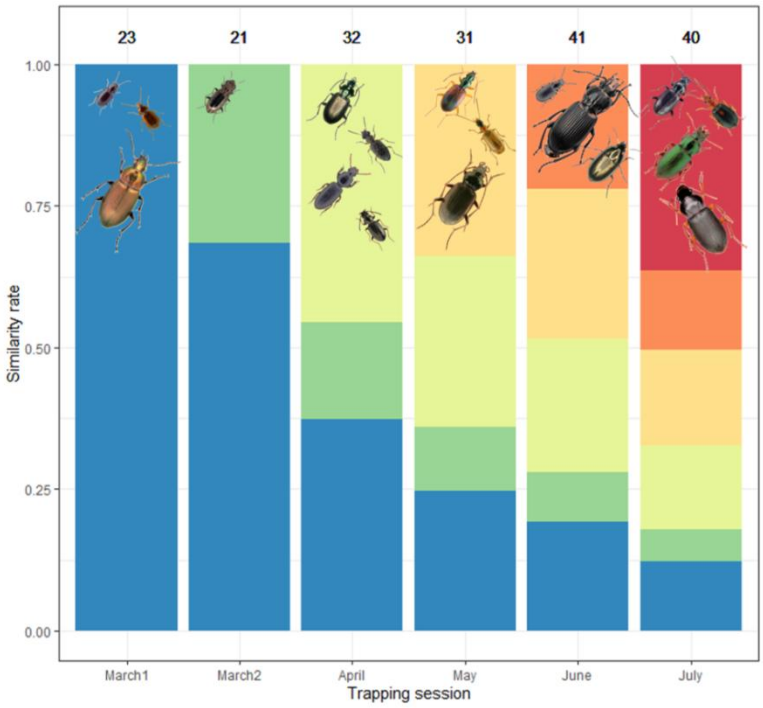


Figure 1 – Similarity rates of carabid communities over time (system 2). Black numbers over the bar plot represent species richness for each session. Each color represents the degree of similarity of a carabid community at a specific session with previous ones (blue: March1, green: March2, yellow: April, light orange: May, orange: June, red: July). Carabid beetles whose abundance increases significantly at a specific session are represented.

2.2. What to sample?

To define the "sampling object", three aspects need to be considered: (i) the type of the sample, (ii) the heterogeneity of the target species/community and conditions and (iii) the right identification of the sampling object.

Regarding the first point, we can discern between two main different types of samples, both represented in the NGB project (Table 1), and following a gradient of precision: from environmental DNA samples (in NGB systems 3, 4, 5, 6), which aim at uncovering species interactions from data on species/OTU abundance/occurrence in individual eDNA samples, to more precise targets (e.g. a single organism involved in the interaction, in NGB systems 1, 2).

In environmental samples, the links between species are not observed but inferred, as for microbiota networks (systems 3, 5 and 6), or known *a priori* in the snail-parasite network (system 4). In the latter, the prevalence of parasites in hosts is so low that screening each host individual is not as efficient as barcoding the water directly. The downside is that the interactions between each individual snail species and each parasite species have to be inferred *a posteriori* from numerous water samples. Conversely, when sampling a precise target (e.g. individual organisms), sampling often focuses on one of the interaction players, for instance hosts (not parasites), pollinators (not plants), or predators (not prey). This structured sampling can affect the way the network will be finally inferred and analyzed, as we will see in sections 4 and 5.

The second challenge of sampling is to choose which individuals to include in the study, in particular when ecological communities of interest are greatly heterogeneous. Super-dominance of one or few species is common, particularly in agricultural systems (Geslin et al., 2017). Dominant species can change over time, e.g. turnover in carabid communities (Fig. 1). As a consequence, one may have to oversample the rare species to have enough information on them, notably to infer interactions between two rare species (e.g. see discussion in Blüthgen, 2010). This issue is connected to the definition of the network addressed by the sampling. It can be the "potential network" (i.e. the network of all non-forbidden interactions), summing details of all possible interactions and its absolute strength, regardless of abundance distribution and population dynamics. As alluded above, this choice implies limiting the number of species, the frequency of sampling and/or the number of sampling locations included in the study because the sampling effort for each species at each sample session/site has to be intense enough. By contrast, one can target the "realized network", yielding a fair view of the system, and in which the strength of interaction depends also on species abundances. Following the latter choice, sampling has to be representative of the real community, and includes the diversity of species observed, even the rarest.

In the agricultural networks studied in NGB, we chose to sample "realized networks" rather than "potential networks", thus including all species at representative numbers, and limiting the sampling effort of the super abundant species. This strategy was motivated by the fact that little is known about the weak interactions (implying less abundance species), which can have a stabilizing effect on networks (Neutel et al., 2002). It can also allow an investigation of functional species, regrouping different species, including the less abundant ones having similar ecological traits or the same function in the system (sensu Dunne et al., 2002, i.e. the same diet at the same time). In this way, rare species could display a bigger impact on the network functioning than expected.

The third challenge of sampling is to assure the genetic identity of the host species (or even genotype) under study. Misidentifying the samples may greatly mislead the analyses focusing on the interactions between the host and its associated organisms, while host-specific factors may control the structure of microbiota. To gain agreement on host identification and classification, molecular labeling can be used to support or correct the local names and characterize the genetic diversity of the ecosystem of interest (Labeyrie et al., 2016). Specifically, in wild or agricultural systems, samplers may face social, cultural and translational barriers to identification/classification. For example, in the Chinese traditional Hani rice terraces system, three distinct rice genetic clusters were categorized using a genotyping by sequencing approach, while two main groups were initially expected based on farmer nomenclature. In addition, two rice fields planted with a variety bearing the same name were genetically distinct (Alonso et al., 2019, 2020).

2.3. How to sample?

Sampling of DNA in order to infer interaction networks through HTS can take different forms, depending on the type of sampled organisms.

In the case of sampling focused on one level of the interaction network, sampling can either be active (e.g. cutting plants to detect their symbionts or using hand nets to capture flying insects) or passive. In the latter case, the use of traps to capture macro-organisms, especially arthropods or molluscs, is a popular technique, but it requires careful preparation as all trapping techniques are not necessarily equivalent in terms of community representativeness (see e.g. Westphal et al., 2008, Prendergast et al., 2020 on different techniques to sample bees). Whether active or passive, targeted sampling needs to account for spatial heterogeneity and for the spatial extent of the sampled unit. In other words, in both cases one needs to sample homogeneous environments, spreading out traps or moving around the field when actively capturing individuals.

For example, most plant-pollinator networks (like system 1) are actively sampled in the field, either by capture of the individual insect (using hand net or similar tools) or picture (close-range

photography). These recordings are limited in their temporal depth – one cannot record interactions that are not directly observed – and hindered by collector experience and often restricted to diurnal pollinators (but see e.g. Walton et al., 2020). Passive techniques (e.g. using UV-bright pan traps), by contrast, are easy to deploy, do not require any experience, and do not have any restriction on temporal depth since they can be deployed for quite long periods, day and night. However, such traps often effectively sample only a portion of the pollinating fauna and do not inform on interactions with the plants, which need to be reconstructed *a posteriori* and with the support of other techniques, such as pollen analysis. Therefore, the pollen transported on pollinator bodies have been recently used to build or to complete and validate plant-pollinator networks (Bosch et al., 2009, Banza et al., 2015, Pornon et al., 2016, Bell et al., 2017, Pornon et al., 2017, Lucas et al., 2018, Macgregor et al., 2019, de Manincor et al., 2020).

Other techniques, for instance sample dissection, can be used to avoid overrepresentation of non-target DNA. Non-target DNA can be known, for example (as in system 2) extracting DNA from a whole organism to assess its prey species would result in an overrepresentation of the focal organism's DNA (and likely miss some rare prey). Alternatively, non-target DNA can be unknown, for example (system 3) bacterial DNA amplified from the whole body of fruit flies would include the gut microbiota together with the microbiota from other organs, thus likely hampering reconstruction of a gut microbiota network.

When sampling targets environmental DNA, the ultimate precision of the data will heavily depend on the quantity of samples because DNA is often quite diluted, and thus nearly undetectable, in most samples (Carraro et al., 2021). The spatial and temporal variation of sampling will also allow capturing hints of species presence depending on species activity periods and preferred habitats.

2.4. Sampling biases and potential issues

In the course of our investigations for the NGB project, we encountered some situations in which sampling biases could be detected, as well as potential issues arising from the techniques used to sample interactions.

In the case of plant-pollinator systems sampled through passive trapping (system 1), one issue is that, contrary to plant-pollinator networks that are actively sampled (e.g. using hand nets), the network obtained through HTS cannot be compared to a network obtained using direct observation. Combined with the difficulty of teasing apart which pollen grains belonged to which insect in a "bee soup" retrieved from pan traps or malaise traps, this can lead to difficulties when inferring the network of interactions between species. By contrast, sampling DNA from pollen grains found on preserved insects caught by hand can lend itself to double checking through microscopic identification of pollen grains. Microscopic identification of pollen grains is time-consuming and relies on palynological experts. However, using local pollen atlases, one can identify pollen grains at the species level, even for some closely related taxa (de Manincor et al., 2020). Only recently the application of deep-learning techniques to identify microscope slides have been successfully tested, but it also needs the compilation of pollen reference libraries and requires improvement (Olsson et al., 2021).

When passively sampling carabid beetles preying on agrosystem pests (system 2), gut content analyses can produce false positives (i.e. apparent prey in unlikely predators, such as spiders eaten by carabids). These events probably occurred because predators can consume prey that are trapped together with them, even ones that could have escaped in natural conditions. In this particular instance, one possible way of preventing these false positives from occurring is to put clay balls into the trap, which provide potential prey with the means to avoid predators in the trap. Another potential issue associated with traps that do not instantly kill animals outright (system 2) comes from the degradation of DNA in predators' guts: the longer the predators stay in a trap before sampling of

DNA, the less prey DNA will be found in their gut since DNA digestion occurs while the predator is trapped.

2.5. Preparation of samples for molecular analyses

Preparation of multiple samples, with rigorous methods to focus on a precise tissue, can be an obstacle. Sample preparation can affect the detection of a target interaction. The main issue of inappropriate sample preparation is the detection of non-target DNA present in the sample, such as microbiota from other tissues (system 3) or DNA of the consumer (system 2), or the non-detection of existing target DNA (and thus, of existing interactions).

A lot of variation exists in protocols for sample preparation. In the course of the NGB project, we investigated the impact of two widely adopted sample-processing procedures preceding library preparation: (i) the extraction of targeted tissues to avoid non-target DNA via dissections or the use of regurgitates, feces, pollen pellets, etc. and (ii) the pooling of individuals within the same sample to increase the number of screened individuals. We used two contrasted systems investigated in the NGB project: the diet composition of a carabid generalist predator of crops (system 2), and the microbiota of tropical fruit flies (system 3).

2.5.1. Comparing whole-body vs. regurgitate extraction for the assessment of carabid diets (system 2)

In system 2, we identified the prey species constituting the diet of generalist predator communities (carabid beetles), found in cereal fields in Brittany, France. We compared the use of whole predator bodies with the gut content obtained with a regurgitation protocol using thermally induced stress (as described in Wallinger et al., 2015). We expected some bias when using whole-body extraction because the universal primers used to reveal prey species can amplify predator's DNA. Our study shows that the use of regurgitates for metabarcoding of generalist arthropods can greatly affect the diversity of preys found in the gut content (Fig. 2).

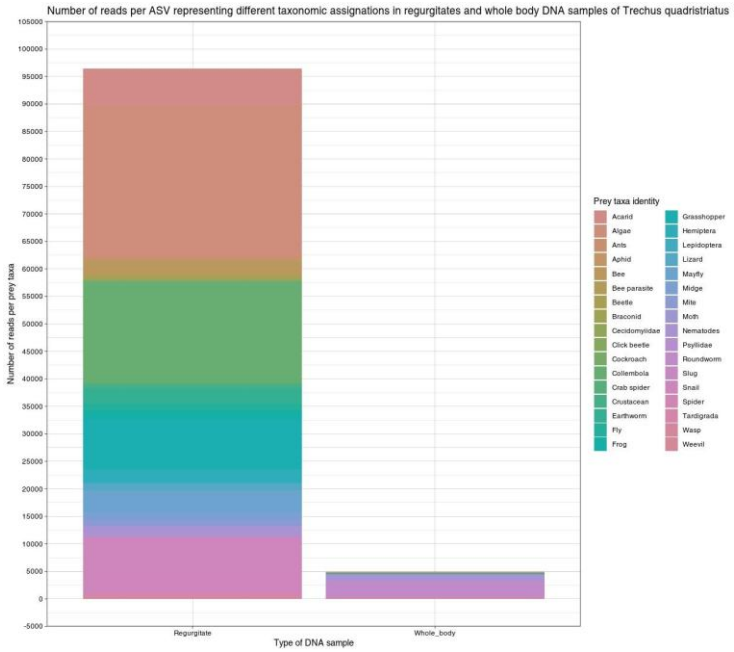


Figure 2 – Number of reads per ASV in regurgitate vs. whole-beetle samples of *Trechus quadristriatus*

2.5.2. Testing the effect of dissection and pooling samples for the assessment of fruit fly gut microbiota (system 3)

The impact of sample pooling on gut bacteria metabarcoding was tested on natural populations of tropical fruit flies sampled on Reunion Island. The effect of dissection was evaluated by comparing 16 samples with either dissected guts or full abdomens of males from three fruit fly species, *Bactrocera dorsalis* (n = 6), *B. zonata* (n = 5), and *Ceratitis quilicii* (n = 5). The effect of sample pooling was evaluated on a set of ten dissected male guts of *B. dorsalis*, by comparing ten individual guts separately, to two pools of five guts and a pool of the ten guts.

Effect of dissection – Treatments did not differ in community composition at the phylum level (PERMANOVA: df = 1, F = 1.2758, p = 0.2757, R^2 = 0.08) and only a marginally significant difference was detected at the family level (df = 1, F = 2.4568, p = 0. 08891, R^2 = 0.15, Fig. 3a). Eleven out of 21 families, including the eight most abundant families, occurred in both treatments. At both taxonomic levels, alpha diversity of dissected guts (phylum level: 1.621 [1.388-1.926], family level: 3.076 [2.538-3.693]) tended to be higher than in whole-abdomen samples (phylum level: 1.260 [1.049-1.576], family level: 2.310 [1.739-2.928]), but with some overlap in confidence intervals (Fig. 3b).

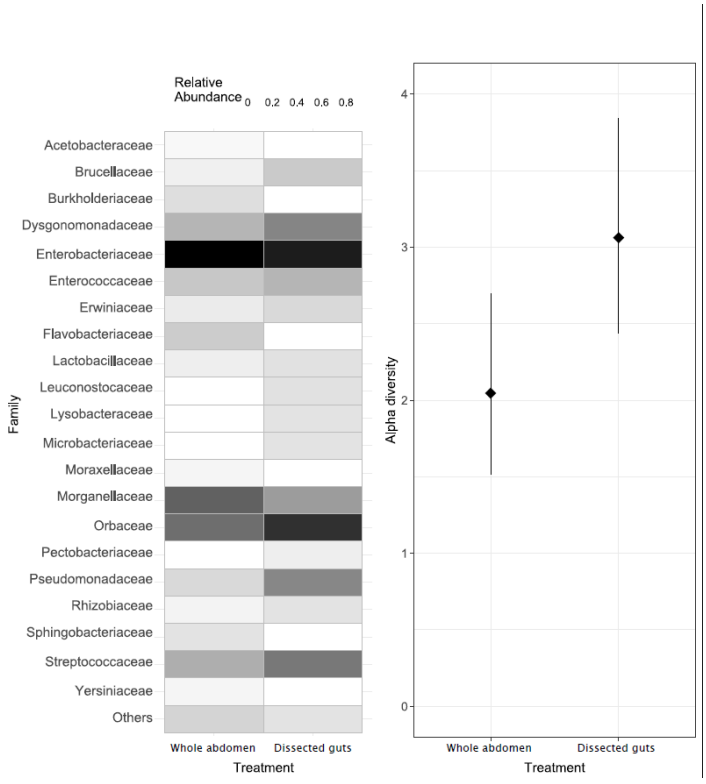


Figure 3 – Effect of dissection vs. whole-abdomen sampling on (a) the relative abundances of families of bacteria and (b) the diversity of these taxa in the gut microbiota of the fruit fly *Bactrocera dorsalis*.

Effect of pooling samples – At the phylum level, the average alpha diversities of the 10 one-individual samples and of the two five-individual samples were equivalent to the alpha diversity of the ten-individual sample (2.161, 2.141 and 2.291 respectively). The gamma (total) diversity of the 10 one-individual samples was 2.272 (sample alpha diversity ranging from 1.479 to 2.749). At the family level, the average alpha diversity of the 10 one-individual samples and of the two five-individual samples were significantly lower than the alpha diversity of the ten-individual sample (3.153 (2.808-3.469), 3.803 (3.362-4.301) and 4.238 respectively, Fig. 4b). The gamma diversity of the 10 one-individual samples was 4.262 (alpha diversity ranging from 2.541 to 4.272).

The four most abundant families (Enterobacteriaceae, Orbaceae, Streptococcaceae and Dysgonomonadaceae) in the ten-individual sample were present in all one-individual samples (except two for Streptococcaceae, Fig. 4a). On the contrary, the least abundant families in the ten-individual sample were absent from several one-individual samples. Finally, two families (Morganellaceae and Ruminococcaceae) were not detected in the ten-individual sample although present in some one-individual samples.

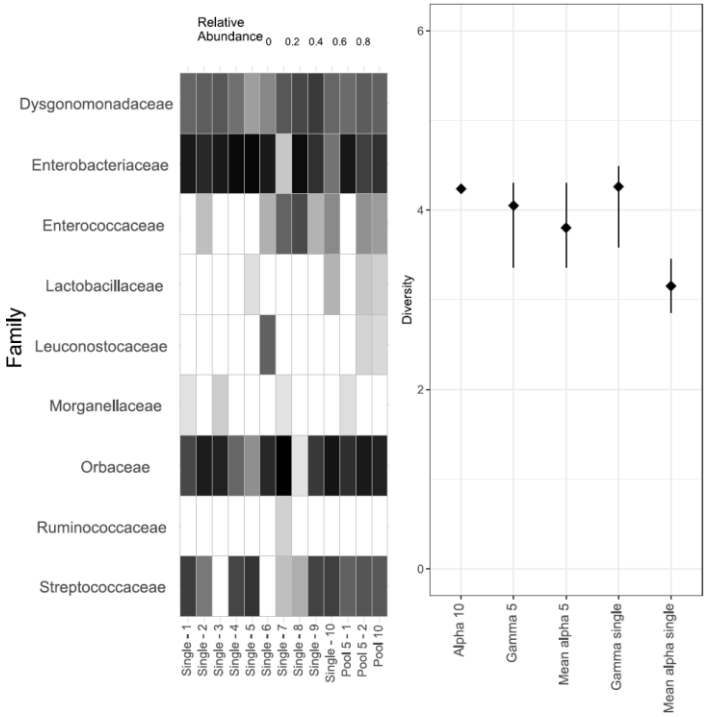


Figure 4 – Effect of sample pooling size on (a) the relative abundances families of bacteria and (b) the diversity of these taxa in the gut microbiota of three species of Tephritidae fruit flies.

Overall, our results showed no major effects of dissection or pooling on the outcome of metabarcoding at the phylum level. At family level, families of bacteria found at low frequency in dissected guts were often not retrieved in whole-abdomen samples. Variability in community composition was observed among individual samples, particularly affecting low-frequency families. As a consequence, the number of guts under study affected community composition and diversity. Pooling *per se* did not affect the results: community diversity could be inferred equivalently as the alpha diversity of a pooled sample or as the gamma diversity of different individual samples. Other studies have also shown effects of sample preservation procedures on inferred gut microbiome (Hammer et al., 2015, Song et al., 2016, De Cock et al., 2019).

These results have important implications for the planning of future studies and when comparing studies that used different sample preparation protocols. Overall, it seems that most differences observed between dissection and pooling treatments were on family-level profiles and more specifically on low-frequency families. As a consequence, the choice of a quick-and-dirty vs. a more time-consuming, expensive, and precautious protocol could be adjusted to the question tackled and to the likelihood that non-target DNA can be amplified. Precisely describing gut community composition or studying the functional roles of bacteria within guts will definitely require accessing low-frequency bacteria and their fine taxonomy. Such research should mandatorily dwell on fine-tuned sample preparation protocols. In contrast, surveillance of large-scale trends at the ecosystem

scale may afford using relaxed sample preparation protocols, which will be sufficient to describe community composition at a high taxonomic rank or community diversity indices.

3. Biomolecular and bioinformatic treatments in NGB

In any metabarcoding project, some of the most important steps are those of biomolecular analysis and sequencing, closely linked to those of bioinformatics processing i.e. transitioning from a database of gene sequences and their abundances to data on species abundances in communities. Some important issues must be dealt with along the way. These include how to deal with unassigned sequences (i.e. not matching those existing in databases) or assignation mistakes, especially when using short sequences. How to validate and estimate the precision of species assignation and how to build a proper local, hand-curated database and whether it is necessary to do so? Such questions and the more general issue of whether different kinds of taxa lead to different answers, call for a general appraisal of the experience we have collectively accrued in the course of the NGB project.

Before retrieving known species or OTUs or ASVs from batches of sequences, an important aspect of NGB protocols is to decide which gene(s) to monitor and for what purpose. In this section, we give an overview of some of the crucial questions that we think ought to be addressed in order to build a metabarcoding approach that allows final inference of interaction networks: (i) whether sequences are obtained using existing primers, already used by other groups and thus probably referenced in existing databases or using primers specifically designed for the focal organisms; (ii) whether the sequencing technology produces short or long reads, and in turn whether existing sequence bases are complete and reliable (short reads being easier to produce, but probably more equivocal on average); (iii) whether the sequence assignment process will make use of public databases or local, individually curated databases.

3. 1. Primer choice

DNA barcoding can reveal interactions where a large number of species are involved. Target taxa can be detected using one or several DNA regions depending on the desirable coverage, accuracy of detection and rate of conservation of the loci, as for mitochondrial (well conserved at intraspecific level) and ribosomal (well conserved at interspecific level) DNA regions. So depending on the purpose of the research and its desired level of accuracy at the taxonomical level, certain DNA regions could aid in primer designing to cover the target organisms.

3.1.1. What is the ideal primer?

The type of question and the taxonomic coverage needed in the study dictate the kind of marker required. While species-specific primers are best applied for the detection of specific targets at the species level, group-specific primers could be used in cases where a certain functional or taxonomic group is the desired target, and finally, universal primers can be employed in cases where the goal is to obtain information about trophic links involving diverse and/or unknown taxa (Leray et al., 2013). In the latter case, difficulties can appear when dealing with environmental DNA samples, where target and non-target DNA are mixed. In system 2 for instance, the DNA of the arthropod preys is mixed with the DNA of the arthropod predator. An ideal primer should then amplify arthropod prey but not the predator DNA, a real challenge when both are phylogenetically close. Similarly, in leaf DNA extracts of system 5, bacterial DNA was mixed with chloroplastic DNA. This is why we used a primer pair that excludes chloroplastic sequences (Table 2). As a wide range of markers is available or possible to design, it is then important to evaluate primer performance and estimate the success in the detection of the desired target taxa, while avoiding the non-target taxa. To evaluate the viability of a putative group of markers, and then choose the ones that could be adjusted for the study, bioinformatic tools and pipelines can help the process.

Table 2 – The variety of primers used in the NGB project

System	Nature of the network	Type of primers used
1	Plant-pollinator network	Primers were tested this way: 1/ extraction of sequences from BOLD (plant and pollinator species that we knew we would find in the dataset); 2/ sequences were truncated to these regions and aligned; 3/ genetic distances were calculated using ABGD (Puillandre et al., 2012). Generic primers CO1: lll_B 5'-CCIGAYATRGCITTYCCICG (Yu et al., 2012) Fol-degen-rev: 5'-TANACYTCNGGRTGNCCRAARAAYCA ITS_S2F: 5'-ATGCGATACTTGGTGTGAAT (Chen et al., 2010) ITS4: 5'-TCCTCCGCTTATTGATATGC (White et al., 1990)
2	Predator-prey network [carabid beetles and their prey]	Forward: fwhF2 (5' - GGDACWGGWTGAACWGTWTAYCCHCC - 3') Reverse: fwhR2n (5' - GTRATWGCHCCDGCTARWACWGG - 3') That primer pair targets a 254 bp fragment of the mitochondrial COI region and was specially designed for degraded DNA (Vamos et al., 2017).
3	Host-gut microbiota network [fruit flies and their gut microbiota]	27F 5'-AGAGTTTGGATCMTGGCTCAG-3' 1492R 5'-GGTTACCTTGTTACGACTT-3' That primer pair targets the 16S rRNA gene.
4	Host-parasite network [trematodes infecting gastropods]	16S rRNA mitochondrial sequences Available gastropod specific marker (Taberlet et al., 2018, in-silico design via on ecoPrimers and ecoPCR tools, very well preserved priming sites) Manual design of a second Thiaridae family-specific marker to amplify and detect environmental DNA from water filtered samples
5	Microbial networks associated with tree leaves under drought	Bacteria: V5-V6 region of the 16S rDNA gene using the primers 799F-1115R (Chelius and Triplett, 2001, Redford et al., 2010) to exclude chloroplastic DNA. To avoid a two-stage PCR protocol and reduce PCR biases, each primer contained the Illumina adaptor sequence, a tag and a heterogeneity spacer, as described in Laforest-Lapointe et al. (2017).

	stress	799F: 5’-CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTxxxxxxxxxxxxHS-AACMGGATTAGATACCCKG-3’ 1115R: 5’- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxxxxxxxHS-AGGGTTGCGCTCGTTG-3’, where HS represents a 0–7-base-pair heterogeneity spacer and “x” a 12 nucleotides tag Fungi: ITS1 region of the rDNA gene (Schoch et al., 2012) (Schoch et al. 2012) using the primers ITS1F-ITS2 (White et al., 1990, Gardes and Bruns, 1993). To avoid a two-stage PCR protocol, each primer contained the Illumina adaptor sequence and a tag. ITS1F: 5’- CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTxxxxxxxxxxxxCTTGGTCATTTAGAGGAAGTAA-3’ ITS2: 5’- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxxxxxxxGCTGCGTTCTTCATCGATGC-3’, where “x” is the 12 nucleotides tag.
6	Plant microbiota	Bacteria: V3-V4 region of the 16S rRNA gene using the primers 341-F (5'CTACGGGNGGCWGCAG3') and 785-R (5'GACTACHVGGGTATCTAATCC3') High coverage and specificity of the bacteria domain for soil and plant associated microbiota based on experimental study and in silico analysis against the SILVA database (Thijs et al., 2017). Fungi: ITS2 region of the rRNA operon using the primers ITS86-F (5'GTGAATCATCGAATCTTTGAA3') and ITS4-R (5' TCCTCCGCTTATTGATATGC3') Based on experiments with plant-associated soils and in silico analyses against sequences downloaded from NCBI, this primer air offers a good coverage of the fungal community and PCR efficiency (Op De Beeck et al., 2014).

3.1.2. Evaluation of primer's efficiency with in silico PCR

In silico tools offer a straightforward procedure for primer evaluation in an early stage of the experimental design, in order to sort and select primer candidates that offer the best performance. In this respect, *in silico* validation of the already available markers as an alternative to primer designing is a worthwhile way of choosing optimized markers for the target taxonomic group. The realization of *in silico PCRs* allows us to investigate the theoretical result of PCRs, using our knowledge about what is probable to affect the success of amplifications. The process compares multiple candidate primer pairs using various factors such as the thermodynamic properties of the barcodes, the variability of the primer-binding sites on the target DNA sequence, and the generated fragment size (Deagle et al., 2014, Elbrecht and Leese, 2017). An accurate analysis of those factors will allow us to avoid primer biases and losing information of some taxa.

For metabarcoding purposes, where the effect of the variability among DNA sequences has to be taken into account, the abovementioned tools are not always efficient in the evaluation of universal primers, molecular markers that are proposed to be used in the amplification of multiple species' DNA sequences. Standardized methods, such as the case of ecoPrimers (Riaz et al., 2011) and PrimerMiner R package (Elbrecht and Leese, 2017), have succeeded in this problem, allowing us to develop and/or evaluate molecular markers especially for metabarcoding purposes with fewer biases. Tools such as ecoPCR, are openly available to find suitable DNA regions to select a barcode depending on the target taxa (Coissac et al., 2012).

Those *in silico* tools, particularly developed for metabarcoding primer evaluation, have the benefit of testing for different combinations of primer pairs at the same time using the same DNA reference databases, reducing significantly the computing time. PrimerMiner, in particular, considers the interspecific and intraspecific variability in the DNA sequences without an overrepresentation of certain taxa. This is achieved by clustering the different sequence variants in OTUs, reducing biases associated with the number of sequences available in the databases. In the case of the PrimerMiner tool, it considers the type of nucleotide that mismatch the target sequences and analyzes the position of the nucleotides in the fragment of the mismatches in the binding sites.

As the primer evaluation is completely based on matching with the reference DNA sequences database, having an unbiased and solid database including all DNA sequences for the target taxa is crucial. In addition, the bias regarding the quality of the DNA samples and laboratory conditions are not accounted for completely, so testing primer candidates *in vitro* is recommended.

3.1.3. Food web example of in silico PCR

PrimerMiner tool has been used to decide on the universal primers targeting invertebrate prey in the predator-prey system of NGB project (system 2). Eight forward and eight reverse published primers have been tested. *In silico* PCR has been realized on a selected database including all arthropod families expected to be present in the agricultural system (Fig. 5), i.e. the target species. Penalty scores were calculated for each primer, using the number and position of mismatches between the primer and the binding sites. Above a value of 120, the penalty score was supposed too high, and the primer was dropped from the analysis. The three forward and the five reverse remaining primers represented 15 potential primer pairs, with 6 pairs automatically discarded because of incompatible primer position on the gene (forward after the reverse). We then chose to select primer pair on the size of amplified fragment (between 254 and 464 bp for the feasible pairs), and the two shorter pairs were kept for lab test *in vivo* conditions.

We did the same analysis on non-target species (carabid DNA, or possible contaminants as mammals), to select primers that also display strong penalty scores on those unwanted sequences, but results did not allow to discriminate between primers (not shown here).

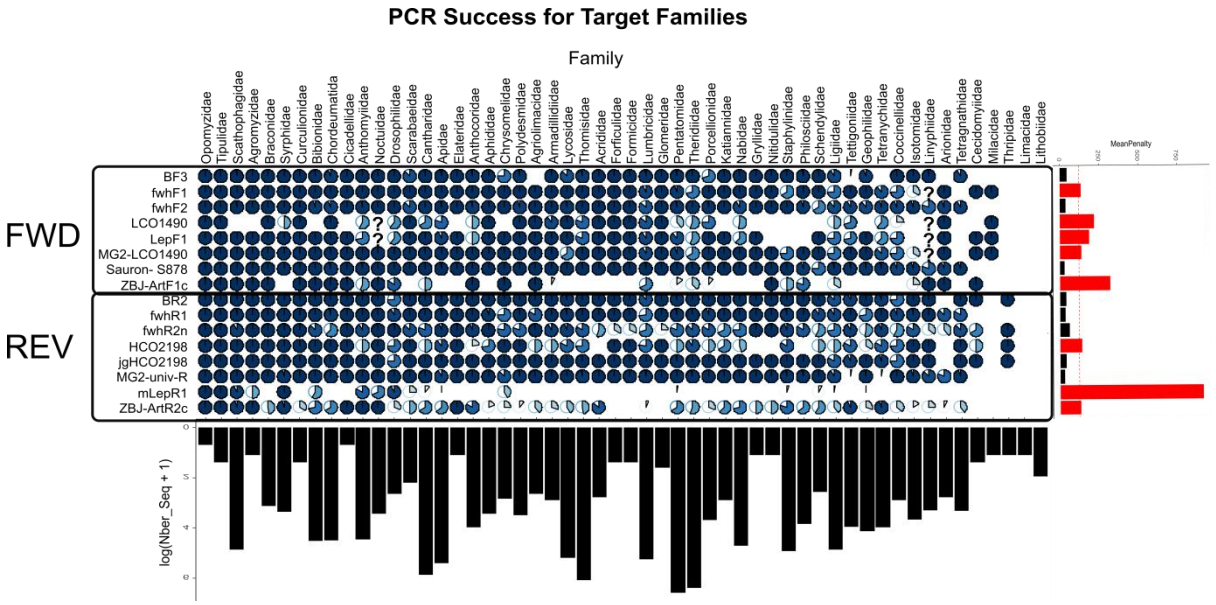


Figure 5 – Evaluation of various forward (FWD) and reverse (REV) universal primers targeting COI gene in arthropod thanks to *in silico* PCR performed with PrimerMiner. Pies represent the score of the binding for each primer on sequences from a given family. Number of sequences used for each family is represented below (it includes only sequences suitable for calculating a penalty score, all incomplete sequences have been discarded). MeanPenalty score is given for each primer, the ones above 120 being dropped from the analysis (in red).

3.2. Sequencing and bioinformatic treatment of sequences

3.2.1. Sequencing technique: the question of long vs. short read

As mentioned above, short reads have advantages for degraded DNA in all systems. Long reads from MinION sequencers, despite their lower sequencing quality, can offer a more accurate species identification, which is based here on much longer fragments (>1kb). Moreover, the portability of the MinION sequencer, and its ability to provide real-time identification, might be considered in the future, despite its high cost per sample. All systems studied in NGB have used paired-end Illumina short reads from a MiSeq sequencer, at Bordeaux PGTB Platform. We also investigated the accuracy of MinION long reads in microbiota systems. More information on this test can be found in section 3.3.

3.2.2. Bioinformatic pipelines used in NGB

Sequencing data needs to be processed through a variety of steps before being analysed and used for network reconstruction. Briefly, sequences need to be filtered based on their quality, chimeric sequences need to be removed, and sequence errors produced during PCR amplification and sequencing need to be detected and cleared. Finally, sequences can be clustered into relevant taxonomic units, depending on the taxonomic resolution and variability needed for the study. Many bioinformatic pipelines exist to achieve this data treatment. They have different pros and cons and may impact the final composition of the community, and thus, the final reconstructed network. This question has not been tackled in the NGB project, and in most cases, the dada2 pipeline (Callahan et al. 2016) has been used as it has previously been proved to be more efficient than other pipelines to retrieve community composition (e.g. Callahan et al. 2017, Pauvert et al. 2019). In this pipeline, sequences are cleaned and exact Amplicon Variant Sequences (ASVs) are inferred, allowing for the correction of sequence errors. Those ASVs can then be used for taxonomic assignment using either public or home-brewed reference databases, to allow further ecological interpretations of reconstructed networks.

3.2.3. Database used to assign sequence: the question of public vs. locally constructed databases

The availability of accurate DNA reference databases for the desirable loci, whether are local or in public servers, is a fundamental factor to maximize the coverage of all target taxa.

In databases that are publically available (c.a. NCBI and BOLD Systems), the number of resources has rapidly increased and it now becomes convenient to use them in most cases. However, the number of sequences available can be drastically different depending on the gene and taxa consider, as well as the diversity of the sequence (coming from various locations or populations). Those gaps can be deleterious for studies since interactions will be missing. Conversely, the presence of very unlikely taxa and especially mistaken sequences can introduce false positive interactions. Indeed, DNA sequences databases are not free of biases, especially in terms of technical mistakes or taxonomical errors concerning organisms that have been poorly studied. In this regard, generating custom databases and/or adding certain taxa in a local database can help enhance coverage while ensuring reliable interactions (e.g. in systems 1 and 3). However, in system 2, the use of a local custom database introduced false assignations due to the lack of some important taxa, and ultimately assignations had to be done using a public database (NCBI).

In the particular case of *in silico* evaluation of primers, as it is completely based on the matching with the reference DNA sequences database, having a variable and curated database becomes essential to succeed in this process. If this step is undervalued, it could determine and limit the quality of the analytical outputs. On the one hand, the variations that could occur in the sequences should be taken into account by not limiting the reference database to few reference sequences. On the other hand, this type of analysis will be particularly sensitive to biases introduced by a heterogeneous distribution of sequence numbers among taxa. Then, using all possible sequences available in public databases for certain taxa could introduce biases because of overrepresentation among the others during the primer evaluation procedure (Elbrecht and Leese, 2017). A solution can be to pool sequences representing given taxa and build a consensus sequence.

Available *in silico* tools can override such problems and have the advantage of taking into account the interactions that have been poorly studied and where few reference sequences have been published. This is particularly useful when the researchers evaluate the candidate primer's performance, in terms of amplification of the DNA from target taxa, with samples where the analysis of trophic links is particularly difficult. This is the case of prey remains in which the DNA degradation in the environment or gut content prevents the amplification of long fragments (often > 150 bp), and in consequence, decreasing the detection of some taxa (Coissac et al., 2012).

3.3. Validation of primer efficiency across various systems studied in NGB through mock experiments

Mock communities commonly serve as controls in metabarcoding studies. Mock communities are pools of known DNA quantities of several species. Their sequencing allows estimating rates of false positive taxa, and relative abundance distortions. For instance, in studies on microbiota, all stages along the production of microbiome data may induce errors and biases in inferred community composition (Brooks et al., 2015). Analysis of mock communities can help assess some of these biases and facilitate the interpretation of results from environmental samples. Mock communities can also be used to tune bioinformatics pipelines. For instance, mock communities provide minimal relative abundances of true positive taxa and maximal relative abundances of false positive taxa that can be used as objective criteria to define thresholds to filter contingency tables prior to diversity analyses. Mock communities were generated for systems 1, 2 and 3 of NGB.

3.3.1 Mock communities in the plant-pollinator system (system 1)

In the context of system 1 (plant-pollinator interaction network), we provide inferred communities profiles to compare the bias on DNA detections when dealing with plant versus insect DNA, and when modifying parameters and databases in the bioinformatic pipeline.

Known quantities of plant and insect species were pooled together and amplified for CO1 (to retrieve insect sequences; primers CO1 Ill_B 5'-CCIGAYATRGCITTYCCICG and Fol-degen-rev 5'-TANACYTCNGGRTGNCCRAARAAYCA) and for ITS2 (to retrieve plant sequences; primers ITS_S2F 5'-ATGCGATACTTGGTGTGAAT and ITS4 5'-TCCTCCGCTTATTGATATGC).

Depending on parameters used for the bioinformatic pipeline (quality threshold, adaptors trimming, reference database) and using an identity threshold of 99.8%, plant's recovering rates range from 50.7% to 65.2% (60% on average) at all taxonomic levels, but false positives range from 0 to 60% (ca. 37% on average at species and genus level, and 25% at family level - Fig. 6a). Although using BDD5 (local database) appears as a good solution on mock samples (low FP rate), other tests show that using local databases reduces recovering rates on other kinds of samples (because all taxon are not represented).

Insect assignment was better than for plants (Fig. 6b). With an identity threshold of 99.6%, recovering rates range from 54 to 76% (65% on average) at all taxonomic levels, and false positives range from 0 to 20% (17% at species level, <1% at genus and family levels).

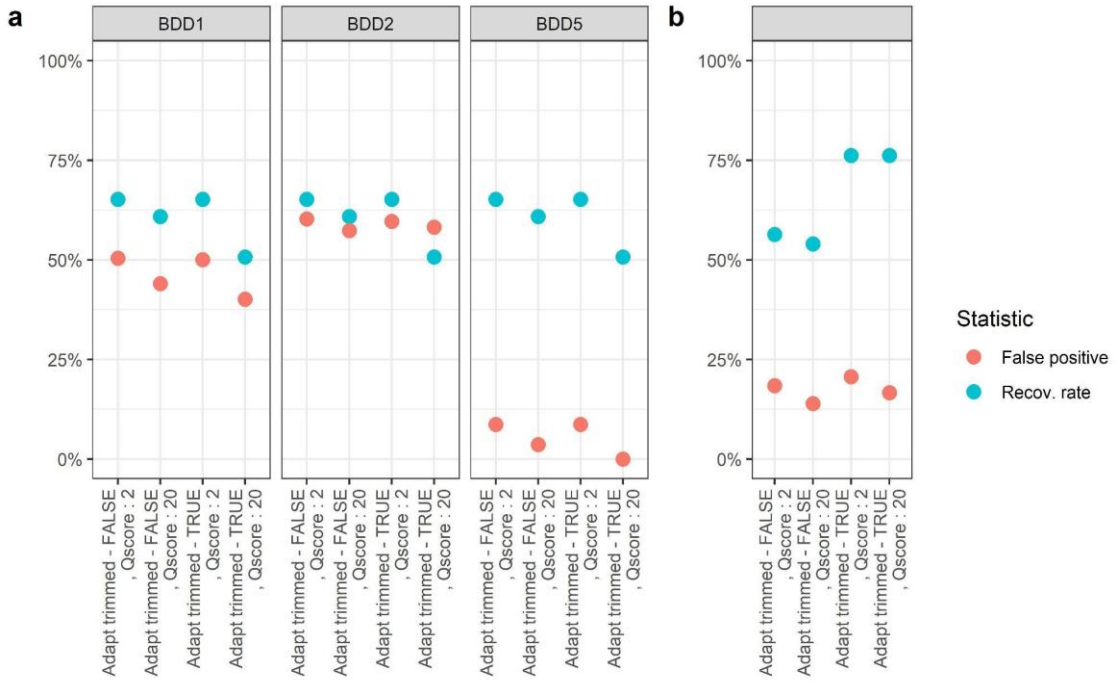


Figure 6 – Recovering and false positive rates (species level) for plants (ITS2 - panel **a**) and insects (COI - panel **b**). Each sub-panel corresponds to a different reference database for plants (BDD1: all expected families retrieved from BOLD System, BDD2: same as BDD1 but reduced to species observed in studied sites, BDD5: local database built using SANGER sequencing). Only one database was used for insects. X-axis indicated whether adaptors were trimmed (TRUE) or not (FALSE), and the quality score used for reads filtering.

Regarding mock compositions (Fig. 7 and 8 for two examples), results were overall better for insects than for plants. In the first example (Fig. 7), numerous unexpected species were found whatever the parameters used (but mistakes are reduced when using the local database). With BDD1, expected *Geranium molle* was never found, *Sonchus oleraceus* was found only with an identity threshold of 99.8% and *Knautia arvensis* was always found except when adaptors are trimmed and a quality threshold of 20 was used. Some unexpected species were consistently found (e.g., *Brassica nigra, Centaurea jacea, Coronilla varia, Lotus corniculatus, Potentilla reptans*). With the local database (BDD5), false positive rate is lower (among the previously unexpected cited species, only *B. nigra* and *C. varia* were found when the identity threshold is lower than 99.6%) and *Sonchus oleraceus* is always retrieved. Regarding insects, better results were obtained when adaptors were trimmed and

using a high quality score. Here, both *Andrena* sp. and *Bombus lapidarius* were retrieved although relative reads counts are strongly distorted in favor of *Bombus sp.*

In the second example (Fig. 8), results are different since almost no false positive was detected for plants (only *C. jacea* but at very low frequencies - not visible on the plot), but *Knautia arvensis* was not retrieved. As in the previous example, *S. oleraceus* was found only with an identity threshold of 99.8% when using BDD1 but always retrieved with BDD5. Relative read counts were strongly distorted in favor of *Medicago lupulina*. For insects, a false positive (*Andrena fragilis*) can be identified when adaptors are trimmed and using an identity threshold lower than 99.6%. *Osmia caerulescens* is consistently identified as *Osmia leaiana*, and *B. lapidarius* is always found. As in the previous exemple, relative reads counts were strongly distorted in favor of *B. lapidarius*.

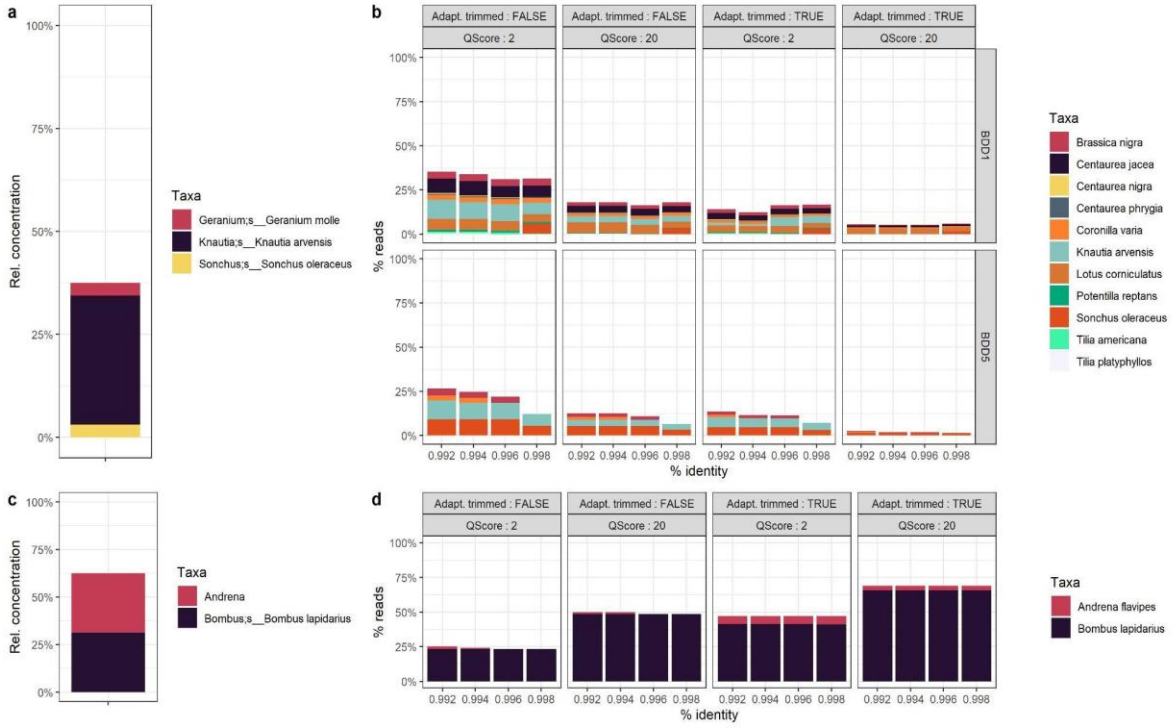


Figure 7 – Relative DNA concentration in mock sample (panel **a** for plants, panel **c** for insects) and relative reads counts observed (panels **b** and **d**) depending on adaptors trimming and quality score (horizontal sub-panels), identity threshold used (x-axis), and reference database (for plants only, vertical sub-panels; BDD1: all expected families retrieved from BOLD System, BDD5: local database built using SANGER sequencing). Note: colors are not necessarily the same for a given taxa in expected vs. observed composition.

Mock results indicate sometimes important discrepancies between expected and observed sample composition. Whether such differences were due to technical aspects or to human errors (especially mixing of sample's tags) was unclear. In order to test the hypothesis of human mistake, we compared sample composition for 110 samples that were amplified and sequenced twice in two different runs. That comparison allows us to show that despite high levels of variability, sample pairs were more similar to each other than random samples (Fig. 9). Results were similar regardless of the reference database used: at the species level, ca. 30% of the taxon identified were retrieved in both samples whereas each sample presents on average 50% of taxon not retrieved in the other. At the genus and family level, *ca.* 65% were retrieved in both sample and *ca.* 35% were not. Thus, mixing of samples' tags seems unlikely here.

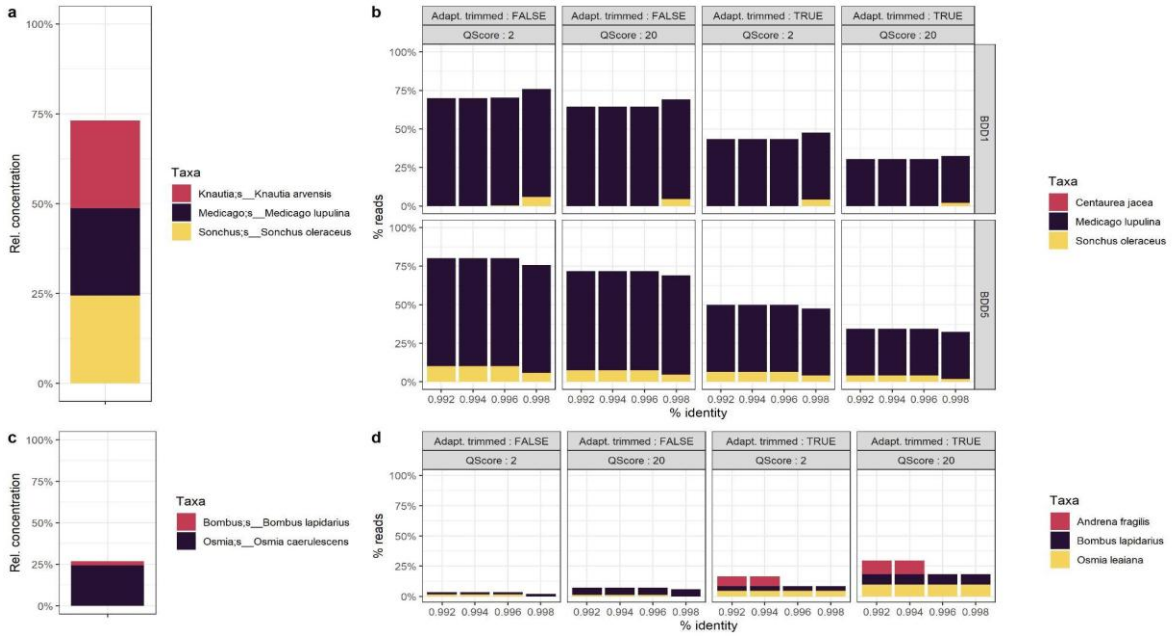


Figure 8 – Relative NA concentration in mock sample (panel **a** for plants, panel **c** for insects) and relative reads counts observed (panels **b** and **d**) depending on adaptors trimming and quality score (horizontal sub-panels), identity threshold used (x-axis), and reference database (for plants only, vertical sub-panels; BDD1: all expected families retrieved from BOLD System, BDD5: local database built using SANGER sequencing). Note: colors are not necessarily the same for a given taxa in expected vs. observed composition.

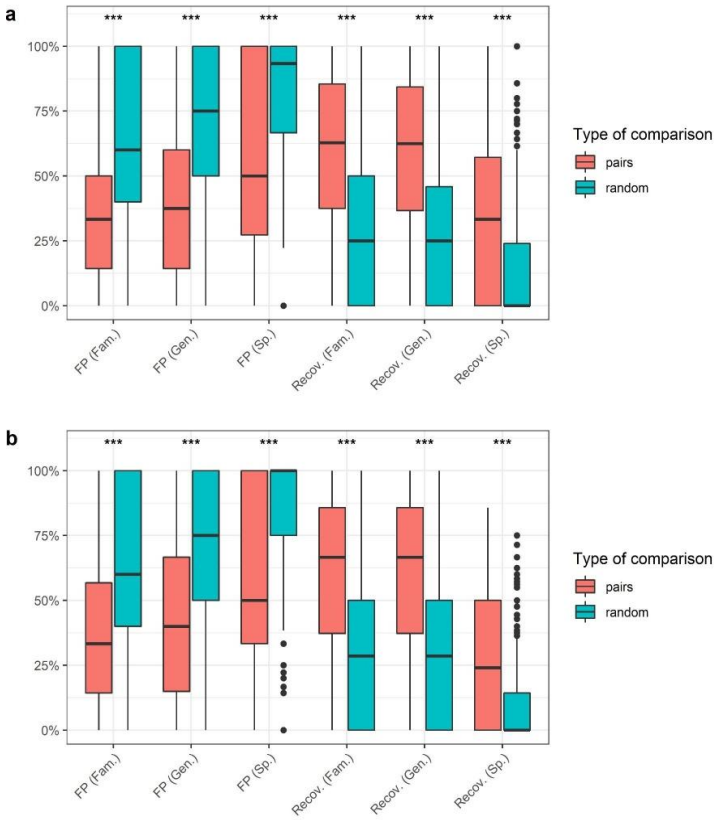


Figure 9 – Comparisons of families/genera/species retrieved from the same sample sequenced twice vs. from two different samples (taken at random) in system 1. "FP": false positive rate, "Recov.": proportion of taxon retrieved. Fam., Gen., and Sp., stand resp. for family, genus and species. Statistical differences were assessed using Wilcoxon-Mann-Whitney tests. (Panel **a**: comparisons made using unfiltered BOLD sequences, panel **b**: comparisons made using the local database).

3.3.2. Mock communities in the predator-prey agricultural network (system 2)

One of the major particularities of the arthropod predator-prey systems is the occurrence of a large quantity of predator DNA itself in the sample, that will be also amplified and acting as source of detection bias for the prey signals. In the context of system 2 (predator-prey interaction agricultural network), we provide inferred community profiles to explore i) if preys can be detected homogeneously, and ii) how the presence of the predator DNA in the gut content samples can affect our abilities to retrieve prey's DNA, when the primer pairs can amplify both. The two sample preparation types for consumer's gut content experienced for the system (i.e. regurgitate or whole body) were compared.

We created 6 mocks mixing increasingly diverse equimolar pools of highly probable preys collected in the winter wheat crop in Brittany (5 to 60 species from up to 11 different orders). We then created more realistic versions of each mock adding predator DNA, in a large quantity (60%, representing the regurgitate gut content for large carabid species, *Poecilus cupreus* in this experiment) or a very large quantity (90%, representing the whole carabid body extracts for small carabid species, here *Trechus quadristriatus*). We amplified the mocks with the primer pairs used for this system (Table 2) and sequence the products with a MiSeq Illumina Run V2 2x250bp (~12 million reads). Sequences have been treated and assigned with Dada2 pipeline using NCBI as a reference database.

Concerning the homogeneous detection of prey, for simple mock compositions, as the mock MPI composed of 5 major preys of agricultural value, we retrieved all prey when the mock was free of carabid DNA (MIP, Fig. 10a), but in biased proportions compared to expectation, with an over representation of earthworm and collembolan sequences compared to aphid and slug sequences. This result suggests that studies investigating the ecosystems service of a predator could underestimate and overestimate the predator's contribution to pest control (aphid and slug) and disservice predation (consumption of beneficial organisms). We also noted that assignations were approximate and do not allow for detection down to species level in most cases. For more complex mock compositions, as the mock MP composed of 60 potential preys species, we observed a similar pattern with high disproportions (MP, Fig. 10b), and with some species that were not found at all (centipede prey), or at the contrary found while they have not been put in the mock (millipede prey). The assignations problem should be investigated to determine if they are caused by a poor quality of reference database, or a primer pair not able to distinguish between all the preys used in the mock.

When adding carabid DNA in the mix, we observed a lower quality of those detections, mainly because of a delay in sequences number assigned to prey (compared to predator, not shown on Fig. 10a, in black in Fig. 10b). This pattern was more important for 90% predator DNA mock (MIP B90 / MP B90) than for 60% predator DNA mock (MIP B60 / MP B60) when considering the predator *P. cupreus*. However, the presence of carabid DNA does not seem to affect the disproportion of each prey's sequence. Interestingly, this result is less strong when considering another predator species, *Trechus quadristriatus* (MIP S90 Fig. 10a and MP S90 Fig. 10b). It could be because the DNA sequence of this species is less amplified by the primer pair used. We also noticed that this sequence is also less well assigned, as MP S90 (Fig. 10b) shows lots of Coleoptera DNA sequence that is probably *T. quadristriatus.*

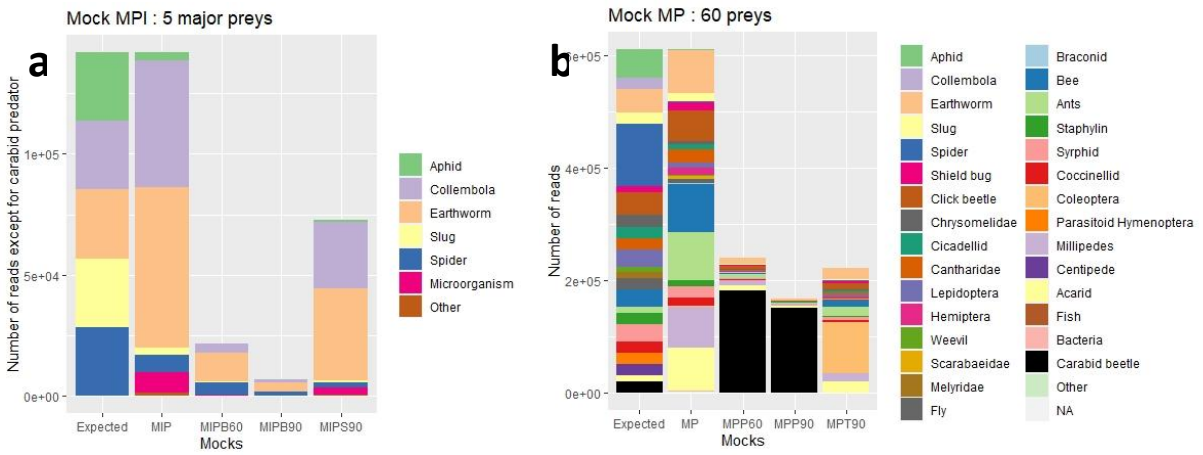


Figure 10 – Number of reads in two mock trials investigating the effect of predator (carabid) DNA on the prey's sequence detection, **a**) a simple mock composed of 5 major preys of agricultural value (MPI), and **b**) a complex mock composed of 60 potential preys species naturally present in the predator environment (MP). Preys are mixed in equimolar proportion in mocks. Predator DNA is added in various proportions and from 2 carabids species: 60% of *Poecilus cupreus* DNA (MIP B60 / MP B60), 90% of *Poecilus cupreus* DNA (MIP B90 / MP B90) or 90% of *Trechus quadriastriatus* DNA (MIP S90 / MP S90).

3.3.3. Mock communities in the microbiota network in fruit flies (system 3)

In the context of system 3 (fruit fly gut microbiota), we provide inferred community profiles on a commercial mock community of 8 bacterial species in non-equivalent proportions (Mock ZymoBIOMICS, Microbial Community DNA standard, ref. D6306) using either Oxford Nanopore Technology MinIONlong-read technology or more classical short-read Illumina MiSeq sequencing.

Short-read mock communities

Sequence production – In two independent experiments, the Zymo mock community was used four times as a control, resulting in eight mock samples, analyzed using the following protocol. ~15ng of extracted DNA was amplified using KAPA HiFi HotStart with specific primers that target a 251 bp portion of the V4 region of the 16S rRNA gene (Rombaut et al., 2017). After bead purification for removal of excess primers, amplification products were indexed using Illumina barcodes. Obtained libraries were checked for fragment size (Tapestation; Agilent, Santa Clara, USA) and concentration (Qubit; Thermo Fisher Scientific, Waltham, USA), prior to multiplexed, paired-end sequencing on the MiSeq platform (2x300 bp) (Illumina Inc., San Diego, USA). Microbiome bioinformatics were performed with QIIME2 2020.8 (Bolyen et al., 2019). Raw sequence data were demultiplexed and quality filtered using the q2-demux plugin followed by denoising and trimming with DADA2 (Callahan et al., 2016) (via q2-dada2). The number of ASVs per sample ranged from 5,156 to 10,523 across samples. Taxonomy was assigned to ASVs using the VSEARCH-based consensus taxonomy classifier on the Silva 138 database (via classify-consensus-vsearch with 97% identity, Rognes et al., 2016). Taxonomic assignation and feature tables were imported as a phyloseq object in R (McMurdie and Holmes, 2013) for further exploration. The percentage of assignments was very high (90.67%) for phylum down to family levels. It dropped to 79.11% for genus and 11.55% for species-level assignments. Hence we only present results at the genus level hereafter.

Results – In all eight samples, only seven of the eight expected genera were retrieved, *Salmonella* being missing from all profiles. Three false positive genera were detected. Two of them were always below 0.003 in relative abundance. The remaining one, *Gilliamella* (Orbaceae), was detected in four samples with a relative abundance ranging from 0.005 to 0.016 (Fig. 11 left panel). Mock profiling was relatively constant across samples. It was, however, less constant, and more different from the expectation than with long-read sequencing. At the family level, these observations were again

verified. In addition, all seven expected families were retrieved in all samples. Orbaceae were again detected in four samples while absent from the mock.

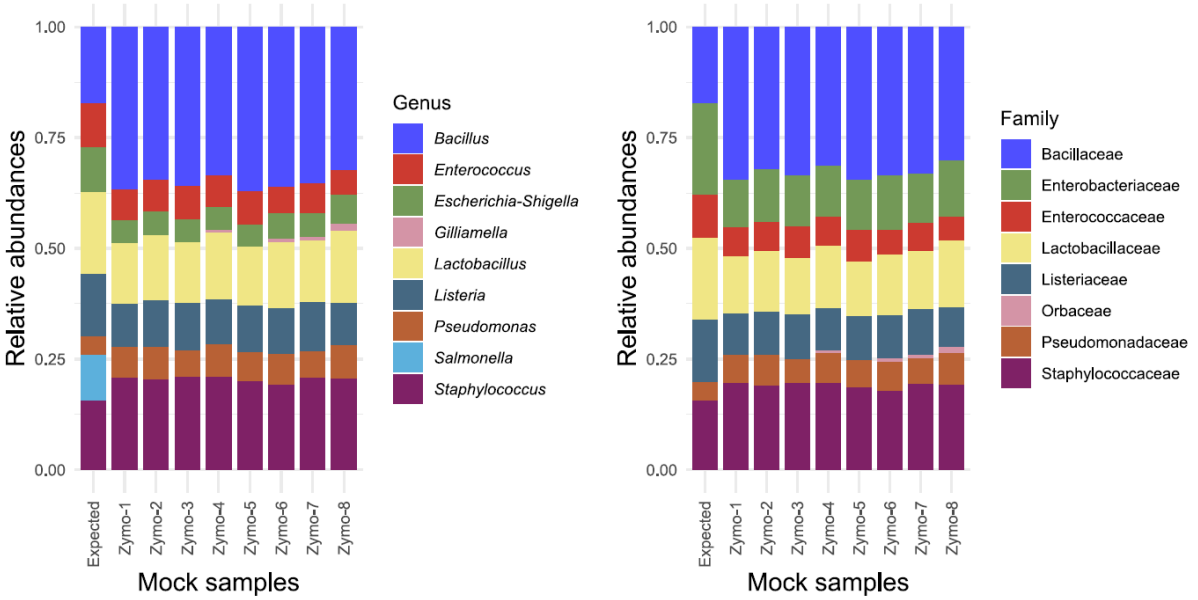


Figure 11 – Relative abundances of the eight bacterial taxa present in Zymo mock communities, as revealed through the seven short-read HTS experiments (a) at the genus level, (b) at the family level.

Long-read mock communities

Sequence production – In seven independent experiments, the aforementioned mock community was used as a control and analyzed using the following protocol. About 15ng of extracted DNA was amplified using specific primers that target the 16S rRNA gene (27F 5'-AGAGTTTGGATCMTGGCTCAG-3'; 1492R 5'-GGTTACCTTGTTACGACTT-3'), as well as subsequent specific barcodes using a 16S Barcoding Kit (SQK-RAB204, Nanopore). After bead purification for removal of excess primers, amplification products were attached to rapid sequencing adapters before loading on a flowcell for sequencing. Basecalling, demultiplexing and chimera removal were performed using Guppy v2.2.3 (https://community.nanoporetech.com) producing a total of 4,100,814 sequences, ranging from 468,516 to 809,834 across samples. Reads were trimmed (60 - 1400 pb) and filtered using Nanofilt (De Coster et al., 2018): only sequences longer than 900 pb and above quality score Q10 were kept, leading to a total of 99,107 sequences (ranging from 7,912 to 25,744 across samples). Taxonomy was assigned by confronting reads to the Silva 138 database (Quast et al., 2013, Yilmaz et al., 2014) using vsearch 2020.8.0 (Rognes et al., 2016) embedded in QIIME2 2020.8 (Bolyen et al., 2019), with perc_identity=0.92, max_accepts = 100, max_rejects = 100 and max_hits = 'all'. A phyloseq object was produced and imported in R (McMurdie and Holmes, 2013) for further exploration. The percentage of assignments ranged from 65.95% from phylum down to genus level to 65.02% for Species-level assignments.

Results – Examining mock samples revealed correct identification of mock taxa at the species level. In particular, *Pseudomonas aeruginosa*, which had the lowest expected frequency (0.04) was always retrieved at a congruent frequency, ranging from 0.02 to 0.04 across samples. The highest relative abundance of a false positive was 0.00328 and was associated with an uncultured bacterium. The highest relative abundance of a false positive associated with a named species was 0.00078 for *Bacillus velezensis* in RUN2C_barcode11. Mock community profiling was both very constant across runs and very close to the expectation (Fig. 12), with one noticeable exception, *Lactobacillus fermentum*, whose relative abundance was consistently underestimated in all samples.

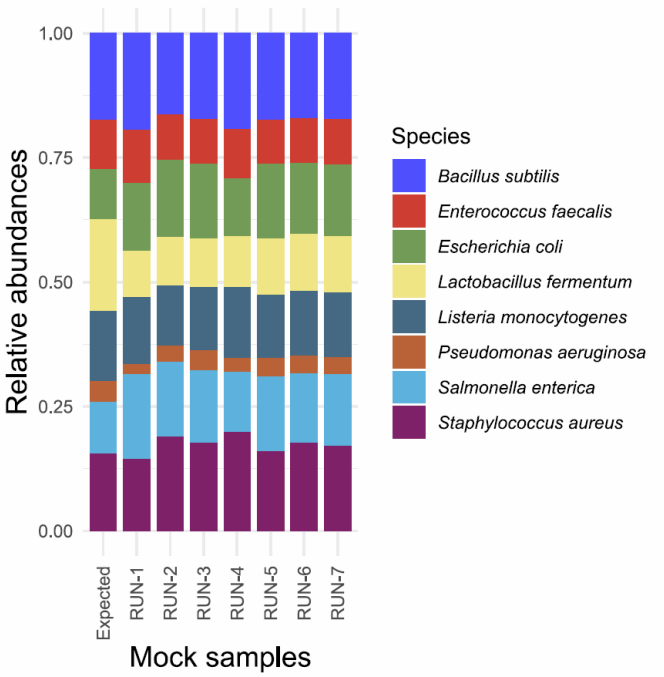


Figure 12 – Relative abundances of the eight bacterial taxa present in Zymo mock communities, as revealed through the seven long-read HTS experiments.

3.5. Conclusions of the mock community experiments

We found that using standard molecular biology and bioinformatics protocols, long-read sequencing resulted in correct identification of community composition at the species level (no false negative and very low-frequency false positive taxa). Relative abundances were very repeatable across experiments and relatively congruent with the expectation. Short-read sequencing did not allow working at the species level. At the genus level, one false negative (*Salmonella*) was observed. Both at genus and family levels one false positive (*Gilliamella*, Orbaceae) was found at non-negligible but lower relative abundance than true positive taxa. Relative abundances were relatively consistent across samples but quite distant from the expectation at both genus and family levels.

Given the still contrasted costs of long-read vs. short-read technologies, opting for one or the other should definitely be based on the question tackled. Short-read sequencing allows multiplying samples paying the risk of small errors of identification and abundance biases. Long-read sequencing seems to be more reliable as for taxa presence/absence, relative abundances especially at low taxonomic levels.

4. Network reconstruction using inference

Two methodologies of inferring networks are being used in NGB, representing the two broad classes of network inference from DNA relative abundances developed to date. These are the Poisson Log-Normal (PLN) network model, which is based upon statistical inference, and Abductive/Inductive Logic Programming (A/ILP) that uses logic-based inference. The aims, benefits and limitations of these two approaches are detailed in the following sections.

4.1. Inferring networks using Poisson log-Normal models

4.1.1. Gaussian Graphical Models

The question of network reconstruction has been a hot topic in applied genomics for about twenty years and the advent of microarray data. In statistical learning, powerful inference procedures have emerged in the framework of graphical modeling, such as Gaussian Graphical Models (GGM) for continuous data and Ising models for binary data (Yuan and Lin, 2007, Banerjee et al., 2008, Ambroise et al., 2009, Ravikumar et al., 2010). GGM have been successfully used to understand complex genetic regulations (Moignard et al., 2015, Fiers et al., 2018), to identify direct contacts between protein subunits (Drew et al., 2017) or to identify functional pathways associated to a disease (Yu et al., 2015). However, these methods have to be rethought to match the characteristic of ecosystem biomonitoring data. Count data do not follow a Gaussian distribution, they vary over many orders of magnitude and are often more dispersed than expected under a simple Poisson distribution. Furthermore, the observed counts may result from different sampling efforts in each sample and/or for each entity, which hampers direct comparison. It is also highly desirable to remove the effects of external covariates describing the environment to avoid spurious edges in the network.

4.1.2. Existing statistical methods for network from counts data

By analogy to the Gaussian graphical setting, many efforts have been devoted throughout the years to develop multivariate Poisson distribution in order to model dependencies between count variables (see Inouye et al., 2017 for a review). Unfortunately, there is no satisfying Poisson counterpart to the multivariate Gaussian: Besag (1974) proved that Poisson Graphical Models (PGM) are limited to negative dependencies to ensure proper joint distribution. Yang et al. (2012) proposed several variants, but all of them fail to have both marginal and conditional Poisson distributions. Also, observed count data often display a variance larger than expected under the Poisson assumption, so that a model that induces over-dispersion is highly desirable.

A different line of work used for microbial ecology in SPIEC-EASI (Kurtz et al., 2015), gCoda (Fang et al., 2017) or BAnoCC (Schwager et al., 2017) addresses the problem by (i) replacing counts with (regularized) frequencies and (ii) taking their log-ratios before (iii) moving back to the GGM framework. A positive side effect of this transformation is to remedy the compositionality problem that counts cannot be compared among samples as they depend on a sample-specific size-factor, which may induce spurious negative correlations. The transformation is simple but prevents integration of heterogeneous data sources and thus discovery of interactions between nodes from different sources (e.g. bacteria and fungi), although important ones have been experimentally documented (Lima-Mendez et al., 2015). In the same spirit but with different statistical tools, Cougoul et al. (2019) rely on copulas to take into account the non-Gaussian nature of the data.

4.1.3. Sparse Multivariate Poisson Lognormal Model

In Chiquet et al. (2019), we adopt a different standpoint by building on the multivariate Poisson log-normal (PLN) model of Aitchison and Ho (1989), a model that belongs to the family of Joint Species Distribution Models (JSDM) which are known in ecology for providing a general multivariate framework to study the joint abundances of all species from a community (see Warton et al., 2015 for a general presentation). The idea of JSDM is to take into account both structuring factors (e.g., environmental gradients, nutrients availability, etc.) and potential interactions between the species (competition, mutualism, parasitism, etc.). Considering both effects at once is instrumental in disentangling meaningful ecological interactions from mere statistical associations induced by environmental drivers and/or habitat preferences. The PLN model relies on the same hierarchical backbone as many JSDM: dependencies are first modeled in a latent layer through the covariance matrix of a multivariate Gaussian vector and counts are then sampled independently conditionally to this latent vector of expected (transformed-)abundances with a Poisson distribution. Such an approach enables arbitrary correlation signs and over-dispersion of the counts. Dependencies between counts are captured by the covariance matrix of the latent vector, whereas environmental effects are accounted for in the vector mean value. This distinction is convenient from a modeling

point of view, as it typically separates a regression part that accounts for abiotic effects from a random part that accounts for dependency between species (biotic effects).

The PLN-network extension that we introduced in Chiquet et al. (2019) is the analog of the graphical-lasso (Banerjee et al., 2008, Friedman et al., 2008) for the inference of interaction networks. Formally, species can be associated but they are in direct interaction only if they are still dependent after conditioning on both the covariates (abiotic effects) and all the other species (biotic effects). In the latent Gaussian layer, this distinction coincides with the difference between correlation and partial correlation. Correlations between pairs of species are captured by the covariance matrix whereas partial correlations are encoded by its inverse - the precision matrix. Because the network is usually supposed to be sparse (i.e., only a few pairs of species are expected to be in direct interaction), we add a so-called sparsity-inducing constraint on the precision matrix by resorting to the l1-norm just like (graphical)-lasso. At the end of the day, the PLN network model can be viewed as a PLN model with a constraint on the coefficients of the precision matrix, or equivalently, on the partial correlation.

Fitting such a model requires the optimization of a penalized likelihood where the likelihood term is not easily tractable. We thus resort to a variational approximation for parameter inference and solve the corresponding optimization problem by alternating a gradient descent on the variational parameters and a graphical-Lasso step on the covariance matrix. We also select the sparsity parameter using the resampling-based StARS procedure. Details are available in Chiquet et al. (2019) and distributed as an R/C++ package in Chiquet et al. (2021).

4.1.4. Caveats and limitations

We want to distill some remarks on the possible uses and the limits in terms of interpretation of this approach. First of all, and this statement is true for all methods relying on an interpretation based on graphical models (graphical-lasso and SPIEC-EASI for example, and most JSDM), the user must keep in mind that the edges of the reconstructed network have a precise meaning only from a statistical point of view: their value is directly proportional to a measure of partial correlation between species abundances. It is tempting to interpret them in terms of interactions between species (trophic for example), but this must be subject to caution. Moreover, the statistical performance of these methods is obviously subject to the convergence of the corresponding estimators, and in particular of the ratio n/p (number of sites/number of species) and of the expected number of interactions per species. It should also be kept in mind that the network finally reconstructed and retained for interpretation is linked to a choice of hyper-parameter that controls the number of interactions, for which a small variation can significantly change the topology of the reconstructed network. Therefore, the statistical analysis of the reconstructed network with indicators such as betweenness, clustering coefficient, or by graph-partitioning, cannot be performed with as much confidence as for a biological network validated by actual experiments. It is generally wiser to consider for the interpretation a set of reconstructed networks corresponding to several values of the hyper-parameter because the conclusions drawn on this set of networks will be more robust and less likely to be due to sampling fluctuations. Finally, in the particular case of latent models such as PLN and most JSDMs, the user must keep in mind that the reconstructed network corresponds to the interactions in the latent layer and not to the observed count vector level. The overlap between the two is not true in general.

4.2. Inferring networks using Abductive/Inductive Logic Programming

The aim of logic-based learning of ecological networks is to examine the exciting possibility that we can extend the reconstruction of networks beyond simple association webs by *directly* identifying the types of interaction that occur between species in any community or ecosystem, including those communities for which little information to guide interpretation of correlational associations is

available. Abductive/Inductive Logic Program (A/ILP), in the language Progol 5, learns or infers hypotheses using abductive logic. This can be described as a reasoning that seeks the simplest, likely conclusion from the set of available observations. A hypothesis produced by A/ILP can therefore be seen as the 'best available' explanation, obliging future testing and validation. A/ILP has been used in many different fields of knowledge, from metabolic network inference (Tamaddoni-Nezhad et al., 2006) to elaborating the processes involved in cow milk production (Sasaki et al., 2019). It has also been used to infer promising results for arthropod trophic interactions in data from farmland system very similar to System 2 (Bohan et al., 2011).

The Progol 5 process of abductive learning uses relatively simple, logical statements to infer interactions that might occur between any two ASVs that are in the sample data. The inference process uses the number of counts of each ASV as a measure of its abundance in a sample, measured as the sequence count. It starts with the creation of a matrix of 'ASV change' of each ASV across the samples, which has the benefit of controlling for some aspects of compositionality in the abundance data noted in section 4.1. ASV change is characterized as a variable for an increase (*up*), decrease (*down*) or no change (*no change*) in abundance for any given ASV between two samples. To calculate these changes, we treat the number of sequence reads in the sample data-set as count data and do a χ^2 test on a 2 x 2 contingency table, where the sequencing depth of a sample is considered the total population count. A non-significant χ^2 statistic indicates that there is *no change* or difference in sequence count of ASV_1. A significant test statistic would indicate that there is a change, and this difference is assigned an appropriate *up* or *down* value in the ASV change matrix.

The A/ILP learning is being used to examine system 5 for the interactions that occur between bacteria and fungal species within the microbiota of the leaf phyllosphere. This is with a view to understanding how the microbiota changes structurally with drought and whether reconstructed networks might be used as an indicator for the biomonitoring of drought stress. The logical statements for interactions are constructed around the idea that past or ongoing interactions between two microbial ASVs will have led to the change in sequence counts that we observe. Conceptually, therefore, ASV_1 and ASV_2 might have or be undergoing an interaction if there is some pattern to the changes of the two ASVs across the dataset. With A/ILP, we relate the abundance change to the presence of species, rendered as symbolic logic to infer those patterns. The models are encoded in Progol 5 using logical statements, such as:

presence(*y*,ASV_2,yes):-
presence1(*x*,ASV_2,no),
abundance(*x*,*y*,ASV_1,up),
effect_up(ASV_2,ASV_1).

presence(*y*,ASV_2,yes):-
presence1(*x*,ASV_2,no),
abundance(*x*,*y*,ASV_1,down),
effect_down(ASV_2,ASV_1).

which state that ASV_2 is causing an effect (either *up* or *down*) on the abundance of ASV_1 when there is an abundance change from sample *x* to sample *y* and ASV_2 is present in sample *y* and not present in sample *x*. A/ILP also allows the inclusion of background, ecological information, where this exists, to improve the learning (Tamaddoni-Nezhad et al., 2013). This might include microbial functional group information from databases like FUNGuild (Nguyen et al., 2016) or FAPROTAX (Louca et al., 2016), coded as logic rules.

The abduction process produces a list of hypothetical effects between pairs of ASVs. For each hypothesis we obtain the following information: (i) the pair of interacting ASVs; (ii) the effect on the abundance; (iii) a compression value. It is this value of compression, which is computed by Progol 5 using the logical statements, which determines whether an effect can be supported as a hypothesis of interaction between two ASVs. Compression is a logical measure of the amount of information that supports a possible inferred interaction by contrast to the total amount of information in the dataset, and is particularly effective for noisy, biological data (Muggleton, 1995, Muggleton and Bryant, 2000).

We adopt the interaction motifs described by Derocles *et al*. (2018) to construct the relation between the significant effects on sequence count and the ecological mechanisms (Table 3). As an example, the logical statement for a competition interaction compares the change values of any two ASVs, say ASV_1 and ASV_2 across two samples, and would compute a competition interaction if both of the ASVs have an effect down caused by the other. Different effect combinations allow the inference of the ecological mechanisms of mutualism, competition, predation commensalism and amensalism (Table 2, *sensu* Derocles et al., 2018).

Table 3 – Relationships between the ecological mechanisms of an interaction and the motif effects observed on the sequence counts or the interacting ASVs. Table follows the description in Derocles et al. (2018)

Type of interaction mechanism	Effect on ASV_1 count	Effect on ASV_2 count	Nature of interaction
Mutualism	*up*	*up*	Mutual benefits to both ASVs
Competition	*down*	*down*	ASVs have negative effect on each other
Predation/Parasitism	*up*	*down*	Predator/Parasite ASV develops at the expense of the Prey/Host ASV
Commensalism	*up*	*no change*	ASV_1 benefits while ASV_2 is unaffected
Amensalism	*down*	*no change*	ASV_2 has a negative effect on ASV_1, but ASV_2 is unaffected

Assessing the predictive accuracy of a network inference tool is done using either computer generated datasets where interactions are known (Röttjers and Faust, 2018) or by measuring other properties of the inferred network (Barroso-Bergadà et al., 2021). This uses 'consensus' networks, where those interactions that exist in several network examples are pooled, maximising the likelihood of the hypotheses being common to the whole system. The ecological veracity of any link can currently only be determined by discussion with expert microbial ecologists and bibliographic searches of the literature, as has previously been done for arthropod networks (Tamaddoni-Nezhad et al., 2013). We can, however, evaluate the methodological performance of the learning across the

consensus networks. As detailed in Tamaddoni-Nezhad et al. (2013), the predictive accuracy of different ILP explanations can be evaluated using *fold* validation, by randomly splitting the change matrix into a number, *n*, of equally sized folds. The predicted interactions and their compression values, inferred from *n*-1 folds, can then be used to predict the abundance change between two samples contained in the excluded fold data as the probability that this prediction is realized (estimated as Accuracy, Fig. 13).

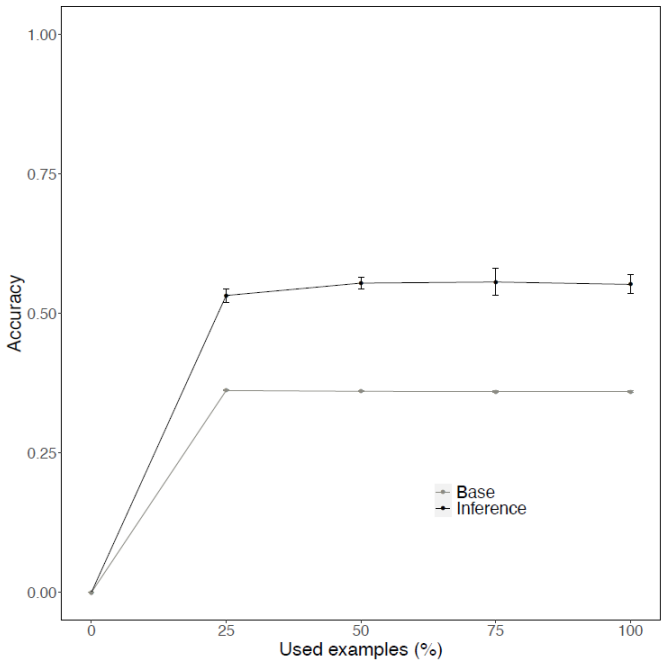


Figure 13 – Change in inference accuracy over a 5-fold validation, using real data. Random subsets of the 30 most abundant ASVs were selected for the folds. Here the 5-fold validation mean and SD are plotted for each combination of the learning and predicted fold examples. Similar predictive accuracy is found for all inferred folds and these are significantly higher than the majority class (grey line). There is some evidence of model overfitting when all folds are used for learning at 100%.

Our initial A/ILP work has been directed towards developing a method of assigning a statistical significance to the value of compression for any given effect using bootstrapping (Barroso-Bergadà et al., 2021). This work uses simulated microbial data-sets generated using the simulation approaches in Weiss et al. (2016), in place of the ASV table. Simulated data-sets were produced for each of different interaction types in Table 3., computed matrices of change, in the manner described above, and learned with A/ILP and a statistical network inference approach benchmarked by Weiss et al. (2016), i.e. SparCC (Friedman and Alm, 2012). This work demonstrated both that the compression of any given link can be assigned a statistical significance using bootstrapping, and also that A/ILP logical statements for specific link types can detect the presence of simulated links at least as well as the statistical approaches. This suggests to us that A/ILP might be used on real ASV matrix data to learn link interaction types directly.

5. Network comparison

5.1. Background and motivations

The main objective of sampling and reconstructing ecological interaction networks is to assess or monitor the influence of the environment on the interactions or to compare ecosystems. The notion of network comparison covers a large number of situations and objectives, each of them requiring a

specific statistical method. For example, networks collected at different places, at different times, in different conditions (wet and dry season for instance) could all be compared. One may also be interested in comparing the organisation of a common group of species when in interaction with several other groups of species, e.g. comparing the interaction strategies of plants with respect to pollinators (a mutualistic relation) and herbivores (an antagonistic interaction).

Obviously, a naive link-to-link comparison is in general not feasible since, when considering HTS data, these links are reconstructed with a certain level of uncertainty that should be taken into account and in general the networks at stake do not involve the same species (at least partially). As a consequence, the comparison must be performed in terms of topological properties (i.e. organisation) of these networks assuming that these structures are a macroscopic representation of their functional organization.

Many techniques exist in the literature to summarize the topology of networks (see Delmas et al., 2019 and references therein) and many of them have been used to assess changes in networks in space or time (Pellissier et al., 2018, Song and Saavedra, 2020, Fortin et al., 2021). Descriptive statistics such as size, connectance, or nestedness may be calculated. Community detection can also be performed to highlight groups of species more connected within their community than without.

However, although widely used in practice, these techniques have reached their limit. First of all, they are well defined for binary interactions but their definition is less clear if the interaction is weighted or if the available information is a probability of connection. In that context, a standard approach to assess the “strength” of a discovered structure is to compare the obtained value on the observed network with its distribution on a population of networks sampled randomly from the original network respecting the same degree sequences for instance. This strategy has the great advantage of being non-parametric (meaning that no assumption is made on the network), but its extension to assess differences between networks reconstructed with uncertainty is far from being easy. Moreover, many of these global statistics are interdependent, but their relationship is complex. For example, nestedness and modularity are known to be correlated, but the nature and the intensity of their correlation depends on the value of the connectance (Fortuna et al., 2010). Hence, even if one could conclude that the differences in observed values are the result of actual differences in the organization of the ecosystems, one can only interpret them with difficulty.

A concurrent approach assumes that the observed networks are the realizations of a parametric probabilistic model and fit the parameters adapted to the observed network. Many probabilistic models have been proposed in the literature to mimic ecological networks. Among them, one can cite Stochastic or Latent Blocks models (Mariadassou et al., 2010), also referred as group models (Allesina and Pascual, 2009), and their degree-corrected versions, i.e. expected degree distribution models (Chung and Lu, 2002, Ouadah et al., 2021) or Latent space models (Hoff et al., 2002). Note that all these models have the property to be generative hierarchical models, and so are flexible enough to handle non-binary interactions, missing interactions, the effect of covariates etc. Moreover, these models do not set any structure a priori: they are agnostic insofar as they will discover any structure present in the network or possibly no structure if all the considered species apply the same non-organized ecological strategy. Although parameter inference requires sophisticated optimization algorithms, the probabilistic framework leads to standard hypothesis testing strategies with theoretically provable asymptotic guarantees.

Once a strategy to decipher the structure of the networks has been chosen, two approaches may be considered. Often the object of interest can be the role of a particular species or a particular group of species in several observed networks, corresponding to conditions or ecosystems. In that case, a multipartite ecosystemic strategy can be considered as it is done in Bar-Hen et al. (2020). By contrast, the labels (i.e. the species) might not be relevant, and the focus would then be the mesoscale structure. In that case, exchangeable models, i.e. probability models which are unchanged if the labels are switched, can be used. From this symmetry, one may identify the limiting distribution of some objects used to capture the properties of a network. For example, U-statistics (described

below) are a class of estimators on the whole network, formed as an estimator on a small sub-network, averaged over all the sub-networks of the network. The distribution of U-statistics are shown to be asymptotically normal for exchangeable models (Le Minh, 2021), which allows the construction of statistical hypothesis tests.

5.2. A novel unlabelled network comparison method

In the following example, we show how U-statistics can be used to compare networks. We consider weighted bipartite networks, e.g. plant-pollinator interaction networks in which the intensity of interactions is measured. Let Y_{ij} be a positive integer representing the number of visits of insects of species i on the plants of species j. Y is the weighted incidence matrix of the network of size $m \times n$, where m is the number of insect species and *n* the number of plant species. We assume that Y is generated by a weighted bipartite version of the expected degree distribution model (WBEDD, Le Minh, 2021).

In this model, each insect species (respectively plant species) draws an expected total number of interactions from some distribution characterized by a monotonic function f (respectively g), and random variables ξ_i and η_j, drawn for all insects and plants, respectively:

$$\xi_i, \eta_j \overset{iid}{\sim} \mathcal{U}[0,1].$$

Then Y_{ij} follows a Poisson distribution determined by the expected number of interactions of species i and j, which is obtained using functions f and g together with random variables ξ_i and η_j:

$$Y_{ij} \mid \xi_i, \eta_j \sim \mathcal{P}(\lambda f(\xi_i) g(\eta_j)).$$

The model is characterized by the density $\lambda > 0$ and the f and g functions, normalized such that $\int f = \int g = 1$. One can measure how these distributions are imbalanced by calculating for some $k > 1$, $F_k = \int f^k$ for the insect species and $G_k = \int g^k$ for the plant species. These quantities can be understood grossly as analogues to statistical moments (i.e. variance, skewness, etc.), but applied to the components underlying the structure of Y_{ij}.

If the asymmetry of connections among insects is the question of interest, one might design a test using F_2, for example $\mathcal{H}_0 : F_2 = 1$ against $\mathcal{H}_1 : F_2 > 1$. In order to estimate F_2, one can use the following U-statistics, named $T_{m,n}$ and $U_{m,n}$:

$$T_{m,n}(Y) = \binom{m}{2}^{-1} \binom{n}{2}^{-1} \sum_{\substack{1 \le i_1 < i_2 \le m \\ 1 \le j_1 < j_2 \le n}} \frac{1}{2} (Y_{i_1 j_1} Y_{i_1 j_2} + Y_{i_2 j_1} Y_{i_2 j_2})$$

and

$$U_{m,n}(Y) = \binom{m}{2}^{-1} \binom{n}{2}^{-1} \sum_{\substack{1 \le i_1 < i_2 \le m \\ 1 \le j_1 < j_2 \le n}} \frac{1}{2} (Y_{i_1 j_1} Y_{i_2 j_2} + Y_{i_1 j_2} Y_{i_2 j_1}),$$

where $\mathbb{E}[T_{m,n}(Y)] = \lambda^2 F_2$ and $\mathbb{E}[U_{m,n}(Y)] = \lambda^2$, i.e. the expectations of the two U-statistics are obtained from network density, λ, and the function organizing the distribution of connections per insect, f.

Let $N = m + n$. If m / N has a finite limit (noted c) when N goes to infinity (i.e. for infinitely large networks, the ratio of insect to plant is assumed to converge to a finite value), then the limiting

distributions of $T_{m,n}(Y)$ and $U_{m,n}(Y)$ are jointly normal. Hence $Z_{m,n}(Y) = T_{m,n}(Y) / U_{m,n}(Y)$ is also asymptotically normal, which allows the construction of acceptance intervals for this test based on the asymptotic confidence intervals of $Z_{m,n}(Y)$.

Now if we take two observed networks and their adjacency matrices Y^A of size $m^A \times n^A$ and Y^B of size $m^B \times n^B$, generated by (λ, f^A, g^A) and (λ, f^B, g^B) (i.e. both weighted networks are assumed to have the same density parameter), we can state that the insect species interactions are equally distributed in the two networks if $f^A = f^B$ (hypothesis $\mathcal{H}_0$). Under this hypothesis, $F_2^A - F_2^B = 0$. This hypothesis can be tested using the previously defined U-statistics: under $\mathcal{H}_0$, $Z_{m^A,n^A}(Y^A) - Z_{m^B,n^B}(Y^B) = T_{m^A,n^A}(Y^A) / U_{m^A,n^A}(Y^A) - T_{m^B,n^B}(Y^B) / U_{m^B,n^B}(Y^B)$ is asymptotically normal.

Since it is built upon asymptotic confidence intervals, the performances of this test are improved for larger networks. This is confirmed by simulation results as shown in Fig. 14.

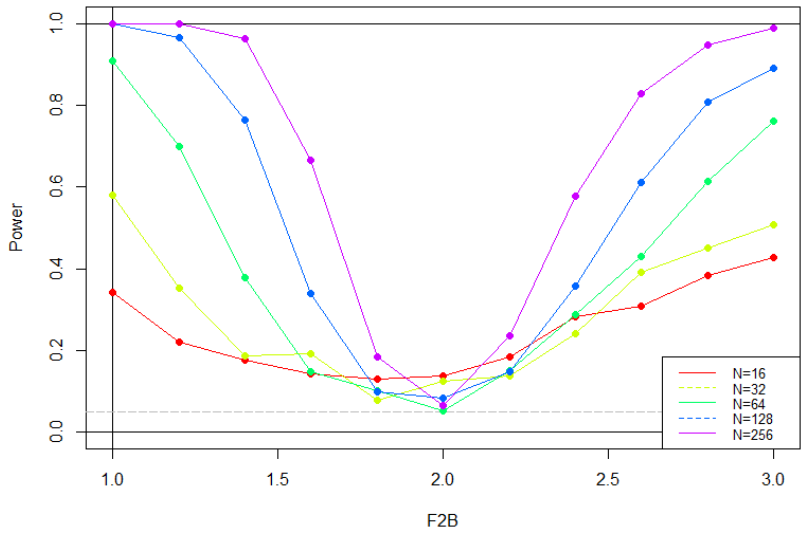


Figure 14 – Results of simulations, in which networks Y_A and Y_B of the same size $N/2 \times N/2$ were generated under the WBEDD model, with f^A and f^B being power functions, e.g. $f(u) = u^\alpha$ with $\alpha \geq 0$, which means $F_2 = (\alpha+1)^2 / (2\alpha+1)$. Set $F_2^A = 2$, this figure shows the reject rate of the test using asymptotic confidence intervals at level 0.95 for different values of F_2^B in a [1,3] range, for different values of N. It is expected from a useful test that the reject rate is equal to 0.05 when $F_2^B = F_2^A$ and 1 when F_2^B is very different from F_2^A.

6. Discussion

We have explored the possibility of using network reconstruction methods, by inference, on HTS DNA data sampled from the environment. This was with a view to creating a next generation of biomonitoring (NGB) of agricultural environmental change at ecosystem, landscape and higher scales. We have shown that there are a number of scientific and methodological opportunities presented by this possibility, which could greatly expand our understanding of ecological functioning and response to perturbation and change. We also highlighted pitfalls that if not avoided could lead to significant data biases that could render the networks learnt for this purpose invalid. Our work to

date does not provide a definitive description of which methodology to adopt to achieve the next-generation of biomonitoring envisioned, but it does show that many of the problems we identified could be solved with further research and development. Importantly, our preliminary results provide no evidence that NGB would not work.

The likelihood that an NGB approach to biomonitoring will not work will be greatly diminished by appropriate choices that we have termed the 'when', 'what' and 'how' of sampling for DNA in the environment. The systems studied within the NGB project were deliberately chosen to represent the widest possible range of situations we could imagine in agriculture. The biological scales of organization ranged from the gut and leaf phyllosphere microbiomes to the macrobiomes of freshwaters, grasslands and arable agriculture, and covered water, soil and aerial biomes. The systems each had, therefore, their own peculiarities that need to be taken into account, particularly with respect to their specific ecology and the appropriate protocols of sampling DNA therein. We believe, however, that the biomonitoring of these systems has similarities that give us some optimism that a generic approach to NGB may be possible, built around asking 'when', 'what' and 'how'.

The *when* of biomonitoring - Ecological networks have until recently been considered as static objects that in effect do not change in time or space (but see e.g. Kaiser-Bunbury et al., 2017). This was due to the often considerable effort, in time and manpower, required to construct a single network that led to a relatively low number of examples or replicates. In principle, NGB with network inference removes this limitation, allowing large numbers of samples to be taken and many example networks to be created. It may be that for a given system a single, composite network will represent its functioning well. For the most part, though, systems are subject to considerable natural variability that with the effects of perturbation imposes a dynamical behaviour that should be reflected in the scales, frequencies and replication of sampling. Considering the network as a dynamic object allows both natural and perturbation-induced changes to be understood and detected, such as the changes due to network rewiring (e.g. based on diversity measures of links among networks, Ohlmann et al., 2019) or species switching between different functional groups (e.g. using time-dependent stochastic block models, Matias and Miele, 2017), but can further exacerbate problems of network reconstruction because it introduces biases into the data, such as those due to spatial and temporal autocorrelation.

The *what* of biomonitoring - It is evident that we cannot sample all biodiversity all the time. The costs of time and human resource are prohibitive and the biological and ecological knowledge necessary to interpret these kinds of data are just not there yet. Pragmatic choices and what to sample need to be made. By choosing to use ecological networks as a methodology for representing an ecosystem, we are clearly making a statement about our understanding of the drivers of change in that system that the sampling approach should again reflect. The approach should therefore respond to whether we expect that the dominant species or links are the ones that lead to change. This would mean that ASVs with relatively low sequence count in the HTS would be excluded. Alternatively, it might be hypothesized that a particular subset of interactions are important because these drive the ecosystem service of interest (e.g. Dee et al., 2017). Those interactions that play this role may not be the most abundant or obvious in the sample, and need to be targeted specifically or risk being swamped and lost in the global picture provided by the HTS. Finally, this global picture itself may be conjectured to be the target of the biomonitoring as a way of evaluating the relative abundances and change in interactions across the ecosystem following perturbation. This global approach can also drive the further scientific discovery of unknown links and functions, and the balance between different types of functions and ecosystem services as part of filling in our currently missing knowledge.

The *how* of biomonitoring – The precision of biomonitoring using DNA present in the environment is in large part determined by the amount of samples taken. DNA is typically very diluted in environments, and thus nearly undetectable, in most samples (Carraro et al., 2021). The spatial and

temporal pattern of sampling needs to be constructed to take this level of dilution into account, thereby allowing species presence, activity periods and preferred habitats to be inferred to answer the questions raised by the when and what of the biomonitoring task. Sampling of DNA in order to infer interaction networks through HTS can take active or passive forms. Trapping can be used to target specific species grouping or functions but with the high overhead of skill and equipment costs that can limit the number of samples taken. Trapping is often situation-specific, which can in turn limit the ability to compare across situations (see e.g. Westphal et al., 2008, Prendergast et al., 2020 on different techniques to sample bees). Passive techniques are by contrast, easier to deploy and have the benefit that they require much lower overheads of expertise and equipment. The signal contained in a sample can also be specifically augmented using techniques such as dissection. This can remove many of the background signals that may not be of interest, such as the overrepresentation of host organisms in whole-body versus gut samples or similarly the signal of the microbiota from other organs when the gut microbiota alone is of interest.

Even with appropriate structuring of the 'when', 'what' and 'how' of sampling for DNA in the environment, NGB will be subject to a myriad of biases. The biases include those associated with HTS and bioinformatics. These biases and analyses to uncover the sources of biases using mock communities are extensively described in Section 3 of this paper. The sampling is also a potential major source of bias, with the potential for inappropriate links being enforced or the strength of links being augmented by the sampling procedure. For example intra-guild trophic interactions between predators, such as carabids and spiders, can be enforced by a concentration effect with the pitfall trap sampling approach that is commonly used. Passive sampling approaches, such as the use of pan traps, produces a sample soup that means that it is difficult to tease out the specific interactions that occur between particular insects and pollen, for example.

It is here that inference procedures might enter into the process of NGB. Our initial expectation was that with appropriate work it would be possible to develop inference methodologies that would both control for the biases sampling and HTS data and appropriately detect and classify the links that occur within the DNA data. The NGB project has used PLN and A/ILP, respectively representing the two main classes of inference methodologies; statistical and logical. These statistical and logical methodologies for recovering network structure are in the process of rapid development. What the work done to date within NGB has demonstrated is that the theoretical framework and statistical underpinnings of network reconstruction are now well established. The PLN approach shows that with appropriate formulation of the biomonitoring question, it is possible to build networks of association that ecologists can interpret in respect of natural variation and environmental perturbation. The A/ILP logical framework is maybe not so far advanced, but shows promise as a method to identify the mechanisms of interactions directly from HTS data, without the process of interpretation, and with a sensitivity of link detection close to competing statistical methodologies.

Once networks are discovered through automated inference procedures, the ultimate step of NGB is to be able to assess the amount of change between networks, both in space and time, e.g. before and after some perturbation. As explained in section 5, some methods are able to evidence changes in network structure, but with a focus on particular groups, e.g. by gauging the diversity of species and links shared or not by two (or more) networks (Ohlmann et al., 2019). By contrast, some other methods such as the U-statistic-based approach detailed in section 5, can compare networks without making explicit reference to species names, and thus can compare the structure of ecological networks sharing no common species. More generally, most network models which are based on the inference of a latent component determining species interactions (e.g. stochastic block models, weighted expected degree distribution models, etc.) could in theory lead to comparisons insensitive to species names through comparisons of the latent components determining node connections rather than the characteristics of the nodes (i.e. through graph embedding and distances between graphs). This area of research is rapidly advancing and might hopefully result in other ways to accomplish the ultimate goal of NGB – comparing networks and assessing whether the observed variation in network structure corresponds to "natural" expectations.

Acknowledgements

This study was funded by the following projects and contracts: ANR NGB (17-CE32-0011), ANR EcoNet (18-CE02-0010), ANR ARSENIC (14-CE02-0012), BCMicrobiome (Consortium Biocontrôle), Nouvelle-Aquitaine Region Athene project (2016-1R20301-00007218). The authors thank the French Ministère de l'Enseignement Supérieur et de la Recherche, the Hauts-de-France Region and the European Funds for Regional Economic Development for their financial support. All sequencing experiments have been done at the PGTB (grants from the Conseil Régional d'Aquitaine n°20030304002FA and 20040305003FA, from the European Union FEDER n°2003227 and from Investissements d'Avenir ANR-10-EQPX-16-01).

References

Aires, T., Moalic, Y., Serrao, E. A. & Arnaud-Haond, S. (2015) Hologenome theory supported by cooccurrence networks of species-specific bacterial communities in siphonous algae (Caulerpa). *FEMS Microbiology Ecology,* **91**.

Aitchison, J. & Ho, C. H. (1989) The multivariate Poisson-log normal distribution. *Biometrika,* **76,** 643-653.

Allesina, S. & Pascual, M. (2009) Food web models: a plea for groups. *Ecology Letters,* **12,** 652-662.

Alonso, P., Blondin, L., Gladieux, P., Mahé, F., Sanguin, H., Ferdinand, R., Filloux, D., Desmarais, E., Cerqueira, F., Jin, B., Huang, H., He, X., Morel, J.-B., Martin, D. P., Roumagnac, P. & Vernière, C. (2020) Heterogeneity of the rice microbial community of the Chinese centuries-old Honghe Hani rice terraces system. *Environmental Microbiology,* **22,** 3429-3445.

Alonso, P., Gladieux, P., Moubset, O., Shih, P.-J., Mournet, P., Frouin, J., Blondin, L., Ferdinand, R., Fernandez, E., Julian, C., Filloux, D., Adreit, H., Fournier, E., Ducasse, A., Grosbois, V., Morel, J.-B., Huang, H., Jin, B., He, X., Martin, D. P., Vernière, C. & Roumagnac, P. (2019) Emergence of Southern Rice Black-Streaked Dwarf Virus in the Centuries-Old Chinese Yuanyang Agrosystem of Rice Landraces. *Viruses,* **11,** 985.

Ambroise, C., Chiquet, J. & Matias, C. (2009) Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics,* **3,** 205-238, 34.

Astegiano, J., Massol, F., Vidal, M. M., Cheptou, P.-O. & Guimarães, P. R., Jr. (2015) The robustness of plant-pollinator assemblages: Linking plant interaction patterns and sensitivity to pollinator loss. *PLoS ONE,* **10,** e0117243.

Balvanera, P., Pfisterer, A. B., Buchmann, N., He, J. S., Nakashizuka, T., Raffaelli, D. & Schmid, B. (2006) Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology Letters,* **9,** 1146-1156.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *the Journal of machine Learning research,* **9,** 485-516.

Banza, P., Belo, A. D. F. & Evans, D. M. (2015) The structure and robustness of nocturnal Lepidopteran pollen-transfer networks in a Biodiversity Hotspot. *Insect Conservation and Diversity,* **8,** 538-546.

Bar-Hen, A., Barbillon, P. & Donnet, S. (2020) Block models for generalized multipartite networks: Applications in ecology and ethnobiology. *Statistical Modelling***,** 1471082X20963254.

Barroso-Bergadà, D., Pauvert, C., Vallance, J., Delière, L., Bohan, D. A., Buée, M. & Vacher, C. (2021) Microbial networks inferred from environmental DNA data for biomonitoring ecosystem change: Strengths and pitfalls. *Molecular Ecology Resources,* **21,** 762-780.

Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., Morozumi, C. & Brosi, B. J. (2017) Applying Pollen DNA Metabarcoding to the Study of Plant–Pollinator Interactions. *Applications in Plant Sciences***,** 1600124.

Beng, K. C., Tomlinson, K. W., Shen, X. H., Surget-Groba, Y., Hughes, A. C., Corlett, R. T. & Slik, J. W. F. (2016) The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. *Scientific Reports,* **6,** 24965.

Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological),* **36,** 192-225.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R. & Willerslev, E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLOS ONE,* **2,** e197.

Blanchard, J. L. (2015) A rewired food web. *Nature,* **527,** 173-174.

Blüthgen, N. (2010) Why network analysis is often disconnected from community ecology: a critique and an ecologist's guide. *Basic and Applied Ecology,* **11,** 185-195.

Blüthgen, N., Menzel, F. & Blüthgen, N. (2006) Measuring specialization in species interaction networks. *BMC ecology,* **6,** 9.

Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A. & Tamaddoni-Nezhad, A. (2011) Automated discovery of food webs from ecological data using logic-based machine learning. *PLoS ONE,* **6,** e29028.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J. & Woodward, G. (2017) Next-generation global biomonitoring: Large-scale, automated reconstruction of ecological networks. *Trends in Ecology & Evolution,* **32,** 477-487.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology,* **37,** 852-857.

Bosch, J., Martín González, A. M., Rodrigo, A. & Navarro, D. (2009) Plant–pollinator networks: adding the pollinator's perspective. *Ecology Letters,* **12,** 409-419.

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Strauss, J. F., Jefferson, K. K., Buck, G. A. & Vaginal Microbiome, C. (2015) The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology,* **15,** 66.

Bush, A., Compson, Z. G., Monk, W. A., Porter, T. M., Steeves, R., Emilson, E., Gagne, N., Hajibabaei, M., Roy, M. & Baird, D. J. (2019) Studying ecosystems with DNA metabarcoding: Lessons from biomonitoring of aquatic macroinvertebrates. *Frontiers in Ecology and Evolution,* **7**.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. & Holmes, S. P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods,* **13,** 581.

Carraro, L., Stauffer, J. B. & Altermatt, F. (2021) How to design optimal eDNA sampling strategies for biomonitoring in river networks. *Environmental DNA,* **3,** 157-172.

Chacoff, N. P., Vázquez, D. P., Lomáscolo, S. B., Stevani, E. L., Dorado, J. & Padrón, B. (2012) Evaluating sampling completeness in a desert plant–pollinator network. *Journal of Animal Ecology,* **81,** 190-200.

Chakraborty, C., Doss, C. G. P., Patra, B. C. & Bandyopadhyay, S. (2014) DNA barcoding to map the microbial communities: current advances and future directions. *Applied Microbiology and Biotechnology,* **98,** 3425-3436.

Chelius, M. K. & Triplett, E. W. (2001) The diversity of Archaea and Bacteria in association with the roots of *Zea mays* L. *Microbial Ecology,* **41,** 252-263.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. & Leon, C. (2010) Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *PLOS ONE,* **5,** e8613.

Chiquet, J., Mariadasou, M. & Robin, S. (2019) Variational inference of sparse network from count data. *36th International Conference on Machine Learning*, pp. 1988-1997.

Chiquet, J., Mariadassou, M. & Robin, S. (2021) The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances. *Frontiers in Ecology and Evolution,* **9**.

Chung, F. & Lu, L. (2002) The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences,* **99,** 15879-15882.

Coissac, E., Riaz, T. & Puillandre, N. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology,* **21,** 1834-1847.

Cougoul, A., Bailly, X. & Wit, E. C. (2019) MAGMA: inference of sparse microbial association networks. *BioRxiv***,** 538579.

Cristescu, M. E. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution,* **29,** 566-571.

Daudin, J. J., Picard, F. & Robin, S. (2008) A mixture model for random graphs. *Statistics and Computing,* **18,** 173-183.

David, P., Thébault, E., Anneville, O., Duyck, P. F., Chapuis, E. & Loeuille, N. (2017) Impacts of invasive species on food webs: a review of empirical data. *Advances in Ecological Research vol. 56 - Networks of Invasion: A Synthesis of Concepts* (eds D. A. Bohan, A. J. Dumbrell & F. Massol), pp. 1-60. Academic Press.

De Cock, M., Virgilio, M., Vandamme, P., Augustinos, A., Bourtzis, K., Willems, A. & De Meyer, M. (2019) Impact of sample preservation and manipulation on insect gut microbiome profiling. A test case with fruit flies (Diptera, Tephritidae). *Frontiers in Microbiology,* **10**.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics,* **34,** 2666-2669.

de Manincor, N., Hautekèete, N., Mazoyer, C., Moreau, P., Piquot, Y., Schatz, B., Schmitt, E., Zélazny, M. & Massol, F. (2020) How biased is our perception of plant-pollinator networks? A comparison of visit- and pollen-based representations of the same networks. *Acta Oecologica,* **105,** 103551.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. *Biology Letters,* **10,** 20140562.

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., Kartzinel, T. R. & Eveson, J. P. (2019) Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology,* **28,** 391-406.

Dee, L. E., Allesina, S., Bonn, A., Eklöf, A., Gaines, S. D., Hines, J., Jacob, U., McDonald-Madden, E., Possingham, H., Schröter, M. & Thompson, R. M. (2017) Operationalizing Network Theory for Ecosystem Service Assessments. *Trends in Ecology & Evolution,* **32,** 118-130.

Deiner, K., Walser, J.-C., Mächler, E. & Altermatt, F. (2015) Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation,* **183,** 53-63.

Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães Jr., P. R., Hembry, D. H., Newman, E. A., Olesen, J. M., Pires, M. M., Yeakel, J. D. & Poisot, T. (2019) Analysing ecological networks of species interactions. *Biological Reviews,* **94,** 16-36.

Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., Plantegenest, M., Vacher, C. & Evans, D. M. (2018) Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis. *Advances in Ecological Research*, pp. 1-62. Academic Press.

Derocles, S. A. P., Evans, D. M., Nichols, P. C., Evans, S. A. & Lunt, D. H. (2015) Determining plant-leaf miner-parasitoid Interactions: a DNA barcoding approach. *PLOS ONE,* **10,** e0117872.

Derocles, S. A. P., Le Ralec, A., Besson, M. M., Maret, M., Walton, A., Evans, D. M. & Plantegenest, M. (2014) Molecular analysis reveals high compartmentalization in aphid–primary parasitoid networks and low parasitoid sharing between crop and noncrop habitats. *Molecular Ecology,* **23,** 3900-3911.

Derocles, S. A. P., Plantegenest, M., Simon, J.-C., Taberlet, P. & Le Ralec, A. (2012) A universal method for the detection and identification of Aphidiinae parasitoids within their aphid hosts. *Molecular Ecology Resources,* **12,** 634-645.

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., Holdaway, R. J., Lear, G., Makiola, A., Morales, S. E., Powell, J. R. & Weaver, L. (2018) Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources,* **18,** 940-952.

Drew, K., Müller, C. L., Bonneau, R. & Marcotte, E. M. (2017) Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. *PLoS Computational Biology,* **13,** e1005625.

Duffy, J. E., Cardinale, B. J., France, K. E., McIntyre, P. B., Thebault, E. & Loreau, M. (2007) The functional role of biodiversity in ecosystems: incorporating trophic complexity. *Ecology Letters,* **10,** 522-538.

Dunne, J. A., Williams, R. J. & Martinez, N. D. (2002) Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters,* **5,** 558-567.

Elbrecht, V. & Leese, F. (2017) PrimerMiner: an R package for development and in silico validation of DNA metabarcoding primers. *Methods in Ecology and Evolution,* **8,** 622-626.

Evans, D. M., Kitson, J. J. N., Lunt, D. H., Straw, N. A. & Pocock, M. J. O. (2016) Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Functional Ecology,* **30,** 1904-1916.

Evans, D. M., Pocock, M. J. O. & Memmott, J. (2013) The robustness of a network of ecological networks to habitat loss. *Ecology Letters,* **16,** 844-852.

Fang, H., Huang, C., Zhao, H. & Deng, M. (2017) gCoda: Conditional Dependence Network Inference for Compositional Data. *Journal of Computational Biology,* **24,** 699-708.

Faust, K. & Raes, J. (2012) Microbial interactions: from networks to models. *Nat Rev Micro,* **10,** 538-550.

Fayle, T. M., Scholtz, O., Dumbrell, A. J., Russell, S., Segar, S. T. & Eggleton, P. (2015) Detection of mitochondrial COII DNA sequences in ant guts as a method for assessing termite predation by ants. *PLOS ONE,* **10,** e0122533.

Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z. & Aerts, S. (2018) Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics,* **17,** 246-254.

Fontaine, C., Guimarães, P. R., Kéfi, S., Loeuille, N., Memmott, J., Van Der Putten, W. H., Van Veen, F. J. F. & Thébault, E. (2011) The ecological and evolutionary implications of merging different types of networks. *Ecology Letters,* **14,** 1170-1181.

Fortin, M.-J., Dale, M. R. T. & Brimacombe, C. (2021) Network ecology in dynamic landscapes. *Proceedings of the Royal Society B: Biological Sciences,* **288,** 20201889.

Fortuna, M. A., Stouffer, D. B., Olesen, J. M., Jordano, P., Mouillot, D., Krasnov, B. R., Poulin, R. & Bascompte, J. (2010) Nestedness versus modularity in ecological networks: two sides of the same coin? *Journal of Animal Ecology,* **79,** 811-817.

Friedman, J. & Alm, E. J. (2012) Inferring Correlation Networks from Genomic Survey Data. *PloS Computational Biology,* **8,** e1002687.

Friedman, J., Hastie, T. & Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics,* **9,** 432-441.

Fuhrman, J. A. (2009) Microbial community structure and its functional implications. *Nature,* **459,** 193-199.

Gardes, M. & Bruns, T. D. (1993) ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Molecular Ecology,* **2,** 113-118.

Geslin, B., Gauzens, B., Baude, M., Dajoz, I., Fontaine, C., Henry, M., Ropars, L., Rollin, O., Thébault, E. & Vereecken, N. J. (2017) Massively introduced managed species and their consequences for plant–pollinator interactions. *Advances in Ecological Research,* **57,** 147-199.

Gibson, R. H., Knott, B., Eberlein, T. & Memmott, J. (2011) Sampling method influences the structure of plant–pollinator networks. *Oikos,* **120,** 822-831.

Hammer, T. J., Dickerson, J. C. & Fierer, N. (2015) Evidence-based recommendations on storing and handling specimens for analyses of insect microbiota. *PeerJ,* **3,** e1190.

Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002) Latent space approaches to social network analysis. *Journal of the American Statistical Association,* **97,** 1090-1098.

Hrček, J. & Godfray, H. C. J. (2015) What do molecular methods bring to host–parasitoid food webs? *Trends in Parasitology,* **31,** 30-35.

Ings, T. C., Montoya, J. M., Bascompte, J., Blüthgen, N., Brown, L., Dormann, C. F., Edwards, F., Figueroa, D., Jacob, U. & Jones, J. I. (2009) Ecological networks - beyond food webs. *Journal of Animal Ecology,* **78,** 253-269.

Inouye, D. I., Yang, E., Allen, G. I. & Ravikumar, P. (2017) A review of multivariate distributions for count data derived from the Poisson distribution. *WIREs Computational Statistics,* **9,** e1398.

Jakuschkin, B., Fievet, V., Schwaller, L., Fort, T., Robin, C. & Vacher, C. (2016) Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen *Erysiphe alphitoides*. *Microbial Ecology,* **72,** 870-880.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B. C. & Yu, D. W. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters,* **16,** 1245-1257.

Joffard, N., Massol, F., Grenié, M., Montgelard, C. & Schatz, B. (2019) Effect of pollination strategy, phylogeny and distribution on pollination niches of Euro-Mediterranean orchids. *Journal of Ecology,* **107,** 478-490.

Jordano, P. (2016) Sampling networks of ecological interactions. *Functional Ecology,* **30,** 1883-1893.

Kaiser-Bunbury, C. N., Mougal, J., Whittington, A. E., Valentin, T., Gabriel, R., Olesen, J. M. & Blüthgen, N. (2017) Ecosystem restoration strengthens pollination network resilience and function. *Nature,* **542,** 223.

Kamenova, S., Bartley, T., Bohan, D., Boutain, J. R., Colautti, R. I., Domaizon, I., Fontaine, C., Lemainque, A., Le Viol, I., Mollot, G., Perga, M. E., Ravigné, V. & Massol, F. (2017) Invasions toolkit: current methods for tracking the spread and impact of invasive species. *Advances in Ecological Research,* **56,** 85-182.

Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L. & McMahon, K. D. (2013) A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *ISME J,* **7,** 680-684.

Kéfi, S., Berlow, E. L., Wieters, E. A., Navarrete, S. A., Petchey, O. L., Wood, S. A., Boit, A., Joppa, L. N., Lafferty, K. D., Williams, R. J., Martinez, N. D., Menge, B. A., Blanchette, C. A., Iles, A. C. &

Brose, U. (2012) More than a meal ... integrating non-feeding interactions into food webs. *Ecology Letters,* **15,** 291-300.

Kitson, J. J. N., Hahn, C., Sands, R. J., Straw, N. A., Evans, D. M. & Lunt, D. H. (2019) Detecting host–parasitoid interactions in an invasive Lepidopteran using nested tagging DNA metabarcoding. *Molecular Ecology,* **28,** 471-483.

Kitson, J. J. N., Warren, B. H., Vincent Florens, F. B., Baider, C., Strasberg, D. & Emerson, B. C. (2013) Molecular characterization of trophic ecology within an island radiation of insect herbivores (Curculionidae: Entiminae: Cratopus). *Molecular Ecology,* **22,** 5441-5455.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., Zaneveld, J. R., Zhu, Q., Caporaso, J. G. & Dorrestein, P. C. (2018) Best practices for analysing microbiomes. *Nature Reviews Microbiology,* **16,** 410-422.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. & Bonneau, R. A. (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol,* **11,** e1004226.

Labeyrie, V., Thomas, M., Muthamia, Z. K. & Leclerc, C. (2016) Seed exchange networks, ethnicity, and sorghum diversity. *Proceedings of the National Academy of Sciences,* **113,** 98-103.

Lafferty, K. D., Allesina, S., Arim, M., Briggs, C. J., De Leo, G., Dobson, A. P., Dunne, J. A., Johnson, P. T. J., Kuris, A. M., Marcogliese, D. J., Martinez, N. D., Memmott, J., Marquet, P. A., McLaughlin, J. P., Mordecai, E. A., Pascual, M., Poulin, R. & Thieltges, D. W. (2008) Parasites in food webs: the ultimate missing links. *Ecology Letters,* **11,** 533-546.

Laforest-Lapointe, I., Paquette, A., Messier, C. & Kembel, S. W. (2017) Leaf bacterial diversity mediates plant diversity and ecosystem function relationships. *Nature,* **546,** 145-147.

Le Minh, T. (2021) Weak convergence of U-statistics on a row-column exchangeable matrix. *ArXiv e-prints*.

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., Cruickshank, R., Dopheide, A., Handley, K. M., Hermans, S., Kamke, J., Lee, C. K., MacDiarmid, R., Morales, S. E., Orlovich, D. A., Smissen, R., Wood, J. & Holdaway, R. (2018) Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology,* **42,** 10-50A.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T. & Machida, R. J. (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology,* **10,** 34.

Lewinsohn, T. M., Prado, P. I., Jordano, P., Bascompte, J. & Olesen, J. M. (2006) Structure in plant-animal interaction assemblages. *Oikos,* **113,** 174-184.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A. M., Coppola, L., Cornejo-Castillo, F. M., d'Ovidio, F., De Meester, L., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S. G., Sunagawa, S., Bork, P., Sullivan, M. B., Karsenti, E., Bowler, C., de Vargas, C. & Raes, J. (2015) Determinants of community structure in the global plankton interactome. *Science,* **348,** 1262073.

Louca, S., Jacques, S. M. S., Pires, A. P. F., Leal, J. S., Srivastava, D. S., Parfrey, L. W., Farjalla, V. F. & Doebeli, M. (2016) High taxonomic variability despite stable functional structure across microbial communities. *Nature Ecology & Evolution,* **1,** 0015.

Lucas, A., Bodger, O., Brosi Berry, J., Ford Col, R., Forman Dan, W., Greig, C., Hegarty, M., Neyland Penelope, J. & Vere, N. (2018) Generalisation and specialisation in hoverfly (Syrphidae)

grassland pollen transport networks revealed by DNA metabarcoding. *Journal of Animal Ecology,* **0**.

Macfadyen, S., Gibson, R., Polaszek, A., Morris, R. J., Craze, P. G., Planqué, R., Symondson, W. O. C. & Memmott, J. (2009) Do differences in food web structure between organic and conventional farms affect the ecosystem service of pest control? *Ecology Letters,* **12,** 229-238.

Macgregor, C. J., Kitson, J. J. N., Fox, R., Hahn, C., Lunt, D. H., Pocock, M. J. O. & Evans, D. M. (2019) Construction, validation, and application of nocturnal pollen transport networks in an agro-ecosystem: a comparison using light microscopy and DNA metabarcoding. *Ecological Entomology,* **44,** 17-29.

Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., Brennan, G., Bush, A., Canard, E., Cordier, T., Creer, S., Curry, R. A., David, P., Dumbrell, A. J., Gravel, D., Hajibabaei, M., Hayden, B., van der Hoorn, B., Jarne, P., Jones, J. I., Karimi, B., Keck, F., Kelly, M., Knot, I. E., Krol, L., Massol, F., Monk, W. A., Murphy, J., Pawlowski, J., Poisot, T., Porter, T. M., Randall, K. C., Ransome, E., Ravigné, V., Raybould, A., Robin, S., Schrama, M., Schatz, B., Tamaddoni-Nezhad, A., Trimbos, K. B., Vacher, C., Vasselon, V., Wood, S., Woodward, G. & Bohan, D. A. (2020) Key Questions for Next-Generation Biomonitoring. *Frontiers in Environmental Science,* **7**.

Makiola, A., Dickie, I. A., Holdaway, R. J., Wood, J. R., Orwin, K. H., Lee, C. K. & Glare, T. R. (2019) Biases in the metabarcoding of plant pathogens using rust fungi as a model system. *MicrobiologyOpen,* **8,** e00780.

Mariadassou, M., Robin, S. & Vacher, C. (2010) Uncovering latent structure in valued graphs: A variational approach. *The annals of applied statistics,* **4,** 715-742, 28.

Matias, C. & Miele, V. (2017) Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **79,** 1119-1141.

McMurdie, P. J. & Holmes, S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE,* **8,** e61217.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S.-I., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J. & Göttgens, B. (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology,* **33,** 269-276.

Montoya, D., Rogers, L. & Memmott, J. (2012) Emerging perspectives in the restoration of biodiversity-based ecosystem services. *Trends in Ecology & Evolution,* **27,** 666-672.

Muggleton, S. (1991) Inductive logic programming. *New generation computing,* **8,** 295-318.

Muggleton, S. (1995) Inverse entailment and Progol. *New generation computing,* **13,** 245-286.

Muggleton, S. H. & Bryant, C. H. (2000) Theory completion using inverse entailment. *Inductive Logic Programming* (eds J. Cussens & A. Frisch), pp. 130-146. Springer, Berlin, Heidelberg.

Muggleton, S. H., Lin, D. & Tamaddoni-Nezhad, A. (2015) Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. *Machine Learning,* **100,** 49-73.

Navarrete, A. A., Tsai, S. M., Mendes, L. W., Faust, K., de Hollander, M., Cassman, N. A., Raes, J., van Veen, J. A. & Kuramae, E. E. (2015) Soil microbiome responses to the short-term effects of Amazonian deforestation. *Molecular Ecology,* **24,** 2433-2448.

Neutel, A. M., Heesterbeek, J. A. P. & de Ruiter, P. C. (2002) Stability in real food webs: weak links in long loops. *Science,* **296,** 1120-1123.

Nguyen, N. H., Song, Z., Bates, S. T., Branco, S., Tedersoo, L., Menke, J., Schilling, J. S. & Kennedy, P. G. (2016) FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology,* **20,** 241-248.

Novak, M., Wootton, J. T., Doak, D. F., Emmerson, M., Estes, J. A. & Tinker, M. T. (2011) Predicting community responses to perturbations in the face of imperfect knowledge and network complexity. *Ecology,* **92,** 836-846.

Ohlmann, M., Miele, V., Dray, S., Chalmandrier, L., O'Connor, L. & Thuiller, W. (2019) Diversity indices for ecological networks: a unifying framework using Hill numbers. *Ecology Letters,* **22,** 737-747.

Olsson, O., Karlsson, M., Persson, A. S., Smith, H. G., Varadarajan, V., Yourstone, J. & Stjernman, M. (2021) Efficient, automated and robust pollen analysis using deep learning. *Methods in Ecology and Evolution,* **n/a**.

Op De Beeck, M., Lievens, B., Busschaert, P., Declerck, S., Vangronsveld, J. & Colpaert, J. V. (2014) Comparison and Validation of Some ITS Primer Pairs Useful for Fungal Metabarcoding Studies. *PLOS ONE,* **9,** e97629.

Ouadah, S., Latouche, P. & Robin, S. (2021) Motif-based tests for bipartite networks. *arXiv preprint arXiv:2101.11381*.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters,* **20,** 561-576.

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., Lesur, I., Vallance, J. & Vacher, C. (2019) Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology,* **41,** 23-33.

Pellissier, L., Albouy, C., Bascompte, J., Farwig, N., Graham, C., Loreau, M., Maglianesi, M. A., Melián, C. J., Pitteloud, C., Roslin, T., Rohr, R., Saavedra, S., Thuiller, W., Woodward, G., Zimmermann, N. E. & Gravel, D. (2018) Comparing species interaction networks along environmental gradients. *Biological Reviews,* **93,** 785-800.

Piñol, J., San Andrés, V., Clare, E. L., Mir, G. & Symondson, W. O. C. (2014) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources,* **14,** 18-26.

Piñol, J., Senar, M. A. & Symondson, W. O. C. (2019) The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology,* **28,** 407-419.

Pocock, M. J. O., Evans, D. M. & Memmott, J. (2012) The robustness and restoration of a network of ecological networks. *Science,* **335,** 973-977.

Poisot, T. & Gravel, D. (2014) When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ,* **2,** e251.

Poisot, T., Stouffer, D. B. & Kéfi, S. (2016) Describe, understand and predict: why do we need networks in ecology? *Functional Ecology,* **30,** 1878-1882.

Pornon, A., Andalo, C., Burrus, M. & Escaravage, N. (2017) DNA metabarcoding data unveils invisible pollination networks. *Scientific Reports,* **7,** 16828.

Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., Pellizzari, C., Iribar, A., Etienne, R., Taberlet, P., Vidal, M., Winterton, P., Zinger, L. & Andalo, C. (2016) Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports,* **6,** 27282.

Porter, T. M. & Hajibabaei, M. (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology,* **27,** 313-338.

Prendergast, K. S., Menz, M. H. M., Dixon, K. W. & Bateman, P. W. (2020) The relative performance of sampling methods for native bees: an empirical test and review of the literature. *Ecosphere,* **11,** e03076.

Pržulj, N., Corneil, D. G. & Jurisica, I. (2006) Efficient estimation of graphlet frequency distributions in protein–protein interaction networks. *Bioinformatics,* **22,** 974-980.

Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology,* **21,** 1864-1877.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research,* **41,** D590-D596.

Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. (2010) High-dimensional Ising model selection using ℓ₁-regularized logistic regression. *The Annals of Statistics,* **38,** 1287-1319, 33.

Redford, A. J., Bowers, R. M., Knight, R., Linhart, Y. & Fierer, N. (2010) The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environmental Microbiology,* **12,** 2885-2893.

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P. & Coissac, E. (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research,* **39,** e145-e145.

Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* **4,** e2584.

Rombaut, A., Guilhot, R., Xuéreb, A., Benoit, L., Chapuis, M. P., Gibert, P. & Fellous, S. (2017) Invasive *Drosophila suzukii* facilitates *Drosophila melanogaster* infestation and sour rot outbreaks in the vineyards. *Royal Society Open Science,* **4,** 170117.

Röttjers, L. & Faust, K. (2018) From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews,* **42,** 761-780.

Sasaki, S., Hatano, R., Ohwada, H. & Nishiyama, H. (2019) Estimating productivity of dairy cows by inductive logic programming. *The 29th ILP conference*. Plovdiv, Bulgaria.

Säterberg, T., Sellman, S. & Ebenman, B. (2013) High frequency of functional extinctions in ecological networks. *Nature,* **499,** 468-470.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W. & Consortium, F. B. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences,* **109,** 6241-6246.

Schwager, E., Mallick, H., Ventz, S. & Huttenhower, C. (2017) A Bayesian method for detecting pairwise associations in compositional data. *PLoS Computational Biology,* **13,** e1005852.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research,* **13,** 2498-2504.

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., Golding, G. B. & Hajibabaei, M. (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports,* **5,** 9687.

Song, C. & Saavedra, S. (2020) Telling ecological networks apart by their structure: An environment-dependent approach. *PLoS Computational Biology,* **16,** e1007787.

Song, S. J., Amir, A., Metcalf, J. L., Amato, K. R., Xu, Z. Z., Humphrey, G., Knight, R. & Dearing, M. D. (2016) Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. *mSystems,* **1,** e00021-16.

Taberlet, P., Bonin, A., Zinger, L. & Coissac, E. (2018) *Environmental DNA: For biodiversity research and monitoring,* Oxford University Press.

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H. & Kawabata, Z. i. (2012) Estimation of fish biomass using environmental DNA. *PLOS ONE,* **7,** e35868.

Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A. & Muggleton, S. (2006) Application of abductive ILP to learning metabolic network inhibition from temporal data. *Machine Learning,* **64,** 209-230.

Tamaddoni-Nezhad, A., Milani, G. A., Raybould, A., Muggleton, S. & Bohan, D. A. (2013) Construction and validation of food webs using logic-based machine learning and text mining. *Advances in Ecological Research* (eds G. Woodward & D. A. Bohan), pp. 225-289. Academic Press.

Thébault, E. & Loreau, M. (2006) The relationship between biodiversity and ecosystem functioning in food webs. *Ecological Research,* **21,** 17-25.

Thijs, S., Op De Beeck, M., Beckers, B., Truyens, S., Stevens, V., Van Hamme, J. D., Weyens, N. & Vangronsveld, J. (2017) Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. *Frontiers in Microbiology,* **8**.

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H. & Trites, A. W. (2016) Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources,* **16,** 714-726.

Thompson, M. S. A., Bankier, C., Bell, T., Dumbrell, A. J., Gray, C., Ledger, M. E., Lehmann, K., McKew, B. A., Sayer, C. D., Shelley, F., Trimmer, M., Warren, S. L. & Woodward, G. (2016) Gene-to-ecosystem impacts of a catastrophic pesticide spill: testing a multilevel bioassessment approach in a river ecosystem. *Freshwater Biology,* **61,** 2037-2050.

Thompson, R. M., Brose, U., Dunne, J. A., Hall Jr, R. O., Hladyz, S., Kitching, R. L., Martinez, N. D., Rantala, H., Romanuk, T. N., Stouffer, D. B. & Tylianakis, J. M. (2012) Food webs: reconciling the structure and function of biodiversity. *Trends in Ecology & Evolution,* **27,** 689-697.

Toju, H., Guimarães, P. R., Olesen, J. M. & Thompson, J. N. (2014) Assembly of complex plant–fungus networks. *Nature Communications,* **5,** 5273.

Toju, H., Sato, H., Yamamoto, S., Kadowaki, K., Tanabe, A. S., Yazawa, S., Nishimura, O. & Agata, K. (2013) How are plant and fungal communities linked to each other in belowground ecosystems? A massively parallel pyrosequencing analysis of the association specificity of root-associated fungi and their host plants. *Ecology and Evolution,* **3,** 3112-3124.

Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., Schwaller, L., Chiquet, J., Smith, M. A., Vallance, J., Fievet, V., Jakuschkin, B. & Bohan, D. A. (2016) Learning ecological networks from next-generation sequencing data. *Advances in Ecological Research* (eds G. Woodward & D. A. Bohan), pp. 1-39. Academic Press.

Vamos, E. E., Elbrecht, V. & Leese, F. (2017) Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics,* **1**.

Wallinger, C., Sint, D., Baier, F., Schmid, C., Mayer, R. & Traugott, M. (2015) Detection of seed DNA in regurgitates of granivorous carabid beetles. *Bulletin of Entomological Research,* **105,** 728-735.

Walton, R. E., Sayer, C. D., Bennion, H. & Axmacher, J. C. (2020) Nocturnal pollinators strongly contribute to pollen transport of wild flowers in an agricultural landscape. *Biology Letters,* **16,** 20190877.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C. & Hui, F. K. C. (2015) So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution,* **30,** 766-779.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J. & Knight, R. (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal,* **10,** 1669-1681.

Westphal, C., Bommarco, R., Carré, G., Lamborn, E., Morison, N., Petanidou, T., Potts, S. G., Roberts, S. P., Szentgyörgyi, H. & Tscheulin, T. (2008) Measuring bee diversity in different European habitats and biogeographical regions. *Ecological Monographs,* **78,** 653-671.

White, T. J., Bruns, T., Lee, S. & Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications* (eds M. A. Innis, D. H. Gelfand, J. J. Snisky & T. J. White), pp. 315-322. Academic Press, San Diego.

Wirta, H. K., Hebert, P. D. N., Kaartinen, R., Prosser, S. W., Várkonyi, G. & Roslin, T. (2014) Complementary molecular information changes our perception of food web structure. *Proceedings of the National Academy of Sciences,* **111,** 1885-1890.

Yang, E., Allen, G., Liu, Z. & Ravikumar, P. (2012) Graphical Models via Generalized Linear Models. *Advances in Neural Information Processing Systems,* **25,** 1358-1366.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W. & Glöckner, F. O. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research,* **42,** D643-D648.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution,* **3,** 613-623.

Yu, X., Zeng, T., Wang, X., Li, G. & Chen, L. (2015) Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. *Journal of Translational Medicine,* **13,** 189.

Yuan, M. & Lin, Y. (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika,* **94,** 19-35.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., Kauserud, H., Orlando, L., Pansu, J., Pawlowski, J., Tedersoo, L., Thomsen, P. F., Willerslev, E. & Taberlet, P. (2019) DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology,* **28,** 1857-1862.