Automatic learning of interaction networks from next-generation sequence data

Didac Barroso-Bergada

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A l'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) de Dijon

Ecole doctorale n°554

« Environnements – Santé » (E-S)

Doctorat de biologie des populations et écologie

Par

Didac Barroso Bergadà

Automatic learning of interaction networks from next-generation sequence data

Thèse présentée et soutenue à Dijon, le 8 décembre 2022

Composition du Jury :

Alex Dumbrell	Professeur à l'Université d'Essex	Président et Rapporteur
Christophe Mougel	Directeur de recherche à l'INRAE de Rennes	Rapporteur
Lucie Zinger	Professeure associée à l'École Normale Supérieure, Université Paris Sciences & Lettres	Examinatrice
Tristan Cordier	Chercheur principal, Molekylær økologi og paleogenomikk - MEP	Examinateur
David A. Bohan	Directeur de recherche à l'INRAE de Dijon	Directeur de thèse

Titre : Apprentissage automatique des réseaux d'interaction à partir des données de séquence next-génération

Mots clés : Apprentissage automatique explicable, Séquençage de nouvelle génération, Réseaux d'Interaction

Résumé : Le changement climatique et d'autres processus induits par l'homme modifient les écosystèmes, à l'échelle mondiale, à un rythme toujours plus rapide. Les communautés microbiennes jouent un rôle important dans le fonctionnement des écosystèmes, en maintenant leur diversité et leurs services. Ces communautés sont façonnées par les différents effets environnementaux abiotiques auxquels elles sont soumises et par les interactions biotiques entre tous les membres de la communauté. Le projet ANR Next-Generation Biomonitoring (NGB) a proposé de reconstruire des réseaux d'interactions à partir de mesures d'abondance obtenues par séquençage de l'ADN environnemental (eDNA) et d'utiliser ces réseaux pour suivre l'évolution des écosystèmes. Dans cette thèse, menée dans le cadre du projet NGB, j'évalue le potentiel de deux outils de reconstruction de réseaux statistiques existants, SparCC et SPIEC-EASI, pour reconstruire des réseaux microbiens afin d'évaluer le changement des écosystèmes. Les communautés microbiennes des feuilles de vigne ont été utilisées pour différencier deux pratiques agricoles différentes, en identifiant les métriques de réseau appropriées pour capturer le changement d'écosystème. Les expériences ont montré que, bien qu'il soit difficile d'obtenir des réseaux répliqués, même dans les mêmes conditions environnementales, il est toujours possible de différencier les réseaux de différentes pratiques agricoles en utilisant certaines métriques de réseau. Bien que les outils de reconstruction de réseaux basés sur des statistiques puissent obtenir des réseaux d'associations entre micro-organismes, avec précision, ces associations statistiques ne sont pas des indicateurs directs des processus écologiques d'interaction sous-jacents.

Pour résoudre ce problème, j'ai développé un nouvel outil de reconstruction de réseau appelé Interaction Inference using Explainable Machine Learning (InfIntE), basé sur Explainable Machine Learning (EML). L'EML est une branche de l'apprentissage automatique qui utilise les connaissances préalables d'un domaine scientifique, tel que l'écologie, pour déclarer des énoncés logiques de concepts (hypothèses) afin de réaliser des inférences compréhensibles par l'homme. InfIntE utilise les règles d'interaction écologiques ainsi que les informations sur l'abondance obtenues par le séquençage de l'eDNA pour reconstruire des réseaux par inférence logique. Contrairement à la reconstruction de réseaux basée sur des méthodes statistiques, l'utilisation de règles d'interaction permet de classer directement les interactions déduites selon leur type (par exemple, mutualisme, compétition), ce qui permet d'obtenir des réseaux d'interaction plus informatifs et objectifs. La performance d'InfIntE a été évaluée en utilisant des données générées par ordinateur ainsi que des ensembles de données obtenus par échantillonnage d'eDNA du microbiome des feuilles de vigne. Mes résultats montrent qu'InfIntE peut détecter des réseaux d'interaction avec une précision similaire à celle des outils statistiques testés, SparCC et SPIEC-EASI, avec l'avantage significatif de la classification directe du type de chaque interaction.

Université Bourgogne Franche-Comté. 32, avenue de l'Observatoire 25000 Besançon

Title : Automatic learning of interaction networks from next-generation sequence data

Keywords : Explainable Machine Learning, Next-Generation Sequencing, Interaction Networks

Abstract : Climate change and other human-induced processes are modifying ecosystems, globally, at an ever increasing rate. Microbial communities play an important role in the functioning ecosystems, maintaining their diversity and services. These communities are shaped by the different abiotic environmental effects to which they are subjected and the biotic interactions between all community members. The ANR Next-Generation Biomonitoring (NGB) project proposed to reconstruct interaction networks from abundance measures obtained sequencing environmental DNA (eDNA) and to use these networks to monitor ecosystem change. In this thesis, conducted as part of the NGB project, I evaluate the potential of two existing statistical network reconstruction tools, SparCC and SPIEC-EASI, to reconstruct microbial networks in order to evaluate ecosystem change. Microbial communities from grapevine leaves were used to differentiate between two different agricultural practices, identifying the appropriate network metrics to capture ecosystem change. The experiments showed that although it is difficult to obtain replicate networks, even from the same environmental conditions, it is still possible to differentiate networks from different agricultural practices using some network metrics. Although statistically-based network reconstruction tools can obtain networks of associations between microorganisms, with accuracy, these statistical associations are not direct indicators of the underlying ecological processes of interaction. To address this issue, I developed a new network reconstruction tool called Interaction Inference using Explainable Machine Learning (InfIntE), based upon Explainable Machine Learning (EML). EML is a branch of machine learning which uses the prior knowledge from a scientific domain, such as Ecology, to declare logical statements of concept (hypotheses) to carry out human-understandable inference. InfIntE uses ecological rules of interaction together with the abundance information obtained from sequencing eDNA to reconstruct networks by logical inference. In contrast to statistically-based network reconstruction, the use of interaction rules allows direct classification the inferred interactions to their type (e.g. mutualism, competition), obtaining more informative and objective interaction networks. The performance of InfIntE was evaluated using computer-generated data as well as datasets obtained by eDNA sampling of grapevine leaf microbiome. My results show that InfIntE can detect interaction networks with similar accuracy to the tested statistically-based tools, SparCC and SPIEC-EASI, with the significant benefit of direct classification of the type of each interaction.

This Ph. D. thesis was funded by the ANR project (ANR-17-CE32-0011) and a demi bourse de thèse from Syngenta AG.

The work was suppported by the Consortium Biocontrôle via the BCMicrobiome project and the Biodiversa BiodivRestore FRESHH project (ANR-21-BIRE-0004).

Remerciements

I want to start by thanking my thesis supervisor, David Bohan for his help and support during the three years of this Ph.D. Dave, these have been three funny years. At different points we have been trapped in England (on the wrong side of la Manche), we had programs that would not compile, Biblical-level floods on the way to a restaurant, results that did not look promising - all in the middle of the global covid-19 pandemic. But, I have had the luck of having a supervisor who helped me in any case and was always there to offer counsel or just listen to me complaining. During the days of writing the thesis you have helped me a lot. You've always been positive, even when I was starting to write nonsense. I hope that when the thesis is over and becomes a memory, we can have a beer to celebrate (maybe at the Wooden Bridge).

I also want to thank Alireza Tamaddoni-Nezhad, Corinne Vacher and Nika Galic for their help and counsel during the thesis. Without you this thesis would have been much more difficult, and it would not be the thesis it is. Alireza, thank you for the interesting discussions we had about EML and your help in learning how to use it. I think that during this thesis we have taken a small step towards making EML useful for Ecologists and Microbiologists, متشكرم.. Corinne, thank you so much for your help to understand the microbial world. Your counsel and suggestions drove me to improve the work I was doing. I still have some revisions to do, so you will be hearing from me..., merci. Nika, thank you for your support and guidance during the thesis, and maybe also bribing the conference organisers for my Best Paper prize. And also for letting me win in the quiz, eventually. I owe you a box of chocolates, hvala vam.

Thanks to Elsa Canard and Manuel Blouin for their advice and guidance as part of my Comité de thèse. I am sorry for stealing a morning from you every year, but the truth is that the thesis committees were really interesting and helped me to shape this thesis, merci.

I haven't yet had the opportunity to meet Alex Dumbrell, Tristan Cordier, Christophe Mougel and Lucie Zinger, but I would like to thank them for accepting the task of evaluating my work. I very much look forward to seeing you at the thesis defense.

Thanks must go to all the members of the BCMicrobiome and NGB projects for their help during the different stages of the thesis. I am especially grateful to Valérie Laval, Frédéric Suffert, Stéphane Robin and Marine Cambon for the interesting discussions we had about my work, merci. It is a great pity that due the covid-19 pandemic we didn't have more of an opportunity to meet in person. Also, I would like to thank Charlie Pauvert for his work and help. Maybe you didn't realized but you have been a reference to follow during the thesis.

I am especially grateful to Dany Varghese. Dany, without you the last part of my thesis would have been close to impossible. Thank you for your work and help in all the parts involving the EML. And also, thank you for the curry (you said it was not spicy, but I have to respectfully *completely* disagree), നന്ദി.

Je tiens à remercier l'ensemble de l'équipe COMPARE pour m'avoir aidé à franchir les étapes de la thèse. Je pense que j'ai eu la chance de me retrouver dans l'équipe avec les personnes les plus gentilles que je pouvais trouver. Merci à Bruno de m'avoir aidée à résoudre les problèmes d'aller et venir et pointage, merci.

Venir dans un pays différent, avec une langue différente, n'est jamais facile. Mais j'ai eu la chance de trouver des compagnons qui m'ont fait me sentir chez moi, que ce soit en voyage, en mangeant ensemble, en allant regarder le football ou surtout en prenant une bière. Merci à Iheb, Eirini, Lucile, Benjamin, Guillaume, Emeric, Yaoyun, Sarah, Audrey, Caro, Thomas et Romain pour avoir fait de ces trois années bien plus qu'une simple thèse, شكرًا, Ευχαριστώ, 謝謝.

I per descomptat no seria aquí sense la meva família. Mama, papa, gràcies pel vostre suport, sempre, sense condicions. Si he arribat fins aquí és perquè sempre m'heu ajudat, fins i tot quan no em volia deixar ajudar. Sense el que he après de vosaltres aquesta tesi no hagués estat possible. Esther, Ramon, gràcies per tot, per ser-hi, per l'ajuda. Marta, Joan, tot i que a vegades em costi dir-ho, us he trobat a faltar. Però no dubteu que sempre he sabut que hi éreu i que podia comptar amb vosaltres (Algunes sent més receptives que d'altres). Avi, àvia, moltes gràcies per sempre ajudar-me, portar-me als llocs o preparar-me carmanyoles. Yaya, yayo, pienso a menudo en vosotros. Tietes, tiets, cosins i cosines, he tingut la sort increïble de tenir una família meravellosa i així tot sempre és més senzill.

Ariadna, han estat tres anys que hem hagut de viure separats. No ha estat fàcil per cap dels dos. Sé que hi ha hagut moments que necessitaves que fos més a prop i no hi era. Espero que a partir d'ara puguem continuar fent més passos cap a la nostra vida junts. Estar al teu costat em fa feliç. Gràcies per aquests anys i per sempre ajudar-me i fer-me somriure. T'estimo.

Liste des productions scientifiques

Articles

Published / In press / Accepted

- Barroso-Bergadà, D., Pauvert, C., Vallance, J., Delière, L., Bohan, D. A., Buée, M., and Vacher, C. (2021). Microbial networks inferred from environmental DNA data for biomonitoring ecosystem change: Strengths and pitfalls. *Molecular Ecology Resources, 21(3):762–780* **Chapter** II

- Dubart, M., Alonso, P., Barroso-Bergada, D., Becker, N., Bethune, K., Bohan, D. A., Boury, C., Cambon, M., Canard, E., Chancerel, E., Chiquet, J., David, P., de Manincor, N., Donnet, S., Duputié, A., Facon, B., Guichoux, E., Le Minh, T., Ortiz-Martínez, S., Piouceau, L., Sacco-Martret de Préville, A., Plantegenest, M., Poux, C., Ravigné, V., Robin, S., Trillat, M., Vacher, C., Vernière, C., and Massol, F. (2021). Coupling ecological network analysis with high-throughput sequencing-based surveys: Lessons from the next-generation biomonitoring project. *Advances in Ecological Research, 65, pages 367–430.*

- Barroso-Bergadà, D., Tamaddoni-Nezhad, A., Muggleton, S. H., Vacher, C., Galic, N., and Bohan, D. A. (2022). Machine Learning of Microbial Interactions Using Abductive ILP and Hypothesis Frequency/Compression Estimation, *Inductive Logic Programming, 13191, pages 26–40.* **Chapter** III

- Barroso-Bergadà, D., Massot, M., Vignolles, N., D'Arcier, J. F., Chancerel, E., Guichoux, E., Walker, A.-S., Vacher, C., Bohan, D. A., Laval, V., and Suffert, F. (Accepted). Metagenomic Next-Generation Sequencing (mNGS) Data Reveal the Phyllosphere Microbiome of Wheat Plants Infected by the Fungal Pathogen *Zymoseptoria tritici*. *Phytobiomes Journal* **Annex** B

- Varghese, D., Barroso-Bergada, D., Bohan, D. A., and Tamaddoni-Nezhad, A. (Accepted). Efficient Abductive Learning of Microbial Interactions using Meta Inverse Entailment. *Proceedings of the 31st International Conference on ILP. Prize of Best Application Paper.*

- Fournier, P., Pellan, L., Barroso-Bergadà, D., Bohan, D. A., Candressed, T., Delmotte, F., Dufour, M., Lauvergeat, V., Le Marrec, C., Marais, A., Martins, G., Masneuf-Pomarède, I., Rey, P., Sherman, D., This, P., Frioux, C., Labarthe, S., and Vacher, C. (Accepted) The functional microbiome of

grapevine throughout plant evolutionary history and lifetime *Advances in Ecological Research.*

Submitted

- Barroso-Bergadà, D., Delmotte, F., Faivre d'Arcier, J., Massot, M., Chancerel, E., Demeaux, I., Guimier, S., Guichoux, E., Bohan, D. A., and Vacher, C. Leaf microbiome data for European cultivated grapevine (*Vitis vinifera L.*) during downy mildew (*Plasmopara viticola*) epidemics in three wine-producing regions in France. (Submitted) *Phytobiomes Journal.* **Chapter** VII

In Prep.

- Barroso-Bergadà, D., Tamaddoni-Nezhad, A., Varghese, D., Vacher, C., Galic, N., and Bohan, D. A. Explainable inference of the diversity of microbial interactions. *Unravelling the Dark Web: explainable inference of the diversity of microbial interactions* (In Prep.) *The ISME Journal.* **Chapter** IV

Communications orales

- Barroso-Bergadà, D., Tamaddoni-Nezhad, A., Muggleton, S. H., Vacher, C., Galic, N., and Bohan, D. A. Machine Learning of Microbial Interactions Using Abductive ILP and Hypothesis Frequency/Compression Estimation *30th International Conference on Inductive Logic Programming, ILP2021 (2021)*

- Barroso-Bergadà, D., Tamaddoni-Nezhad, A., Varghese, D., Vacher, C., Galic, N., and Bohan, D. A. Biomonitoring and the inference of microbial interactions *Colloque INRAE Genomics 2022*

- Barroso-Bergadà, D., Tamaddoni-Nezhad, A., Varghese, D., Vacher, C., Galic, N., and Bohan, D. A. Inference of microbial interactions using explainable machine learning *Syngenta Biological Sciences and Product Safety collaborations event (2022). Prize of Best Talk, Product Safety.*

- Varghese, D., Barroso-Bergada, D., Bohan, D. A., and Tamaddoni-Nezhad, A. Efficient Abductive Learning of Microbial Interactions using Meta Inverse Entailment. *31st International Conference on Inductive Logic Programming, ILP2022 (2022). Prize of Best Application Paper.*

Contents

I Introduction **13**

1 Ecosystem monitoring 14

2 Microbial communities 15

2.1 The study of microorganisms 15

2.2 Composition of microbial communities 16

2.3 Microbial services 16

2.4 Microbial ecological interactions 17

3 Metabarcoding 20

3.1 Sampling and DNA extraction 22

3.2 DNA amplification and Sequencing 23

3.3 Sequence clustering 24

3.4 Taxonomic assignment 26

3.5 Metabarcoding data and Community analysis 27

4 Interaction network inference 27

4.1 Cross-sectional network inference 29

4.2 Longitudinal network inference 30

4.3 Considerations on correlation tools for biomonitoring 30

5 Explainable machine learning 32

5.1 Machine learning and the black box 32

5.2 Being explainable 33

6 What does explainable mean? 33

6.1 Abductive/Inductive Logic Programming 35

6.2 Inference Implementation 36

6.3 Interaction network validation 36

7 Network Visualisation 37

8 A case of study: Grapevine foliar microbial communities 38

8.1 The foliar microbial communities 39

8.2 Biomonitoring grapevine leaves' microbial communities . . . 39

II Correlation-based approaches for biomonitoring using DNA **41**

1 Introduction 42

2 Materials and methods 44

2.1 Study site and sampling design 44

2.2 DNA extraction and sequencing 46

2.3 Bioinformatic analysis 47

2.4 Statistical analyses 48

3 Results 50

3.1 All community α-properties detected system change 51

3.2 All community β-properties detected system change 51
3.3 None of the network α-properties detected system change . . 53
3.4 Half of the network β-properties detect system change . . . 53
3.5 Network replicates within each system were highly variable but shared links 56
4 Discussion 58
5 Acknowledgements 64
6 Data Accessibility Statement 65
7 Author Contributions 65

III Using a Logic-based approach to infer interactions from simulated DNA data 66
1 Introduction 67
2 Background and related work 68
3 Methods 70
3.1 Logical description of microbial interactions 70
3.2 Bootstrapping 72
3.3 Simulated data-sets 73
3.4 Compositionality and bias 74
4 Experimental evaluation 75
4.1 Experiment 1 75
4.2 Experiment 2 77
5 Discussion and conclusion 78
6 Acknowledgements 79

IV Logic-based inference of ecological interactions from environmental DNA data 81
1 Introduction 82
1.1 The Dark Web of microbial communities 82
1.2 Inferring microbial correlation networks 83
2 Materials and Methods 86
2.1 Hypothesis framework for learning microbial ecological interactions using abductive logic 86
2.2 Experiment 1: Generating synthetic, ecological-like data for verification 89
2.3 Experiment 2: Inferring networks from real data 92
2.4 Statistical Analysis 93
3 Results 93
3.1 Experiment 1: Generating synthetic, ecological-like networks 93
3.2 Experiment 2: Inferring complex networks, the Dark Web, from real data 95
4 Discussion 97

V InfIntE: a generic, logic-based inference tool for learning networks in R 103
1 R package 104
2 Network visualization 113

VI Discussion **116**
1 General Discussion . . . 117
1.1 Can metabarcoding data be used for biomonitoring? . . . 118
1.2 Can explainable machine learning be used to infer interactions? 119
1.3 Can explainable machine learning be improved? . . . 120
1.4 It is possible to perform a direct classification of interaction types using explainable machine learning? . . . 121
1.5 Considerations on InfIntE . . . 122
1.6 Which network reconstruction tool should I choose for biomonitoring? . . . 122
2 Future perspectives . . . 125
2.1 InfIntE testing . . . 125
2.2 InfIntE improvement . . . 126
2.3 Interaction testing . . . 128
2.4 Final considerations . . . 129

VII Constructing ecological, microbial community data-sets from DNA data **131**
1 Introduction . . . 132
2 Methods . . . 133
2.1 Sampling . . . 133
2.2 DNA extraction . . . 133
2.3 Fungal ITS amplification . . . 134
2.4 Bacterial 16S amplification . . . 135
2.5 MiSeq sequencing . . . 135
2.6 Quantification of *P. viticola* . . . 136
2.7 Bioinformatics . . . 136
2.8 Descriptive statistics . . . 137
3 Results . . . 137
3.1 Abundance of *P. viticola* . . . 137
3.2 Fungal community . . . 139
3.3 Bacterial community . . . 139
4 Future directions . . . 141
5 Availability of Data and Materials . . . 141
6 Author contributions . . . 141
7 Acknowledgments . . . 143
8 Funding . . . 143

Bibliography **144**

A Supplementary Figures and Tables **175**

B Metagenomic next-generation sequencing (mNGS) data reveals the phyllosphere microbiome of wheat plants infected by the fungal pathogen *Zymoseptoria tritici* **194**
1 Introduction . . . 195
2 Methods . . . 196
2.1 Sampling . . . 196

2.2 DNA extraction . . . 197
2.3 Bacterial 16S amplification . . . 197
2.4 Fungal ITS amplification . . . 198
2.5 Sequencing . . . 199
2.6 Bioinformatic treatment . . . 199
2.7 Quantification of *Z. tritici* by qPCR . . . 199
2.8 Analysis . . . 200
3 Results . . . 200
3.1 Fungal communities . . . 201
3.2 Bacterial communities . . . 201
4 Conclusions . . . 202
5 Availability of Data and Materials . . . 203
6 Author contributions . . . 203
7 Acknowledgments . . . 203

C List of Figures and Tables 207

D Glossary 221

E Résumé Français 223

Chapter I

Introduction

1 Ecosystem monitoring

All happy families are all alike; each unhappy family is unhappy in its own way. This is the opening statement of Leo Tolstoy's novel, Anna Karenina. While he refers to the protagonist's family, this statement could be extended by analogy to many other subjects which operate through the interactions of their constituent members. This includes the tiny mechanism of a watch, where the correct counting of time comes only from the perfect coordination of its parts, with each faulty watch having its own particular problems. The Anna Karenina principle can also apply to our world as a whole, where the different ecosystems interact to support life. And our world is changing... *Fast.* Climate change has known effects worldwide and others that we expect but have yet to manifest (Araújo *et al.* 2005; Bohan *et al.* 2022). Despite the efforts of many different institutions and international agreements like Paris Accord of 2015 we are facing continued biodiversity and ecosystem loss (Bellard *et al.* 2012) that could lead to our planet not being capable of supporting life. Monitoring how human beings are changing the environment can help us to understand why changes have occurred and develop management strategies to mitigate or prevent further detrimental change from happening. Here, invoking the Anna Karenina principle demands an holistic view of monitoring (Moore 2001), with partial approaches no longer being an option. Each ecosystem needs to be monitored systematically, and our belief is that the monitoring needs to be done not by focusing on particular bio-indicators or on charismatic species, but rather as a complete, whole picture.

This thesis was developed in the context of the French Agence Nationale de la Recherche (ANR) Next-Generation Global Biomonitoring (NGB) project. NGB developed around the idea of obtaining, systematically, near complete information about the community structure of an ecosystem in order to monitor ecosystem change (Bohan *et al.* 2017). To date, biomonitoring surveys have largely been limited to classical physical or ecological sampling approaches, requiring considerable human effort and time costs, and being limited to life forms that could be directly observed. Next generation sequencing (NGS) of environmental DNA (eDNA) has opened up new opportunities to systematically collect information about most life-forms present at reduced levels of human labor (Cordier *et al.* 2021). The NGB project sampled eDNA from soil, water and air biomes. The source of the eDNA genetic material present in these environmental samples is either whole organisms (such as microbes) or cells excreted by whole organisms or from tissues (Thomsen & Willerslev 2015). Thus, the eDNA sequencing produced a list of sequences present that corresponded to the 'great majority' of the organisms - the whole community - present in an ecosystem. Where the sampling and sequencing process is carried out systematically, it is therefore possible to obtain a near complete overview of the taxa (species, genus, etc.) present in an ecosystem in order to evaluate change in the ecosystem (Makiola *et al.* 2020). Following the Anna Karenina analogy, we can now begin to get to know the members of the family and their interactions, thanks to eDNA.

Knowing all the taxa present in an ecosystem is a first step to global scale biomonitoring (Cordier *et al.* 2021). However, as previously noted, ecosystems, are also shaped by the interactions between their constituents (McCauley *et al.*

2012). We define an interaction as any action or influence of one taxa on another that changes the abundance of one or both taxa, thereby shaping the communities inhabiting an ecosystem (Collins 2022). Given that all the members of a community share the same space, interactions between the constituent taxa cannot be understood out of context. Different interactions have different consequences for a focal taxa and it is important to know all the interactions between community members to predict community-scale effects. Indeed it is the mixture of all these different interactions, acting between all taxa in the community simultaneously, which determines the species richness, diversity patterns, functions and dynamics of the microbial community (see Ohlmann *et al.* 2018). Interaction networks have been used to visualize the interactions in an ecosystem. These show all the members of the communities as nodes of the network, with the different interactions presented as edges joining the nodes. Networks have been used to understand plant-pollinator relations (Lopezaraiza-Mikel *et al.* 2007), fish communities (Ushio *et al.* 2018) and microorganisms (Nicolaisen *et al.* 2014). The NGB project proposed to automatically sample, sequence and infer interaction networks to monitor ecosystems, and thereby characterise system-level responses to environmental change (Blanchard 2015), with the goal of determining the key drivers of ecosystem change.

2 Microbial communities

2.1 The study of microorganisms

Microorganisms are ubiquitous, being present in all ecosystems and participating in a massive number of ecological processes (Handelsman 2007). By biomass, microbes are the most abundant taxa globally (Bar-On *et al.* 2018). Microorganisms have highly flexible (adaptable) lifestyles, allowing them to live in almost any environment, being found in extreme environments like the deep sea (Franco *et al.* 2020), high temperature zones (Merkel *et al.* 2019) and hot springs, as three examples. Microbes also inhabit other living beings, from plants to humans, parasitising them or creating mutually-dependent relationships (Bang *et al.* 2018). This makes them key constituents of any ecosystem monitoring approach. The common characteristic of these life forms, their microscopic size, limits their perception by humans and complicates approaches to their study. In 1664, Robert Hooke was the first person to observe the fruiting structures of moulds, using a primitive microscope (Hooke 1664). This was the first recorded observation of a microbe, and the genesis of the science of microbiology. In 1677, Antonie van Leeuwenhoek wrote the 'letter on the protozoa', where he gave the first detailed description of 'animalcules' - protists and bacteria - living in different environments (Lane 2015). This was possible thanks to the improved microscope he designed. However, for the following centuries, the diversity and functions microorgansims remained a mystery. It was not until the end of the nineteenth century, thanks to Louis Pasteur and his *Théorie des Germes* (Pasteur *et al.* 1878), that the first solid theories about microbial populations and functions were proposed, opening the door on modern microbiology. Joseph Lister proposed the sterilization of surgery equipment (Pitt & Aubin 2012) and Robert Koch described the causal agents of cholera and tuber-

culosis (Koch 1884). In the early twentieth century, Alexander Fleming described the first antibiotic compound (Fleming 1929), and thereby the first microbial interaction. Most of this early microbiological work was based on microscopy and culture studies, thanks to the not well-known contribution of Fanny Hesse (Mortimer 2001). The difficulties of directly observing microorganisms in their natural habitats greatly hindered the study of microbial communities. Microbial cultures, as artificial habitats and substrates, are not always an option since there is a great proportion of microorganisms that are impossible to culture in a controlled media, making their study with conventional culture techniques impossible (Stewart 2012). The rapidity of microbial processes also did not facilitate understanding of their dynamics. Improved culturomics and the invention of electron microscopy boosted the progress of modern microbiology, but the greatest impetus to microbiology has come with the advent of molecular techniques and, in particular, DNA sequencing. One of the goals of the NGB project is, precisely, to: monitor microbial communities from diverse ecosystems using NGS; and, identify appropriate pipelines to sample, sequence and infer interaction microbial networks without the limitations of isolation and culturing of different microorganisms.

2.2 Composition of microbial communities

The microorganisms comprise a large number of taxonomic groups. It includes all bacteria and archaea taxa as well as many eukaryotic fungi and all protists (Hug *et al.* 2016). This brings a great degree of complexity to microbial studies, particularly when this taxonomic diversity is mixed together in a microbial communities. Communities can be defined as multi-species assemblages within which organisms live and interact, in a contiguous environment (Konopka 2009). For microbial communities, the dimensions of the continuous environment can vary greatly, from meters, in case of some fungal structures, to micrometers. The different microbial taxa also vary in abundance within in a community. A few taxa may make up the majority of the communities' individuals while the a large part of the taxanomic richness of the community will have a reduced abundance (Fuhrman 2009). This large taxonomic richness is reflected in the great genetic variability of microbial communities, and high rates of mutation facilitated by the rapid reproduction of many microorganisms. Moreover, mechanism of horizontal transfer of nucleic acids allow different taxa to exchange or share genetic material, even where they have a different taxonomy (Gogarten & Townsend 2005; Saak *et al.* 2020).

2.3 Microbial services

Microorganisms provide ecological functions or services to all ecosystems, and take part in most of the biological processes that sustain life (Mace *et al.* 2012). This makes microbial communities an important subject of study. The ecosystem services provided by microbes can be split in three main groups following the Millennium Ecosystem Assessment classes (World Health Organization 2005; Saccá *et al.* 2017):

- **Regulating** the ecosystems by controlling the development of pathogens and reducing the levels of pollutants;

- **Supporting** other biological processes by stimulating the growth of other organisms or taking part in the nutrient cycle;
- **Provisioning** ecosystems with different compounds, from nutrients to secondary metabolites like antibiotics.

Microbial services also play a major role in sustaining human societies, starting with the human body itself. Humans, and all other macrobial organisms, live in constant symbiosis with a large spectrum of microorgansims, forming a complex that has been called the holobiont (Margulis *et al.* 1991). This complex, and the interdependencies of the host and microbiota, is conserved along time and even inherited, reaching the point where the holobiont itself has been considered a unit of evolution (Skillings 2016). Microorganism process support nutrient uptake (digestion) and regulate infections (Alemao *et al.* 2021; Gao *et al.* 2014). Microorganism may also have an important role in regulating many metabolic processes (Nieuwdorp *et al.* 2014). As a consequence, knowledge of the microbial communities inhabiting a host can lead to a better understanding of health and improve medical treatments (Costello *et al.* 2012). Microorganisms are also basic for the production of multiple goods, including agricultural products. Plants release organic compounds into the soil, enhancing the development of microbial communities, and, in exchange, these microbial communities deliver multiple services to the plant (Riva *et al.* 2019). Microorganisms provide inorganic nutrients necessary for plant development, by fixing nitrogen (Xiong *et al.* 2021; Moreau *et al.* 2019) or decomposing organic matter (Kuzyakov & Cheng 2001; Han *et al.* 2020). Plant development, whether in favorable conditions or under abiotic stress, can not be understood without consideration of the microbial communities living in the rhizosphere (Kumar & Verma 2018). Agricultural and human health are just two of the most relevant examples of the direct impact of microbial communities on human societies, but research into microbial community effects now extends to aquatic ecosystems, energy production and industrial processes(Stulberg *et al.* 2016). Knowledge of the dynamics of microbial communities and of drivers of change (biotic interactions and abiotic factors) is key not only for ecosystem management but also for many fields in medicine and food production.

2.4 Microbial ecological interactions

An understanding of microbial interactions can help us to anticipate changes in ecosystems and to make decisions to mitigate or adapt to change (Blanchard 2015). Ecological networks, encompassing all microbial taxa and their interactions present, might be used as a system-wide indicator of ecosystem stability (Bohan *et al.* 2017). Efficient monitoring would therefore require knowledge of the diversity of interactions affecting the abundance of the microbial taxa, as well as the services these taxa and their interactions might provide.

To detect microbial interactions, the first imperative is to define clearly and explicitly what a microbial interaction is. As referred to in Section 1, an interaction is any action or influence of one taxa on another that leads to a change in abundance of one or both taxa. Intuitively, therefore, the first prerequisite for a microbial interaction to happen is the co-occurrence of both interacting taxa in

the same environment. However, co-occurrence is not equivalent to interaction. We also require an action or influence. The action or influence that one microbial may cause on others is highly diverse; potentially being as varied as the microbial world itself. Some interactions are caused by direct contact of individuals, like predation performed by protists (Leander 2020), while in most cases the action or influence of microorganisms on each other is mediated by metabolites excreted into the environment (Tshikantwa *et al.* 2018; Schmidt *et al.* 2015; Pierce & Dutton 2022). Microorganisms can directly kill other taxa in this manner, but also can transmit information (Mencher *et al.* 2021) or even genetic material (Friesen *et al.* 2006). Monitoring requires us to focus on the interactions that shape the structure of the ecosystem, modifying the abundance of different taxa. Thus, to monitor an ecosystem, an interaction is considered to be any effect on the abundance of a microorganism caused by another (Faust & Raes 2012). This consideration leads to a logical definition of an interaction that is not based on an interaction mechanism followed by a causal action, but rather as the realised result of a mechanism and action; an *effect* of abundance change.

Considering a given pair of interacting taxa, the possible interaction effects between them can be classified as a function of an increase (↑), decrease (↓) or no change (0) in their abundances. The different types of effect of abundance change that can be produced by different interaction types are depicted in Table I.1). Interactions can also involve more than two taxa, leading to more complex combinations of effects on taxa abundances. For example, the building of some biofilms can require the collaboration of more than two different taxa (Liu *et al.* 2016). This might means that interactions cannot be studied in isolation, pairwise, but must be treated at higher orders to have a complete picture of the interactions taking place in an ecosystem. Such biotic complexity is beyond the scope of this thesis, and may even be beyond our computing power. Abiotic conditions, such as temperature, humidity etc., can induce change in taxa that has the appearance of interaction. Taxa that develop in same conditions, and respond similarly to the abiotic conditions, can appear to have a mutualistic interaction. In the same way, microbial taxa adapted to different abiotic conditions can seem to prey upon or compete with one another (Derocles *et al.* 2018). To avoid interpreting spurious correlations as interactions, it is therefore necessary to account for abiotic conditions.

To produce a real interaction, we require that two taxa must co-occur and their interaction must result in a realised effect on one or both of their abundances. There is, however, a temporal delay from the moment where the interaction starts to the time point where there is an effect on the abundance of the taxa. The temporal scales of microbial interactions can be very variable. Effects on abundance can sometimes be observed in fractions of a second or minutes, but, depending on the interactions and ecosystems, the temporal scale can increase to days and even months (Fuhrman *et al.* 2015). For example, it takes some time for an antagonistic molecule excreted by one microorganism to contact another, produce holes in its cell wall and then kill it. This amount of time may be considerably different to the delay required to form a bio-film, that could increase the abundance of other microorganisms. Interactions can also have contrasted dynamics over time. The change in abundance of taxa involved in an interaction can stabilize even as the

Table I.1: Interaction types as a function of the effect on the involved taxa. Inspired by Derocles *et al.* 2018; Faust & Raes 2012.

Type of Interaction	Effect on taxa A abundance	Effect on taxa B abundance	Description
Amensalism	0	↓	Taxa A causes a decrease on the abundance of taxa B without suffering any effect on the abundance
Commensalism	0	↑	Taxa B increases its abundance thanks to the effect of Taxa A
Competition	↓	↓	Both taxa abundance decreases by the effect of the other. This can be caused by direct competition (directly harming the other taxa) or exploitation competition (they need the same resource an in consequence there is less available)(Fredrickson & Stephanopoulos 1981).
Mutualism	↑	↑	The abundance of both taxa increases by the effect of the other.
Neutralism	0	0	Both taxa co-occur but there is no effect on their abundance, and therefore, no interaction.
Parasitism or Predation	↑	↓	Taxa A develops at the expense of taxa B.

interactions continues (Gerber 2014). This is the case for exploitation competitive interactions, for example (Figure I.1A). Here, the appearance of a new taxa starts a competition for a resource, decreasing the abundance of the established taxa to a point where the abundance of both taxa stabilize even as they continue to interact. Interactions can also lead to linear or exponential increases or decreases of the taxa involved. In the case of a decrease, the abundance of at least one taxa may eventually be reduced to zero, causing its exclusion from the environment, and thereby, ending the interaction (Figure I.1B). However, not all the effects on abundance produced by an interaction are continuous in time. Figure I.1C shows the asymptotic behaviour of the abundance of the interacting taxa, over time, common in predator-prey interactions (Goel *et al.* 1971). Here, the effect in the abundance expected for predation ($\uparrow$ predator, $\downarrow$ *prey*) only occurs for the fraction of time that the population of the prey is large enough to be a food resource for the predator population (marked on the Figure in blue). For the remainder of the time, the predator population decreases because there is not enough prey to sustain it, and the prey population increases because it has escaped predation. If biomonitoring sampling is not performed on the correct time scales, therefore, the cyclic behaviour of the abundance greatly hinders interaction detection because it is effectively masked by the dynamics (Derocles *et al.* 2018). Thus, to follow the temporal development of the set of interactions happening in an ecosystem, it is necessary to take into account the different potential interactions that could occur and their relevant time scales.

3 Metabarcoding

As stated above, microbial interactions should be a key component of ecosystem biomonitoring (Bohan *et al.* 2017), and this requires information about the occurrence and abundance of microorganisms in an ecosystem (Faust & Raes 2012). Microorganisms cannot be observed with naked eye and the use of microscopes and other optical devices for the evaluation of occurrence necessitates ecological sampling and subsequent culturing of the microbial community, with its inherent costs and biases. Quantifying the abundance of different microbial taxa using a microscope is also difficult and lacks accuracy. It is in this context, that DNA based techniques have the greatest potential. An environmental sample taken from water, soil, air or other biological material potentially includes all the microorganisms inhabiting that environment, and as a consequence their DNA. The process of identifying the taxa present in eDNA samples is called metabarcoding, making reference to the use of DNA as a unique, identifying barcode for each taxa.

Since the discovery of the structure of DNA by Watson and Crick (Watson & Crick 1953), numerous DNA technologies have been developed to better understand and explore the biology of the living world. Two of these techniques are necessary prerequisites for the development of metabarcoding: Polymerase Chain Reaction (PCR) and DNA sequencing. PCR is a molecular biology technique, developed by Kary Mullis in 1983 (Mullis *et al.* 1986). It uses a thermal resistant polymerase to produce copies of DNA regions between given DNA sequences. These regions are detected by a pair of adaptors called primers. PCR is used to

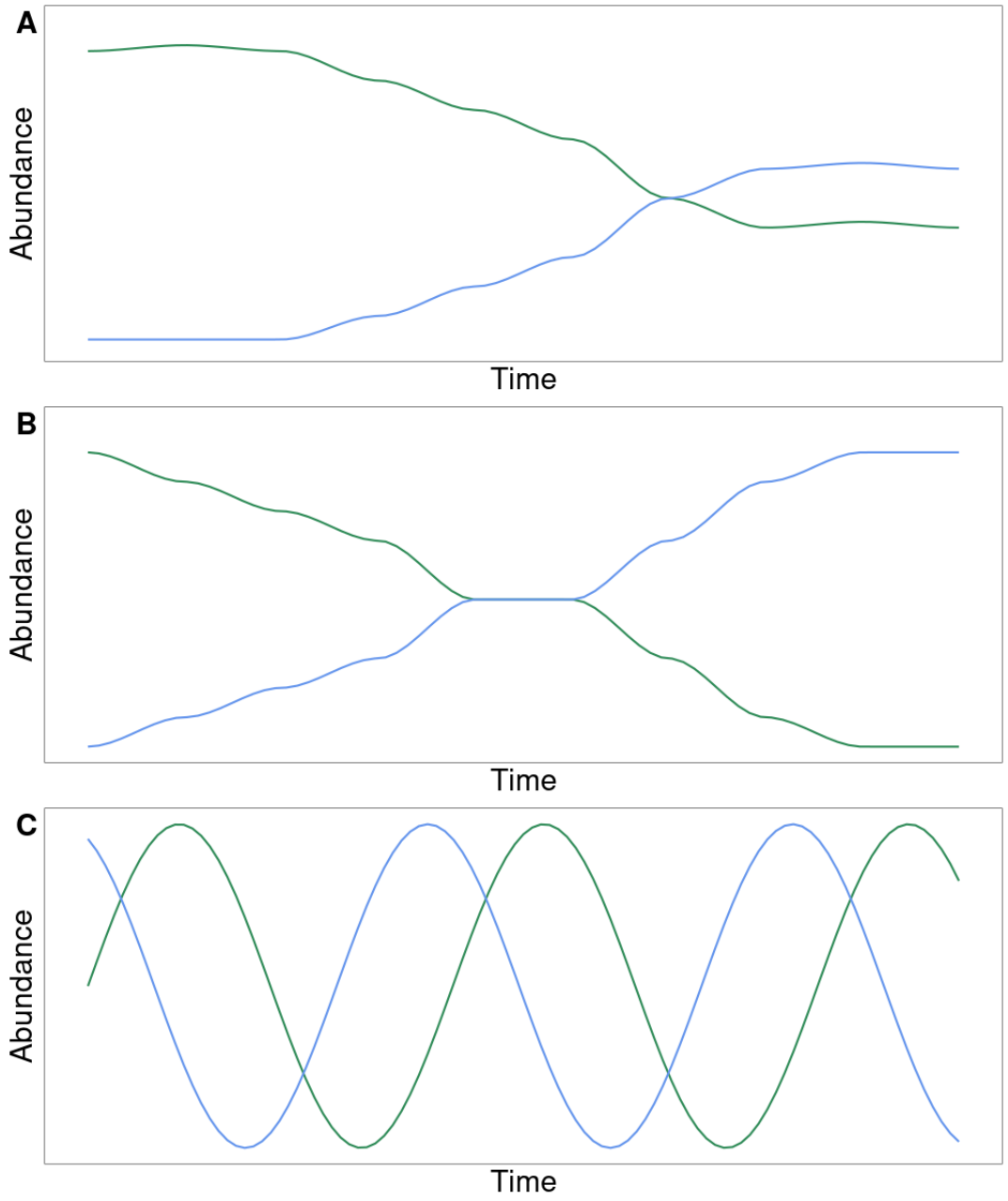


Figure I.1: Change in the abundances of a taxon following the establishment of a second taxon in the media. **A**: Exploitative competition interaction where both taxa arrive at an abundance equilibrium. **B**: Competitive exclusion process where one taxa is able to exclude the other from the media. **C**: Cyclical dynamics of abundance in predator-prey interaction.

generate a large number of copies of the region of the DNA used for barcoding. It is then possible to obtain the exact sequence of this region by DNA sequencing. In 1972, Walter Fiers sequenced a DNA sequence for the first time(Jou *et al.* 1972), but it was a few years later that Fredrick Sanger developed the DNA sequencing technique that was going to be widely adopted (Sanger *et al.* 1977). Sanger sequencing starts by copying the targeted DNA region using a mix of normal and fluorescent nucleotides (one color for each type of nucleotide). When a fluorescent nucleotide was added to the DNA chain, the copying reaction stopped. This produced a set of different length DNA fragments with a fluorescent nucleotide at the last position. Then, these fragments were sorted by their size, initially using gel electrophoresis. However, following further development, a capillary was used to both sort the sequences by size and then detect the fluorescent nucleotide at the end of the sequence. Since the DNA fragments were ordered by size and the color of the fluorescent nucleotides detected indicated which nucleotide was the final one for a given size, it was then possible to reconstruct the whole sequence from start to the end. Sanger sequencing was used for many years, but it only allowed the processing of a single sequence at a time. In 2005, Next Generation Sequening (NGS) was invented as an improvement to Sanger sequencing that allowed multiple sequences to be processed simultaneously (Mardis 2013). This was achieved by adding synthetic adaptors to the DNA sequences, allowing specific sequences to be identified and thereby permitting millions of DNA fragments from different samples to be processed through sequencing at the same time (Hu *et al.* 2021). Each adaptor identifies the sample origin of the sequence and serves as a binding point for the sequencer. NGS lead to a significant decrease in the time and cost of sequencing, extending the use of DNA sequencing to many fields of study, including biomonitoring of environmental samples.

3.1 Sampling and DNA extraction

The first step in sampling eDNA is to define the sampling 'universe', which is either the the ecosystem or environment that the samples represent (Dickie *et al.* 2018). It is important to clearly define the sampling universe in order to properly select the sampling procedure to make the samples representative. For example, if the intention is to study the microbes and their interactions in a pond, it will be necessary to sample the different parts of the pond (perimeter, center) at different water depths (surface, bottom, etc.). If only the surface of the pond is sampled, the study will not reflect all the pond's microbial communities. Once the sampling universe is defined, it is possible to design the sampling procedure. While the size of samples (e.g. volume of water), and their location in space and time may depend upon the precise research question, the number of replicate samples is a critical decision common to all eDNA surveys focused on interactions. The lack of appropriate sample replication can lead to non-significant results. NGS may have reduced the cost of sequencing, but these techniques are still expensive and sampling effort has to be considered carefully in order to keep costs to effective levels (Smart *et al.* 2016).

Microbial eDNA used for metabarcoding is mostly contained inside the microorganisms of the sample, thus requiring a cell lysis process to extract it. The

great differences in cell walls and cytoplasmic membranes of across the members of microbial communities means that a general or universal lysis procedure needs to be adopted. These typically combine physical, chemical and enzymatic disruption of microbial cell walls (Teng *et al.* 2018). DNA purification requires that any waste biological material or inorganic traces in the lysate have to be removed, but also it is necessary to assure that any compounds present in the sample that could damage the DNA are removed (Pearman *et al.* 2020). Extraction kits and protocols have now been developed to extract eDNA from a wide range of environmental samples, greatly simplifying this previously difficult step in metabarcoding.

3.2 DNA amplification and Sequencing

The purified DNA obtained from environmental samples is amplified, increasing the numbers of copies, using PCR. Specific primers are used to discriminate the DNA derived from the microbial community from that of macro-organisms. Generally, DNA regions coding for ribosomal RNA are used in metabarcoding microbes (Abdelfattah *et al.* 2018). The taxonomic heterogeneity of microbial communities also adds a degree of difficulty to metabarcoding, in that there is no single pair of primers that can be used to barcode all microorganisms. Rather it is necessary to perform different PCR amplifications, with different primers, for each taxonomic group studied (Francioli *et al.* 2021). Some of the commonly used primers for the most important groups are shown in Table I.2.

Sequencers typically cannot sequence the whole ribosomal region, as these are too long. Consequently, each taxonomic group has specific primer pairs that are used to target smaller sub-regions of the ribosomal region (Beckers *et al.* 2016). The resulting amplified DNA fragments typically have a length of between 100 and 550 nucleotide pairs. These fragments are then sequenced using one of the NGS platforms available in the market. The most widely used NGS platforms for metabarcoding are 454 pyrosequencing, various Illumina based technologies and Ion Torrent (Aragona *et al.* 2022). These platforms use different protocols to obtain the DNA sequence of each fragment, but produce similar end results; a list of all the DNA sequences present in the amplified DNA from the ecological sample. Each DNA region is normally sequenced in two directions, forward and reverse. Bioinformatic processing is then done to merge the two reads to obtain a higher resolution sequence. A measure of the quality of each sequenced nucleotide is also obtained during sequencing. The quality measure provides information for the likelihood that an error in identifying the correct nucleotide has occurred. Low quality sequences can then be filtered out of the data to reduce the number of introduced errors in the dataset (Pauvert *et al.* 2019).

The different NGS platforms produce similar results, and share a number of limitations. There is a maximum amount of DNA fragments that can be processed, called the sequencing depth. This means that the DNA sequences obtained for each ecological sample are a sub-sample of the total available sequences, introducing a compositionallity bias into the dataset (Gloor *et al.* 2017). While the first generation of NGS platforms could not process sequences longer than around 1000 base pairs, the more recently introduced third generation sequencing technologies allow the sequencing of much longer DNA fragments (Heeger *et al.* 2018; Tedersoo *et al.*

2021). Longer DNA sequences provide better barcodes, at the expense of reduced sequencing depth and other potential sequencing issues (Latz *et al.* 2022). Metagenomics studies can now also be perfromed to study the presence and abundance of different taxa in a sample. Metagenomics uses a sequencing process called shotgun sequencing to obtain random reads of any part of the genome of different taxa (Quince *et al.* 2017). This skips the PCR amplification step and gives information about the whole genome of those taxa present. However, the sequencing depth retrieved is even lower and the method demands comprehensive databases to be able to identify where each sequence comes from in each taxon's genome.

Table I.2: DNA region amplified to metabarcode the different taxonomic groups present in microbial communities.

Taxonomic group	**Amplified region**	**Taxonomy Database**	**Literature**
Bacteria	16S ribosomal unit	SILVA ribosomal RNA; Ribosomal database project (RDP); Greengenes	(Marchesi *et al.* 1998; Quast *et al.* 2012; Cole *et al.* 2014; McDonald *et al.* 2012)
Fungi	Internal transcribed spacer (ITS) between small and large ribosomal DNA	UNITE; Ribosomal database project (RDP)	(Gardes & Bruns 1993; Nilsson *et al.* 2019; Cole *et al.* 2014)
Archea	16S ribosomal unit	SILVA ribosomal RNA; Ribosomal database project (RDP); Greengenes	(Takai & Horikoshi 2000; Quast *et al.* 2012; Cole *et al.* 2014; McDonald *et al.* 2012)
Protists	18S ribosomal unit	The Protist Ribosomal Reference database (PR^2)	(Dollive *et al.* 2012; Guillou *et al.* 2013)

3.3 Sequence clustering

Species are the basic unit of ecology (Mayr 1982). The definition of a species is variable depending on the particular species-concept invoked (De Queiroz 2007). However, there is broad agreement that individual members of the same species will share a highly similar genotype and phenotype, and that they will perform and

undergo similar ecological functions and processes (Ward *et al.* 2008). Study of the ecological functions in an ecosystem typically use microbial species (or genera in the case of some prokaryotes) as the functional unit. The list of DNA sequences present in each eDNA sample represents the different species present, across the taxonomic group targeted for sequencing (bacteria, fungi, etc.). Depending on the sequencing depth of the platform used, the count of sequences per sample can vary between thousands to hundreds of thousands of copies. We posit, therefore, that sequences from individuals of the same species should be similar (or even identical), while sequences from different species will have more differences. Grouping the sequences by similarity produces an approximate measure of how abundant each species is in a sample, as a function of the sequence counts (Schloss *et al.* 2009; Shelton *et al.* 2022). There is considerable, ongoing debate about how approximate such a measure of abundance is, due to the many different biases inherent in the molecular biology processes of sequencing (Gloor *et al.* 2017; Lamb *et al.* 2019; Zinger *et al.* 2019). Normally, sequencing reads are grouped or clustered to a reference sequence that is representative of the whole taxonomic group of interest. The reference sequence could be the longest sequence in the group, the most abundant sequence in the group or sequence selected at random (Links *et al.* 2013). The group clustered around a sequence is called an Operational Taxonomic Unit (OTU), and it is expected to be representative of a predefined taxonomic level, such as a species (Caron *et al.* 2009).

There are two main strategies to cluster the DNA sequences into OTUs: reference-based; and, *de novo*. Reference-based clustering compares the DNA sequences with a database of previously classified sequences and sequence groups that share the same taxonomy. While this method has proved to be accurate in some cases, it relies heavily on the quality of the sequence databases (Cline *et al.* 2017). Thus, reference-based clustering algorithms may fail to identify and group unknown or poorly studied taxa of which there may be many in a biomonitoring situation. The alternative to reference-based clustering is *de novo* clustering. This methodology does not require external taxonomic reference sources, such as databases. *De novo* clustering compares the different sequences and groups them as function of an arbitrary similarity threshold. Typically, the homology threshold is established at 97%. This means that all sequence reads clustered to an OTU share at least 97% of sequence homology (Brown *et al.* 2015). This criterion allows clustering to be applied to any metabarcoding situation, irrespective of the quality of prior knowledge of the microbial community. Nowadays, there are numerous algorithms that perform *de novo* clustering and have been widely used (Caporaso *et al.* 2010; Schloss *et al.* 2009; Edgar 2013). It remains a matter of debate whether a common, possibly arbitrary, similarity threshold can correctly delineate the differences between microbial species (Nguyen *et al.* 2016a). This is because different microbial species have great variability in the same genome region, with some species having 99% homology, while others may be more variable, having 95% homology.

An improvement to OTU clustering using homology thresholds is amplicon denoising. The denoising process is based on the premise that sequences containing simple sequencing errors are less likely to be observed than 'biologically correct' sequences (Callahan *et al.* 2017). Sequences can therefore be clustered together if the difference in nucleotide composition can be caused by a simple error of sequenc-

ing. The pool of metabarcoding sequences is used to estimate the rate at which one nucleotide can be substituted by another in error (Prodan *et al.* 2020). The OTUs obtained by denosising algorithms are called Amplicon Sequence Variants (ASVs) or Exact Sequence Variants (ESVs). Tools like DADA2 (Callahan *et al.* 2016), Deblur (Amir *et al.* 2017) or Unoise (Edgar 2016) can perform amplicon denoising from metabarcoding data. DADA2, which is the most widely used of the tools, has been shown to recover the composition of mock communities better than clustering and other denoising tools (Pauvert *et al.* 2019; Prodan *et al.* 2020). Nevertheless, given that the ASV inference is based on errors during the sequencing process, different ASVs may still be from the same species, and some manual or semiautomatic curation of the obtained ASVs may be necessary (Frøslev *et al.* 2017). Henceforth in this thesis, I treat the terms OTU and ASV as synonyms.

The OTU inference processes provide a list of representative sequences and the number or count of DNA reads for each clustered sequence. Usually, the abundance information of OTUs is organized as a p x n matrix, where p is the number of OTUs, n the number of samples in the metabarcoding survey and each cell contains the number of sequence counts, for each OTU in each sample. Posterior filtering processes may be done to delete chimeric sequences, generated by the mixed amplification of two or more DNA fragments (Wang & Wang 1997). Filtering can be performed to delete low abundance OTUs, which may be non-informative of the ecosystem or even be errors (Cao *et al.* 2021).

3.4 Taxonomic assignment

Each OTU has a representative nucleotide sequence that can be compared to existing databases to gain knowledge of the taxonomic level (Phylum, Order, Genus, Species, etc.). For each taxonomic group selected during the PCR process (Table I.2), a manually curated reference database is then created. These databases contain the list of DNA sequences corresponding to the amplified region of the DNA and taxonomic information related to each sequence. This is normally standardised to genus level information for prokaryotes and species level for eukaryotes (Nilsson *et al.* 2019; Quast *et al.* 2012; Guillou *et al.* 2013). It is also possible to construct custom databases for taxonomic assignment. This option is useful when the microbial community studied is well known (for example a mock community in *in vitro* studies) or in the case of studying rare microbial taxa that are rarely found in the existing databases (Lennard *et al.* 2018). The taxonomic assignment is performed either: 1) by aligning the OTU sequence to the most similar sequence in the database, to a given threshold (Alonso-Alemany *et al.* 2011); or, by 2) estimating the probability (using Bayesian or other methods) that some small section of the the OTU sequence (typically 8 nucleotides) is part of the database sequence, and then evaluating the confidence of the estimation (Wang *et al.* 2007).

Assigning an OTU to a taxonomy facilitates some understanding of those species (or genera) that are present in the ecosystem. The taxonomic assignment also acts as a link between the OTU and the information contained in the literature and in the different relevant databases of functions and processes, allowing attribution of meta-data such as functional traits (Djemiel *et al.* 2022). The accuracy of the taxonomic assignment at genus or species level is variable, how-

ever, and rarely reaches the standard of complete independence of the database or classifier used for assignment (Edgar 2018; Rohwer *et al.* 2018; Joos *et al.* 2020). This would mean either that the OTU was assigned to an incorrect taxon, was incompletely assigned to some taxonomic level other than genus/species or was not assigned at all. The sources of errors in taxonomic assignment are diverse, including sources from sequencing and clustering errors to incomplete and ambiguous taxonomic and sequence databases (Mathon *et al.* 2021).

3.5 Metabarcoding data and Community analysis

The metabarcoding process produces an OTU matrix containing the abundance of the different OTUs in each sample, as detailed above. This matrix may also have associated matrices of abiotic variables, for each of the different samples and for the taxonomy of the different OTUs. Further information that can be attached to the matrix are the alignment trees of the OTU sequences, obtaining information on how each of the OTUs are related phylogenetically (Otu & Sayood 2003). Matrices can be stored in standardised file formats, such as a phyloseq object (McMurdie & Holmes 2013). Prior to studying the microbial interactions, the data obtained from metabarcoding can be used to perform multiple analyses to characterize the microbial communities in the different samples. Measures and metrics of OTU richness, diversity (Simpson 1949) and the evenness of relative OTU abundance (Pielou 1966) are typically computed. These measures indicate how the distribution of OTUs varies across the samples. Measures of microbial community dissimilarity can be analysed to identify those samples that share more similar microbial communities or are outliers (Chao *et al.* 2005; Morris *et al.* 2020). Knowledge of how the microbial communities of different environmental samples diverge can help us to comprehend the role of abiotic factors in shaping these microbial communities, as a complement to the study of microbial interactions within each community.

4 Interaction network inference

Interactions between microbial OTUs can not be understood in isolation. Microorganisms live in communities shaped by their constituent members' interaction network. The objective of the NGB project is to automatically reconstruct these interaction networks from eDNA to monitor ecosystems (Bohan *et al.* 2017). Interaction networks can be graphically represented as a set of nodes, representing the OTUs, connected by a set of edges, representing the interactions (Shannon *et al.* 2003). The size of the nodes can be homogeneous or vary with OTU abundance. The edges can have a different thickness, representing the strength of interaction, or color, depicting different types of interactions. Edges can also have a direction, they are directed, which indicates that the interaction has a contrasted effects on the taxa involved. As noted in section 2, microbes and their interactions are difficult to observe with the naked eye and, furthermore, the mechanisms and high rate of interaction limits the possibility of observing interactions or their effects under the microscope. Indirect techniques have to be used, therefore, to reconstruct the interaction network of a microbial community.

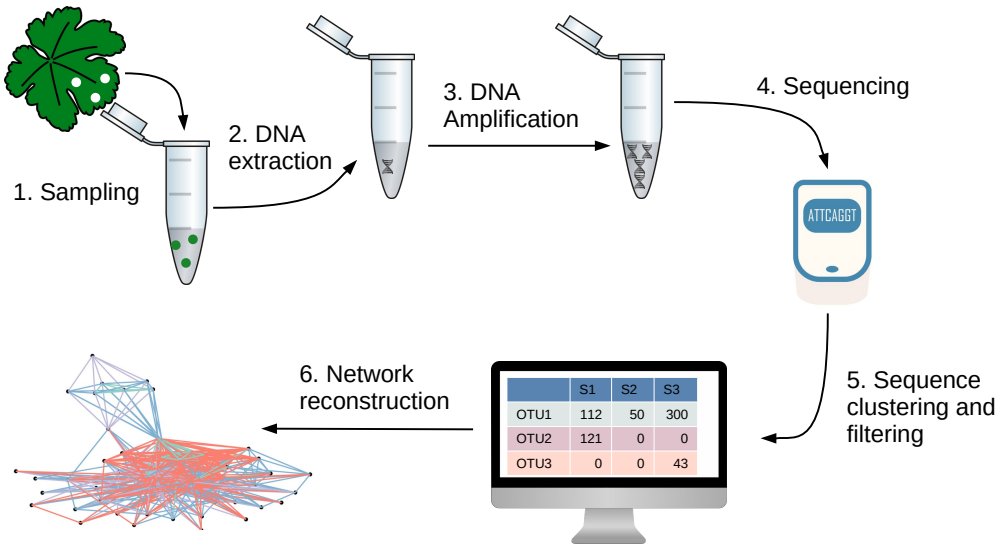


Figure I.2: Steps to reconstruct an interaction network using eDNA: 1. Ecological samples are taken from the environment; 2. eDNA is extracted from the ecological sample; 3. More copies of the DNA are obtained by amplification using primers specific for a taxonomic group; 4. DNA is sequenced; 5. A bioinformatic pipeline is followed to cluster the sequences in OTUs and obtain a measure of OTU abundance; 6. Interaction networks are reconstructed using OTU abundance data.

The realised effect (result) of an interaction is a change in the abundance of at least one OTU involved (Faust & Raes 2012, section 2.4). It is therefore possible to imagine that interaction network reconstruction can be done by studying the changes in OTU abundance. Correlation measures (e.g. Spearman's rank correlation coefficient (Spearman 1904); Pearson correlation coefficient (Pearson 1895)) between abundance of OTU pairs between different samples have all been interpreted as a measure of the likelihood of an interaction (Faust *et al.* 2012). However, interaction networks reconstructed from correlations between abundance metrics have to take into account a number of biases and other problems. First, as detailed in the section 3.2, abundance data obtained by sequencing is compositional. Sequencers are only able to process a given number of DNA sequences and, consequently, the OTU counts are relative to the maximum number of sequences processed (Gloor *et al.* 2017). If abundance information obtained from sequencing is treated as absolute, the predicted interaction network could have spurious edges caused by variation in sequence depth across OTUs and samples (Li *et al.* 2016). Spurious edges might also be introduced to the reconstructed network by abiotic sample conditions. Preference for the same conditions (humidity, temperature, etc.) can result in a positive correlation, all without the action of an interaction. In addition, the effect on two focal OTUs undergoing an interaction, of the abundance of a third OTU, could induce spurious positive or negative correlations depending on the type of interaction (Carr *et al.* 2019). Finally, OTU presence across samples can also be subject to considerable variation caused by the probability of colonisation of the sample (Gotelli 2000; Zhang *et al.* 2016). This variation means that the OTU matrix will be sparse, with many samples

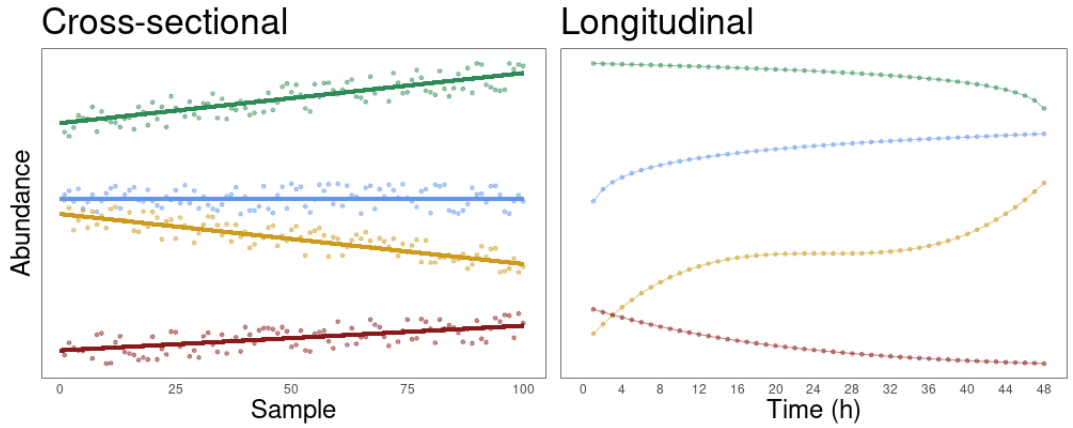


Figure I.3: Cross-sectional and longitudinal abundance measures of four OTUs. Cross-sectional network reconstruction uses the change in abundance of samples at the same point in time, but different point in space for each OTU. Longitudinal network reconstruction uses samples from the same point in space, but different points in time for each OTU.

having absent OTUs that can affect the measure of correlation. There are two main strategies to reconstruct interaction networks from metabarcoding data, using correlation, which have been proposed as solutions to circumvent some of these biases and problems: cross-sectional inference and longitudinal inference (Figure I.3):

4.1 Cross-sectional network inference

Cross-sectional network inference uses the correlation of abundances between different samples taken at the same point in time, in order to infer associations between pairs of OTUs (Dohlman & Shen 2019). Where the between pair correlation measure is significant, an association between the OTUs can be inferred. The development of NGS techniques has stimulated the development of numerous inference approaches focused on reconstructing microbial interactions using cross-sectional metabarcoding data. One of the initial methodologies, proposed by Faust et al. (2012) (Faust *et al.* 2012) used an ensemble of four measures of correlation: Pearson and Spearman correlation, and Bray-Curtis and Kullback-Leibler dissimilarity. At the first step, a general overall score for each network edge was obtained by combining these four measures. Then, at the second step, the significance of this score was assessed by bootstrapping. This ensemble, among others, was later implemented in a network inference tool called CoNet (Faust & Raes 2016). The first widely used network inference tool offering a complete pipeline, from sequence counts in eDNA samples to interaction inference, was SparCC (Friedman & Alm 2012). Published in 2012, SparCC attempted to avoid some of the biases of compositional data by using log-transformation (Aitchison 1982) to obtain linear Pearson's correlations. The significance of the correlations was then computed using a bootstrapping procedure. SparCC kick-started a decade of network reconstruction tool development, facilitated by cheap and accurate metabarcoding datasets. While SparCC deals explicitly with sparse (zeros in the OTU matrix)

and compositional data, it still reports many spurious associations between OTUs due to indirect effects. Newer tools have therefore used measures other than linear correlation. CCLasso (Fang *et al.* 2015) and SPIEC-EASI (Kurtz *et al.* 2015) used correlation of latent variables and Inverse Covariance respectively to discard spurious associations introduced by abiotic factors and third-party, indirect OTU effects. SPIEC-EASI adopted the StARS model for selecting true associations in place of bootstrapping (Liu *et al.* 2010). The management of the sampling effect is improved in HMSC (Ovaskainen *et al.* 2017) and PLN (Chiquet *et al.* 2019), enhancing inference accuracy by taking into account the sample environmental covariate data as offsets. Efforts have been made to introduce sources of information other than the OTU abundance matrix. MPLasso (Lo & Marculescu 2017b), for example, introduces the use of prior microbiological knowledge through data-mining or other external sources, weighting the edges of the network as a function of this information. Many other cross-sectional network inference tools have been developed using different correlation based strategies, and these have recently been extensively reviewed (Röttjers & Faust 2018; Dohlman & Shen 2019; Matchado *et al.* 2021).

4.2 Longitudinal network inference

Longitudinal network inference is based on taking samples from the same microbial community at different time points to infer associations as a function of the correlations of OTU abundance over time (Dohlman & Shen 2019). The dynamics of OTU abundances over time gives a more complete better picture of the effects of interactions and their dynamics (Faust *et al.* 2015). Longitudinal data are also compositional and sparse, like cross-sectional data, but since the samples are taken repeatedly from the same environment, the possibility of introducing spurious associations into the learning via contrasted abiotic effects is greatly reduced. Traditionally, time series of ecological abundance have been modelled using the Lotka-Volterra equations (Volterra 1926; Mounier *et al.* 2008; Fisher & Mehta 2014), but more recently other models have been proposed for longitudinal studies (Kodikara *et al.* 2022). As with cross-sectional analysis, specific tools have been developed to infer interactions from sequence count time series. For example, TIME (Baksi *et al.* 2018) identifies abundance variations over time, by OTU, and clusters taxa by similar abundance behaviours to evaluate those abundance patterns that might be caused by an interaction. An exhaustive listing of the available tools has been reviewed by Dohlman and Shen (2019) (Dohlman & Shen 2019).

4.3 Considerations on correlation tools for biomonitoring

Biomonitoring using interaction networks is based on the premise that OTU abundance data will be obtained from eDNA samples automatically. Automating of all the metabarcoding process would produce this abundance information with a predetermined periodicity (Bohan *et al.* 2017; Cordier *et al.* 2021). Intuitively, one could think that longitudinal inference tools would fit better with this periodic, time series data. The objective of network based monitoring is study how the different interaction networks evolve with time and to detect and/or antici-

pate ecosystem change (Kortsch *et al.* 2015), however. Cross-sectional network inference tools would therefore be a better option for biomonitoring given that they provide an interaction network at each time point, where sufficient samples are taken. Longitudinal network inference also presents two particular challenges that limits its application to biomonitoring. First, microbial interaction temporal scales can be variable, making it difficult to set appropriate sampling rates (Lugo-Martinez *et al.* 2019). Second, longitudinal inference relies on obtaining samples of the same microbial community at different times. While this may readily be possible for communities in aquatic environments, or even for the gut microbiome, it can be difficult to sample the same microbial community inhabiting the soil or plant material more than once. This hinders the automation of sampling and network reconstruction for a wide range of ecosystems, using longitudinal approaches.

Cross-sectional network inference tools based on correlation are extremely robust to noise in experimental data and can be run rapidly on even quite large datasets. They infer interaction networks at a given point in time and most of them have an implementation in R or other commonly used environments that facilitate their use. There is no clear agreement, however, on the minimum number of samples required to build a single network using network inference tools (Berry & Widder 2014; Hirano & Takemoto 2019). The variation of the amount of information needed to build a network leads to difficulties in obtaining reproducible networks from the same ecosystem. In addition, the often enormous variation observed between networks for purportedly the same ecosystem further underscores the requirement for a larger number of samples to detect ecosystem change using networks. Thus, it is imperative to evaluate the reproducibility of networks predicted by current network inference tools and implement statistical methods to obtain reliable network metrics. In this thesis, I address these issues in chapter II, using different network metrics to study the properties of interaction networks produced by the most widely used tools, SparCC and SPIEC-EASI.

Correlation-based tools are useful to explore associations between OTUs found using metabarcoding. They produce interaction networks that display the different OTUs in a microbial community and the associations between them. However, these tools define no clear link between the positive and negative associations they find, and the ecological mechanisms that actually generate them (Röttjers & Faust 2018; Carr *et al.* 2019). Associations, whether positive or negative, can be produced by a variety of different types of interaction mechanism or even be produced spuriously, as noted at the beginning of this section. Some types of interactions may not result in detectable correlations due to their complex nature (Pacheco & Segrè 2019; Weiland-Bräuer 2021). Microbial ecologists therefore need to interpret the associations proposed by the inference tools, and to assign the association to an interaction that can produce that type of correlation. This interpretation is done by applying their knowledge about the microbial OTUs involved and the information available in the literature and databases, and then be tested in culture assays to validate the proposed types of interactions (Pauvert *et al.* 2020; Hromada *et al.* 2021). Such interpretation is clearly subjective, being subject to the ecologist's knowledge of the literature or even their particular prejudices for a given type of interaction. Testing the inferred and interpreted interactions using culture based studies (culturomics) also has its limitations. Some microbial species can not be

isolated, because either they are unknown or they do not grow in synthetic culture media (Lagier *et al.* 2018). Culturomics is also expensive in time and money, and it is possible that the culture conditions will not promote the inferred interactions, which may be mediated by the precise availability of certain nutrients and other metabolites in the environment (Tshikantwa *et al.* 2018).

An automated reconstruction of interaction networks depicting explicitly the different interactions types would be more informative, allowing better monitoring of an ecosystem, testing and decision making. To date, few studies have attempted to classify microbial interactions by their type from eDNA metabarcoding data, as there is no reliable network inference tool able to perform the direct detection and automatic classification of interaction mechanisms (Dohlman & Shen 2019). This is in part because the tool needs to introduce prior ecological knowledge, which establishes a relation between the interaction type and the metabarcoding data. Currently, no correlation based tool is able to do this. Lo et al. (2017) used prior bibliographical data, but this simply weighted the correlations during the network inference process and cannot do automatic classification. Explainable machine learning, which uses logical approaches, is the only approach we have found that might directly detect and classify network links to their interaction type.

5 Explainable machine learning

5.1 Machine learning and the black box

Statistical machine learning and, more recently, Deep Learning have been widely and successfully used in many fields of science and engineering. They are used for pattern recognition in images, language translation and disease prediction (Mohan *et al.* 2019). Machine learning also is starting to have applications in the microbial world. Lee et al. (2020) used deep learning to infer interactions from spatial patterns in microscopic images and MInter is an automated text-mining tool of microbial interactions (Lim *et al.* 2016).

El Naqa and Murphy (2015) define a machine learning algorithm as a "computational process that uses input data to achieve a desired task without being literally programmed to produce a particular outcome". Machine learning algorithms are normally described as being either supervised or unsupervised. Supervised algorithms use labelled data, previously classified by a human, to train a model that is later used to classify new input data. Unsupervised machine learning, by contrast, uses non classified data to construct a model that clusters the examples using patterns in the data (Qu *et al.* 2019). Certain correlation based, statistical methods to infer interactions might also be described as unsupervised statistical learning. To infer microbial interactions, these correlation based methods would use abundance data to train an unsupervised model that clusters pairs of interacting OTUs. These approaches do not directly classify and suffer with problems of explainability, in that they produce inferences from models that are not readily understandable to humans.

Statistical machine learning algorithms receive the input data and construct a model used to process the data and produce inferences as output. The model constructed by the algorithm automatically relates the different features of the data,

potentially providing considerable predictive power. However, the complexity of these models does not allow the user to understand how the model produces the output. The model is essentially a black box model (Samek *et al.* 2017). In the case of microbial interactions, the different correlation measures between abundances will produce a network associating the OTUs, but there is no way to know what mechanisms producing the change in the OTUs abundances are at work (Carr *et al.* 2019). To tackle this black box problem for interaction inference, it is necessary to produce explainable models that establish human understandable associations between the different features of the data (Tonekaboni *et al.* 2019).

5.2 Being explainable

To make the interaction models explainable it is necessary to: 1) introduce scientific domain knowledge to the inference; and, 2) make the inference process transparent and understandable (Roscher *et al.* 2020). If it is desired that a networks represents interactions as network edges, it is necessary to state what interaction is and to define an appropriate mechanism (a rule or hypothesis) for each interaction type. These definitions can then be used to produce a model that infers interactions from interaction effects data. Other ecological knowledge related to interactions needs to be defined. For example, how are changes in OTU abundance to be considered, and which features of different microbial OTUs makes them interact, etc.? Once this domain knowledge is delimited and defined, the knowledge and OTU abundance data need to be expressed in a logical, human readable format. This human readable information will then produce a human understandable and transparent process for obtaining inference output that is both comparable with domain knowledge and expressed in a human understandable format (Beckh *et al.* 2021). The inference process is considered transparent if the user is able to understand all of its component parts and even repeat them at small scale if necessary (Belle & Papantonis 2021).

For machine learning processes to be human understandable they must mimic human logic and reasoning (Dai *et al.* 2019). Depending on the available information, there are different logical inference processes that produce new knowledge (Figure I.4). In the case of microbial interaction inference, the prior information available to us are eDNA derived OTU abundance changes and domain knowledge expressed as rules of interaction. This combination, and our interest to infer interactions denotes that the logical process we follow is one of abduction.

6 What does explainable mean?

Terms like human comprehensible and explainable, which can be linked to still other terminology such as understandable, logical, rules and transparency, are something we can probably all agree are things that we would want in science. However, they are still rather unclear. What do we mean by human explainable and what is human logic? Are there any concrete, simple examples? One example might come from the world of translation of language. The online translation tools, like Google Translate and DeepL, use various forms of correlation and statistical learning to build models linking the existing corpus (data) of written English,

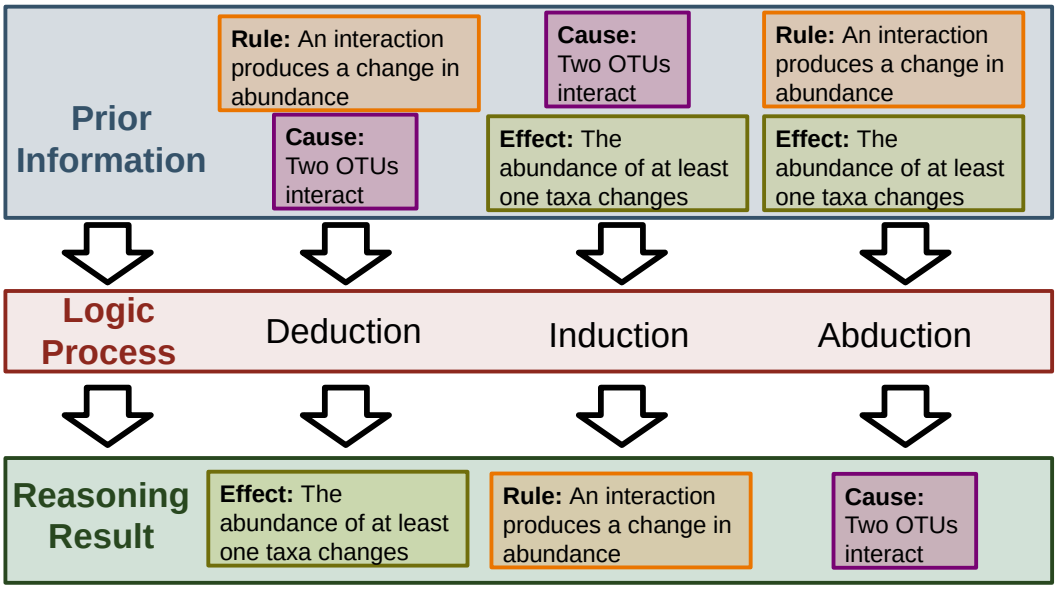


Figure I.4: Logic inference processes as a function of the input information. Deduction uses rules and causes to determine the exact effect. Induction uses causes and effects to determine the rules that drive them. Abduction proposes plausible causes for effects based on rules.

French, Latin, etc. As an example of use, these models might use correlations between the frequencies of particular words in known, comparable sentences in English and French. The model that is built can then make a translation (inference) into French from English.The same process could then be repeated for English and German, but would involve a different model.

There are other ways of imagining a translation being done, however. At school we learn the logical rules for the use of verbs, such as *to be*. These rules determine how the verb is used and also determine the placement, order and frequency of the appearance of the verb, *to be*, in English. In effect the rules of use and the patterns in the data are directly related. The rules and patterns are also readily human comprehensible and explainable – indeed that is the job teachers do. We might also expect that the use of the verb to be in other languages is broadly similar to that in English and the patterns in those languages similar also. What this means, at least in principle, is that we might use our rules for the verb *to be* and our knowledge of pattern to search data of other languages for similar patterns and thereby infer/predict words in those languages that may be similar to *to be*. We might thereby discover the verbs *être* in French or *sein* in German. Conceivably, with sufficient known rules in any one language it might be possible to infer specific meaning to translate from one language to any other. All that is required is an existing, known rule in one language and data with human understandable pattern. The logical, rules-based inference process is potentially generic, with the same model being applied to any language. Indeed, this rule / pattern approach was at the heart of the method used by Champollion to advance the translation of Demotic and hieroglyphic Ancient Egyptian from Ancient Greek written on the Rosetta Stone.

Another example of explainable machine learning application is human face

recognition. Statistical machine learning is able to compare two photos and decide if there is the same person in both. The algorithm is trained using many pairs of photos and it identifies different points on these photos that allow to detect a person. However, humans can not understand why these points are chosen and how these points define a person's face. By contrast, using explainable machine learning, we can define the traits that differentiate faces like nose size, separation between eyes, lip thickness, etc. Then, train the explainable machine learning algorithm to use these traits to compare the faces in the photos. That way, the reason why the machine learning algorithm detects the same face in two photos is "explainable". It might also explain why the identification is wrong or biased, potentially improving the currently low public trust in AI.

It is a process inspired by these descriptions which we propose to use in microbial ecology. We use hypothetical rules for different types of ecological interactions that are published and explainable. We apply these to patterns in our abundance data which we believe are caused by interactions between microbial OTUs. The approach is at its core logical and human explainable, and scientific because this approach has the hallmarks of 'hypothesis and test'. It is moreover, at least potentially, generic with application to any appropriate ecological data-set.

6.1 Abductive/Inductive Logic Programming

Abductive/Inductive Logical Programming (A/ILP) implements the abductive reasoning using and producing information expressed as symbolic logic. Rules based on symbolic knowledge are an expression of human-like reasoning. Thus, the knowledge transfer between the machine learning and the user is facilitated. A/ILP has been used in many different fields of knowledge, including metabolic network inference (Tamaddoni-Nezhad *et al.* 2006), elaborating the processes involved in cow milk production (Sasaki *et al.* 2019), and inferring trophic relations from observational data of arthropods (Bohan *et al.* 2011).

Abductive Logic Programming is typically applied to problems that can be separated into two disjoint sets of predicates: the observable predicates and the abducible predicates (Kakas & Papadopoulos 1996). In practice, observable predicates describe the empirical observations that we are trying to model, such as OTU abundance information. The abducible predicates - here the interactions we infer - describe underlying relations in our model that are not observable directly but can, through the theory, bring about observable information. We may also have background predicates (prior knowledge), which are auxiliary relations that help us link the observable and abducible information (Tamaddoni-Nezhad *et al.* 2021). The abduction process generates hypotheses that compress a set of experimental observations (abundance changes). The amount of information compressed can be understood as a measure of the likelihood of an hypothesis being true, in this case an interaction. Depending on the abduction process, information compression can be used in different ways to decide which abducted cases are more likely to be true.

In this thesis, the potential of A/ILP to infer microbial interactions is explored across Chapters III and IV. In Chapter III, basic interaction rules are used to reconstruct networks from computer-generated abundance data. Chapter III also shows

how the domain standard A/ILP program, Progol, and its compression measures are used to infer interactions (Muggleton 1995). In Chapter IV the interaction rules are extended by introducing the concept of exclusion. The new rules are used to abduce interactions of both computer-generated and real metabarcoding datasets. In chapter IV the program used for the abduction is PyGol, a novel, fast implementation of A/ILP (Varghese *et al.* 2022).

6.2 Inference Implementation

Network inference tools, whether correlation or logic based, perform complex and computationally intensive processes. The potential users of these tools are microbial ecologists and other ecologists wanting to explore the interactions in their metabarcoding datasets. These ecologists do not necessarily have a wide knowledge of computer science. Consequently, they require network reconstruction tools following understandable and robust inference processes, but also implementations that are easy to use and able to offer results on reasonable time scales, using readily-available desktop computers. For the specific case of biomonitoring, simplicity of use may not be important where the inference is included in automated pipelines. However, the large amount of metabarcoding data collected for biomonitoring also requires efficient and fast network reconstruction tools. Thus, any newly developed tool should take into account these needs. In Chapter V, I detail an implementation of the abduction based network inference, called InfIntE, in the commonly used R language (R Core Team 2022).

6.3 Interaction network validation

There is no 'golden standard' dataset or methodology to test interaction network inference tools (Röttjers & Faust 2018). The lack of a metabarcoding dataset for which a significant number of interactions are known and understood, and for which the accuracy of the network reconstruction can be evaluated as function of the interactions detected, means that other strategies are necessary to evaluate any method of inference. Of the many different tools that have been developed to reconstruct interaction networks (see section 4), each publication followed its own unique strategy to demonstrate performance. Typically, validation is carried out using computer-generated datasets, where interactions are simulated with specified general properties, in order to compute test networks that look like those that might be obtained from real metabarcoding data (Friedman & Alm 2012; Chiquet *et al.* 2019). Computer-generated data simulates the abundance of interacting OTUs over a set of samples. It has the advantage that any interaction that appears in a test network is one that was specified at simulation; the test interactions are known and can be treated as true values for the test. As consequence, the accuracy of the network reconstruction can be automatically computed and compared with the performance of other tools (Berry & Widder 2014). The problem with this approach is that there is still considerable debate over how to simulate network data. Each mechanism of interaction, which produces an effect on the abundance of the OTUs, is quite distinct and there is no consensus on how to simulate them (Weiss *et al.* 2016). In contrast, we believe that real metabarcoding

data encompasses abundance changes that are the result of real interactions, but with the cost that we do not know which interactions are at play. This means that either the use of simulated data or real data for the demonstration of performance comes with its own specific set of problems. We propose to follow an indirect strategy that uses an ensemble of different methods to attempt to build a complete and convincing picture of the perfromance of the A/ILP method.

The basis for this ensemble method is that we posit that two networks, reconstructed from samples sharing the same conditions, should be more similar than two networks reconstructed from samples with different conditions. Comparisons of different measures of the network structure might then be used to estimate network reconstruction performance. The main issue of this approach, and where care must be taken, is in the choice of measures to show differences between networks (Poisot *et al.* 2012). It is also important to bear in mind that similar network structures, for similar groups of samples, does not guarantee that the edges represented in the network are indicative of true interactions. A second method in the ensemble approach is to validate the network inference tools by reconstructing networks from real metabarcoding data and using the literature, databases and domain expert knowledge as a source of information for validation (Lim *et al.* 2016; Li *et al.* 2016). As I have stated in section 3, the OTU and external information in the literature, databases and the knowledge of experts is related via the taxonomic assignment of the OTUs. It is therefore not possible to obtain validation information for OTUs that have no or incomplete taxonomic assignment (e.g. taxonomy only assigned to genus level). This issue also applies to information such as microbial traits, the majority of which have yet to be discovered. This general incompleteness means that literature, databases and domain expert based validation of interactions can only be performed in very concrete cases, and microbial network ecology suffers from significant difficulties of obtaining general accuracy measures.

Over the different chapters of this thesis, my ensemble approach will be used to assess the performance of statistical and logical interaction reconstruction tools. In Chapter II, different network structure and similarity measures will be tested to find the differences between interaction networks inferred statistically from grapevine leaves samples grown under different cropping systems. Computer generated data will used in Chapters III and IV to assess the performance of the A/ILP based interaction inference. In Chapter IV, I will also use bibliographical references from the grapevine pathobiome literature to test the performance of the A/ILP based tool with real metabarcoding data.

7 Network Visualisation

Interaction networks have been be treated as graphical constructs to visualize interaction data, in scientific domains from the study of social media to gene regulation (Min *et al.* 2022; Ko & Brandizzi 2020). Network visualisations all illustrate a set of features, represented as nodes, related or linked by a set of edges each of which show the interactions happening between the nodes. In the networks studied in this thesis, therefore, each OTU has a specific node and any interactions

that are inferred are represented as edges joining the nodes of the OTUs undergoing the interaction. In the case of quantitative microbial interaction networks, we can expand this representation of a network to include edges of thickness reflecting the strength of an interaction, with colours that differentiate interaction types.

The visualisation of the network is a key step in the exploration of the reconstructed results that helps ecologists learn ecological information from the inferred network. In the case of automated biomonitoring by interaction networks, network visualisation could identify those features of the ecosystem and network that are undergoing change to formulate management to mitigate or adapt to the 'unwanted' changes. However, metabarcoding datasets can contain a large number of OTUs with a considerable number of potential interactions, leading to complex interaction networks, which have been described as hairballs (Röttjers & Faust 2018). Yoghourdjian et al. 2020 showed that humans fail to correctly interpret a network with more than 50 nodes, which is a relatively small number for many metabarcoding datasets. It is important to provide ecologists with manageable, graphical representations for analysis. One critical part of every network visualisation is the positioning of the different nodes. Algorithms, such as Fruchterman-Reingold (Fruchterman & Reingold 1991) and Kamada-Kawai (Kamada & Kawai 1989), provide network layouts that avoid the superposition of edges and produce edges with a uniform length that can greatly simplify the visualisation (Salter-Townshend *et al.* 2012). These layouts algorithms have been implemented in numerous software tools that allow users to easily obtain the network image. Igraph (Csardi & Nepusz 2006), for example, is widely used due to its versatility and power. It can be use in different programming environments, including R, Python and C++, and offers multiple layout generation algorithms and plotting features. Igraph offers few options for interaction with the network graphic, limiting direct user interaction to move nodes or to select those edges to visualise. Gephi (Bastian *et al.* 2009) and Cytoscape (Shannon *et al.* 2003) offer more interaction, allowing the user to manage the graphs. These tools also allow user-developed plugins to be installed, extending their feature sets (Saito *et al.* 2012; Kauffman *et al.* 2014). None of the available tools met the specific requirements of network visualisation for my PhD; that of analyzing the networks obtained using either statistical or explainable machine learning reconstruction tools. It was therefore necessary to develop a custom tool for network visualisation and manipulation. This tool is detailed in chapter V, and is based on igraph and the interactive R package Shiny (Chang *et al.* 2021). The tool offers many interactive features to evaluate the networks and interaction types, as a function of compression values. The tool was specifically developed in R and Shiny so that it could be included in an automated biomonitoring pipeline to evaluate reconstructed networks and enhance decision making.

8 A case of study: Grapevine foliar microbial communities

The first evidence of wine production, found in Iran, are around 7500 years old (McGovern *et al.* 1996). Since then, wine making and consumption has gone hand-in-hand with the development of Eurasian civilisation, being produced in all

subsequent societies through history (This *et al.* 2006). Today, wine remains one of the most important agricultural products. The total production of grapes for wine in the European Union was 24.1 million tonnes in 2020 (Directorate-General for Agriculture and Rural Development 2021), occupying 3.2 million hectares of agricultural land. Grapevine (*Vitis vinifera*) is a subject of study of great ecological, cultural and economic importance.

8.1 The foliar microbial communities

Microbial communities are ubiquitous throughout the tissues of grapevine plants. Microorganisms are found associated with many plant parts including the roots, trunk, leaves and grapes (Zarraonaindia & Gilbert 2015). Microorganisms inhabiting the grapevine leaves are one of the key living components of the crop, as this community is a reservoir of the microbes that colonize the grapes, having a significant influence on the quality of wine produced (Pinto *et al.* 2014). Vine leaves play a major role in plant development, being the primary site of photosynthesis and regulating the plant's water potential. The microbiota associated with grapevine leaves is highly diverse and can vary with grapevine variety, geographic location and season (Vionnet *et al.* 2018). Communities are made up of bacteria, fungi and oomycetes (Fort *et al.* 2016), inhabiting both the exterior surfaces of leaves (epiphytes) and interior tissues (endophytes) (Zarraonaindia *et al.* 2015). Many members of the microbial community, and especially the endophytes, support plant responses to both abiotic and biotic stresses (Bettenfeld *et al.* 2021). This leaf microbiota helps the plant to resist abiotic stresses, such as drought and high temperatures, by detecting and signaling the stress and facilitating rapid metabolic responses from the plant (Pacifico *et al.* 2019). The microbes may also produce secondary metabolites able to reduce water loss and enhance the plant development (Lakshmanan *et al.* 2017). Microbial communities also play a major role in the biotic defense of the plant. Many grapevine diseases are caused by microorganisms and the leaves are a main target site of infection. Powdery mildew, black-rot and downy mildew are the three main economically-important diseases affecting grapevine in Europe (Armijo *et al.* 2016; Molitor & Beyer 2014). They are respectively caused by the fungi, *Erysiphe necator* and *Phyllosticta ampelicida*, and the oomycete *Plasmopara viticola*. The microbiome present upon the leaves of vines acts as a first layer of defence against such infections. Evidence suggests that microorganisms inhabiting the leaf can prevent the pathogens from establishing by occupying the physical space required by the pathogen or killing it through secondary metabolite production (Musetti *et al.* 2006). Some foliar microorganisms can also help the plant to detect the pathogen and promote plant immune responses (Hacquard *et al.* 2017).

8.2 Biomonitoring grapevine leaves' microbial communities

Understanding and monitoring microbial communities inhabiting the grapevine leaves could lead to better cropping practices. As noted before, microbial communities are an important part of abiotic stress resistance. Microbial communities change when the plant is under stress (Cambon *et al.* 2022), and it might be possi-

ble to use interaction networks to monitor plant stress because microbes and their interactions yield information on processes that are protective against stress and of stress tolerance. This information would allow a better understanding of vine stress management for agricultural decision making. Understanding how to mitigate and manage stress might be done by identifying and promoting specific microorganisms in the microbiome that modulate the plants water needs and enhances their growth. Microbial interaction biomonitoring gains even more importance in the case of disease management. Microorganisms are both the cause of diseases and the first layer of defense against microbial disease pathogens.The basic premise for the part of this thesis, focussed on real data is that reconstructing the interaction networks of the grapevine foliar microbiome, would lead to the identification of potential biocontrol agents of pathogens or even potential microbial metabolites with anti-pathogen activity (Pauvert *et al.* 2020). This is especially interesting in a context where the use of chemical pesticides is being reduced, due to their side-effects on the environment. This makes the microbial communities inhabiting grapevine leaves an interesting case study ecosystem.

In chapter VII, I describe the pipeline (from sampling to the filtered OTU tables) for obtaining metabarcoding datasets from grapevine leaves. In chapter II, grapevine metabarcoding data is used to reconstruct interaction networks using two existing statistical network inference tools. Then, different metrics, obtained from these networks, are used to evaluate the network inference performance. In chapter IV, the grapevine metabarcoding data is used to test the performance of the A/ILP based network inference. In particular the grapevine pathobiome is reconstructed, detecting and directly classifying potential antagonists of *P. viticola*, the grapevine pathogen causing downy mildew that is a major economic problem, by their type of interaction.

Chapter II

Correlation-based approaches for biomonitoring using DNA

Microbial networks inferred from environmental DNA data for biomonitoring ecosystem change: Strengths and pitfalls

Didac Barroso-Bergadà, Charlie Pauvert, Jessica Vallance, Laurent Delière, David A. Bohan, Marc Buée, Corinne Vacher

Abstract:
Environmental DNA contains information on the species interaction networks that support ecosystem functions and services. Next-Generation Biomonitoring proposes the use of this data to reconstruct ecological networks in real-time and then compute network-level properties to assess ecosystem change. We investigated the relevance of this proposal by assessing: (1) the replicability of DNA-based networks in the absence of ecosystem change; and, (2) the benefits and shortcomings of community- and network-level properties for monitoring change. We selected crop-associated microbial networks as a case study since they support disease regulation services in agroecosystems and analyzed their response to change in agricultural practice between organic and conventional systems. Using two statisticalmethods of network inference, we showed that network-level properties, especially β-properties, could detect change. Moreover, consensus networks revealed robust signals of interactions between the most abundant species, that differed between agricultural systems. These findings complemented those obtained with community-level data, that showed, in particular, a greater microbial diversity in the organic system. The limitations of network-level data included (i) the very high variability of network replicates within each system; (ii) the low number of network replicates per system, due to the large number of samples needed to build each network; and, (iii) the difficulty in interpreting links of inferred networks. Tools and frameworks developed over the last decade to infer and compare microbial networks are therefore relevant to biomonitoring, provided that the DNA metabarcoding datasets are large enough to build many network replicates and progress is made to increase network replicability and interpretation.

Keywords *Environmental DNA, Metabarcoding, Community ecology, Ecosystem services, Microbial networks, Network inference, Network comparison*

1 Introduction

Interactions among organisms and with their abiotic environment regulate the ecological processes underlying ecosystem services (Mace *et al.* 2012). Ecological interactions among organisms (e.g. predation, mutualism, parasitism) at a single point in space and time are usually represented as a network, with the organisms as nodes and the interactions as links (Pocock *et al.* 2012). Current challenges focus on understanding how and why these networks vary in space and time (Pellissier *et al.* 2018; Pilosof *et al.* 2017), and which network properties should be conserved or enhanced to sustain ecosystem services (Montoya *et al.* 2012; Raimundo *et al.*

2018; Tylianakis *et al.* 2010). Next-Generation Biomonitoring (NGB) proposes the reconstruction, automatically and in real-time, of ecological networks using the Next-Generation Sequencing (NGS) of environmental DNA (eDNA) data, and the analysis of network and community variation in space and time to detecting and explaining changes in ecosystem functions and services (Baird & Hajibabaei 2012; Bohan *et al.* 2017; Derocles *et al.* 2018; Makiola *et al.* 2020). However, before implementing NGB approaches on a large scale, we need more case studies demonstrating the utility of DNA-based networks and the meaning of their derived network metrics (Compson *et al.* 2019). The goal of the present study is to fill this gap. NGB requires the reconstruction of replicated networks of ecological interactions as well as the development of statistical tools for their comparison and analysis. Theoretical frameworks have been developed for the comparison of ecological networks between contrasted environmental conditions or along environmental gradients (Delmas *et al.* 2018; Pellissier *et al.* 2018; Poisot *et al.* 2012; Tylianakis & Morris 2017). By analogy with the α- and β-diversity of ecological communities, these frameworks define α- and β-properties for ecological networks as whole-network metrics (e.g. connectance) and dissimilarities between pairs of networks, respectively (Pellissier *et al.* 2018). Community- and network-level metrics can be used to assess the impact of environmental changes on the number, identity and abundance of the species forming ecological communities, and on the structure, type and strength of their interactions, respectively. They have for instance been used to evaluate the impact of agricultural practices (Morriën *et al.* 2017), that are a key driver of global change (Tilman *et al.* 2002), on species diversity (Tuck *et al.* 2014) and on pest and disease regulation services supported by species interactions (Ma *et al.* 2019; Macfadyen *et al.* 2009; Tylianakis *et al.* 2007).

Networks of interactions among microorganisms appear as suitable tools for NGB for at least three reasons: NGS techniques are the current rule for studying microbial communities (Bálint *et al.* 2016); microorganisms are present in all Earth ecosystems; and, microbial interactions are crucial to ecosystem functioning, human life and well-being (Gilbert & Neufeld 2014; Zhu & Penuelas 2020). Network ecology, which originates from the study of trophic links between macroorganisms (Ings *et al.* 2009), initially ignored interactions with and among smaller organisms (Lafferty *et al.* 2006). But, with increasing evidence of the contribution of microbial interactions to biogeochemical cycles (Falkowski *et al.* 2008), plant diversity and productivity (van der Heijden *et al.* 2008), and disease regulation in soils (Berendsen *et al.* 2012), plants and animals (Brader *et al.* 2017; Hacquard *et al.* 2017; Vayssier-Taussat *et al.* 2014), microbial networks are now considered key to the understanding of ecosystem functioning (de Vries *et al.* 2018; Karimi *et al.* 2017; Wagg *et al.* 2019). However, given that microbial networks inferred from eDNA data only represent hypothesized interactions among microbial species but rather statistical associations among molecular units that only represent putative signals for microbial interactions (Faust & Raes 2012; Röttjers & Faust 2018; Vacher *et al.* 2016), it is crucial to evaluate the relevance of the derived network properties to the assessment of change in ecosystem functioning. In this study, we analyzed the relevance of microbial network properties to NGB, by assessing (1) the replicability of microbial networks inferred from eDNA data in the absence

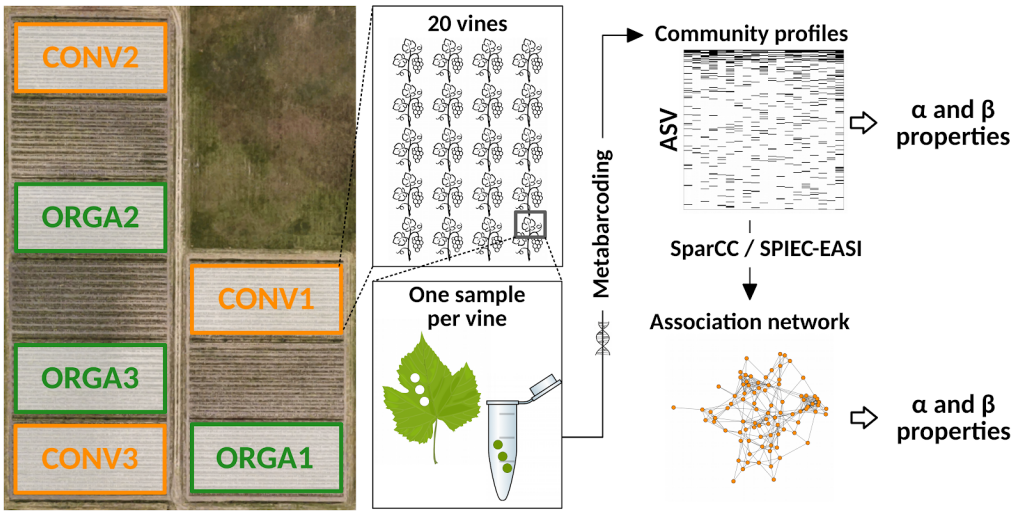


Figure II.1: Experimental design. Foliar fungal communities were characterized in three conventional (CONV) and three organic (ORGA) vineyard plots by a metabarcoding approach. We analyzed 20 foliar samples per plot. For each plot, we obtained 20 community profiles (described in terms of amplicon sequence variants (ASV)) and one association network (inferred either with the SparCC software developed by Friedman & Alm, 2012 or with the SPIEC-EASI software developed by Kurtz et al, 2015). More networks were then obtained by varying network reconstruction parameters (Figure A.3). The effects of cropping system (CONV versus ORGA) on the grapevine foliar microbiota were assessed with both community and network α- and β-properties.

of ecosystem change, and (2) the benefits and shortcomings of community-level and network-level properties for detecting change. We focused on crop-associated microbial networks since they support disease regulation services in agroecosystems (Toju *et al.* 2018), and analyzed their response to change in agricultural practice (conventional vs organic farming). We inferred microbial networks from eDNA sampled from replicated agricultural plots by using two classical methods of network inference, SparCC (Friedman & Alm 2012) and SPIEC-EASI (Kurtz *et al.* 2015). We then computed α- and β-diversity metrics at the community- and network-level to identify the level that best captures change in agricultural practice, by using grapevine and its foliar microorganisms as the case study. These results are then used to discuss those tools and frameworks that are best adapted to NGB approaches.

2 Materials and methods

2.1 Study site and sampling design

Grapevine leaf samples were collected on September 10, 2015, from an experimental vineyard (Figure II.1) located near Bordeaux (INRA, Villenave d'Ornon, France; 44°47'32.2"N 0°34'36.9"W). The experimental vineyard was planted in

Table II.1: List of community-level and network-level α- and β-properties analyzed in the study. The number of independent observations (N) and the size of corresponding dissimilarity matrices (S) are indicated. The last column indicates if the property varied significantly (Yes/No) with change in the cropping system (CS).

Property	Definition	Reference	N	CS
Community α-properties				
Richness	Total number of amplicons sequence variants (ASVs)		N=112	Y
Diversity (Inverse Simpson)	Effective number of ASVs	Simpson, 1949	N=112	Y
Evenness (Pielou's J')	Evenness in ASV relative abundance	Pielou, 1966	N=112	Y
Community β-properties				
Compositional dissimilarity (binary Jaccard)	Dissimilarity of composition due to ASV turnover	Jaccard, 1900	S=6216	Y
Compositional dissimilarity (quantitative Jaccard)	Dissimilarity of composition due to variations in ASV relative abundance	Chao et al, 2006	S=6216	Y
Network α-properties				
Number of links (L)	Total number of links	-	N=6	N
Connectance (C)	Fraction of the total number of possible links actually realized	Coleman & Moré, 1983	N=6	N
Number of connected components (CC)	Number of groups of nodes connected together	Martinez, 1992	N=6	N
Diameter (DIA)	The longest of all the shortest paths between two nodes	Barabási et al, 2000	N=6	N
Mean node degree (DEG)	Mean number of links per node	Martinez, 1992	N=6	N
Proportion of negative links (NLR)	Proportion of links for which the SparCC correlation is negative	Faust et al, 2015	N=6	N
Network β-properties				
Topological dissimilarity (Schieber's D)	Dissimilarity of global and local network structure	Schieber et al, 2017	S=15	N
Association dissimilarity (βWN)	Overall dissimilarity of associations	Poisot et al, 2012	S=15	Y
Association dissimilarity (βOS)	Dissimilarity of associations between shared ASVs	Poisot et al, 2012	S=15	Y
Association dissimilarity (βST)	Dissimilarity of associations due to ASV turnover	Poisot et al, 2012	S=15	N

2011 and was designed to compare three cropping systems: sustainable conventional agriculture (CONV), organic farming (ORGA) and pesticide-free farming (RESI) (Deliere *et al.* 2014). The *Vitis vinifera L.* cultivar Merlot noir grafted onto a 3309 C rootstock was used in both the CONV and ORGA cropping systems. Only the CONV and ORGA systems, that used the same cultivar but different phytosanitary treatments, were compared in the present study to avoid multiplying the sources of variation between systems. RESI used a resistant cultivar, which has 2 Quantitative Trait Loci of partial resistance to downy mildew and total resistance to powdery mildew. The experiment had a randomized block design (Schielzeth & Nakagawa 2013) consisting of three blocks, each composed of three plots, one for each of the cropping systems tested. Each plot covered an area of 2100 m^2 and was composed of 20 rows of 68 vines each, with 1.60 m between rows and 0.95 m between vines in a single row. CONV plots were managed according to the general principles of integrated pest management (IPM), as listed in Appendix III of the 2009/128/EC Directive (European Commission 2009). ORGA plots were managed according to European Council Regulation (EC) No 834/2007 (Council of the European Union 2007). ORGA plots were treated with copper and sulfur-based products, whereas additional phytosanitary products were allowed in CONV plots (Table A.1). The cropping systems differed in terms of the types of pesticides applied and the timing of applications, but not in terms of doses (Table A.1). All products and active ingredients were applied between the end of April and mid-August of 2015. Grapes were harvested on September 10, 2015. The disease incidence and severity at harvest were higher in CONV plots than in ORGA plots for both powdery mildew (caused by the fungal pathogen *Erysiphe necator*) and black rot (caused by the fungal pathogen *Guignardia bidwellii*). Downy mildew symptoms (caused by the oomycete pathogen *Plasmopara viticola*) did not differ significantly between the cropping systems (Table A.2). Grapevine leaves were collected in the two hours prior to grape harvest, from 20 vines per plot in the CONV and ORGA plots (Figure II.1). We attempted to avoid edge effects by selecting the 20 vines from the center of each plot. The third leaf above the grapes was collected from each vine, placed in an individual bag and immediately transported to the laboratory. In total, 120 leaves, corresponding to 1 leaf $\times$ 20 vines $\times$ 3 plots $\times$ 2 cropping systems, were collected. Leaves were processed on the day of collection, with sterilized tools in the sterile field of a MICROBIO electric burner (MSEI, France). Three contiguous discs of 6 mm diameter were cut from the center of each leaf, approximately 2 cm from the midrib. They were placed in the well of a sterile DNA extraction plate. The leaf disks were then freeze-dried overnight (Alpha 1-4 DA Plus, Bioblock Scientific).

2.2 DNA extraction and sequencing

Leaf disks (Figure II.1) were ground with a single-glass ball mill (TissueLyser II, Qiagen) and DNA was then extracted with a CTAB chloroform/isoamyl alcohol (24:1) protocol. A dozen "empty" wells (i.e. containing nothing but extraction reagents) were included on each plate as negative control samples for DNA extraction. Three of these negative control samples were randomly selected and pooled before sequencing. Three replicates of a fungal mock community, each

consisting of an equimolar pool of DNA from 189 pure fungal strains, were also included as positive control samples (Pauvert *et al.* 2019). The nuclear ribosomal internal transcribed spacer (ITS) region, which is considered to be the universal barcode region for fungi (Schoch *et al.* 2012), was then amplified with the ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3', (Gardes & Bruns 1993)) and ITS2 (5'-GCTGCGTTCTTCATCGATGC-3', (White *et al.* 1990)) primer pair, which targets the ITS1 region. PCR was performed in an Eppendorf thermocycler (Eppendorf), with a reaction mixture (25 µl final volume) consisting of 0.04 U Taq polymerase (SilverStar DNA polymerase, Eurogentec), 1X buffer, 2 mM MgCl2, 200 µM of each dNTP, 0.2 µM of each primer, 1 ng.µl-1 bovine serum albumin (New England BioLabs) and 2 µl DNA template. A pseudo-nested PCR protocol was used, with the following cycling parameters: enzyme activation at 95°C for 2 min; 20 (1st PCR with regular primers; Table A.3) and then 15 (2nd nested PCR with pre-tagged primers; Table A.3) cycles of denaturation at 95°C for 30 s, 53°C for 30 s, 72°C for 45 s; and a final extension phase at 72°C for 10 min. "Empty" wells (i.e. containing nothing but PCR reagents) were included on each plate as a negative control for PCR. Three negative control samples were randomly selected and pooled before sequencing. In addition, the PCR product of one sample per plot was split in two, with each half of the sample sequenced independently to serve as technical replicates for sequencing, hence forming six pairs of technical replicates (one per plot). We checked the quality of all the PCR products by electrophoresis in 2% agarose gels. A total of 123 samples were sent to sequencing, corresponding to 112 well-amplified leaf samples, 6 technical replicates, 1 pooled negative extraction control, 1 pooled negative PCR control and 3 mock community replicates. PCR products were purified (CleanPCR, MokaScience), multiplex identifiers and sequencing adapters were added, and library sequencing on an Illumina MiSeq platform (v3 chemistry, 2×250 bp) and sequence demultiplexing (with exact index search) were performed at the Get-PlaGe sequencing facility (Toulouse, France).

2.3 Bioinformatic analysis

Based on the mock community included in the sequencing run, we found that analyzing single forward (R1) sequences with DADA2 (Callahan *et al.* 2016) was a good option for fungal community characterization (Pauvert *et al.* 2019). This pipeline fully exploits the resolution of molecular barcodes (Callahan *et al.* 2016), which is a desired feature in microbial network inference. Indeed, the taxonomic resolution of the nodes should be fine enough to discern the variation in ecological interactions between microbial strains (Röttjers & Faust 2018). Using DADA2 v1.6, we retained only R1 reads with less than one expected error (based on quality scores; (Edgar & Flyvbjerg 2015)) that were longer than 100 bp, and we then inferred amplicon sequence variants (ASV) for each sample. Chimeric sequences were identified by the consensus method of the removeBimeras function. Taxonomic assignments were performed with RDP classifier (Wang *et al.* 2007), implemented in DADA2 and trained with the UNITE database v. 7.2 (UNITE Community 2017). Only ASVs assigned to a fungal phylum were retained. The ASV table was then filtered as described by Galan et al. (2016) with a custom script (github.com/cpauvert/1ba6a97b01ea6cde4398a8d531fa62f9) that

removed ASVs from all samples for which the number of sequences was below the cross-contamination threshold, defined as their maximum number in negative control samples. Finally, we checked the compositional similarity of the six pairs of technical replicates, in terms of both ASV occurrence and relative abundance (Figure A.1), and we removed for each pair of technical replicates the replicate with the lowest number of sequences. We also removed the controls. Therefore, the final ASV table contained 1116 ASVs, 112 leaf samples and 4,760,068 high-quality sequences.

2.4 Statistical analyses

Statistical analyses were performed with R software v3.4.1 (R Core Team 2022), with the packages lme4 v1.1-19 (Bates *et al.* 2015), vegan v2.5-5 (Oksanen *et al.* 2022), permute v0.9-5 (Simpson 2022), phyloseq v1.24.2 (McMurdie & Holmes 2013) including the DESeq2 extension v1.20.0 (Love *et al.* 2014), NST v2.0.4 (Ning *et al.* 2019), and igraph v1.2.4.1 (Csardi & Nepusz 2006). Data were manipulated and plots were created with reshape2 v1.4.3, plyr v1.8.4 and ggplot2 v3.2.0 (Wickham 2016), cowplot v0.9.4 (Wilke 2020), ggraph v1.0.2 (Pedersen 2022) and VennDiagram v1.6.20 (Chen 2022).

Effect of cropping system on community α-diversity

Three community α-diversity properties were computed for each sample: richness, diversity and evenness of fungal communities (Table II.1). Generalized linear mixed models (GLMMs) were then used to test the effect of the cropping system on these properties. The models included the cropping system as a fixed treatment effect and the sampling depth (defined as the total number of raw sequences per sample) as an offset Bálint *et al.* 2015; McMurdie & Holmes 2014). For every property, we compared the likelihood of a full model, including the block and its interaction with the cropping system as random effects and a simplified model, including only the block factor as a random effect. Community richness was defined as the number of ASVs per sample. We used a logarithmic link function to model these count data, assuming a negative binomial distribution to deal with overdispersion (Zuur *et al.* 2009). Community diversity was measured with the Inverse Simpson index (Simpson 1949) and modeled with a Gaussian distribution and the logarithmic link function. Evenness was estimated with Pielou's index (Pielou 1966) and modeled with a Gaussian distribution and the logarithmic link function. The offset was transformed according to the link function. The significance of the fixed treatment effect was finally assessed with the Wald χ^2 test (Bolker *et al.* 2009). Moreover, to investigate whether foliar fungal pathogens of grapevine were responsible for variations in community α-diversity properties, we fitted the models by including the relative abundance of sequences assigned to the *Erysiphe* genus (which includes *Erysiphe necator*, the causal agent of powdery mildew; Armijo *et al.* 2016) and the *Guignardia* genus (which includes *Guignardia bidwelli*, the causal agent of black rot) as fixed additive effects.

Effect of cropping system on community β-diversity

Two community β-diversity properties were calculated for each pair of samples: the quantitative Jaccard dissimilarity and the binary Jaccard dissimilarity (Table II.1). Permutational analyses of variance (PERMANOVAs; Anderson 2011) were then used to evaluate the effect of the cropping system on these compositional dissimilarities. The models included cropping system, sampling depth (log-transformed), block and their interaction as fixed effects. ASVs differing in abundance between cropping systems were identified with DESeq2 (Love *et al.* 2014), by calculating the likelihood ratio between a full model including block and cropping system as fixed effects and a simplified model including only the block factor. The estimated fold-changes in abundance were considered significant if the p-value was below 0.05 after Benjamini and Hochberg adjustment. Moreover, to understand better the processes shaping community structure, the relative contribution of deterministic and stochastic processes in community assembly was assessed by following the framework defined by Ning et al. (2019). This method provides statistics for each sample, named the Normalized Stochasticity Ratio (NST), that ranges from 0 to 100, where 0 means a completely deterministic assembly process and 100 a completely stochastic assembly process. NST was calculated using the tNST function with the binary and quantitative Jaccard dissimilarity indices, the FE null model, and other parameters by default values. We used the FE null model (SIM2 in Gotelli 2000) because it is the most appropriate for comparing standardized samples that have been collected in areas of homogeneous habitat, such as vineyards. This null model reshuffles ASV occurrences among samples by considering that all samples are equally probable. NST values were calculated for each cropping system and then compared using permutational analysis of variance with the nst.panova function.

Network inference

Fungal association networks were inferred at plot level (Figure II.1) with two widely-used methods of microbial network inference: the SparCC algorithm (Friedman & Alm 2012) implemented in FastSpar (Watts *et al.* 2018) with default SparCC values; and, the SPIEC-EASI method (Kurtz *et al.* 2015) using the MB procedure of edge selection. Both methods try to deal with the compositional nature of metabarcoding data. In a metabarcoding dataset, the total number of sequences per sample is arbitrary, imposed by the sequencer. Sequence counts contain only relative abundance information for species. Methods that do not take this feature into account can result in the identification of artifactual associations (Gloor *et al.* 2017). Both SparCC (Friedman & Alm 2012) and SPIEC-EASI (Kurtz *et al.* 2015) attempt to overcome this bias using log ratios of counts. For each method of network inference, ten networks per plot were constructed by varying the percentage P of the ASVs included in the network (with P ranging from 10% to 100% of the most abundant ASVs in the plot). We varied P because we expected that it would influence the replicability of the networks. We expected, in particular, the networks built from only the most abundant ASVs to be more replicable. For the same reason, networks were also inferred after aggregating ASVs at the genus level and removing ASVs that were not taxonomically assigned at

this level. In all cases, the inferred microbial networks had ASVs as nodes and a positive or negative link between ASVs in cases of significant associations between abundance.

Effect of cropping system on network α-properties

Six network α-properties were calculated for each inferred network: number of links, network density, number of connected components, diameter of the largest component, mean node degree and proportion of negative links (Table II.1). The effect of the cropping system on these properties was investigated by performing Wilcoxon rank-sum tests for all values of P. The Benjamini-Hochberg procedure was used to correct p-values for multiple testing.

Effect of cropping system on network β-properties

Four network β-properties were calculated for each pair of inferred networks (Table II.1). The topological distance between networks was calculated with the D index defined by Schieber et al. (2017). Schieber's D, when applied to binary networks (i.e. with unweighted links) captures global and local structural dissimilarities between networks, by comparing node connectivity patterns across scales. The dissimilarity of associations between networks, βWN, according to the framework described by Poisot et al. (2012), was then calculated for all pairs of networks with the binary Jaccard dissimilarity index. βWN was then partitioned into two components (Poisot *et al.* 2012): the dissimilarity of associations between ASVs common to both networks (βOS) and the dissimilarity of associations due to species turnover (βST). In contrast to the Shieber's D index that evaluates how nodes are connected to neighboring nodes and to more distant nodes, these three metrics compare lists of pairwise associations between nodes. PERMANOVA was used to evaluate the effect of the cropping system on the topological distance between networks (D) and the dissimilarity of associations between networks (βWN, βST and βOS). The models included cropping system, the percentage of ASV, P, and their interactions as fixed effects. The permutations (n=999) were constrained within blocks. Finally, for each network inference method and every value of P, consensus networks containing only the shared associations between the three network replicates within a cropping system were built to identify robust associations that could indicate ecological interactions between fungal strains. The number of shared associations between the three network replicates were compared to those obtained between three random networks simulated with the same nodes and the same number of links. The significance of shared associations was evaluated with a pseudo p-value, estimated from 999 simulations and defined as the probability that the three random networks shared more associations than the three inferred networks (Morlon *et al.* 2014).

3 Results

Among the 15 community- and network-level properties computed (Table II.1), 7 indicateddifferences between the organic (ORGA) and the conventional (CONV)

system.

3.1 All community α-properties detected system change

All three community α-diversity properties - richness, diversity and evenness (Table II.1) - were significantly higher in ORGA than CONV plots (Figure II.2 A-C and Table A.4). Community richness, for example, equaled on average 39.69 fungal ASVs per sample in ORGA plots vs 36.40 in CONV plots, with each sample representing 0.85 cm^2 of a single leaf tissue. Including the interaction between the cropping system and the block did not significantly increase the likelihood reported by the GLMM models, indicating that changes in community α-diversity properties due to the cropping system were consistent across blocks. In contrast to our expectations, none of the community α-diversity properties was influenced by pathogen relative abundance (Table A.5). Pathogen abundance within each sample was estimated as the proportion of sequences assigned to the *Erysiphe* genus and ranged between 0% and 36.34%, with an average of 1.12% per sample. No ASV was assigned to the *Guignardia* genus and this variable was therefore not included in the models.

3.2 All community β-properties detected system change

The two community β-diversity properties analyzed in this study - the quantitative and binary Jaccard indices (Table II.1) - detected significant differences in community composition between systems (Table II.2). The cropping system was a major driver of both ASV relative abundance (Figure II.2D) and ASV presence-absence (Figure II.2E), as indicated by the quantitative and binary Jaccard indices, respectively. It explained 7.6% of the variance in ASV relative abundance and 4.5% of the variance in ASV presence-absence (Table II.2). The block effect was also significant, indicating that there were spatial variations in community composition at the scale of the experiment. The block explained 4.3% of the variance in ASV relative abundance, and 2.6% of the variance in ASV presence-absence (Table II.2). There were also large differences in composition among samples within a plot, as indicated by the high percentage of unexplained variance (78.2% for the quantitative Jaccard index and 85.7% for the binary Jaccard index) (Table II.2). In line with these results, we found that the stochasticity in ASV presence-absence was very high in both the ORGA and CONV systems (NST=78.4% and 94.8%, respectively). Nevertheless, it decreased markedly when the relative abundance of ASVs (NST=29.3% and 33.6%, respectively) was taken into account (Table A.6), probably because the ASV, assigned to *Aureobasidium sp.* (Table II.3) was the most abundant, represented more than half of the total number of sequences and was highly abundant in all samples. Stochasticity in ASV presence-absence was significantly higher in CONV plots (Table A.6). A similar trend, although non significant, was observed for ASV relative abundance, suggesting that communities in ORGA plots were more stable, in addition to being richer (Figure II.2A). Overall, the foliar fungal communities were dominated by Ascomycota in both ORGA (87.2% of sequences) and CONV (96.8%) plots. About one-fourth of ASVs (249 over 1116) were shared between cropping systems. These shared ASVs were the

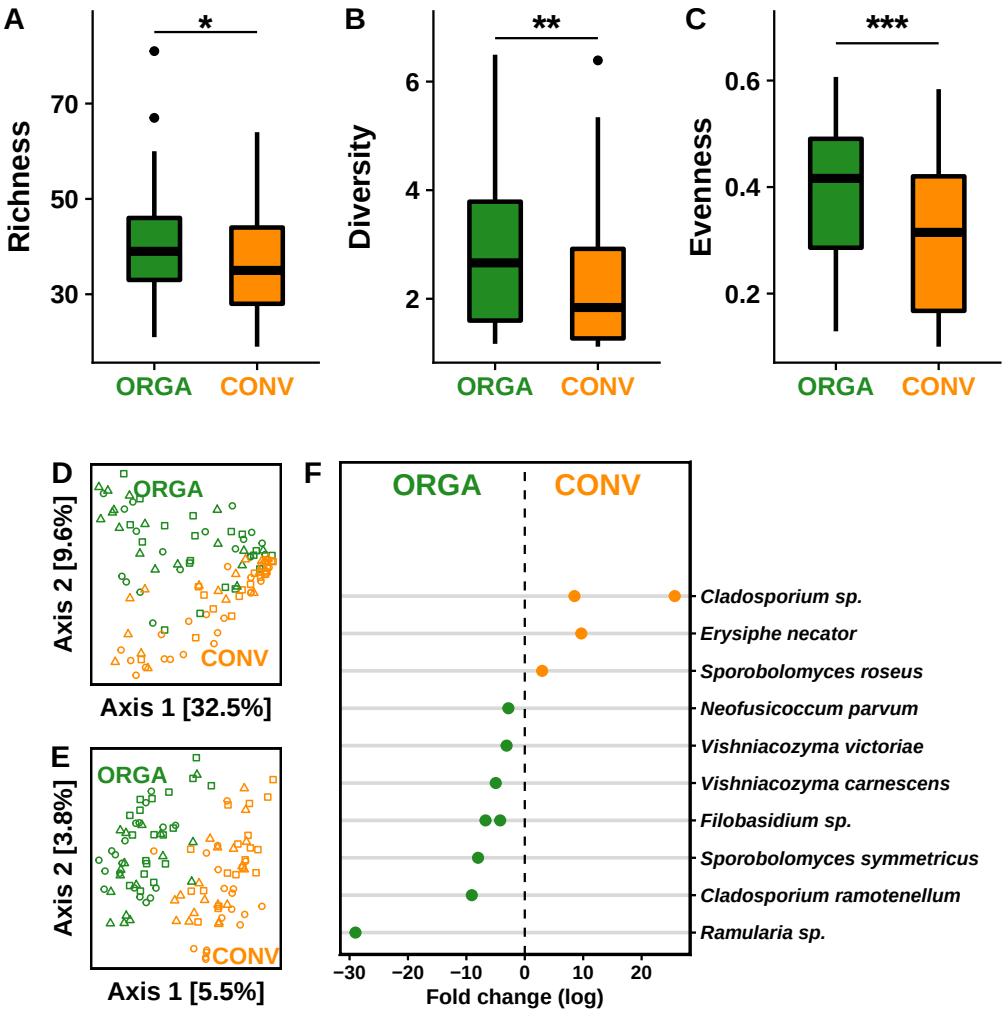


Figure II.2: Effect of cropping system —conventional (CONV) versus organic (ORGA) — on the α-diversity and β-diversity metrics of grapevine foliar fungal communities. **A** Community richness, defined as the number of ASVs. **B** Community diversity, measured with the inverse Simpson index. **C** Community evenness, measured with Pielou's index. Differences in α-diversity metrics between cropping systems were significant (Table S4; * $p<0.05$; **$p<0.01$; ***$p<0.001$). **D** Principal coordinate analysis (PCoA) was used to represent dissimilarities in composition between samples, as assessed with the quantitative and **E** binary Jaccard indices. The effect of the cropping system on both β-diversity metrics was significant, as a single effect for the quantitative Jaccard index and in interaction with block for the binary index (Table II.2). Green circles, squares and triangles correspond to samples collected in the ORGA1, ORGA2 and ORGA3 plots, respectively. Orange circles, squares and triangles correspond to the CONV1, CONV2 and CONV3 plots, respectively (Figure II.1). **F** Log-transformed ratio of ASV relative abundance in CONV plots over that in ORGA plots, for 14 ASVs identified as differentially abundant between cropping systems by DESeq2 analysis followed by Benjamini-Hochberg adjustment (Love *et al.* 2014).

most abundant, representing 98.97% of the total number of sequences. Fourteen ASVs differed significantly in abundance between the cropping systems according to differential abundance analysis performed with DeSeq2 (Figure II.2F). For instance, the causal agent of grapevine powdery mildew, *Erysiphe necator*, which was among the 10 most abundant fungal species, was significantly more abundant in CONV than in ORGA plots (Figure II.2F), according to both the visual records of disease symptoms (Table A.2) and metabarcoding data (2% versus less than 0.1%; Table II.3). The highest abundance of this major grapevine pathogen in samples of CONV plots was however not responsible for their lower α-diversity (Figure II.2 A-C and Table A.5). Differential abundance analysis also revealed that three other ASVs were significantly more abundant in CONV plots, whereas 10 other ASVs, including several yeast species (from the genera *Vishniacozyma*, *Sporobolomyces* and *Filobasidium*), were significantly more abundant in ORGA plots (Figure II.2F).

3.3 None of the network α-properties detected system change

For each method of network inference, we obtained sixty fungal association networks (SparCC: Figure II.3A; SPIEC-EASI: Figure A.2A), each corresponding to one of the six vineyard plots (Figure II.1) and one of the ten values of the percentage P of most abundant ASVs included in the network. Whatever the network inference method, none of the six network α-properties (Table II.1) differed between cropping systems (Table A.7 and A.9), but all were significantly correlated with P (Table A.8 and A.10). Four network α-properties had consistent variations with P between the two methods: the total number of links (L), the number of connected components (CC), the network connectance (C) and the average node degree (DEG) (Tables A.8 and A.10). Increasing the number of ASVs included in the network increased the total number of links, linked the connected components (hence reducing CC) and increased the average node degree. This consistent increase in average node degree with P, however, masked some differences between methods. With SPIEC-EASI, node degree increased more in abundant ASVs, yielding a significant, positive relationship between ASV relative abundance and node degree at P=100% (Figure A.3). This was not the case in SparCC (Figure A.3). Despite this difference, the network connectance decreased with both methods of network inference, consistent with their sparsity assumption (Friedman & Alm 2012; Kurtz *et al.* 2015).

3.4 Half of the network β-properties detect system change

Only two network β-properties, of the four computed (Table II.1), differed significantly between cropping systems whatever the network inference method. As with the network α-properties , the topological dissimilarity between networks, measured with the Shieber's D index (Schieber *et al.* 2017), did not differ between cropping systems but was influenced by P, irrespective of the network inference method (Table II.4 and Table A.11). These results are consistent with the results obtained for node degree and network connectance, which are components of the D index and also vary with P but do not differ between cropping systems (Tables

Table II.2: Effect of cropping system — conventional versus organic — on the β-diversity metrics of grapevine foliar fungal communities. Dissimilarities in community composition between samples were assessed with both the quantitative and binary Jaccard indices. The effects of sequencing depth (SD, log-transformed), cropping system (CS) and block (B) on compositional dissimilarities between communities were evaluated using permutational analysis of variance (PERMANOVA), with the number of permutations set to 999.

Dissimilarity index	**PERMANOVA**				
Quantitative Jaccard	**Variable**	**Df**	**F.Model**	**R2**	**Pr(>F)**
	log(Sampling_Depth) (SD)	1	4.6601	0.0365	**0.002**
	Cropping_System (CS)	1	9.7767	0.0765	**0.001**
	Block (B)	2	2.7462	0.043	**0.001**
	SD x CS	1	1.1651	0.0091	0.278
	SD x B	2	1.0514	0.0165	0.328
	CS x B	2	1.0999	0.0172	0.308
	SD x CS x B	2	1.1698	0.0183	0.246
	Residuals	100		0.7829	
	Total	111		1	
Binary Jaccard	**Variable**	**Df**	**F.Model**	**R2**	**Pr(>F)**
	log(Sampling_Depth) (SD)	1	1.0606	0.0091	0.274
	Cropping_System (CS)	1	5.2676	0.0452	**0.001**
	Block (B)	2	1.5403	0.0264	**0.001**
	SD x CS	1	1.0279	0.0088	0.37
	SD x B	2	0.9425	0.0162	0.754
	CS x B	2	1.1959	0.0205	**0.022**
	SD x CS x Bk	2	0.97	0.0166	0.642
	Residuals	100		0.8572	
	Total	111		1	

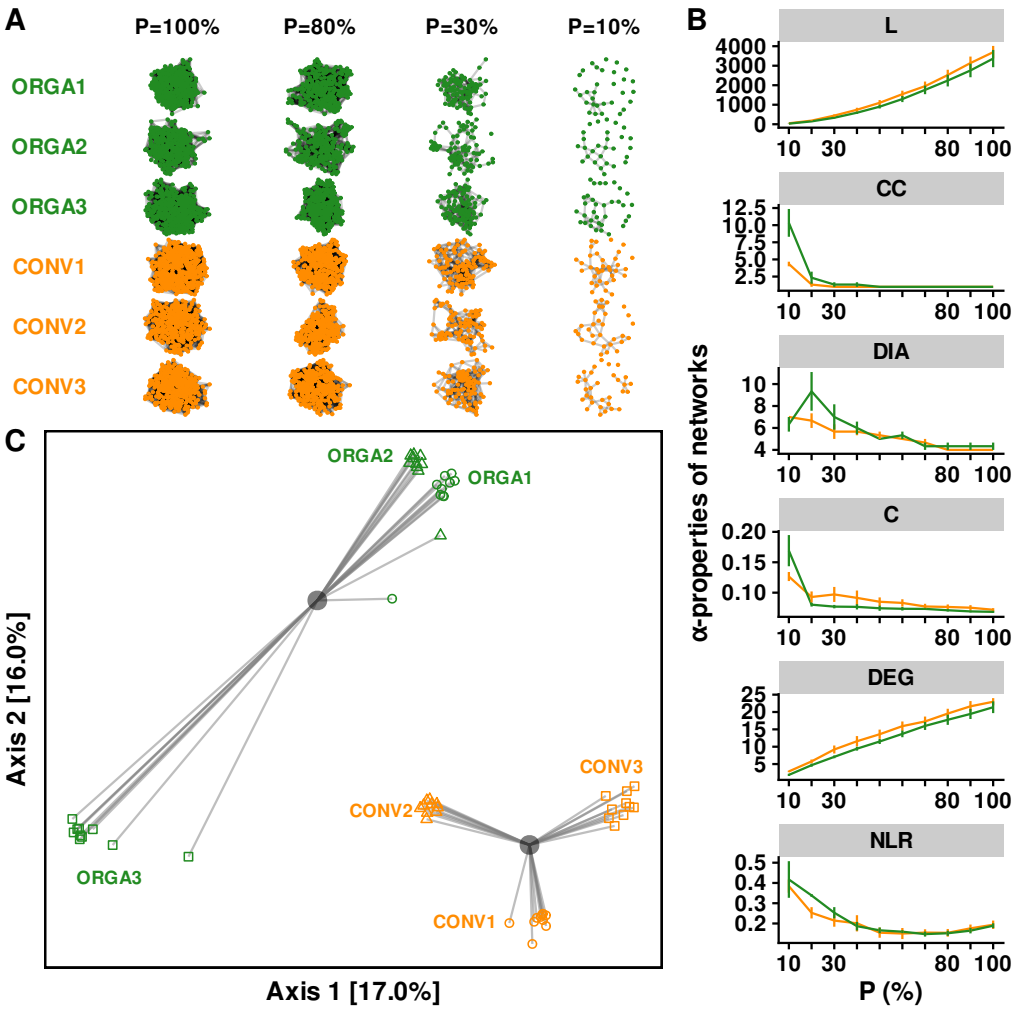


Figure II.3: Effect of cropping system — conventional (CONV) versus organic (ORGA) — on the α-properties and β-properties of grapevine foliar fungal networks. **A** Association networks inferred from fungal metabarcoding data with SparCC (Friedman & Alm 2012). A total of 60 networks were inferred, corresponding to 2 cropping systems × 3 replicates (blocks) × 10 P values, with P the percentage of most abundant ASVs used for network inference. Only four values of P are shown on the Figure. **B** Variations in network α-properties. The following properties (Table II.1) were calculated for each network: the number of links (L) and connected components (CC), the network diameter (DIA) and connectance (C) and the mean degree (DEG) and negative link ratio (NLR). The percentage P of ASVs used for network reconstruction had a significant influence on all properties (Table A.8), whereas the cropping system did not (Table A.7). **C** Principal coordinate analysis (PCoA) represents dissimilarities between networks, measured with the βOS index (Poisot *et al.* 2012) calculated with the binary Jaccard index. βOS measures the dissimilarity between two networks in terms of the presence-absence of associations between shared ASVs. The centroids for each cropping system are represented by gray circles. The effect of the cropping system on βOS was significant (Table II.4). Networks were inferred with SparCC (Friedman & Alm 2012).

Table II.3: Most abundant amplicon sequence variants (ASVs) in grapevine foliar fungal communities according to the cropping system. The relative abundances (RA, in %) and ranks of ASVs were calculated for all leaf samples (TOTAL; n = 112) and for samples collected from organic (ORGA; n = 55) and conventional plots (CONV; n = 57).

ASV taxonomic assignment	TOTAL		ORGA		CONV	
	Rank	RA	Rank	RA	Rank	RA
Aureobasidium sp.	1	61.4	1	55.8	1	66.7
Cladosporium delicatulum	2	6.3	4	6.9	2	5.8
Filobasidium sp.	3	5.1	2	9.7	9	0.7
Alternaria sp.	4	4.4	5	3.9	4	5
Epicoccum nigrum	5	4.1	7	2.7	3	5.4
Cladosporium ramotenellum	6	3.5	3	7	46	<0.1
Mycosphaerella tassiana	7	3.3	8	1.8	5	4.8
Didymella sp.	8	1.4	6	2.7	33	0.1
Erysiphe necator	9	1.1	38	<0.1	6	2
Vishniacozyma victoriae	10	0.9	9	1.6	17	0.3

A.7 to A.10). By contrast, cropping system had a significant effect on the overall dissimilarity of associations (βWN) and the dissimilarity of associations between shared ASVs (βOS) for both SparCC networks (Table II.4 and Figure II.3C) and SPIEC-EASI networks (Table A.11 and Figure A.2C). Cropping system also had a significant effect on the dissimilarity of associations due to ASV turnover (βST), but only in SparCC networks and only in interaction with P (Table II.4 and Table A.11). These findings suggest that network variation between cropping systems are due to the turnover in associations (captured by βOS), rather than the turnover in ASVs (captured by βST), and show that the network β-properties defined in the theoretical ecology framework by Poisot et al. (2012) can be used to detect differences between cropping systems.

3.5 Network replicates within each system were highly variable but shared links

Network replicates varied considerably within a cropping system, whatever the network inference method (Figures II.4 and A.4). When all ASVs were used for network construction with SparCC (P=100%), only 3 associations were common to all three network replicates of the ORGA system, although 80 ASVs were shared between the three network replicates (Figure II.4). Only 5 were common to all three network replicates of the CONV system, although 81 ASVs were shared between the three network replicates (Figure II.4). Similar results were obtained with SPIEC-EASI, with 1 and 5 shared associations, respectively (Figure A.4).

Table II.4: Effect of cropping system — conventional versus organic — on the β-properties of grapevine foliar fungal networks inferred with SparCC. The D index quantifies the topological dissimilarity between networks (Schieber *et al.* 2017) whereas the other three metrics (βWN, βOS and βST), which were calculated with the binary Jaccard index, quantify differences in associations between networks (Poisot *et al.* 2012). The effect of the percentage P of the most abundant ASVs used for network inference, and the effect of cropping system (CS) on the dissimilarities between networks were evaluated in permutational analysis of variance (PERMANOVA). The number of permutations was set to 999 and permutations were constrained by block.

Dissimilarity index	**PERMANOVA**				
Topological dissimilarity (Schieber's D)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	57.75	0.5	**<0.01**
	Cropping_System (CS)	1	1.72	0.01	0.19
	P × CS	1	0.65	0.01	0.51
	Residuals	56		0.48	
	Total	59		1	
Overall dissimilarity of associations (βWN)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	2.41	0.04	**<0.01**
	Cropping_System (CS)	1	5	0.08	**<0.01**
	P × CS	1	2.21	0.03	**<0.01**
	Residuals	56		0.85	
	Total	59		1	
Dissimilarity of associations between shared ASVs (βOS)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	0.53	0.01	0.61
	Cropping_System (CS)	1	11.07	0.16	**<0.01**
	P × CS	1	0.56	0.01	0.57
	Residuals	56		0.798	
	Total	59		1	
Dissimilarity of associations due to ASV turnover (βST)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	1.3	0.02	**<0.01**
	Cropping_System (CS)	1	0.27	<0.01	1
	P × CS	1	1.3	0.02	**<0.01**
	Residuals	56		0.95	
	Total	59		1	

High variability of network replicates within a cropping system was observed for all values of P and was not reduced by the aggregation of ASVs at the genus level nor by the consideration of only the most abundant ASVs, in contrast with our expectation.The networks inferred from only the most abundant ASV or the most abundant genera (P=10% or 20%) barely shared any associations (Table II.5). These low numbers of shared associations between the three networks replicates were, however, generally significantly higher than expected from three random networks having the same number of nodes and links (Table II.5). The number of shared associations between the three network replicates ranged between 0 and 7, depending on P, and the network inference method (Table II.5), while the average number of shared associations between the random networks ranged between 0 and 1.2, suggesting that consensus networks within a cropping system (Figure A.5) do contain robust associations but these are few in number. Five of nine consensus associations were also found by both methods of network inference.The SparcCC and SPIEC-EASI consensus networks obtained for P=100% in the ORGA system shared a negative association between the dominant ASV, assigned to *Aureobasidium sp.*, and the third most abundant ASV in the ORGA system, assigned to *Cladosporium ramotenellum* (Figure A.5 and Table II.3), as an example The consensus networks obtained for the CONV system also shared a negative association between the dominant ASV, assigned to *Aureobasidium sp.*, and the third most abundant ASV in the CONV system, assigned to *Epicoccum nigrum* (Figure A.5 and Table II.3). Three positive associations were also shared by the SparCC and SPIEC-EASI consensus networks in the CONV system (Figure A.5). No association was shared between the two cropping systems, whatever the network inference method and despite 44 ASVs being shared (Figures II.4 and A.4), confirming the significant turnover in associations detected with βOS (Table II.4 and Table A.11).

4 Discussion

The functioning of ecosystems, like that of all complex systems, emerges from the interaction links between its components, and cannot be deduced from a simple listing of organisms (Newman *et al.* 2006). The concept of Next-Generation Biomonitoring (NGB) builds on this property of complex systems and proposes the use of networks of species interactions, rather than a simple list of species, to monitor changes in ecosystem functioning. It also proposes that this could be done via the automatic reconstruction of ecological networks from DNA metabarcoding data (Bohan *et al.* 2017). In the present study, we focused on microbial association networks as a tool for ecosystem monitoring because microbial networks are present in all ecosystems, contribute to ecosystem functioning, and many methods exist to reconstruct them from DNA metabarcoding data (Weiss *et al.* 2016; Dohlman & Shen 2019). We assessed the relevance of microbial networks for NGB approaches using two criteria: (1) their replicability in the absence of environmental change; and, (2) their ability to better detect environmental change than properties at the microbial community level. We focused on a major driver of environmental change, agricultural practices (conventional versus organic agriculture). Our results demonstrated that: (1) microbial network replicates were highly

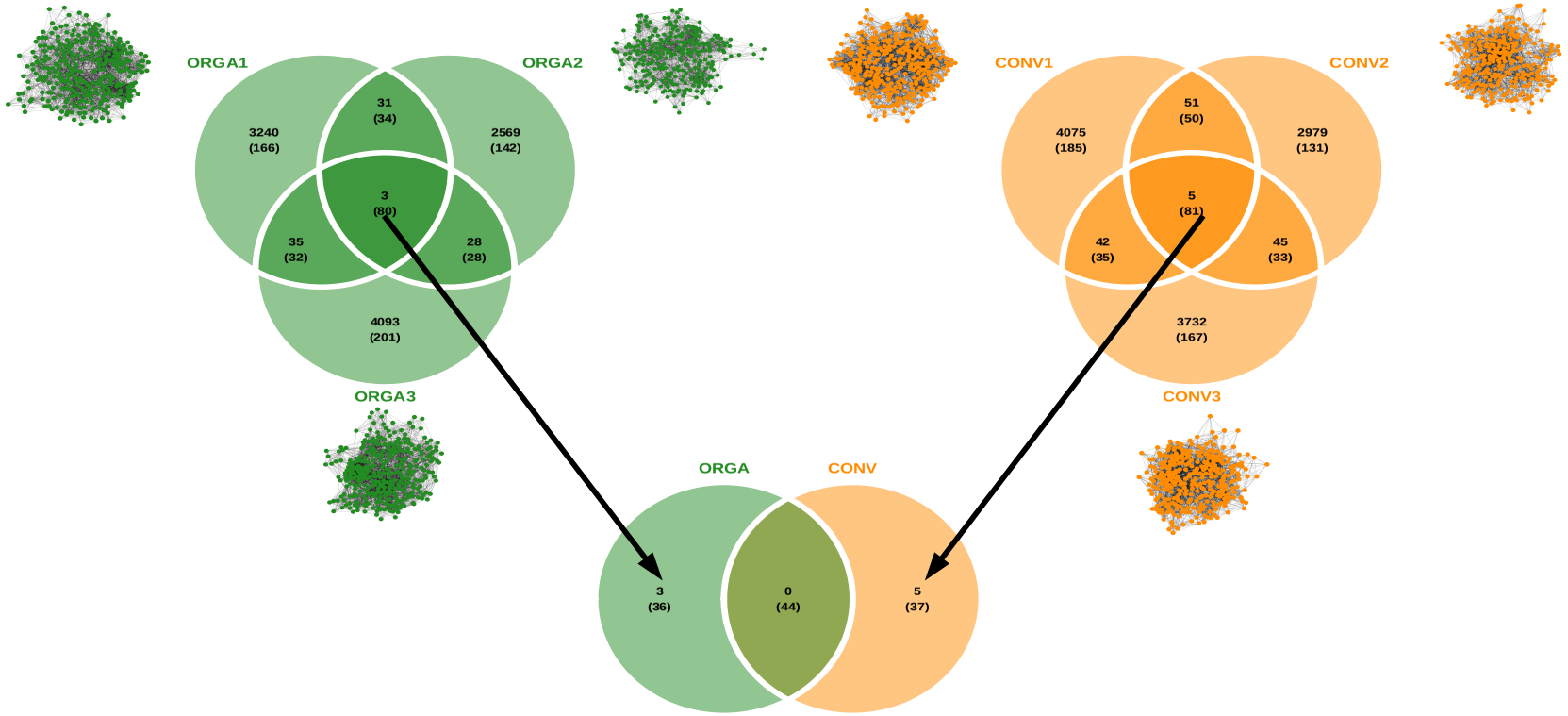


Figure II.4: Venn diagrams showing the number of fungal associations common to network replicates. A Associations common to the three network replicates inferred for the organic cropping system (ORGA1, ORGA2, ORGA3) and **B** the three network replicates inferred for the conventional cropping system (CONV1, CONV2, CONV3), regardless of the sign of the association, in the situation in which all ASVs were used for network construction (P=100%). **C** Associations common to the six networks. Networks were inferred with SparCC (Friedman & Alm 2012). The number of nodes shared by the network replicates is indicated into brackets.

variable within the same set of environmental conditions; and, (2) some network-level metrics, but not all, could detect environmental change. By contrast, all community-level metrics revealed clear-cut changes in the microbial communities in response to environmental change (Table II.1).

The high variability of network replicates within an environmental condition (i.e., in our study, the same cropping system) is the most surprising result of our study. When the whole metabarcoding dataset was used to build the networks, each network replicate was composed of about 160 nodes (fungal ASVs, in our study) and about 3500 links between these nodes (corresponding to co-occurrence or co-exclusion relationships between these fungi). The three network replicates shared half of their nodes but less than 5 links (Figure II.4). Four non-mutually exclusive hypotheses can be put forward to explain this result. First, the variability in microbial associations may reflect real ecological variability. Different assemblages of fungal taxa could play the same role in the ecosystem because of the functional redundancy of the taxa (Louca *et al.* 2016). There would thus be several assemblages, involving different associations of taxa (and thus different networks), adapted to the same cropping system. Second, the relative abundances of fungal taxa, from which the networks are built, could vary within the same environmental condition because of ecological stochasticity. The fungal communities were, like most ecological communities (McIntosh 1962), composed of a small number of ubiquitous species and a large number of rare species whose presence varied greatly, probably because of the large degree of stochasticity in the deposition of fungal spores (Peay & Bruns 2014). This high stochasticity in the composition of the rare microbiome may be responsible for the large number of associations that are unique to each network replicate and explain why the few shared associations involved abundant taxa. Third, the relative abundances of fungal taxa, upon which the networks are built, could vary within the same environmental condition because of methodological biases. Distortions in taxa abundance may be generated at each step of the DNA metabarcoding process, from the collection of samples to their sequencing, and at each step of the bioinformatic processing of the sequences (Ruppert *et al.* 2019). The fungal ITS region, which was used here as a barcode (Schoch *et al.* 2012), is highly variable in terms of length, sequence and number of copies (Nilsson *et al.* 2008; Lofgren *et al.* 2019), and these features could have increased the variability in the sequence data. Metabarcoding data are inherently noisy and this noise may explain why many associations are unique to a network replicate. Fourth, environmental conditions, which we consider homogeneous within a culture system, may not be homogeneous for microorganisms. Our experimental system and sampling protocols were designed to limit environmental variations within a cropping system. The vineyard plots were adjacent to each other and planted with grapevine clones. Moreover, we collected all leaves in less than two hours and controlled for the position of the sampled leaf on the vine. Nonetheless, the significant block effects in community composition indicate that the fungal communities were spatially structured at the scale of the experiment, which could account for spatial variations in networks. This poses a fundamental problem for biomonitoring approaches. The changes we want to monitor, which are generally large-scale changes in ecosystem functioning induced by human activities, may not necessarily be those to which microbial communities and networks

Table II.5: Number of associations shared between network replicates within each cropping system — conventional (CONV) and organic (ORGA) — depending on the method of network inference. Networks were inferred with SparCC (Friedman & Alm 2012) or SPIEC-EASI (Kurtz *et al.* 2015), by aggregating or not the ASVs at the genus level, and by including various percentages P of the most abundant ASVs or genera in the network. The number of shared ASVs or genera between the three network replicates is given into brackets. For every combination of parameters, three random networks having the same number of nodes and links than the three inferred networks were simulated. The pseudo p-value is the probability, estimated with 999 simulations, that the three random networks shared more associations than the three inferred networks (* p<0.05; ** p<0.01; *** p<0.001).

P (%)	Network inference at the ASV level				Network inference at the genus level			
	SPARCC		SPIEC-EASI		SPARCC		SPIEC-EASI	
	ORGA	CONV	ORGA	CONV	ORGA	CONV	ORGA	CONV
10	0 (17)	2*** (17)	0 (17)	0 (17)	0 (8)	0 (6)	0 (8)	0 (6)
20	1** (25)	2*** (23)	0 (25)	0 (23)	0 (13)	0 (13)	0 (13)	0 (13)
30	1* (36)	2* (30)	0 (36)	0 (30)	1** (14)	0 (16)	0 (14)	0 (16)
40	1* (42)	3** (44)	0 (42)	1*** (44)	1*(21)	0 (19)	0 (21)	0 (19)
50	1 (48)	3** (53)	1*** (48)	2*** (53)	1* (27)	0 (25)	0 (27)	0 (25)
60	2* (55)	3** (57)	0 (55)	4*** (57)	1* (31)	1* (28)	0 (31)	0 (28)
70	1(60)	3** (63)	1*** (60)	5*** (63)	2** (37)	0 (33)	0 (37)	0 (33)
80	1(63)	7*** (73)	1*** (63)	5*** (73)	3*** (38)	1 (36)	0 (38)	0 (36)
90	0 (71)	4** (75)	1*** (71)	6*** (75)	2* (43)	1 (42)	0 (43)	0 (42)
100	3* (80)	5** (81)	1*** (80)	5*** (81)	2* (47)	1 (47)	0 (47)	0 (47)

respond.

Our study also highlighted a major pitfall of network comparison analyses, which is the lack of statistical power due to the low number of network replicates. To evaluate the effect of the cropping system, we had 56 replicates per system at the community level, but only 3 at the network level. Indeed, several communities are needed to build a single network. This could explain why all community-level α-properties, but no network-level α-property, detected changes triggered by the cropping system. Despite this lack of statistical power, β-properties of microbial networks differed significantly between cropping systems, revealing a difference in microbial associations between organic and conventional systems. These differences were significant when network pairwise comparisons were based on shared taxa only, suggesting that the differences between organic and conventional networks were not only due to the turnover of taxa between cropping systems, but to re-associations of taxa. Overall, these results show that microbial networks inferred from DNA metabarcoding data can be used to detect changes in ecosystems if they are analyzed with network comparison tools defined by theoretical ecology (Pellissier *et al.* 2018; Poisot *et al.* 2012). They also suggest that β-properties of networks are better indicators of change than α-properties.

Our study also allowed us to compare two microbial network inference methods, SparCC (Fisher & Mehta 2014) and SPIEC-EASI (Kurtz *et al.* 2015). The results obtained with the two methods were, overall, encouragingly consistent. The variability of network replicates within a culture system was very high, regardless of the inference method used. The number of associations per network was lower with SPIEC-EASI than with SparCC (about 800 vs. 3500), probably because SPIEC-EASI infers partial correlations, discarding the indirect associa-

tions retained by SparCC (Kurtz *et al.* 2015). However the number of associations shared between network replicates was very low in both cases (less than 5) (Figure II.4 and A.4). SPIEC-EASI found slightly fewer shared associations than SparCC, especially when the number of nodes were reduced by filtering on taxa abundance or taxonomic aggregation. However, the shared associations detected by SPIEC-EASI had a higher level of significance (Table II.5). These results are in line with previous benchmarking studies showing the lower performance of SparCC compared to other methods of network inference, including SPIEC-EASI (Röttjers & Faust 2018; Hirano & Takemoto 2019) even though SparCC seems to work in low diversity communities (Weiss *et al.* 2016). Both methods, however, revealed very similar consensus associations within each cropping system. Nine associations, in total, were shared by the network replicates and 5 were found by both methods. Although they involved ubiquitous fungal species that have been frequently detected on grapevine, such as *Aureobasidium pullulans*, *Epicoccum nigrum* and *Cladosporium ramotenellum* (Martini *et al.* 2009; Bensch *et al.* 2015; Setati *et al.* 2015; Swett *et al.* 2016; Dissanayake *et al.* 2018), these associations were difficult to interpret due to a lack of knowledge of microbial interactions in natura. Nevertheless, these results show that the combination of network replicates and inference methods permits the identification of apparently robust associations between abundant species, which could be indicative of ecological interactions.

In our study, community-level analyses were found to be more informative, from an ecological perspective, than network-level analyses. We found that the richness, diversity and evenness of fungal communities were significantly higher in organic than conventional vineyards, consistent with the recent findings of Kernaghan *et al.* 2017 (but see Castañeda *et al.* 2018). The cropping system also significantly affected the composition of grapevine foliar fungal communities, as reported in previous studies (Castañeda *et al.* 2018; Kernaghan *et al.* 2017; Pancher *et al.* 2012; Schmid *et al.* 2011; Varanda *et al.* 2016). For instance, *Erysiphe necator*, the causal agent of grapevine powdery mildew, was significantly more abundant in conventional than in organic plots according to DNA metabarcoding data. These results were consistent with visual assessments of disease symptoms, indicating that, despite their numerous biases, metabarcoding data do contain some quantitative information useful for monitoring plant disease development (Jakuschkin *et al.* 2016; Makiola *et al.* 2020; Sapkota *et al.* 2015). The cause for such contrast in the pathogen abundance is possibly the nature and timing of phytosanitary treatments, but not the dose or number of applications that was similar in the two systems (Table A.1). Phytosanitary treatments also influenced several yeast strains, assigned to the genera *Vishniacozyma*, *Sporobolomyces* and *Filobasidium*, that were significantly more abundant in organic plots. These yeast genera are frequently detected on leaf surfaces due to their tolerance of irradiation and they might influence plant growth by producing plant hormone-like metabolites (Kemler *et al.* 2017). In addition, *Vishniacozyma victoriae* (ex *Cryptococcus victoriae*) was reported as a biocontrol agent of postharvest diseases (Lutz *et al.* 2013). Other yeasts possess valuable features of biocontrol agents inlcuding killer activities for some *Sporobolomyces* yeasts (Klassen *et al.* 2017). The yeasts *Vishniacozyma victoriae* and *Filobasidium wieringae* (ex *Cryptococcus wieringae*) were also reported as moderate antagonists of several filamentous fungi (Hilber-Bodmer *et al.* 2017).

Future research should investigate the interactions between these yeast species and grapevine foliar pathogens, including powdery mildew.

In the future, we envision that the analysis of microbial interaction networks in the phyllosphere (i.e. the microbial habitat formed by plant leaves (Vacher *et al.* 2016; Vorholt 2012)) will serve the prediction of foliar disease risk in crop plants. Plant-associated microbial interaction networks can protect plants against disease (Hassani *et al.* 2018; Hacquard *et al.* 2017; Kemen 2014). Resistance to pathogens is mediated by direct antagonistic interactions between the resident microbiota and the invading pathogen species (i.e. the barrier effect;Arnold *et al.* 2003; Kamada *et al.* 2013; Kemen 2014; Koch & Schmid-Hempel 2011; Laur *et al.* 2018) and by indirect interactions due to the activation of the host immune system by the resident microbiota (i.e. the priming effect;Hacquard *et al.* 2017; Kamada & Kawai 1989; Perazzolli *et al.* 2012; Ritpitakphong *et al.* 2016; Vogel *et al.* 2016). The subset of the host-associated microbial network, consisting of a pathogen and its interacting partners has been termed the pathobiome (Brader *et al.* 2017; Vayssier-Taussat *et al.* 2014). To better understand and predict disease risk, we should identify the microbial interactions forming pathobiomes (Durán *et al.* 2018) and the intrinsic network properties that hinder invasion by pathogens (Agler *et al.* 2016; Murall *et al.* 2017; Poudel *et al.* 2016). NGB will require the monitoring in real-time of these properties, based on the automated sequencing on leaf DNA. However, our study shows that statistical network inference, as currently based on a limited sampling effort, generates very few robust hypotheses for microbial interactions, limiting its use to monitoring the disease regulation services provided by the microbiota.

To conclude, here we have demonstrated that microbial networks, automatically inferred from DNA metabarcoding data at the ASV level (Callahan *et al.* 2016) with classical methods of statistical network inference such as SparCC (Friedman & Alm 2012) or SPIEC-EASI (Kurtz *et al.* 2015), and then compared using frameworks defined by theoretical ecologists (Pellissier *et al.* 2018; Poisot *et al.* 2012), can detect ecosystem change and therefore have a role to play in NGB approaches. Our results suggest that network β-properties were better indicators of change than network α-properties and should be preferred in future developments of NGB. We also showed that keeping the sequence data at the ASV level, rather than aggregating them at higher taxonomic levels, was preferable because it increased the replicability of the networks within a system. In our study, however, inferred networks were highly variable within a system whatever the method of network inference. Network replicates shared more associations than random networks of the same size, but the few shared associations involved only the most abundant ASVs and contained little ecological information on the functioning of the ecosystem. Future research in microbial network inference should therefore improve the replicability and interpretability of networks by, for instance, inferring ecological interaction types rather than positive and negative associations between microorganisms. Mutual information approaches, based on maximal information coefficients (MIC; Reshef *et al.* 2011) could overcome this dichotomy although these approaches have not stood out in the inference benchmarkings done to date (Hirano & Takemoto 2019; Weiss *et al.* 2016). All functional and ecological knowledge available on microorganisms needs to be gathered in databases (Louca *et al.*

2016; Nguyen *et al.* 2016a; Větrovský *et al.* 2020and integrated into network inference processes. In a study of trophic networks, Bohan et al.(2011) showed that logic-based machine learning is a promising tool to integrate background knowledge to network inference. Future research should investigate the relevance of this approach to microbial network inference. In our study, community-level analyses of DNA metabarcoding data were more statistically powerful than network-level analyses, because many samples were needed to build each network, and this reduced the number of network replicates by comparison with community replicates. The number of samples recommended in the literature for building a single network varies, from 25 (Berry & Widder 2014) to 200 (Hirano & Takemoto 2019). Our study shows that networks built from fewer samples (20 in the present case) can nevertheless detect ecosystem change, although we would advise more samples to increase the robustness of the inferred networks. In contrast to network-level properties, all community-level properties detected ecosystem change and provided information important for our understanding of the ecosystem functioning, such as for instance the higher microbial diversity and lower pathogen abundance under organic farming. Community-level analyses should therefore not be discarded in future developments of NGB, that will have to rely on very large DNA metabarcoding datasets combined with functional databases to fully benefit from network-level approaches.

5 Acknowledgements

We thank Lucile Muneret, Andreas Makiola and all members of the ANR NGB Consortium (ANR-17-CE32-0011) for helpful scientific exchanges on next-generation biomonitoring and for their comments on the manuscript. We also thank Gregory Gambetta, Guilherme Martins, Frédéric Barraquand, Isabelle Lesur, Adrien Rush, Lucie Zinger and four anonymous reviewers for very helpful comments on the results. We thank the Genotoul sequencing facility (Get-PlaGe) for sequencing and the Genotoul bioinformatics facility (Bioinfo Genotoul) for providing computing and storage resources. We also thank Julie Sappa from Alex Edelman & Associates for English language revision of the first version of the manuscript. We thank the INRA MEM metaprogram (Meta-Omics of Microbial Ecosystems) for financial and scientific support. Sequencing was funded by the INRA MEM MetaBAR project and bioinformatic and statistical analyses were performed as part of the INRA MEM Learn-biocontrol project. Additional funding was received from the LABEX COTE (ANR-10-LABX-45), the LABEX CEBA (ANR-10-LABX-25-01) and INRA EcoServ metaprogram on ecosystem services (IBISC project) and the Aquitaine Region (Athene project, n°2016-1R20301-00007218). CP's PhD grant was funded by the INRA and Bordeaux Sciences Agro (BSA). DBB's grant was funded by the ANR (ANR-17-CE32-0011) and SYNGENTA CROP PROTECTION AG (TK527180). The management of the experimental site was partly funded by the AFB (French Agency for Biodiversity) within the DEPHY network.

6 Data Accessibility Statement

The raw sequence data were deposited in Dataverse and are available in the FASTQ format at https://doi.org/10.15454/3DPFNJ while the filtered ASV table is available at https://doi.org/10.15454/WOICSE. The code is available as an archive at https://doi.org/10.15454/ZWDFJK.

7 Author Contributions

CP and DBB performed the bioinformatic and statistical analyses. JV and CV performed the sampling and processed the leaf samples. LD managed the sampling site and provided data on phytosanitary treatments and disease symptoms. JV performed the DNA extractions and amplifications. MB coordinated the work on the fungal mock community and the sequencing of all samples. DB helped interpreting the results in the context of next-generation biomonitoring. CV conceived the study, supervised the analyses and wrote the manuscript with the help of CP and DBB. All authors discussed the results and revised the manuscript.

Chapter III

Using a Logic-based approach to infer interactions from simulated DNA data

Machine learning of microbial interactions using Abductive ILP and Hypothesis Frequency/Compression Estimation

Didac Barroso-Bergada, Alireza Tamaddoni-Nezhad, Stephen H. Muggleton, Corinne Vacher, Nika Galic, David A. Bohan

Abstract:
Interaction between species in microbial communities plays an important role in the functioning of all ecosystems, from cropland soils to human gut microbiota. Many statistical approaches have been proposed to infer these interactions from microbial abundance information. However, these statistical approaches have no general mechanisms for incorporating existing ecological knowledge in the inference process. We propose an Abductive / Inductive Logic Programming (A/ILP) framework to infer microbial interactions from microbial abundance data, by including logical descriptions of different types of interaction as background knowledge in the learning. This framework also includes a new mechanism for estimating the probability of each interaction based on the frequency and compression of hypotheses computed during the abduction process. This is then used to identify real interactions using a bootstrapping, re-sampling procedure. We evaluate our proposed framework on simulated data previously used to benchmark statistical interaction inference tools. Our approach has comparable accuracy to SparCC, which is one of the state-of-the-art statistical interaction inference algorithms, but with the the advantage of including ecological background knowledge. Our proposed framework opens up the opportunity of inferring ecological interaction information from diverse ecosystems that currently cannot be studied using other methods.

Keywords */Inductive Logic Programming (A/ILP), Interaction Network Inference, Machine learning of ecological networks, Hypothesis Frequency Estimation (HFE)*

1 Introduction

Networks of interactions between species of microbes are believed to drive many of the biological functions that determine effects as diverse as soil health, crop growth, and plant and human disease. Next generation sequencing of DNA samples taken from microbial communities can produce lists of those species present and metrics for their abundance, by treating the number of each sequence type in the sample either as absolute or relative counts. Inferring networks from these data could yield important results, improving our ability to manage these systems and issues (Vacher *et al.* 2016). For example, learning interactions of competition or predation of a disease-causing microbial agent could be used to identify species for biological control, and the chemistry that is involved could lead to the development of new drugs (Golubev 2006). Current approaches to reconstructing ecological networks of interaction between microbial species use statistical learning to infer the presence of

an interaction via correlation. Human experts subsequently interpret whether the correlation indicates an interaction between the two correlated microbial species, such as competition or predation.

Abductive/Inductive Logic Programming (A/ILP) was previously used to automatically generate plausible and testable food webs from ecological census data (Tamaddoni-Nezhad *et al.* 2012). The approach in Tamaddoni-Nezhad et al. (2012) (Tamaddoni-Nezhad *et al.* 2012) also included a probabilistic approach, called Hypothesis Frequency Estimation (HFE) for estimating probabilities of hypothetical trophic links based on their frequency of occurrence when randomly sampling the hypothesis space. Through a review of the literature, it was found that many of the learned trophic links are corroborated by the literature. In particular, links ascribed with high probability by machine learning are shown to correspond well with those having multiple references in the literature. In some cases novel, high probability links were suggested, some of which were subsequently tested and confirmed in empirical studies (Tamaddoni-Nezhad *et al.* 2013).

In this paper we extend the A/ILP and HFE approaches in Tamaddoni-Nezhad et al. (2012) (Tamaddoni-Nezhad *et al.* 2012) for the purpose of learning microbial interactions. We will describe the existing context on interaction inference, detail the A/ILP based inference method and evaluate this method with a benchmark dataset. We also compare our results with SparCC, which is a state-of-the-art statistical interaction inference algorithm.

2 Background and related work

Microbial ecologists have clear criteria for interactions between species that can readily be transcribed into logical statements. In effect, past or ongoing interactions between two microbial species will have led to changes in the abundance of one or both species. Conceptually, therefore, two species might have undergone or might be undergoing an interaction if there is some pattern to the changes of the two species across a data-set. Thus, if one of the species always increases or decreases in abundance in the presence of the other, microbial ecologists might hypothesize an interaction between the two species. The ecological mechanisms of these interactions, along with their expected changes in abundance of the two species, have previously been described in Derocles et al. (2018) (Derocles *et al.* 2018) as shown in Table III.1.

In this paper we extend the A/ILP approach in Tamaddoni-Nezhad et al. (2012) (Tamaddoni-Nezhad *et al.* 2012) with logical statements for putative microbial interactions included as background knowledge, to infer ecological interactions directly, with less or even without the intervention of humans at the interpretation step. This direct approach would be particularly valuable for reconstructing microbial networks in previously unstudied ecosystems where human knowledge for interpretation may effectively be non-existent (Figure III.1).

In this paper we also extend the Hypothesis Frequency Estimation (HFE) approach introduced in Tamaddoni-Nezhad et al. (2012) (Tamaddoni-Nezhad *et al.* 2012). Microbial ecologists rely on statistical probability estimates, typically at the conventional 5% significance level, to evaluate the importance of a correla-

Table III.1: Type of interactions in function of the changes in abundance following Derocles *et al.* 2018.

Type of interaction	Effect on Species 1 abundance	Effect on Species 2 abundance	Nature of interaction
Mutualism	Up	Up	Mutual benefits of the species
Competition	Down	Down	Species have negative effect on each other
Predation; Parasitism	Up	Down	Parasite develops at the expense of the host
Commensalism	Up	Null	Species 1 benefits while Species 2 is not affected
Amensalism	Down	Null	Species 2 has a negative effect on Species 1, but Species 2 is not affected

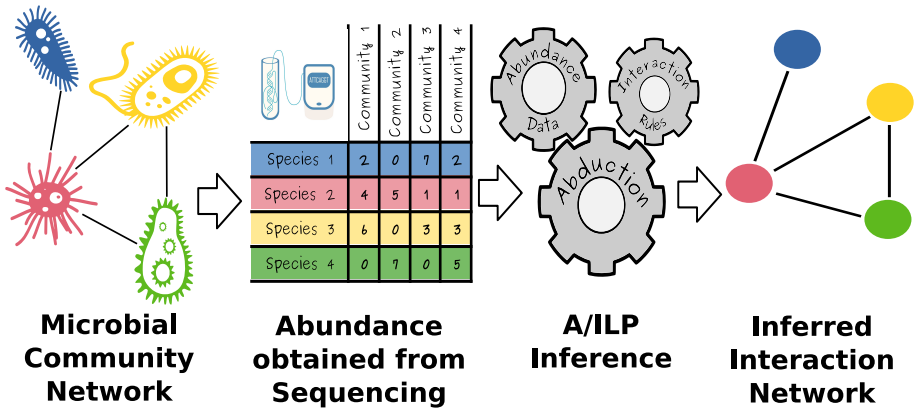


Figure III.1: Description of the interaction inference process. Microbial communities are shaped by the interaction between their members. DNA sequencing together with bioinformatic processes allow estimation of the abundance of the different microbes present in the communities. Using the abundance information from different communities as training examples, and the rules of interaction as background knowledge, it is possible to infer an interaction network that generalizes the interactions between microbes.

tional link between any two microbial species (Röttjers & Faust 2018). Most ILP approaches, including HFE rely on coverage based measures such as 'compression' for selecting hypotheses.

The problem we try to address here is whether compression can be evaluated within a statistical framework that meets the needs of microbial ecologists, to a degree that might be sufficient to convince them of the statistical importance and veracity of any learned interaction. In particular, we explore an extension of HFE where both the frequency and the compression of the hypotheses are considered within an statistical framework.

Benchmarking statistical learning approaches for inferring correlational links have used simulated data-sets. For example, Weiss et al. (2016) (Weiss *et al.* 2016) produced simulated microbial data-sets to benchmark the ability of different statistical methods, such as SparCC (Friedman & Alm 2012) and CoNet (Faust & Raes 2016), to detect different interaction types via correlation. In this paper we use the method of Weiss et al. (Weiss *et al.* 2016) to simulate ecological-like replicated data-sets of interactions, of given interaction strengths. We then use ILP to evaluate the presence of the simulated interactions, as a known set of expectations. Our specific goals are to: determine the most sensitive parameter of compression for recovering an interaction, given a discrete number of permutations; and, evaluate the probabilistic significance of the compression parameter using a form of bootstrapping.

3 Methods

3.1 Logical description of microbial interactions

A microbial interaction can be defined as a conserved effect on the abundance of one microbial species caused by the presence of another microbial species. Thus, the aim of the abductive procedure is to infer interactions, following ecological theory to explain the observed changes in the abundance of the species. To do this, the first step is to reflect the abundance changes between communities of each species using logical statements, following the form: abundance(C1, C2, S1, Dir). Here C1 and C2 symbolize two different community samples where species S1 is present and Dir the change in direction of abundance. To calculate the change in direction, the abundances of a species in the two different samples are compared using a Pearson Chi-square test. The test uses the total, summed abundance of all species in a community as the total population and checks the independence of the abundances of the species between the two samples. Where the species counts are found to be independent an abundance change is deemed to exist. An increase is symbolized as an up (↑) and a decrease as a down (↓). Where the species abundances are not independent between the two samples, a no abundance change condition is symbolized as zero (0). The presence of each species is also converted to logical clauses with the structure: presence(C1, S2, yes/no) where C1 refers to a sample community, S2 to a species and yes/no describes if S2 is present in C1 or not.

The abundance change and presence logical statements are used as observations in an abduction process conducted using the A/ILP system Progol 5.0 (Muggleton

1995). The effect on species abundances, either up or down, is described as the change in abundance of one species, S1, due to a second species, S2, when they co-occur in a community, C2. To ensure that the change is caused by S2 it is necessary to evaluate the abundance changes observed in communities where only S2 is present, C1, to communities where both co-occur, C2.

$$\begin{aligned} abundance(C1, C2, S1, up) :- \\ presence(C2, S2, yes), \\ presence(C1, S2, no), \\ effect\_up(S2, S1). \\ \\ abundance(C1, C2, S1, down) :- \\ presence(C2, S2, yes), \\ presence(C1, S2, no), \\ effect\_down(S2, S1). \end{aligned} \tag{III.1}$$

Progol5.0 uses a standard covering algorithm to conduct the abduction process where each observation is generalised using a multi-predicate search. This search is carried out over all the predicates associated with 'modeh' declarations, or abducible predicates, effect_up and effect_down. These two abducible predicates limit the possible variations in abundance that a species can experience due to the effect of another species. The search for the best hypotheses is guided by an evaluation function called 'compression' which is defined as follows:

$$f = p - (c + n) \tag{III.2}$$

where p is the number of observations (training examples) correctly explained by the hypothesis (positive examples), n is the number incorrectly explained (negative examples) and c is the length of the hypothesis (in this study, always 1 because the hypothesis is a single fact).

At the end of the abduction process, a list of ground hypotheses with the form effect_up/down(S2,S1) is returned, each hypothesis being supported by a compression value f. Implementations of A/ILP usually consider hypotheses with positive compression values. However, compression also offers a quantitative measure of information that can be used to discriminate between true and false interactions. For this purpose, first it is necessary to normalize compression values to a common scale. This is because while some species may not be present in all communities due their different random distribution. It is also possible that negative interactions reduce the abundance of a species to zero. Hence, each species will experience uneven combinations of abundance change mechanisms that require normalization. The normalization is performed using the logarithmic co-occurrence/occurrence ratio of the interacting species.

For an interaction between S1 and S2 to exist, there must be a consistent and constant effect, either up or down, on at least one of the species over all communities. Hence, we use the probabilistic estimator I supporting the interaction between S2 and S1 as defined below:

$$I_{S2,S1} = |f_{up}(S2, S1) - f_{down}(S2, S1)| \quad \text{(III.3)}$$

Compression is dependent on the order in which abundance clauses used as observations are supplied, due to the predicate search process that uses observations as seeds. To obtain reliable compression values, it is necessary to perform the inference several times, permuting randomly the order of examples to obtain different sampling of the hypothesis space. The permutation process will produce a set of possible effects and a corresponding compression value for each pair of species. Effects can be present in all the samples of the hypothesis space or just one part. Thus, it is necessary to define an approach to use the output of the abduction process, after sampling the hypothesis space, as a probabilistic measure. The HFE approach (Tamaddoni-Nezhad *et al.* 2012) estimates probabilities for hypothetical links in ecological networks, based on their frequency of occurrence when randomly sampling the hypothesis space. In this approach, the compression value was not taken into account to obtain a probabilistic measure of interaction. We propose a different method here that extends HFE to compute a probabilistic estimator I from the values of compression values f. In place of using the frequency of hypotheses with positive compression over all re-samples, here a function $func$ is applied to the f values to obtain an estimator I which summarizes the information contained in all the samples.

$$I_{S2,S1} = |func(f_{up}(S2, S1)_{1,\ldots,n}) - func(f_{down}(S2, S1)_{1,\ldots,m})| \quad \text{(III.4)}$$

In the experiments in this paper, we examined the following $func$ functions to obtain the estimator, I:

- **Frequency** = HFE is computed for each effect.
- **Independent permutations** = Where there is more than one compression value in a permutation, the sum is computed. Maximum values for each interaction among all permutations are retained.
- **Maximum** = Compression values from all permutations are pooled. Then, maximum compression is selected for each effect.
- **Sum** = Compression values from all permutations are pooled. Then, compression is summed for each effect.

3.2 Bootstrapping

Having a probabilistic measure of the likelihood of an interaction is critical for ecologists to interpret the networks resulting from interaction inference. This should allow the selection of those interactions that are realistic and might then be tested in cost- and time-expensive laboratory experiments. The most intuitive selection method would establish a threshold for the estimator value. However, most of the ecological systems where A/ILP interaction inference could help are poorly described and there are no references to guide selection of such a threshold (Röttjers & Faust 2018).

It is a common assumption that the interaction networks shaping microbial communities are sparse. This means that the number of interactions of each species is only a small fraction of the total set of interactions that are possible. Thus, where the estimator value of the observed interactions, I, is greater than the values of non existing interactions, it is possible to assess the statistical significance of an interaction using a bootstrapping procedure. Statistical bootstrapping is a method of re-sampling a dataset to create new simulated datasets (Efron & Tibshirani 1993). Let d be all the compression values for effect up and effect down, involving at least one of the potential interacting species S1 and S2, the real final estimator value, I_0, is obtained by applying equation III.4 to n compression values that support an effect up of S2 on S1 and m compression values that supports and effect down of S2 on S1. The bootstrapping procedure re-samples compression values in d to obtain two new sets of values d^{up*} and d^{down*} of n and m lengths, respectively. Then, an alternative estimator value I_a is obtained applying equation III.4 to d^{up*} and d^{down*}. If the re-sampling process is repeated B times, a pseudo p-value can be computed for the potential interaction between S1 and S2 averaging the simulated values I_a that are bigger than I_0 (Li *et al.* 2009).

$$\begin{aligned} I_0 &= |func(f_{up}(S2, S1)_{1,\dots,n}) - func(f_{down}(S2, S1)_{1,\dots,m})| \\ I_a &= |func(d^{*up}_{1,\dots,n}) - func(d^{*down}_{1,\dots,m})| \\ p - value &= \sum_{b=1}^{B} \{(I_{a_b} \geq I_0)\}/B \end{aligned} \tag{III.5}$$

3.3 Simulated data-sets

The aim of ILP based network inference is to use logical descriptions of interactions to detect and classify those interactions between species as a function of the ecological mechanism that drives them. Hence, a simulation model to generate test-datasets should follow the different ecological mechanisms that they are simulating (Faust & Raes 2012). Information required for network inference is structured in tables, where each row contains the information for a species and each column contains the information for a microbial community. Each cell summarizes the count of individuals of each species in each community (abundance). Weiss et al. (2016) (Weiss *et al.* 2016) proposed a simulation model to create computer-generated tables including the effects of ecological-like, linear interactions. The model uses the log-normal distribution to simulate the abundance of non-interacting species in a set of microbial communities or samples. The log-normal distribution has been shown to appropriately model the abundance distributions of microbial communities (Shoemaker *et al.* 2017). Interactions are then introduced by modifying the abundance of species in accordance with the different ecological mechanisms (Faust & Raes 2012). For any two species, say S1 and S2, the abundance modifications only happen in communities where the species co-occur. The abundance modification is based on the effect that S2 has on S1. If the effect is positive, the abundance of S1 increases as a function of the abundance of S2, modulated by a strength of the interaction. If the effect is negative, the abundance of S1 is

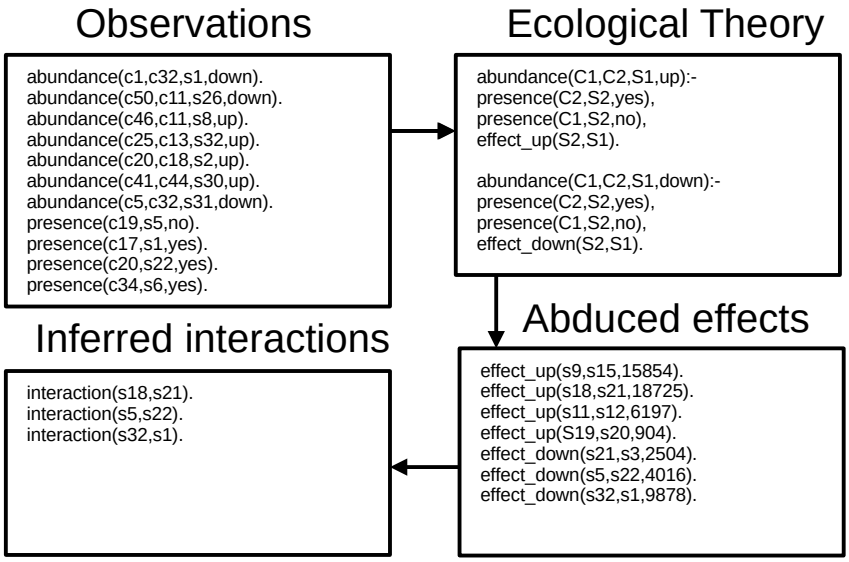


Figure III.2: Summary of the inference process of microbial interactions using A/ILP. Observations are obtained assessing the abundance change between OTUs. The ecological theory describes how the presence of an OTU can affect the abundance of a second OTU. Abduction is performed using the observations and theory. Significant interactions are assessed by bootstrapping the compression value from different permutations of observations

decreased following a similar mechanism. In the case that the interaction affects both species, their abundance is modified in parallel.

Using the method proposed in Weiss et al. (2016) (Weiss *et al.* 2016) we generated three tables containing the abundances of 16 pairs of interacting species in 100 communities. The tables were simulated using interactions of different strength values (2, 3 and 5), and four different ecological mechanisms: amensalism, commensalism, competition and mutualism (Derocles *et al.* 2018).

3.4 Compositionality and bias

Modern sequencing technologies allow us to recover information about microbial communities from samples of environmental DNA. As noted in Section 1, the number of times that a DNA sequence from a species is 'read' in a sample can be used as a measure of abundance or count. A sequencer can only read a limited number of sequences in a sample, and these are shared amongst species, imposing a compositional bias on the data (Gloor *et al.* 2017). Thus, to generate ecological-like microbial tables it is necessary to re-introduce compositionality into the simulated data-sets. To do this, we normalized the sequencing depth as probabilities in a multinomial distribution and then sampled the distribution to obtain the simulated counts across a common sequencing depth.

4 Experimental evaluation

The performance of the A/ILP based microbial inference (Figure III.2) is evaluated using the computer-generated datasets. First, it is tested the number of samples of the hypothesis space and the different functions used to obtain the I statistic. Then the best setting found in the first experiment is used to asses the performance of the bootstrapping procedure compared with a threshold for I and SparCC. The simulated data and the code used to perform the experimental evaluation have been included in a public repository[1].

4.1 Experiment 1

Null Hypothesis 1:

Using the estimator I as defined in (III.4) using different functions does not lead to higher accuracy over the frequency-based approach HFE for predicting microbial interactions.

Materials and Methods:

The three computer-generated tables described in section 2.1, computed using the methodology of Weiss et al. (2016) (Weiss *et al.* 2016), are used to test the performance of functions used to obtain the estimators. 100 abductions of possible effects are performed for each table. The observations produced from the tables are randomly permuted at each execution. The logical description of effect is used as background knowledge. Then the estimators are obtained using the different functions described previously.

Since the interactions that drive the abundances of the computer-generated tables are known, it is possible to treat interaction inference as a classification problem. Interactions can be classified between existing and non existing and the estimator values obtained using the different functions are the classification accuracy. Thus, the area under the curve (AUC) of the true positive rate against the false positive rate (ROC curve) can be used as a measure of performance. AUC is computed for all functions at n permutations = 1, 5, 10, 25 and 50. An ANOVA test is performed together with a Tuckey's range test to asses the significance of differences of AUC values between all functions.

Results and Discussion:

AUC values for the different methodologies to obtain estimators and number of permutations are displayed in Figure III.3. As expected, values of AUC increase as the number of permutations used for the inference increases. These stabilize at around n = 50 permutations. AUC values are similar where the strength of interaction is reduced, being significantly lower at the highest strength. This can be explained by the low performance of the logical model in describing the specific case of a negative interaction reducing the abundance of a given species to 0, the likelihood of which increases with stronger interactions. This ecological process,

[1] https://github.com/didacb/Machine-learning-of-microbial-interaction

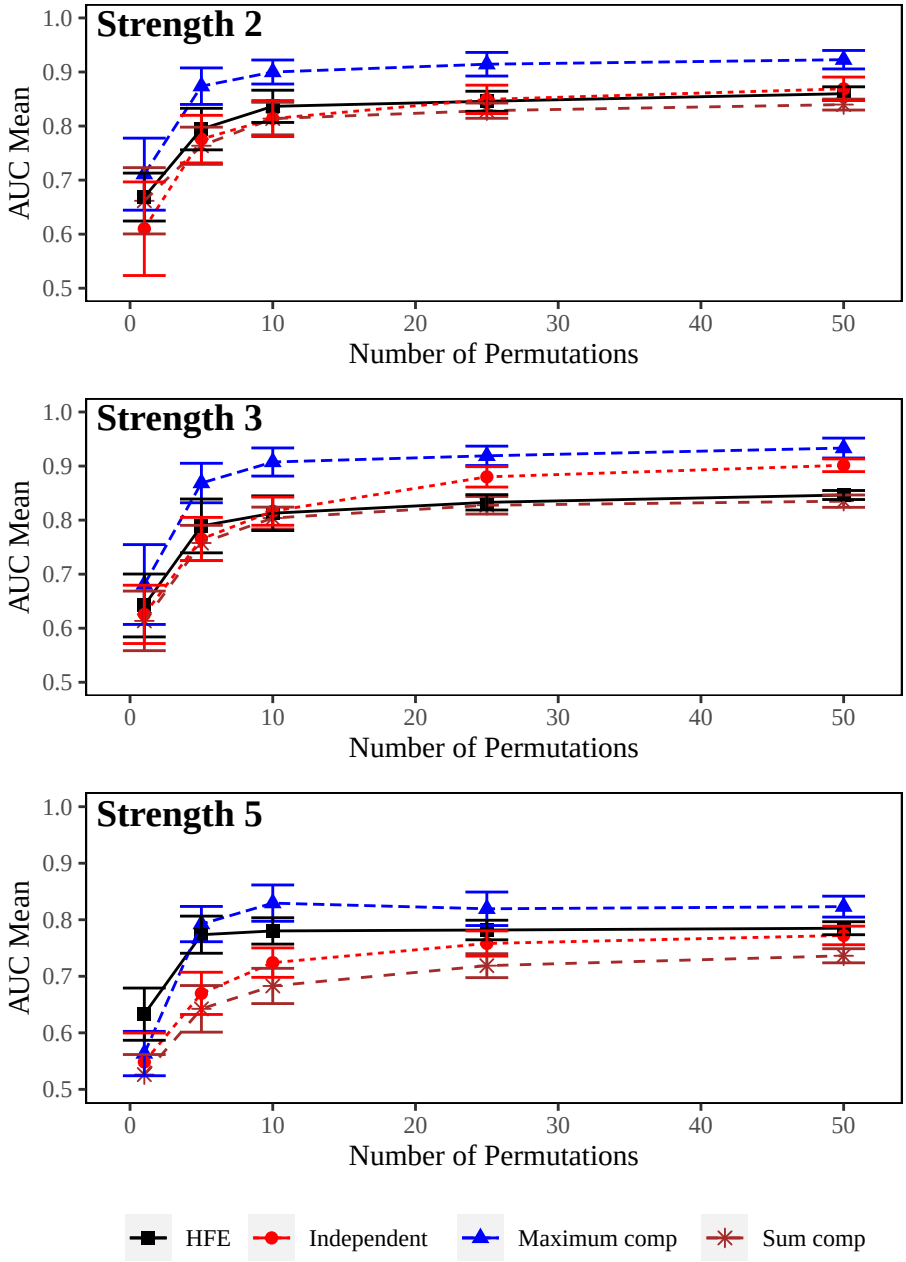


Figure III.3: Area under the ROC curve values (AUC) obtained using different number of permutations. Each plot shows the AUCs obtained for interactions of different strengths. Each line represents a method used to obtain the estimators. Error bars show the standard deviation of the means.

called exclusion, greatly reduces the co-occurrence between species and, as a consequence, the information available to infer an interaction. Maximum compression used to obtain I is the metric that gives the highest values of AUC, for any given number of permutations and interaction strengths. HFE, Sum and independent permutations have similar AUC values at strength 2 and 5 while the independent permutation method performs best at strength 3. The ANOVA shows that all functions have significantly different AUC values, except independent permutations and HFE at strength 2. Consequently, the null hypothesis can be rejected because the method using the maximum compression values to obtain I performs better than HFE in all cases.

4.2 Experiment 2

Null Hypothesis 2:

The bootstrapping procedure described in Sec 3.2 leads to lower accuracy compared to the optimal threshold and other statistical methods for interaction inference.

Materials and Methods:

The Bootstrapping procedure is conducted using the three computer generated tables used in the preceding experiment. The procedure uses the maximum compression to obtain the I estimator. Two different bootstrapping techniques are evaluated: ordinary and strata. Ordinary bootstrapping performs the bootstrapping independently on all compression values while the strata method constrains the bootstrapping to compression values by effect. Interactions with p value < 0.05 are considered to exist. Bootstrapping accuracy is compared with the accuracy of prediction using an optimal threshold for I estimator. The optimal threshold metric is obtained automatically from the ROC curves of the preceding test using the pROC R package best threshold method (Robin *et al.* 2011). To have a reference for comparing the performance of A/ILP inferring interactions, the interaction inference was also performed using SparCC (Friedman & Alm 2012), a widely used statistical inference tool. The process was performed using the FastSpar 1.0 implementation (Watts *et al.* 2018) with default parameters.

Results and Discussion:

Accuracy measures are displayed in Table III.2. Ordinary bootstrap presents better accuracy than strata bootstrap at strength = 2 and 3, while strata performs better at strength = 5. However, ordinary bootstrap allows the detection of a larger number of true positives in contrast to strata. We believe this to be the better option to use for detecting interactions, therefore. In all cases bootstrap accuracy is higher than the optimal threshold accuracy. Bootstrapped A/ILP sensitivity values are significantly lower than SparCC at all strengths. However, the specificity values are slightly higher. Thus, SparCC has a greater number of false positives, while A/ILP generates a higher number of false negatives. This produces similar accuracy measures for SparCC and bootstrapped A/ILP, independent of the interaction strength. We therefore reject the null hypothesis.

5 Discussion and conclusion

This work proposes a framework to infer ecological-like microbial interactions that allows us to use abundance information and descriptions of interactions as logic statements for obtaining a probabilistic measure of the significance of a given interaction, based on compression values. By using logical descriptions of interactions, microbial ecologists can apply their knowledge not only to the interpretation of results but also to the inference process itself.

Tamaddoni-Nezhad et al. (2013) (Tamaddoni-Nezhad *et al.* 2013) showed how the introduction of ecological expertise (in the form of background knowledge) to the interaction inference can lead to interesting results for invertebrate food webs, and we believe that the work presented here will facilitate similar results for microbial networks.

Interactions between species can be driven by different mechanisms, thus it is necessary to obtain a common quantitative measure of these interactions for appropriate ecological interpretation. Statistical relation learning (SRL) has been used to obtain quantitative measures using ILP-like representations and inference (Getoor & Taskar 2007). However, most of the cases where SRL has been used requires probabilistic data, where each observation has an associated probability. However, the observations obtained from NGS data are purely deterministic; a species is either present or not in a given community, and its abundance in this community is also an invariable number. Other authors have proposed different methods to perform probabilistic approaches to deterministic data, such as using a binary matrix obtained from a deterministic process to obtain a support vector machine (Amini *et al.* 2007). The idea of using compression as a probabilistic estimation was also used by Bryant et al. (2001) (Bryant *et al.* 2001) in their implementation of ASE-Progol. ASE-Progol uses compression to select between contradictory candidate hypothesis. Tamaddoni-Nezhad et al. (2012) (Tamaddoni-Nezhad *et al.* 2012) developed the Hypothesis Frequency Estimation approach for sampling and estimating the probability of abductive hypotheses. We extended this idea to use the value of compression as a measure for estimating the likelihood of any given interaction. To do this, it is necessary to sample the hypothesis space enough times to ensure that the distribution of compression values obtained for each interaction is representative of all the possible values. Our first experiment showed that a re-sampling of 50 times is enough in a setting involving 32 species and 50 communities, given that the AUC values obtained using a larger sampling were not significantly different. This experiment also showed that retaining the maximum compression values among all hypothesis space samples has greater accuracy than using the HFE, or the other numeric metrics of compression tested, independent of the strength of interactions. This is consistent with the predicate search algorithm of Progol5.0 which selects the hypothesis with the maximum compression from all possible hypotheses (Muggleton & Bryant 2000). Lastly, it is important to note that the AUC values decreased in all cases where the interactions were strong enough to cause exclusion (Derocles *et al.* 2018). In future applications of A/ILP-based interaction inference, it will be important to incorporate logical rules describing exclusion in the learning.

Bootstrapping is a statistic technique used in many areas of knowledge dis-

covery. It has also been applied in statistical inference of interactions (Friedman & Alm 2012). We showed that the bootstrapping procedure has better accuracy values than the optimal thresholds obtained using ROC curves. Thus, it is possible to use this procedure for real data, where the interactions are unknown and the ROC curves cannot be used. Even though bootstrapping offers good accuracy and specificity measures, the sensitivity of inference is insufficient to detect all true interactions. As noted previously, this is in part related to the effect of interaction-derived exclusion of one or both species. It is also due to a restrictive effect of the bootstrapping procedure. Where the bootstrapping is constrained by the effect of abundance, leading to a low number of examples, the sensitivity is low. It is expected that, in real cases where each species interacts with more than one species providing more high compression values for the bootstrapping, the sensitivity will increase.

Weiss et al. (2016) (Weiss *et al.* 2016) used their method of generate ecological-like datasets, as described in section 3.3, to benchmark many of these interaction inference tools. Comparing the results obtained by SparCC in Weiss et al. (2016) (Weiss *et al.* 2016) and in this work, a reduction in the number of interacting species reduced specificity and increased sensitivity. However the accuracy values remained similar. A/ILP inference using bootstrap obtained accuracy measures in the same range as SparCC, using the same computer-generated data. This accuracy can be further improved by expanding the range of logical descriptions to other ecological effects and interactions, such as exclusion. Also, it makes it possible to include other sources of biological and ecological information from existing databases as background knowledge.

Our work shows that A/ILP can be used to infer ecological interactions accurately from computer-generated datasets, using an estimator obtained from compression as a numeric measure of interaction and a bootstrap procedure to detect true interactions. Hence, A/ILP interaction inference has the potential to become a valuable tool for microbial ecologists for the inference of ecological interactions.

6 Acknowledgements

This work was supported by the Agence Nationale de la Recherche, Grant/ Award Number: ANR-17-CE32-0011, and SYNGENTA CROP PROTECTION AG. Corinne Vacher and David A. Bohan acknowledge the support of the Learn-Biocontrol project, funded by the INRAE MEM metaprogramme, and the BCMicrobiome project funded by the Consortium Biocontrôle. Alireza Tamaddoni-Nezhad and Stephen Muggleton were supported by the EPSRC Network Plus grant on Human-Like Computing (HLC).

Table III.2: Performance of the bootstrapping estimator compared with optimal threshold obtained from the ROC curve and SparCC. The three datsets used for the interaction inference have 16 real interactions over 496 possible interactions.

Strength 2

	Optimal threshold	Ordinary Bootstrap	SparCC
Total	40	13	26
TP	13	9	12
FP	27	4	14
TN	453	476	466
FN	3	7	4
Sensitivity	0.812	0.562	0.75
Specificity	0.944	0.992	0.971
Accuracy	0.94	0.978	0.964

Strength 3

	Optimal threshold	Ordinary Bootstrap	SparCC
Total	69	7	31
TP	14	6	11
FP	55	10	20
TN	425	479	460
FN	2	10	5
Sensitivity	0.875	0.375	0.688
Specificity	0.885	0.998	0.958
Accuracy	0.885	0.978	0.95

Strength 5

	Optimal threshold	Ordinary Bootstrap	SparCC
Total	50	27	40
TP	12	10	13
FP	38	17	27
TN	442	463	453
FN	4	6	3
Sensitivity	0.75	0.625	0.812
Specificity	0.921	0.965	0.944
Accuracy	0.915	0.954	0.94

Chapter IV

Logic-based inference of ecological interactions from environmental DNA data

Explainable inference of the diversity of microbial interactions. *Unravelling the Dark Web: explainable inference of the diversity of microbial interactions*

Didac Barroso-Bergada, Alireza Tamaddoni-Nezhad, Dany Varghese, Corinne Vacher, Nika Galic, David A. Bohan

Abstract:
The functional diversity of microbial communities emerges from a combination of the great number of species and the many interaction types, such as competition, mutualism, predation or parasitism, in microbial ecological networks. Understanding the relationship between microbial networks and the services and functions delivered by the microbial communities is a key challenge for Microbial Ecology, particularly as so many of these interactions are difficult to observe and characterize. We believe that this 'Dark Web' of interactions could be unravelled using an explainable machine learning approach, called Abductive/Inductive Logic Programming (A/ILP) in the R package InfIntE, which uses mechanistic rules (interaction hypotheses) to infer directly the network structure and interaction types. Here we attempt to unravel the dark web of the plant microbiome using metabarcoding data sampled from the grapevine foliar microbiome. Using synthetic, simulated data, we first show that it is possible to satisfactorily reconstruct microbial networks using explainable machine learning. Then we confirm that the dark web of the grapevine microbiome is diverse, being composed of a range of interaction types consistent with the literature. This first attempt to use explainable machine learning to infer microbial interaction networks advances our understanding of the ecological processes that occur in microbial communities and allowed us to infer specific types of interaction within the grapevine microbiome that could be validated through experimentation. This work will have potentially valuable applications, such as the discovery of antagonistic interactions that might be used to identify potential biological control agents within the microbiome.

Keywords *Explainable AI, Abduction, Microbial Interactions, Networks*

1 Introduction

1.1 The Dark Web of microbial communities

The high taxonomic, morphological and functional diversity of microbial communities (Konopka 2009) emerges from a combination of the great number of species and the many interaction types in the ecological networks of the microbiome. These ecological interactions, typically involving metabolites produced or used by the interacting species (Tshikantwa *et al.* 2018), result in economically and socially important ecosystem services and functions, including biofilm formation (Magana

et al. 2018), nitrogen fixation (Shrestha *et al.* 2007) and disease regulation (Ishaq 2017; Hacquard *et al.* 2017), and structure the communities of microbial organisms observed in nature (Faust & Raes 2012). The abundance of any two species in a community will be determined by whether they participate in a pairwise interaction and the precise ecological interaction type, as well as being influenced by other biotic and abiotic factors (de Vries *et al.* 2012). Where the abundance of both species is observed to decline, this might be hypothesised to come about as a result of a competitive interaction. Predation can lead to an increase in abundance of one species at the expense of the other (Faust & Raes 2012), although the dependence of the predator on the prey might mask this effect (Derocles *et al.* 2018). Mutualistic interactions might result in both species increasing in abundance, while amensalism and commensalism will cause a differential benefit or cost to the abundance of only one of the species. It is the mixture of all these different interactions, acting between all species in the ecological network simultaneously, which determines the species richness, diversity patterns, functions and dynamics of the microbial community.

Understanding the relationship between microbial interactions and the services and functions delivered by the microbiome is a key challenge for Microbial Ecology. We expect, for example, that interactions such as competition or predation that might result in declines in the abundance of species will be associated with ecosystem services, such as biological control of microbial pathogens (Musetti *et al.* 2007; Poveda *et al.* 2021). The great challenge is that these interactions can be difficult to evaluate directly, due to the complexities of observing microbial species *in-situ* or to the difficulties of culturing microbial species in the laboratory (Wu *et al.* 2019; Crhanova *et al.* 2019), rendering estimation of species presence and abundance difficult. Many microbial taxa remain unknown to science, forming what has been termed a microbial "Dark Matter" (Marcy *et al.* 2007). As a consequence, many of the interactions of microbial ecological networks are unobserved; they are 'Dark Webs' that with the additional problems of observing microbial species contribute to our poor understanding of these systems. This limits our ability to advance the science of the microbiome. In this paper, we attempt to unravel the dark web of the microbiome using direct inference of specific types of ecological interactions from DNA metabarcoding data.

1.2 Inferring microbial correlation networks

Analysis of microbial species and communities *in situ* has been greatly facilitated by metabarcoding surveys of environmental DNA samples amplified using generic primers for particular taxonomic groupings (Thomsen & Willerslev 2015; Ruppert *et al.* 2019). The process of sequencing the sampled eDNA yields information on the number of DNA copies of each sequence that with care can be treated as quantitative information for the count of the different taxa in the microbial community. Pipelines like VSEARCH (Rognes *et al.* 2016) cluster the counts into operational taxonomic units (OTU) in each sample by their similarity (Pauvert *et al.* 2019), under the assumption that sequences with the greatest similarity represent phylogenetically similar organisms (He *et al.* 2015). Other pipelines, like Dada2 (Callahan *et al.* 2016), find amplicon sequence variants (ASVs), identical

Table IV.1: Interactions types as described in Derocles et al. 2018. (Derocles *et al.* 2018)

Type of Interaction	Effect on taxa A abundance	Effect on taxa B abundance	Description
Amensalism	0	↓	Taxa A causes a decrease on the abundance of taxa B without suffering any effect on the abundance
Commensalism	0	↑	Taxa B increases its abundance thanks to the effect of Taxa A
Competition	↓	↓	Both taxa abundance decreases by the effect of the other. This can be caused by direct competition (directly harming the other taxa) or exploitation competition (they need the same resource an in consequence there is less available)(Fredrickson & Stephanopoulos 1981).
Mutualism	↑	↑	The abundace of both taxa increases by the effect of the other.
Neutralism	0	0	Both taxa co-occur but there is no effect on their abundance, and therefore, no interaction.
Parasitism or Predation	↑	↓	Taxa A develops at the expense of taxa B.

sequences which could vary due possible sequencing errors.

DNA sequencers can only process a given number of DNA sequences in any given metabarcoding run and the environmental samples within a run are usually equimolarly pooled before sequencing. As a consequence, each sample theoretically receives the same number of sequences, whatever the initial abundance of the species composing it. The sequencing data produced is therefore compositional (Gloor *et al.* 2017), reflecting the relative abundance of the species within each sample but not their absolute abundance. The inherent biases of compositional data are typically controlled for by applying log-transformations to the sequence count data. SparCC (Friedman & Alm 2012) infers networks using linear Pearson correlation between the log-transformed components, for example. CCLasso uses Lasso to infer the correlation network (Fang *et al.* 2015) and SPIEC-EASI uses inverse covariance or neighbourhood selection and StARS to obtain the most stable network (Kurtz *et al.* 2015; Liu *et al.* 2010). PLN-network uses the Poisson-LogNormal model, where sequence counts follow Poisson distributions, and introduces sequencing depth as an offset. Then, uses different model selection methods like StARS or EBIC (Chiquet *et al.* 2019).

Statistical associations between counts has been used to infer possible interactions between OTUs, and to hypothesise microbial networks (Faust & Raes 2012; Weiss *et al.* 2016; Röttjers & Faust 2018). The frameworks for network inference, such as those described above, have robust theoretical statistical foundations, and the tools are typically flexible, fast and robust to noise in the metabarcoding sample data (Dohlman & Shen 2019; Weiss *et al.* 2016). However the taxa (nodes) and interactions (edges) of the networks that are produced need considerable interpretation, post-analysis, as to their ecological importance. Positive or negative statistical assocations do not indicate causality and are not specific indicators of a type of interaction. This interpretation problem resembles the black box problem found in many fields of knowledge where statistical machine-learning is applied (Castelvecchi 2016). The tool may detect that something is happening, from the available data, but it cannot provide a mechanistic understanding of the underlying process. A mechanistic interpretation can only be provided *post-hoc* using expert knowledge or the literature (Tamaddoni-Nezhad *et al.* 2021). Explanatory machine learning (XML) has been proposed as one possible solution to the black box problem (Ai *et al.* 2021; Gilpin *et al.* 2018). Here, we use a form of XML, called Abductive/Inductive Logic Programing (A/ILP), to detect and then classify ecological interactions directly from microbial metabarcode samples. A/ILP infers interactions by searching the data using ecological rules for each type of interaction, defined *a priori* by the user, thereby providing explanatory or comprehensible output of networks with the edges that are a direct expression of the result of the rules.

Abductive ILP (A/ILP) has previously been used to provide a symbolic, mechanistic explanation of metabolic regulation (Tamaddoni-Nezhad *et al.* 2004) and to reconstruct food webs of trophic interactions from invertebrate abundance data (Bohan *et al.* 2011). To our knowledge it has never been used on microbial community sequence data or to infer numerous types of interaction simultaneously. This has in part been due to the intensive computational requirements of XML approaches and their lower robustness to ecological noise. It may also be due to

the much more limited experience of logical approaches amongst ecologists, by comparison with the more familiar approaches of statistics. The A/ILP approach we detail proceeds through two steps: first to demonstrate an ability to detect links of a particular type; and, second, to asses the validity of the ecological interaction networks that are produced. The first step of this process uses synthetic data for the different types of ecological interactions, generated as described in Weiss et al., 2016 (Weiss *et al.* 2016). This allows us to verify whether we can both detect and classify interactions, by their correct type, with appropriate levels of significance. The A/ILP is then used in the second step to unravel the dark webs of the different interaction types from real microbial eDNA sequence data, sampled from healthy and symptomatic leaves of grapevine, *Vitis vinifera L.* during downy mildew epidemics, caused by *Plasmopara viticola*. Interactions reducing the abundance of *P. viticola* (e.g. competition, amensalism) could guide future experimental research on biological control of mildew.

2 Materials and Methods

2.1 Hypothesis framework for learning microbial ecological interactions using abductive logic

Explainable approaches to inferring ecological interactions start with a clear declaration of the rules for an ecological interaction. The mechanisms of ecological interactions between any two, or possibly more, OTUs might be described in a multitude of different ways (Faust & Raes 2012; Tshikantwa *et al.* 2018). We posit that the minimum common facts for all hypothesised interactions are that: the two OTUs undergoing an interaction should be present together in at least one sample; and, at least one of the OTUs involved in an interaction undergoes a change in abundance. Here, abundance is understood as a measure of the size of an OTU population in a sample and is derived from the number of OTU sequence reads found in each sample. Thus, to evaluate change in abundance of all OTUs, across all samples, the sequence counts of an OTU and the total sequence depths in any two samples collected in the same biotic and abiotic conditions, are used to construct a contingency table, with the significance of the change in OTU abundance between the samples being evaluated by a χ^2-test of independence. Significant changes are then classified either as an increase, *up*, or as a decrease, *down*, compared to the relative abundance of the OTU, in the two samples. Symbolically, this can be expressed as the logic clause $abundance(s1, x, y, up/down)$. Here, $s1$ is any given OTU, (x, y), are two given samples sharing the same conditions and $up/down$ describes the direction of the significant abundance change. The abundance changes are computed in this way across all OTUs in all samples. This avoids many of the compositional biases inherent in treating DNA sequence data as counts since only the counts from the same OTU are compared and the total sequence depth is taken into account by the χ^2-test of independence. The presence, *yes*, or absence, *no*, of an OTU in a sample, x, can be expressed as the clause: $presence(s1, x, yes/no)$.

The abductive logic process uses these clauses to find possible explanations (ef-

fects) for the observed changes in abundance and presence using *a priori* hypotheses for ecological interactions that reflect the existing state of ecological knowledge, present in the literature. In this case, we hypothesise that an interaction will have occurred where the presence of at least one OTU ($s1$) produced a consistent effect on the abundance of another OTU ($s2$) in the samples. The logical relationships for the effect of such an interaction can be described from the abundance and presence clauses as:

$$\begin{aligned} effect\_up(s1, s2) \quad \text{if:} \quad & \begin{cases} abundance(x, y, s2, up) \\ presence(s1, x, no) \\ presence(s1, y, yes) \end{cases} \\ effect\_down(s1, s2) \quad \text{if:} \quad & \begin{cases} abundance(x, y, s2, down) \\ presence(s1, x, no) \\ presence(s1, y, yes) \end{cases} \end{aligned} \tag{IV.1}$$

Where OTU $s2$ has a greater abundance in sample y than in the sample x, due to the presence of OTU $s1$ in sample y and its absence in sample x, then this might suggest the beginnings of a pattern. Should this pattern be consistent across different sample pair combinations, then the abduction process would infer an *up* effect of $s1$ on $s2$. A pattern is considered consistent if the number of observations following such pattern is considerably larger than the number of observations contradicting it. Effect_down would be abduced from consistent *down* effect of $s1$ on the abundance of $s2$.

The abduction process computes a compression value for each possible effect between two OTUs in the dataset. Compression is a numerical measure representing the amount of observations that support each abduced effect (Muggleton 1995), and therefore indicates how likely an interaction is to have occurred. We require that an interaction should have consistent effect across different sample combinations. This means that to be important an interaction should give a greater effect in one direction than the other, all other things being equal. We therefore compute an overall statistic for the likelihood of interaction, I, as the difference between the compressions for the *up* and *down* effects (Barroso-Bergada *et al.* 2022).

Detection of significant interactions

For each pair of OTUs that are considered, the value of I is treated as the weight of a directed edge in an ecological interaction network. Setting a threshold, λ, for the absolute value, I, selects a list of inferred edges for a network. $\lambda = 0$ would select all possible edges, while a $\lambda = max(I)$ would deliver an empty network with no edges selected. $max(I) = nobservations$ and it is dependent on the number of observations in a dataset, however, and it is not possible to establish a common λ value for the reconstruction of any network. We therefore select the significant interaction edges empirically using a subsampling methodology called StARS (Liu *et al.* 2010). The StARS procedure subsamples 80% of the samples, multiple times, and performs the abduction of network edges. The most stable network of interactions is then identified using the frequency that the edges appear at

different values of λ. Here we use 50 resamplings of the data and 50 λ values linearly increasing from 0 to $max(I)$. The number of sub-samples and length of lambda path are selected following the recommendations of Muller et al., 2016 (Müller *et al.* 2016), with a restrictive stability threshold of 0.01 so as to minimise the number of false positive interactions.

Classification of interaction types

The StARS procedure selects those interactions of a network that have a consistent direction of effect of $s1$ on $s2$ and $s2$ on $s1$, and may therefore be treated as significant. The direction of these detected interactions, *up*, *down* or no effect detected, are characteristic of particular interaction types and can be used to classify the types of interaction directly (Faust & Raes 2012; Derocles *et al.* 2018). Thus, where the abundance of $s2$ increases in the presence of $s1$, $effect\_up(s2, s1)$ and $s1$ also consistently goes up in the presence of $s2$ $effect\_up(s1, s2)$ it becomes possible to classify the interaction that is inferred as one of mutualism that benefits both OTUs. $effect\_down(s2, s1)$ and $effect\_down(s1, s2)$, by contrast, could be assigned to a competition interaction, due to the interaction costs to the abundance of both OTUs. Across all possible inferred pairwise combinations of up, down and no change between the OTUs in the dataset it becomes possible to classify directly the ecological interactions of mutualism, predation, competition, commensalism and amensalism (Table, (Faust & Raes 2012), many of which cannot readily be observed or measured in classical microbial experiments.

Modelling ecological exclusion

Ecological interactions that cause the abundance of an OTU to decrease can lead, *in extremis*, to its exclusion from a sample. Some OTUs may also depend upon the presence of a second OTU in order to exist within a sample. The hypothesis of interaction described thus far does not take into account the possibility of such exclusion or mutual dependence that could affect markedly the numbers of zeros in the sample data and our ability to detect interactions. We therefore expand the hypothesis of interaction framework to entail these exclusion and mutual dependence cases. This can be described by logical clauses with the form $abundance(x, y, s1, app/dis)$, where *app* symbolises a change from 0 to a numbers of counts of OTU $s1$ between the samples x and y, and *dis* symbolises a change from a positive number of counts to 0. Significance of change is equally assessed using a χ^2-test of independence. The theory is expanded as:

$$
\begin{aligned}
effect\_up(s1, s2) \quad \text{if:} \quad & \begin{cases} abundance(x, y, s2, up) \quad or \quad abundance(x, y, s2, app) \\ presence(s1, x, no) \\ presence(s1, y, yes) \end{cases} \\
effect\_down(s1, s2) \quad \text{if:} \quad & \begin{cases} abundance(x, y, s2, down) \quad or \quad abundance(x, y, s2, dis) \\ presence(s1, x, no) \\ presence(s1, y, yes) \end{cases}
\end{aligned}
\tag{IV.2}
$$

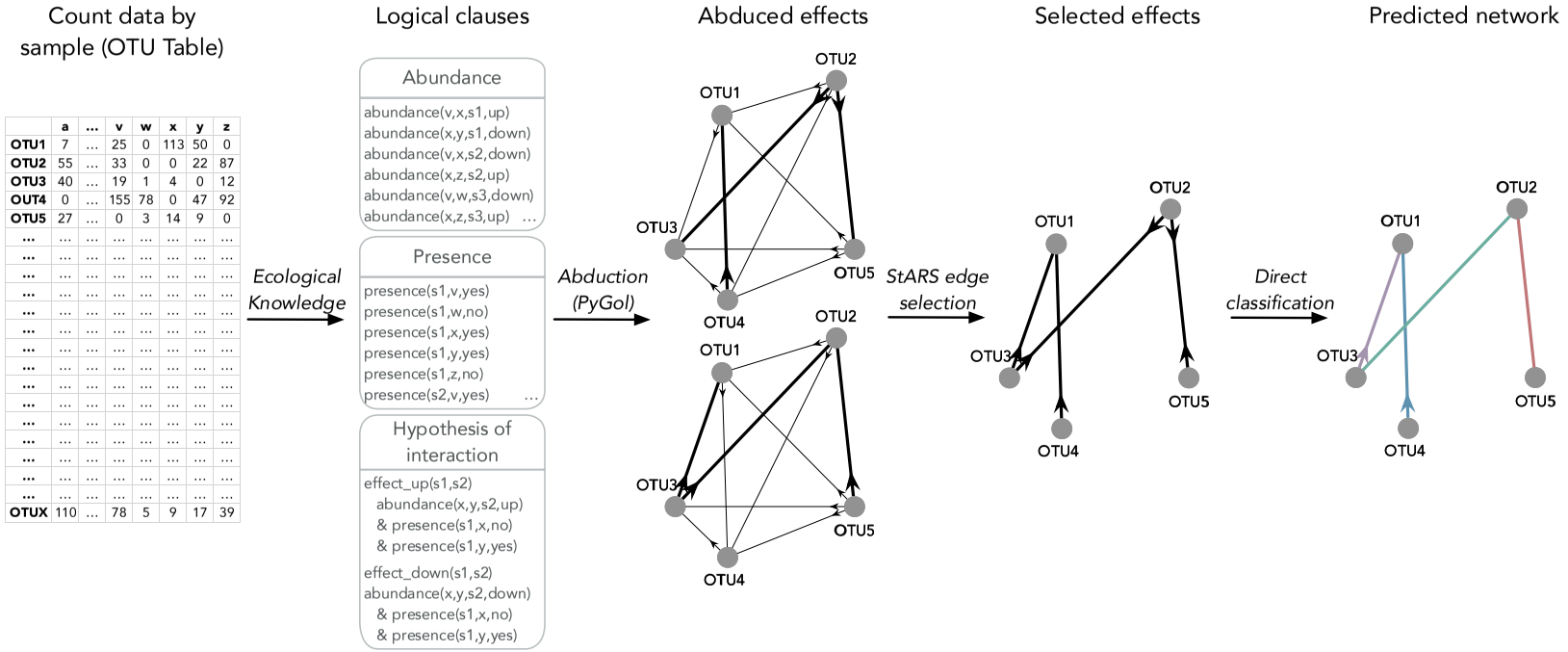
It is important to note that this 'with exclusion' formulation of the theory considers that the effect caused by $s1$ on $s2$ or vice versa, is consistent irrespective of whether the result is an up/down or app/diss. This explicitly includes the possibility that an interaction can cause both a reduction in the abundance of an OTU and ultimately its exclusion. This formulation of the theory with exclusion also allows the I statistic to be computed as previously described.

Implementation of the abductive process for inferring ecological interactions

A full description of the logical process of abduction and of A/ILP is not given here, but can be found in Kakas et al. (1992) and Muggleton (1995)(Kakas *et al.* 1992; Muggleton 1995). Rather we detail the specific implementation of abductive network inference in a new R package, InfIntE (INFerence of INTeractions using Explainable machine learning, Figure IV.1. InfIntE parses the OTU count data into logical clauses using the R base package tools (R Core Team 2022). The logical clauses and hypotheses of interaction are run in PyGol, which is a high-performance implementation of A/ILP newly developed for Python, which has very short execution times (Varghese *et al.* 2022; Varghese & Tamaddoni-Nezhad 2022). InfIntE uses the R package, reticulate (Ushey *et al.* 2022), to provide the logical clauses to PyGol and then retrieve the abduced compression values and I statistic. PyGol generates bottom clauses, as seed examples, and finds the abducible effects on OTU abundance covering the given example with maximum compression. Model selection by StARS is then conducted in the R pulsar package (Müller *et al.* 2016). A custom pulsar function uses the bottom clause to abduce subsets of the OTU table and then retrieves the interaction networks along the λ path. This whole pipeline is performed in InfIntE as a single R function and typically takes 3 to 4 hours to execute for a dataset consisting of 80 OTUs and 60 samples.

2.2 Experiment 1: Generating synthetic, ecological-like data for verification

Many methodologies have been proposed to simulate the effect of interactions in metabarcoding data, based on the assumption that interactions cause a change in the sequence counts of OTUs involved. We use the broadly-accepted models for generating ecological interaction data proposed by Weis et al. (2016)(Weiss *et al.* 2016; Tackmann *et al.* 2019). These models simulate changes in counts in an OTU table, caused by specific interaction types that we expect to be able to detect and classify with InfIntE. The Weiss et al. (2016) models produce the OTU tables for the counts of p non-interacting OTUs over n samples using a lognormal distribution (Shoemaker *et al.* 2017). These counts are then forced to increase or decrease as a function of the counts of the interacting OTUs, modulated by a strength of interaction, s. The generated OTU tables consist of simulated abundances of the p OTUs simulating either amensalism, commensalism, competition or mutualism interactions following the abundance modifications proposed in Faust et al. 2012 (Faust & Raes 2012). Each simulated interaction type will have a different effect



	a	...	v	w	x	y	z
OTU1	7	...	25	0	113	50	0
OTU2	55	...	33	0	0	22	87
OTU3	40	...	19	1	4	0	12
OUT4	0	...	155	78	0	47	92
OTU5	27	...	0	3	14	9	0
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
OTUX	110	...	78	5	9	17	39

Figure IV.1: Schematic diagram representing the InfIntE pipeline. The pipeline performs: conversion of abundance data contained in the OTU table to logical clauses, based upon our ecological knowledge; abduction of interaction effects from the logical clauses using PyGol; selection of important edges using StARS; and, the direct classification of interaction types depicted by the different edge colors. Arrows show the direction of effects or interactions. Edges without arrow in the network represent non-directed interactions.

on the abundance of the OTUs involved in the interaction. For example, if a mutualism interaction is simulated, the abundance of both OTUs will be increased in the samples where they co-occur. To introduce compositionality to the data, the relative counts of an OTU in a sample are used as a probability to sample a multinomial distribution at a common sample size.

The number of samples is an important variable in network inference (Berry & Widder 2014). As a consequence, we generate OTU tables with different number of samples, n, to assess the effects of sampling effort on InfIntE performance. For each sample size, $n = 20, 30, 40, 50, 60, 70, 80, 90$, we create three OTU tables with strengths of interaction, $s = 2$, 3 and 5, and $p = 80$ OTUs to obtain a total of 72 OTU tables mixing the four types of interactions. The number of $p = 80$ OTUs was chosen to reflect the number of abundant OTUs typically observed in our metabarcoding microbial datasets from agriculture.

Inference and detection of interactions from simulated data

InfIntE was used to infer interaction networks for each of the simulated OTU tables for the hypotheses of interaction both with and without exclusion. The area under the receiver operating characteristic curve (AUC)(Fan *et al.* 2006) was then evaluated. The AUC was treated as a measure of how well the tool detected interactions in the simulated datasets that we knew to be real, i.e. present in the dataset, or false, i.e. not present in the dataset. The interaction inference was also done using the statistical inference tools, SparCC (Friedman & Alm 2012) and SPIEC-EASI glasso (Kurtz *et al.* 2015), to provide a performance comparison for interaction detection between our logical inference approach and current statistical networks inference tools that have been broadly adopted in microbial ecology. The SparCC inference was done in FastSpar v1.0.0 (Watts *et al.* 2018) and SPIEC-EASI glasso inference was run in the R package SpiecEasi v1.1.2, both with their respective default settings. The I statistic was computed for the with and without exclusion hypotheses in InfIntE. SparCC correlations were obtained directly and SPIEC-EASI correlations were obtained from the inverse covariance matrix at $\lambda = 0$. Given that these three different tools produce different kinds of interaction networks, either classified interaction or correlational networks, the largest value of I or correlation, obtained for each pair of OTUs, was used for comparing the performance of the statistics used by the different tools.

Evaluating the accuracy of interaction detection and classification in simulated datasets

For the two hypotheses of interaction, without and with exclusion, the accuracy of InfIntE and the StARS procedure to detect simulated interactions, was computed using the function:

$$Accuracy = \frac{TP + TN}{N} \tag{IV.3}$$

where true positives, TP, are the true real simulated interactions detected by the StARS selection, TN are the true non-interacting pairs of OTUs within the simulated dataset, and N is the total number of possible interactions that might

exist in a fully saturated network. The evaluation of the interaction selection of SparCC was performed using the default bootstrapping procedure with 999 permutations. The SPIEC-EASI pipeline uses the StARS procedure to select important interactions as a function of edge stability. The default parameters of StARS model selection in SPIEC-EASI are 20 subsamples and a stability threshold of 0.05.

2.3 Experiment 2: Inferring networks from real data

We used InfIntE to reconstruct the ecological interaction networks on leaves sampled from the grapevine (*Vitis vinifera* in 9 vineyards in France. Our goal was to detect and classify all interaction types amongst amplicon sequence variants (ASVs) within the leaf phyllosphere. ASVs are a high resolution type of OTUs that cluster the sequences as a function of possible sequencing errors (Callahan *et al.* 2017). Fungal ASV tables were constructed from samples of healthy and symptomatic parts of vine leaves. In total, 60 leaves were sampled in each vineyard, giving a total of 534 samples in the dataset. The symptomatic part of the leaves were identified as having lesions caused by the downy mildew, *Plasmopara viticola*. Sequencing was performed using pairs of primers to barcode the fungal communities. A bioinformatic pipeline, including dada2 (Callahan *et al.* 2016) and posterior filtering, was applied to obtain the final ASV tables. The filtered fungal dataset contains abundance information for 650 ASVs. Automatic taxonomic assignment using the UNITE all eukaryotes v8.3 database (Abarenkov *et al.* 2021) did not assign all ASVs to taxa. As a consequence, taxonomic assignment of non-assigned ASVs was done in the BLAST utility of the NCBI platform (Altschul *et al.* 1990). The ASV sequence alignment with highest score was selected. Where an ASV sequence had the same maximum score for more than one species in the database, taxonomic assigment was done to the genus level. The abundance of *P. viticola* was assessed by qPCR as an absolute, non compositional count. Change in the abundance of *P. viticola* was then computed using the count values obtained across the different samples. The *up/down* was considered significant if the logarithmic absolute amount of *P. viticola* DNA differed by 0.05 between samples. A full explanation and description of the sampling design and bioinformatic pipeline production of the ASVs is given in Barroso-Bergada et al. (2022).

Inference using InfIntE of the grapevine phyllosphere data was performed independently for each vineyard. The network inference was computed for only the top 80 most abundant ASVs and the PCR abundance data of *P. viticola*. All 60 samples from each vineyard were used for the inference.

Evaluating the significance of interactions using predictions of change

Most microbial community interactions are unknown or poorly understood. There is therefore no complete and understood microbial network that might be used to evaluate the accuracy of the interactions detected and classified by InfIntE, of which we are aware. It is necessary, therefore, to evaluate the significance of inferred interactions using the data itself. Changes in OTU abundance, caused by an ecological interaction, can be used to compute predictive accuracy. InfIntE does this using a k-fold cross-validation that predicts abundance changes using the

I statistic. Abundance change observations are randomly divided into 5 equal size, 20%, folds of the dataset. Interaction inference is then performed using InfIntE on 4 randomly selections of these folds, leaving one fold for validation. The I statistic for the effects of OTUs on, for example, $s1$ in sample y, are retained and the $up/down$ is predicted from the I sum value of effect up and effect down. Predictive accuracy is then tested by computing the number of correctly predicted abundance changes across the validation fold.

2.4 Statistical Analysis

All statistical analyses were conducted in the R v4.1.3 (R Core Team 2022). Plots were made using ggplot2 v3.3.5 (Wickham 2016) and cowplot v1.1.1 (Wilke 2020). The AUC of each inferred network was measured using the using the pROC package v1.18.0 (Robin *et al.* 2011).

3 Results

3.1 Experiment 1: Generating synthetic, ecological-like networks

Modelling exclusion increases I statistic predictive power

The InfIntE AUC for the I statistic was higher when interactions were inferred with exclusion than without exclusion (Figure IV.2A). This difference became ever greater for datasets with higher strengths of interaction and for smaller sample sizes. Sample size had an important effect on the AUC, independent of the strength of interaction and the hypothesis of interaction used, plateauing at about 60 samples. The InfIntE AUC varied by up to 20% between hypotheses without and with exclusion and from 20 to 90 samples. The AUC obtained for the InfIntE I statistic with exclusion was similar to the AUC obtained by the correlation values computed in SparCC or in SPIEC-EASI (Figure IV.2B). The AUC for SparCC and SPIEC-EASI, varied with the number of samples. The strength of simulated ecological interaction had a negligible effect. No significant difference was found in AUC values between the three network inference tools.

Accuracy of InfIntE detection of interactions increases with sample size

The accuracy of InfIntE with exclusion using StARS model selection varied significantly with the sample size (Figure IV.2C), increasing to a plateau at approximately 60 samples. InfIntE accuracy did not change with the strength of simulated interaction. SparCC and SPIEC-EASI showed levels of accuracy that did not depend on the number of samples, but SPIEC-EASI accuracy was higher than SparCC. InfIntE accuracy was lower than SPIEC-EASI and SparCC at small sample sizes, but was comparable to both at higher sampling effort.

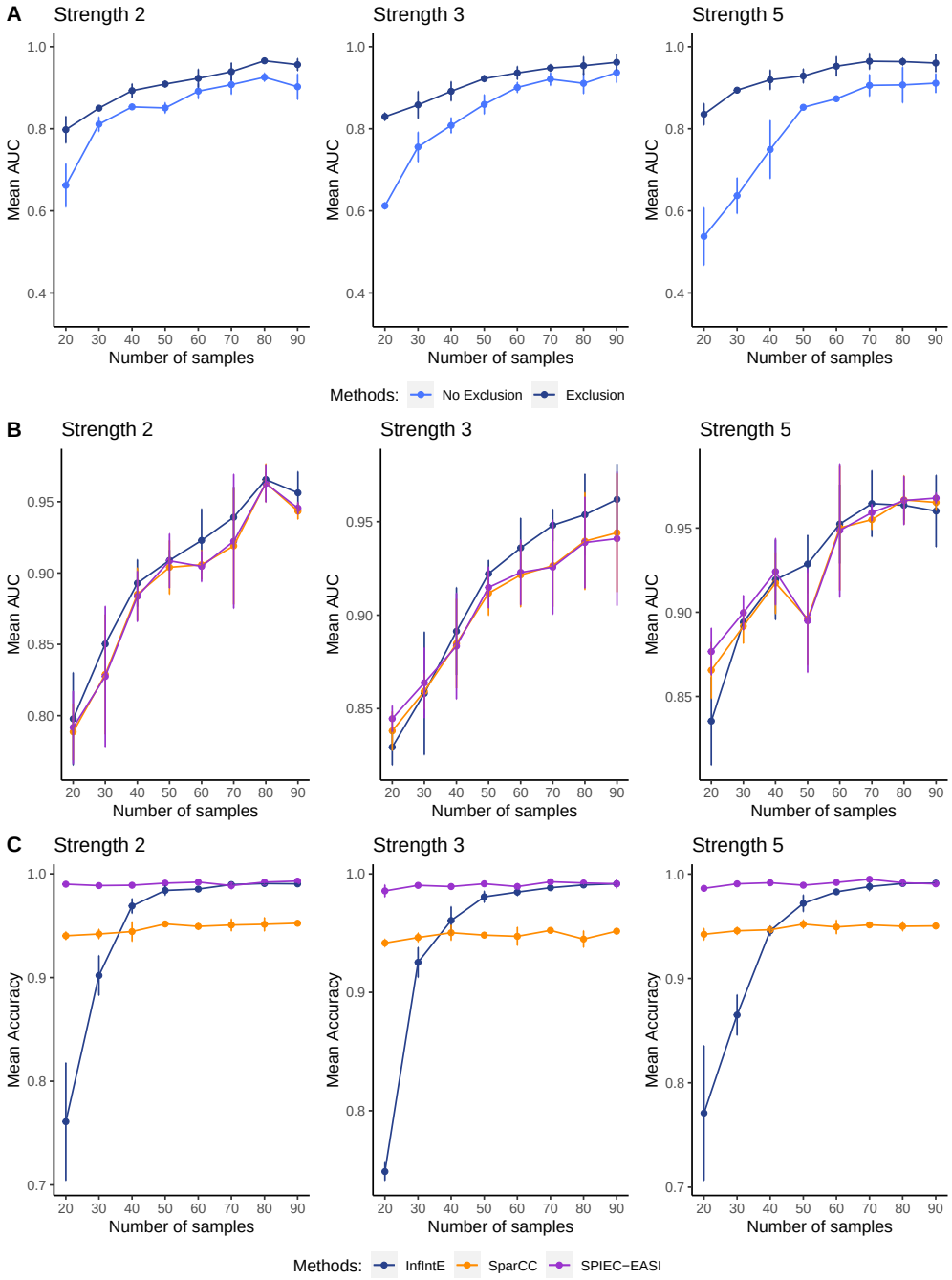


Figure IV.2: Relationship between number of samples and interaction inference performance for different strengths of interaction. Datasets were computer-generated simulating four different interaction types: amensalism, commensalism, competition and mutualism. **A:** Area under the roc curve values (AUC)(Fan *et al.* 2006) obtained by I statistic with and without exclusion. Larger AUC values represent better specificity and sensitivity in interaction detection. I statistic is used by InfIntE as a numeric measure of interaction. **B:** Area under the roc curve values (AUC) obtained by InfIntE's I statistic and SparCC and SPIEC-EASI correlation like measures. InfIntE used the hypothesis of interactions including exclusion. SparCC and SPIEC-EASI were executed with default settings. **C:** Accuracy of interaction detection computed as described in section 2.2. InfIntE used the hypothesis of interactions including exclusion. SparCC and SPIEC-EASI were executed with default settings.

InfIntE classification of simulated interactions

The classification of interactions performed by InfIntE at a fixed sample size of 60, was dependent on the interaction type simulated (Figure IV.3). More than 90% of the synthetic commensalism interactions were detected, and of all of these were then correctly classified. Most mutualism interactions where also detected by InfIntE, and the majority of these were correctly classified. Those that were wrongly classified were classed as commensalism interactions. Slightly less than 20% of competition interactions were detected by InfIntE. The great majority of these were classified correctly, when detected. Synthetic amensalism interactions were not detected by InfIntE. The InfIntE inference produced low numbers of false positives. Where these occurred, all detected links were classified as commensalism.

SparCC detected almost all synthetic interactions simulated as mutualism, commensalism or competition. About 25% of the simulated ammensalism interactions were also detected. This detection came at the cost of an elevated rate of false positive detections compared to InfIntE. SPIEC-EASI showed detection performance similar to InfIntE, detecting the majority of synthetic commensalism and mutualism interactions. It had greater difficulty in detecting synthetic competition and still poorer performance in detecting amensalism. SPIEC-EASI produced very low rates of false positives.

3.2 Experiment 2: Inferring complex networks, the Dark Web, from real data

Structure of interaction types in the foliar networks of grapevine

The interaction networks of each of the nine vineyards (Figure IV.5) were predicted to comprise all the different types of ecological interaction (Derocles *et al.* 2018), Table IV.1). The interaction networks did not show unconnected sub-networks, and the total number of interactions varied from 551 to 1410, depending on the vineyard. The predominant interactions in all 9 vineyards were commensalism and competition, each comprising at least 30% of the total interactions in most networks. Commensalism accounted for 74.5% and competition 13.7% of the interactions in one vineyard. Amensalism interactions never made up more than 17% of interactions and mutualism varied from 2.5 to 8.7%. Interactions classified as predation were not found at all in 2 vineyards and never represented more than 0.7% of the total inferred interactions. When the capacity of the inferred networks to predict abundance changes was evaluated using k-fold cross-validation, predictive accuracy was estimated to be approximately 75% of the observations in the test fold (Figure IV.4), when at least 50% of the sample dataset was used for inference. No difference in accuracy was found between the vineyards for a given standard percentage of the sample dataset.

Identifying potential biological control agents using InfIntE inference

A total of 20 fungal ASVs, corresponding to 12 different species and 2 genera, were identified as potential antagonists of *P. viticola* (Table IV.2). Of these, five fungal

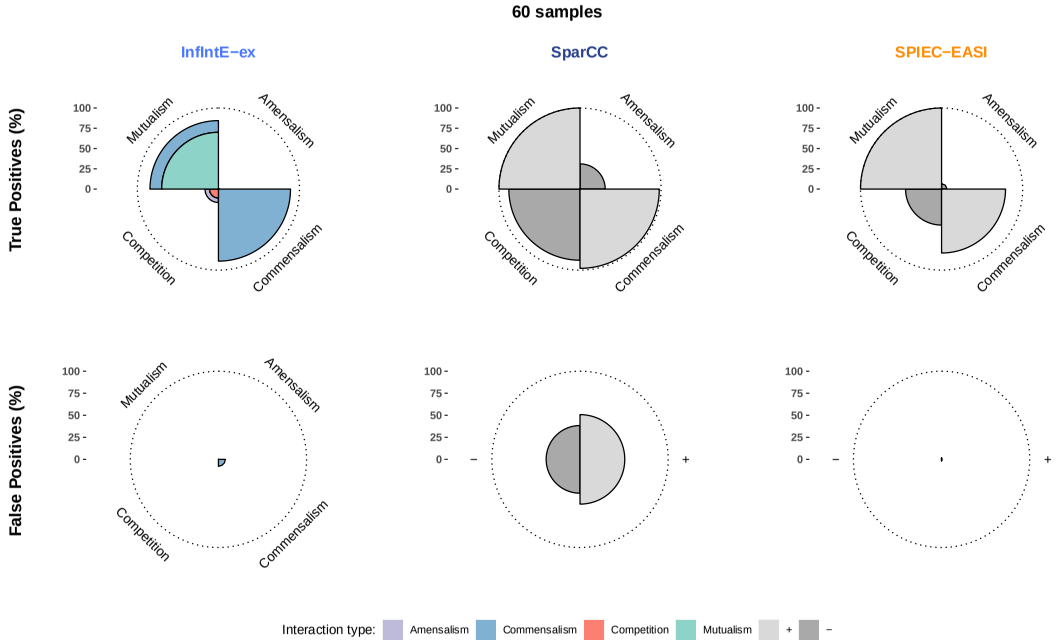


Figure IV.3: Nightingale rose charts comparing the percentage of correct interaction classification by types. OTU tables were synthetically generated simulating groups of 60 replicated samples mixing four different types of interactions: amensalism, commensalism, competition and mutualism. Inference of interactions was performed using InfIntE, SparCC and SPIEC-EASI. Charts in the top row show the percentage of each interaction type correctly detected by each tool. Bottom row charts show the percentage of false positives proposed by each tool over the total possible false positives. Each petal of a rose chart is colored as a function of the classification of the detected interaction type given for each inference tool. InfIntE automatically classifies interactions as amensalism, commensalism, competition and mutualism while SparCC and SPIEC-EASI return positive (+) or negative (-) associations. InfIntE correctly detects and classifies most mutualism and commensalism as well as around 20% of competition interactions with few false positives. SparCC detects most interactions at the expense of a large amount of false positives. SPIEC-EASI has a similar performance to InfIntE, but without interaction type classification.

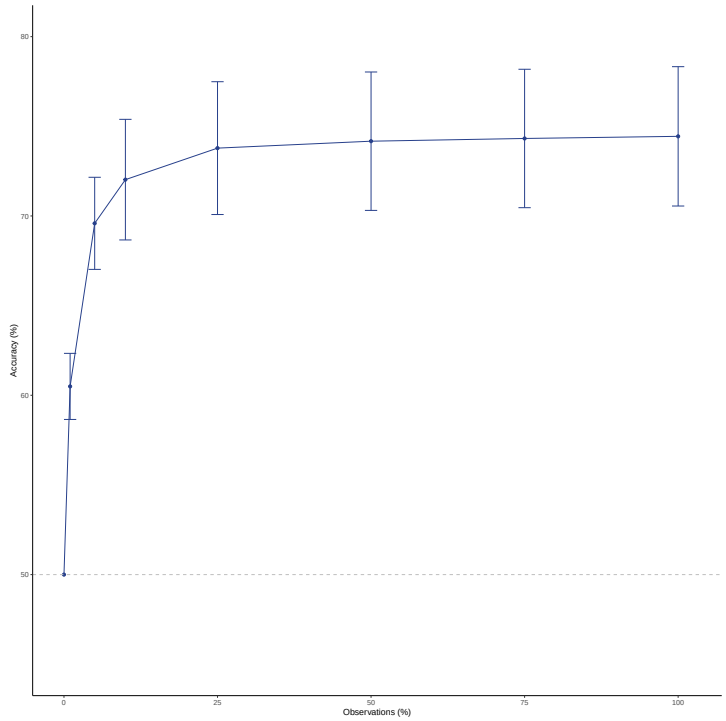


Figure IV.4: Accuracy of abundance change prediction in grapevine metabarcoding data as a function of the number of observations used for the inference. Each dataset consisted of 60 samples. Each point presents the mean accuracy of prediction for different fold combinations of the 9 vineyards

ASVs from the genera *Alternaria* and *Fusarium genus* have already appeared in the literature as *P. viticola* antagonists (Musetti *et al.* 2006; Ghule *et al.* 2018). These ASVs were classified to have competition interactions with *P. viticola* in at least one of the vineyard networks. The fungal species, *Aureobasidium pullulans*, was also inferred to have a competition interaction with the pathogen, and has been described antagonist of *P. viticola*(Harm *et al.* 2011). ASVs corresponding to the genera *Cladosporium*, *Phlebia*, *Sporobolomyces* and *Vishniacozyma* have not previously been described as antagonists of *P. viticola*, but have been identified as antagonist of pathogens in other foliar systems. ASVs assigned to *Mycosphaerella tassiana* and two species from the *Filobasidium* genus were also classified as antagonistic to *P. viticola* by inference, but these are not described as antagonists in the literature and represent new potential biocontrol agents.

4 Discussion

Our work shows that it is possible to reconstruct and then explore the dark web of the grapevine foliar microbiome, through the direct classification of a diversity of interactions using explainable machine learning. The combination of A/ILP and simple hypotheses for ecological interactions, defined *a priori* and coded within the InfIntE R package, works satisfactorily. It accurately detects and classifies computer generated interactions when enough samples are available. We use it to detect, classify and thereby infer interaction types from real DNA sequence data

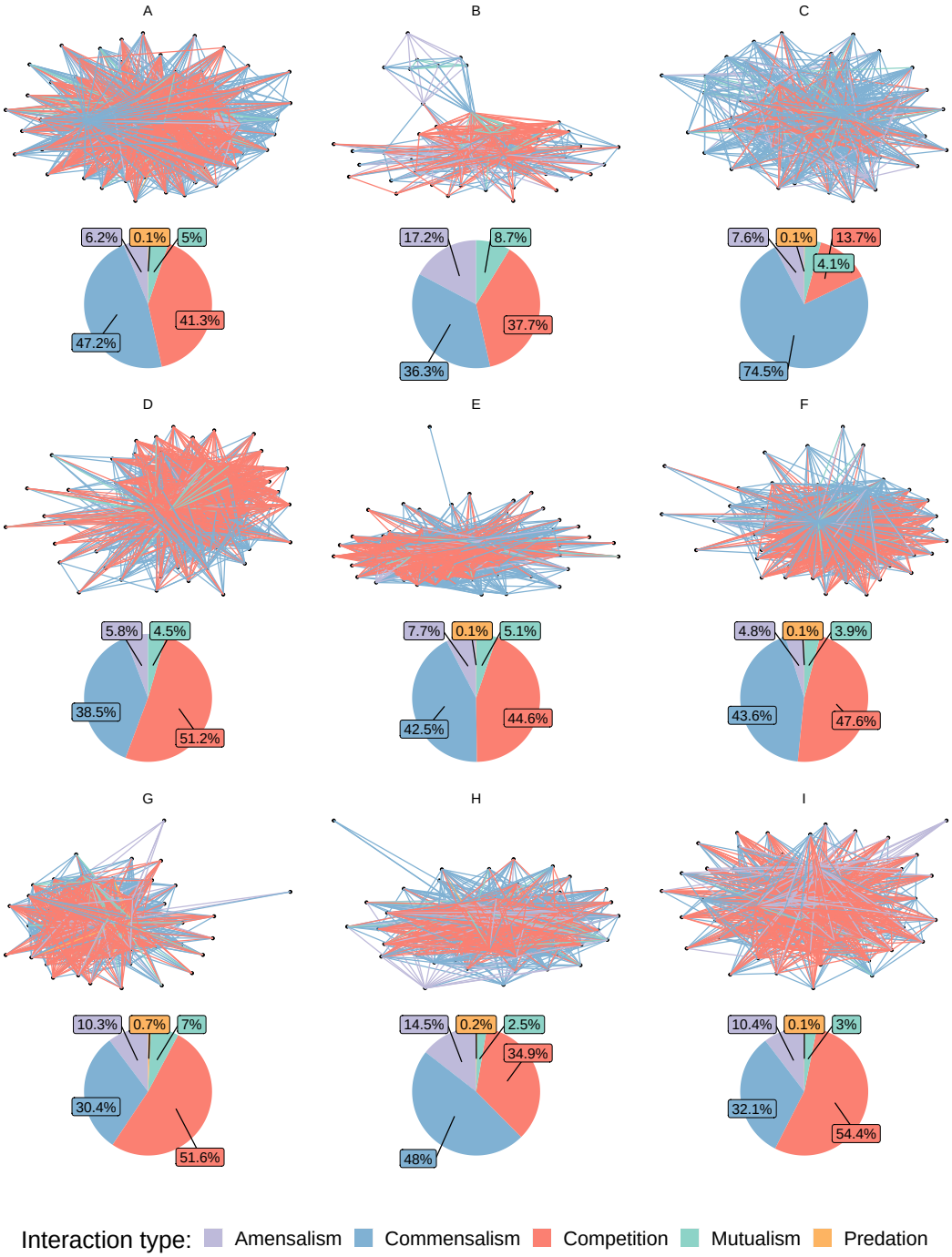


Figure IV.5: Interaction networks predicted for each of the 9 different vineyards in the dataset, inferred using InfIntE. Each vineyard dataset was composed by 60 samples. The edge colours follow the interaction typology. The pie chart associated with each network indicates the relative percentage of each interaction type in the network

for validation against the literature, reflecting well the interactions that have previously been proposed (Table IV.1). This first attempt to use explainable machine learning to infer microbial interaction networks is a first step towards making the link between generic ecological hypothesis that scientists define and the inference of specific types of interaction within the microbiome that can be validated through experimentation. This showed that the dark web of the microbiome is diverse, being composed of a range of interaction types consistent with the hypotheses of interaction we have used. These include antagonistic interactions that might be used to identify potential biological control agents of *P. viticola* within the foliar microbiome of grapevines.

The work also shows potential for improvement to the learning by improving the description of the existing hypotheses and the definition of new hypothesis for types of interaction we have not considered here. The explainable approach of A/ILP emphasises the description of ecological interactions. The interaction hypotheses we have defined describe the process by which two species interact, contingent on the way that their respective abundances change, using simple ecological descriptions (Faust & Raes 2012; Derocles *et al.* 2018). In our work we extended the definition of these hypotheses to include the ecological process of exclusion, whereby the action of one species can cause the exclusion of the other, and found an improvement in the learning with a marked increase in the power to discriminate true interactions. This serves as a good indication that our simple generic interaction hypothesis are only one representation of ecological interaction and could be further improved to infer interaction networks which reflect better the ecological reality of microbial interactions. The work of Bohan et al. (Bohan *et al.* 2011) showed that A/ILP predation interaction hypothesis, which included species traits, produced explainable food web networks for invertebrates. The networks that are inferred by A/ILP are therefore dependent upon and reflect the quality of our ecological knowledge and theory for the hypothesis of interaction, making InfIntE a potentially valuable tool for unravelling the Dark Webs of different ecosystems.

There is no standard dataset in microbial ecology, of which we are aware, for evaluating network inference tools (Röttjers & Faust 2018). In this paper, we used the models of Weiss et al. (Weiss *et al.* 2016) to generate synthetic datasets of ecological-like interactions to evaluate and test the combined performance of InfIntE and our simple ecological rules. While these datasets could be criticised for being simplified approximations of the different typologies of interaction, as changes in the abundance of simulated microbial taxa (Faust & Raes 2012), our analysis suggests that the I statistic, computed in InfIntE, was able to discriminate interactions of different types accurately where sufficient samples were available. All approaches to inferring networks are sensitive to sample size (Hirano & Takemoto 2019) and the appropriate number of samples is a matter of considerable debate in the literature. For the simulated datasets, we found that InfIntE network inference and accuracy became stable at around 60 samples per netwrok sharing the same abiotic conditions. The strength of the simulated interactions seems not to have an important effect on the performance of the A/ILP inference. A change in abundance of around 20% of the total OTU abundance was sufficient to detect an interaction, and greater changes in abundance did not have an

important effect on the probability of detection.

We find that where the number of samples was large enough, InfIntE and our interaction hypotheses could detect simulated interactions with an accuracy comparable to SparCC and SPIEC-EASI, that use correlation-like statistics. Both SparCC and SPIEC-EASI can detect interactions to this level of accuracy at 20 samples, suggesting that they may be a better option for network inference at reduced sampling effort, but with the consequence that link classification will need to be performed by experts based on their knowledge in microbiology and pathology. Within each interaction type, the InfIntE approach detects the simulated interactions with a probability similar to those that SparCC and SPIEC-EASI infer to be a possible interaction of unknown type. Moreover, the InfIntE approach does this with many fewer false positives than SparCC and a similar level of false positive to SPIEC-EASI. Subsequent classification by InfIntE is dependent upon the type of interaction simulated, however. Competition and amensal interactions, which can cause the detrimental exclusion of the interacting OTUs, present difficulties both of detection and of classification. This result is consistent with the findings of Weiss et al. (Weiss *et al.* 2016), who have found that the detection of simulated detrimental interactions is difficult for all inference procedures due to the loss of information that comes with the exclusion of one or both interacting taxa. InfIntE detects detrimental interactions less well than SPIEC-EASI, and while incorporating exclusion as part of the hypothesis of interaction does improve the accuracy of detection, our InfIntE approach has still lower detection performance. We would note, however, that the competition interactions that are detected by the InfIntE approach are classified correctly to a high probability and with a near zero rate of false positives. This suggests that where competition interactions are detected and classified in a dataset, we would expect this classification to be accurate even as some competition interactions are not detected. In general, given this early stage in the development of A/ILP approaches for learning microbial interactions, the InfIntE approach detects interactions to an accuracy similar SparCC and SPIEC-EASI. It also does so with the benefit of direct, automatic classification, which it achieves with good accuracy for the simulated datasets.

The application of the InfIntE approach to real metabarcoding data for the fungal microbiome of grapevine leaves yielded networks that varied across the nine vineyards of the dataset, only being shared a single interaction by the nine vineyards. This result is consistent with the few consensus interactions found by Barroso-Bergada et al. (2021) along different vineyard plots. Each network showed a diversity of interaction types, including interactions that are typically dark and difficult to observe using classical ecological approaches. The most frequently found interaction types were classified as commensalism and competition interactions that may be based on energy tranfer chains (Tshikantwa *et al.* 2018) or the exploitations of resources, such as space (Lloyd & Allen 2015). Predation interactions were rarely classified, and this may be due to problems of detection with this type of interaction being masked by other interaction types (Derocles *et al.* 2018). The inclusion of qPCR data for the grapevine pathogen, *P. viticola*, allowed the prediction of subnetworks centred on this disease-causing agent, with a view to understanding its ecology and the potential for management. While chemical pesticides are commonly used to protect grapevine plants from *P. viti-*

Table IV.2: Potential *Plasmopara viticola* antagonists found by InfIntE. The table shows the fungal species found to have a potential interaction able to reduce the abundance of *P. viticola*. A bibliographical search in Google Scholar, Pubmed and Science Direct was conducted to identify whether potential antagonists have previously been described as biocontrol agents of *P. viticola* or other pathogens in the literature. The keywords used for the search were the name of the potential antagonist, "*Plasmopara viticola*", "biocontrol" and "antagonist". Those taxa identified with an asterisk were not automatically assigned to a taxonomic grouping in UNITE and required manual curation and assignment using BLAST. When there was more than one OTU assigned to the same species having the same interaction it is noted with xn, where n is the number of OTUs

Name	**Vineyard**	**Interaction**	**Bibliography against plasmopara**	**Bibliography biocontrol**
Cladosporium delicatulum	I	competition		Kohl et al. 2019; Baharvandi et al. 2015; Becker et al. 2020
Mycosphaerella tassiana	I	competition		
*Alternaria sp.**	A	amensalism	Mussetti et al. 2006	
*Alternaria alternata**	I	competition	Mussetti et al. 2006, 2007	
Alternaria brasicae	B	competition	Duhan et al. 2021	
*Aureobasidium pullulans**	I	competition	Harm et al. 2011	
Filobasidium chernovii	Ix2	competition		
*Filobasidium magnu*m*	D	competition		
*Fusarium sp.**	A, B, E	competition	Ghule et al. 2018; Bakshi et al. 2001	
Phlebia rufa	E	amensalism		White and Boddy 1992
Sporobolomyces roseus	Ix3	competition		Janisiewicz et al. 1994; Filonow et al. 1996;
*Sporobolomyces pararoseus**	A,G	competition		Li et al. 2017 (in grapes)
Vishniacozyma victoriae	B,C	amensalism, competition		Gramisci et al. 2018; Lutz et al. 2020
Vishniacozyma carnescens	D	amensalism		Becker et al. 2020

cola in conventional vineyards, the potential for developing resistance to fungicides and the needs of organic growers has driven a search for alternative and or complementary methods of pest control. Our expectation for this disease is that *P. viticola* interacts with the microbiome during infection of grapevines, with some members of the microbial community facilitating the invasion of the pathogen and still others acting as antagonists (Thambugala *et al.* 2020). We used InfIntE and the ecological hypotheses of interaction to identify those taxa that might act as *P. viticola*-antagonists and might therefore represent candidate biological control agents. A total of 14 taxa were hypothesised to be antagonistic to *P. viticola*, with the predominant interactions inferred being of competition and more rarely amenalism. The list of biocontrol candidates included five fungi taxa already identified in the literature as *P. viticola* antagonists (Bakshi *et al.* 2001; Duhan *et al.* 2021; Ghule *et al.* 2018; Harm *et al.* 2011; Musetti *et al.* 2006; Musetti *et al.* 2007). Six of the remaining taxa have also been proposed as biocontrol agents of different pathogens in other foliar systems. We also propose *Mycosphaerella tassiana* and two species from the *Filobasidium* genus as potential biological control agents. These literature validations give us some confidence that the InfIntE approach correctly classifies interactions, allowing us to unravel an approximation of the dark web of microbial interactions and the pathobiome.

Our InfIntE approach introduces the use of explainable machine learning to microbial interaction inference. Explainable machine learning has had a limited use in Ecology due, at least in part, to its often-intensive computational requirements and lower robustness to noise in comparison to statistical learning. The duration of an explainable approach increases exponentially with the size of the dataset, limiting the interactivity and scale of the learning that can be done (Varghese *et al.* 2022). Moreover, explainable machine learning approaches have been largely limited to researchers in the field of logic that has a distinct theoretical framework and suite of methodologies with which Ecologists are unfamiliar. InfIntE uses a new implementation of A/ILP called PyGol, written in Python. Ecologists are more familiar with Python, facilitating its use. PyGol is also much faster than previous A/ILP implementations and appears more robust to noise in datasets. This has seen the time required to run the explainable machine learning we use here decrease from several days to a few hours, for a dataset with 60 samples and 80 OTUs. Most importantly, this promotes a much more interactive, experimental approach to the inference, because the time costs of mistakes in coding are much lower, which has seen our work advance markedly. This increase in speed of computation has also opened up new avenues of research into the microbial Dark Web. Other sources of information, such as functional information from databases like Funguild (Nguyen *et al.* 2016b) or Faprotax (Louca *et al.* 2016), can readily be included in the logical hypothesis in InfIntE to infer interactions with greater descriptive precision, potentially further improving the inference process and our understanding of these systems. Future developments should also consider testing between alternative descriptions of interactions, such as those based on ecological functions or taxonomy. New hypotheses for ecological interactions might also extend to rules that describe effects on more than the two interacting taxa considered here (Weiss *et al.* 2016).

Chapter V

InfIntE: a generic, logic-based inference tool for learning networks in R

1 R package

Explainable Machine Learning (EML) is an alternative approach to the more widely used statistical machine learning, with potential application to a broad range of learning situations (Roscher *et al.* 2020). EML uses domain knowledge to discover new understanding, from data, in a human understandable process (Beckh *et al.* 2021). There has been a significant increase in the use of EML in recent years, as part of an initiative to understand what increasingly complex machine learning models do and thereby increase confidence and trust in the learning. There are two main reasons why EML has not been much used in microbiology and ecology to date. The first was the requirement to have algorithms able to perform computation in a reasonable amount of time while dealing with the inherent noise of biological data. These limitations are beginning to be solved, particularly with programs like PyGol that implements a new approach to hypothesis search, drastically decreasing execution times and offering significant robustness to noise (Varghese *et al.* 2022). The second reason why EML use has not been widely adopted is the specialism necessary to work with logic. Microbial and ecological hypotheses and domain knowledge need to be expressed as logical statements, in abstruse programming languages like Prolog (Wielemaker *et al.* 2012). The concepts, structure and know-how of logic programming represents a steep learning curve that is a significant barrier to adoption. The documentation of logic programming tools is also conspicuous by its absence. Microbiologists and ecologists tend to be analytical generalists, preferring applications that require little programming skill and that can be run using a few commands. Most correlation based network reconstruction tools have an implementation in the statistical software environment R (R Core Team 2022) (e.g. Sparcc, CCLasso, SPIEC-EASI, HMSC, PLN, MPLasso (Kurtz *et al.* 2015; Fang *et al.* 2015; Ovaskainen *et al.* 2017; Chiquet *et al.* 2019; Lo & Marculescu 2017b), implemented as an easy-to-install R package. The learning can normally be launched by a single command, to perform the network reconstruction from an OTU table. These R packages also have documentation detailing both how it is used and the process implemented in each function.

The application of EML to reconstruct interaction networks is described in Chapter IV. To make InfIntE an option that ecologists and microbiologists could consider for inferring interaction networks, it is necessary to provide them with an easy to use tool. InfIntE was therefore encoded as an R package from the start (R Core Team 2022), because it is well used and accepted among all potential users. The package has been developed to have a single function to perform network reconstruction in a single run. This saves users from dealing directly with the logical programming language since InfIntE converts the data to logic statements automatically and these are then parsed to PyGol (Varghese *et al.* 2022) to perform the logical abduction process of possible effects on OTU abundances. The InfIntE R package and pipeline is detailed in Figure IV.1. In order to help any potential user, a vignette (shown in the following pages) was written to detail how to install and run InfIntE. The vignette illustrates the use of InfIntE as a single function, detailing the whole pipeline as a series of steps. The vignette uses the metabarcoding dataset described in the Annex B (Barroso-Bergadà *et al.* 2022).

The InfIntE package can be download from a private GitHub repository with permission, but we will make this repository public once the articles describing InfInE are published.

Vignette InfIntE

What is InfIntE?

InfIntE stands for Inference of Interactions using Explainable machine learning. This package uses abundance data to directly infer ecological interactions using PyGol, an Abductive/Inductive logic program. The interactions inferred are directly classified by its interaction type.

Table of contents

1. Installation
2. Example Data
3. Interaction Inference
4. Step by Step
5. Use of absolute data

Installation

InfIntE and required packages are installed using devtools

```
library(devtools)
if(!"InfIntE" %in% rownames(installed.packages())){
  install_github("didacb/InfIntE")
}
library(InfIntE)
```

One of the steps of the interaction inference is the abduction. Abduction is performed using PyGol. PyGol is written in c. To compile PyGol and obtain the functions for abduction run:

```
load_PyGol()
```

For now PyGol only works in linux environments. If there is nay problem with the compilation be sure that the following linux packages are installed:

- cython
- python-dev

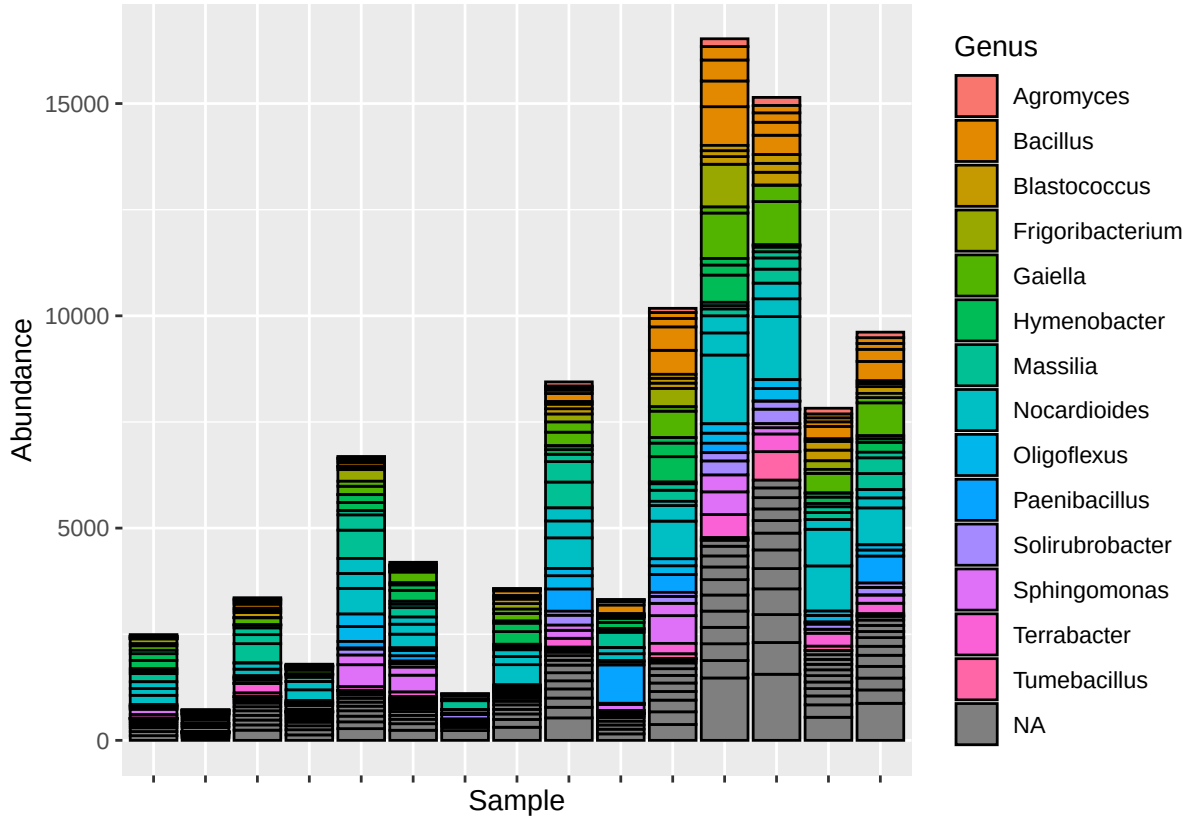
Example Data

To show how InfIntE works we will use wheat foliar fungal ASV data. The data characteristics are detailed here. The ASV data is in phyloseq format. First, let's import and subset the data to obtain a manageable size.

```
#Import data
library(phyloseq)
data("BCM_16S_wheat_phyloseq_filtered_lulu")
wheat_metadata<- sample_data(BCM_16S_wheat_phyloseq_filtered_lulu)

#Keep only green samples from march
selected_samples<- wheat_metadata$Date == "03_18" &
                   wheat_metadata$Specie == "wheat" &
                   wheat_metadata$Variety == "Apa" &
                    wheat_metadata$Tissue == "G"

asv_subset<- prune_samples(selected_samples, BCM_16S_wheat_phyloseq_filtered_lulu)

#Keep only the most abundant ASVs
asv_subset<- prune_taxa(taxa_sums(asv_subset)>1000, asv_subset)
```

The wheat fungal community has many different fungal genus represented

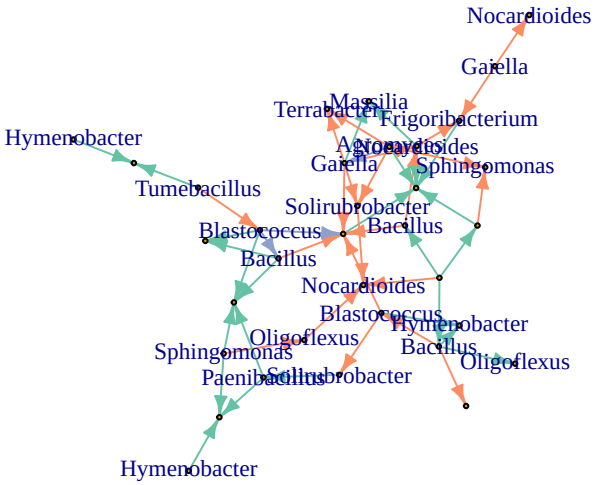
```
library(ggplot2)
plot_bar(asv_subset, fill = "Genus")+theme(axis.text.x = element_blank())
```



Interaction Inference

To infer interactions, InfIntE offers a homonymous function to perform the whole pipeline in a single step

```
library(igraph)
#Infer interactions
interactions<- infinte(otu_tb = otu_table(asv_subset, taxa_are_rows = T),
                       exclusion = TRUE, ncores = 25)

#Get network
network_graph<-graph_from_data_frame(interactions$selected_interactions)

#Change ASV names to genus
V(network_graph)$name<- data.frame(tax_table(asv_subset))[V(network_graph)$name,]$Genus

#Add color to different interactions
library(RColorBrewer)
colors_edges<- brewer.pal(5, "Set2")

E(network_graph)$color<- sapply(E(network_graph)$lnk, function(x){
                                  colors_edges[which(unique(E(network_graph)$lnk)==x)]
                                })
#Plot
set.seed(123)
lay <- layout.kamada.kawai(network_graph)
plot(network_graph, layout=lay, vertex.size=2,
     vertex.label.cex = 0.75, edge.arrow.size=0.5 )
```



Step by step

But what InfIntE exactly does? It uses an hypothesis of interaction written as a logical relation between presence, abundance and effects in abundance.

```
hypothesis<-
c("abundance(C1,C2,S1,up):-presence(C2,S2,yes)&presence1(C1,S2,no)&effect_up(S2,S1)",
  "abundance(C1,C2,S1,app):-presence(C2,S2,yes)&presence1(C1,S2,no)&effect_up(S2,S1)",
  "abundance(C1,C2,S1,down):-presence(C2,S2,yes)&presence1(C1,S2,no)&effect_down(S2,S1)",
  "abundance(C1,C2,S1,dis):-presence(C2,S2,yes)&presence1(C1,S2,no)&effect_down(S2,S1)")
```

Then transforms the ASV matrix into logic clauses related by the hypothesis

```
# Join absolute and compositional data in a table
otu_data <- join_abundances(otu_tb=otu_table(asv_subset, taxa_are_rows = T),
                            absolute_abundance = NULL, depth = NULL)

# All possible pairs of samples
comparisons <- get_comparsions(length(otu_data$samp_names))


# Get head logic clauses
head_clauses <- lapply(rownames(otu_data$otu_tb), function(otu) {
    pos <- which(rownames(otu_data$otu_tb) == otu)
    abundances <- do.call(
      what = otu_data$abundance_function[pos],
      args = list(
        "otu_abundance" = otu_data$otu_tb[pos, , drop = FALSE],
        "comparisons" = comparisons, "depth" = otu_data$depth, "exclusion" = TRUE
      )
    )
    return(abundances)
  })

head_clauses <- unlist(head_clauses)

# Get Body logic clauses
body_clauses <- get_presence(otu_data)

head(body_clauses)
```

```
## [1] "presence(c1,s1,yes)." "presence(c2,s1,no)."  "presence(c3,s1,no)."
## [4] "presence(c4,s1,no)."  "presence(c5,s1,yes)." "presence(c6,s1,no)."
```

It uses PyGol to generate the bottom clause and abduce the effects on the OTU abundance caused by other ASVs. InfIntE renames the ASVs during the abduction optimize process.

```
# Produce bottom clause
bottom_clauses <- get_bottom_clause(otu_data = otu_data,
                                    head_clauses = head_clauses,
                                    body_clauses = body_clauses)

# Abduce effects
```

```
abduced_effects <- abduce(bottom = bottom_clauses, hypothesis = hypothesis)

# Get I values
abduced_effects <- get_I_values(abduced_effects)#Infer interactions


head(abduced_effects)
```

```
##   sp1 sp2         lnk comp
## 1  s1  s1   effect_up 5506
## 2  s1 s10 effect_down  560
## 3  s1 s11   effect_up  417
## 4  s1 s12   effect_up  120
## 5  s1 s13   effect_up  446
## 6  s1 s14   effect_up  430
```

To select the interactions InfIntE uses the pulsar package to run the StARS model selection

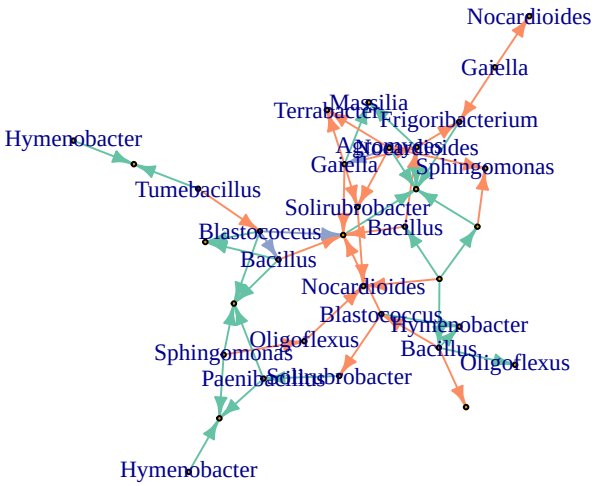
```
# Length observations
mx <- length(bottom_clauses$head)

# Lambda distribution
lambda <- pulsar::getLamPath(max = mx, min = 0, 50, FALSE)

# Pulsar execution
pulsar_output <- pulsar::pulsar(t(otu_data$otu_tb),
  fun = pulsar_infinte,
  fargs = list(lambda = lambda, bottom_clauses = bottom_clauses,
               hypothesis = hypothesis, exclusion = TRUE, otu_data = otu_data),
  rep.num = 50, lb.stars = TRUE, ub.stars = TRUE, thresh = 0.01, ncores = 25,
)

# Format output to dataframe
fitted_model <- pulsar::refit(pulsar_output, criterion = "stars")
interactions <- data.frame(igraph::get.edgelist(
                    igraph::graph_from_adjacency_matrix(fitted_model$refit$stars)))

head(interactions)
```

```
##    X1  X2
## 1 s38  s1
## 2  s9  s1
## 3 s23 s12
## 4 s37 s12
## 5 s30 s14
## 6 s30 s18
```

As a last step InfIntE classifies the interactions by its type

```
# Take values from abduced effects dataframe
interactions <- abduced_effects[paste0(abduced_effects[, 1], abduced_effects[, 2])
                                %in% paste0(interactions[, 1], interactions[, 2]), ]
```

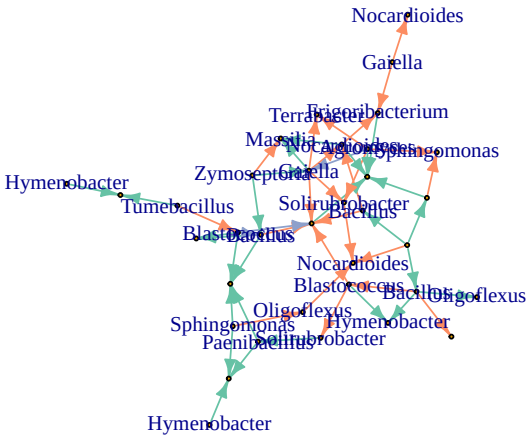
```
# Classify and give back original names
interactions <- classify_interactions(interactions)
interactions <- return_names(interactions, otu_data$otu_names)


#Get network
network_graph<-graph_from_data_frame(interactions)

#Change ASV names to genus
V(network_graph)$name<- data.frame(tax_table(asv_subset))[V(network_graph)$name,]$Genus

#Add color to different interactions
library(RColorBrewer)
colors_edges<- brewer.pal(5, "Set2")

E(network_graph)$color<- sapply(E(network_graph)$lnk, function(x){
                                  colors_edges[which(unique(E(network_graph)$lnk)==x)]
                                })

#Plot
set.seed(123)
lay <- layout.kamada.kawai(network_graph)
plot(network_graph, layout=lay, vertex.size=2,
     vertex.label.cex = 0.75, edge.arrow.size=0.5 )
```



Use of absolute data

InfIntE can also use absolute abundance data, complementing the compositional data obtained from eDNA. In this example we will use the qPCR measurements of the pathogen *Z. tritici* available in the metadata.

```
#Retrieve absolute abundance
absolute_abundance<- t(data.frame(sample_data(asv_subset))[,7,drop=FALSE])
absolute_abundance<- ifelse(is.na(absolute_abundance),0,absolute_abundance)
#Infer interactions
interactions<- infinte(otu_tb = otu_table(asv_subset,
                                          taxa_are_rows = T), ncores = 25,
                       absolute_abundance = absolute_abundance, exclusion = TRUE)
#Get network
network_graph<-graph_from_data_frame(interactions$selected_interactions)
#Change ASV names to genus
zymo.pos<- grep("Zymoseptoria", V(network_graph)$name)

V(network_graph)$name<- data.frame(tax_table(asv_subset))[V(network_graph)$name,]$Genus
V(network_graph)$name[zymo.pos]<- "Zymoseptoria"
#Add color to different interactions
library(RColorBrewer)
colors_edges<- brewer.pal(5, "Set2")
E(network_graph)$color<- sapply(E(network_graph)$lnk, function(x){
                                  colors_edges[which(unique(E(network_graph)$lnk)==x)]})
#Plot
set.seed(123)
lay <- layout.kamada.kawai(network_graph)
plot(network_graph, layout=lay, vertex.size=2,
     vertex.label.cex = 0.75, edge.arrow.size=0.5 )
```



2 Network visualization

Ecological interaction networks are complex representations of an ecological reality, with many different kinds of interactions and taxonomic units potentially being displayed. This complexity means that it can be difficult to explore and obtain appropriate information from ecological networks. Indeed, complex networks have been described as 'hairballs' that are impossible to disentangle. In chapters II, III.5 and IV, several existing tools were examined as part of network analysis. But, analysis is only one part of data exploration. Visualization is also an necessary way to obtain information from networks. Cytoscape is the most widely used tool for network visualization (Saito *et al.* 2012). It is an open source application that offers many plugins to adapt the application and output to user needs. Nevertheless, we have found that the solutions offered in Cytoscape are not suitable for visualizing the networks obtained using EML, for which it is necessary to assess network structure across modified compression thresholds, interactively. We therefore decided to develop an interactive application for EML network visualisation, based on the shiny R package (Chang *et al.* 2021). The application uses the output obtained from InfIntE or other network reconstruction tools, to produce a network graph that can be customised, facilitating network interpretation. The app is intuitive and can be used on any computer with R installed. It can also be used to share constructed networks for feedback from ecologists, thereby further improving network inference by incorporating domain knowledge.

Figure V.1 is a screenshot of the interactive tool. The left-hand panel includes all the visualisation options of the tool. The user can select:

- The interaction network file to visualise. The file has to contain a table with four columns and one interaction per row. The first two columns must be the two interaction OTUs, the third column must be the numeric measure of interaction (e.g. compression) and the fourth the interaction type.
- The taxonomic table to apply to the OTU names. This is specifically designed to use the taxonomies obtained from phyloseq objects.
- The layout to apply to the network. New layout functions can be readily be included to extend network visualisation.
- Whether to show the direction of the interactions. For some interaction types (e.g commensalism), the effect caused to the OTU by the interaction differs. As a consequence, it may be interesting to see the direction of the interaction.
- The threshold of compression (or other metric of interaction importance) to be applied, in order to select those interactions to be shown in the network.
- The interaction types to be shown.
- The thickness of the edge, so as to display the interaction important metric (e.g compression).
- The label size.

- A second network for comparison. Edges will then be colored as a function of the coincidence between the edges of the two networks, in place of their interaction type.

The right-hand panel shows the image of the network, showing the OTU name of each node. The interactions are shown as lines of different color depending on the interaction type. A single-click selects an interaction or OTU, and highlights all the other OTUs and interactions linked to it. A double click on a node of the network can be used to move it around, changing the conformation of the network. The OTU is then deselected by double-clicking. The network image can be saved using the dialogue box under the network image.

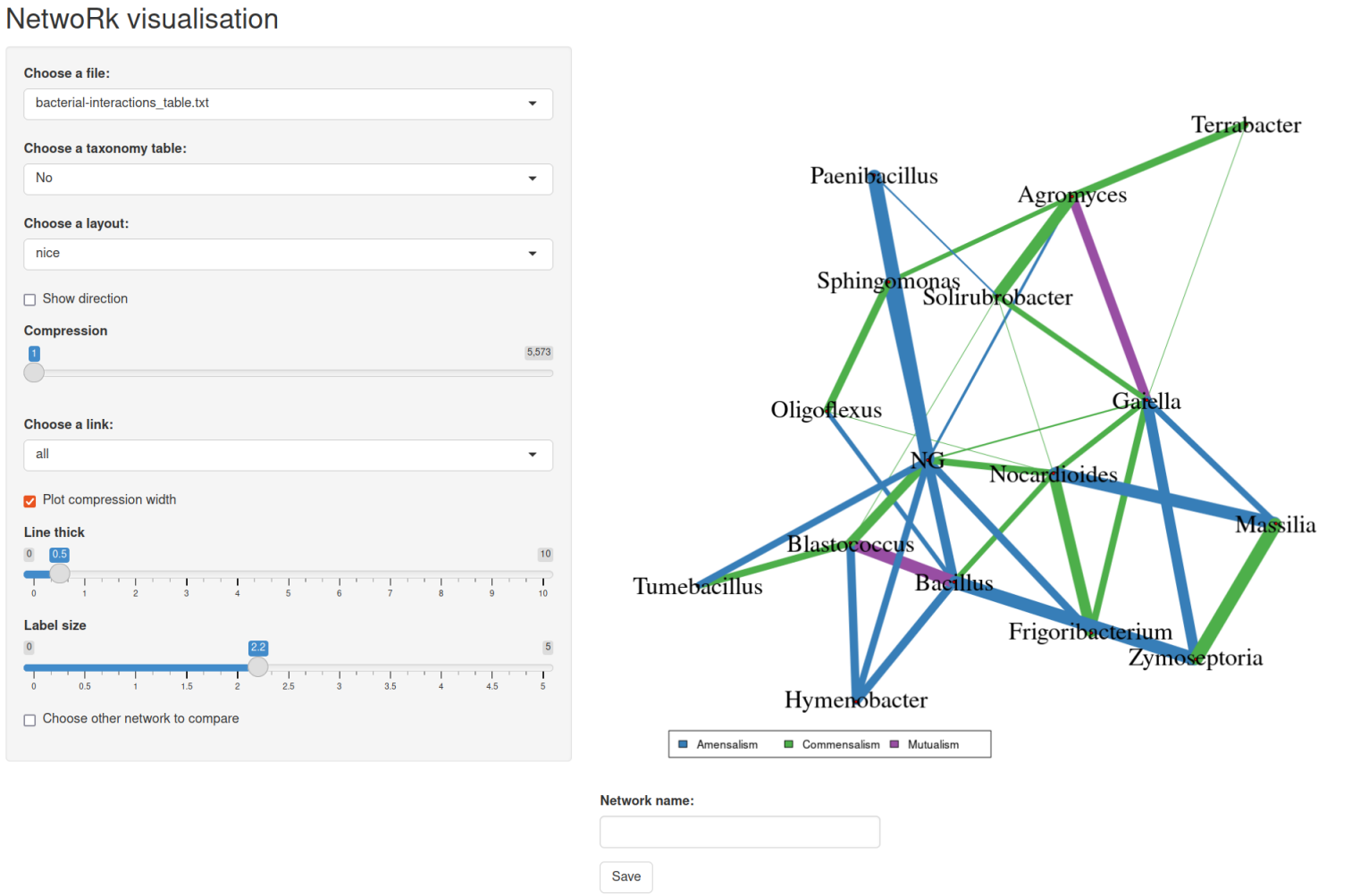


Figure V.1: Screenshot image showing the network visualisation tool. Interactive visualisation options are available in the left-hand panel. The image of the network is displayed at the right-hand panel. The network displayed was predicted in the vignette example. The network image shows the interactions between bacterial OTUs inhabiting wheat leaves

Chapter VI

Discussion

1 General Discussion

This thesis was done as part the Agence National de la Recherche (ANR) Next-Generation Global Biomonitoring (NGB) project. As a consequence, both the project and this Ph.D. share a similar objective of developing a framework for the use of interaction networks for automated monitoring of ecosystems (Bohan *et al.* 2017). This framework was envisioned to have three pillars: 1) the automated sampling and sequencing of DNA; 2) the use of bioinformatic pipelines to obtain OTU abundances; and, 3) network reconstruction and analysis. For the first pillar of the approach, the different research groups in the NGB consortium identified potential monitoring-interesting ecosystems and developed appropriate pipelines for their sampling and DNA sequencing. This work is detailed in Dubart et al., (2021). This thesis was developed to focus on doing the second and third pillars, using data-sets created in the first pillar. Among all the available ecosystems and taxa that were biomonitored in NGB, I focused on microbial communities for two main reasons. Microorganisms are ubiquitous, being present in all ecosystems, and they are difficult to monitor using any technique other than metabarcoding. With that in mind, the main question to answer was whether interaction networks could be used to monitor microbial communities? At first glance, this seems a pretty straightforward question. There are many statistically-based, network reconstruction tools available, and these seem pretty robust (Röttjers & Faust 2018; Dohlman & Shen 2019). It therefore appeared that the answer would come from building correlation networks, using statistical approaches, and comparing them, and thereby the monitoring would be done. But, when one starts to consider this question in detail, things get more complicated.

The first complication appeared around the use of the term interaction. What is a microbial interaction? Plenty of examples and definitions of the diversity of ecological interactions can be found in the literature (Derocles *et al.* 2018; Tshikantwa *et al.* 2018). The complication is whether the networks of correlations produced by statistical approaches capture the evident diversity of ecological interactions in nature. Given the aim of biomonitoring is to take decisions to manage the ecosystems, any reconstructed interaction networks should report interaction change and give enough information to decision-makers for appropriate decision-making. Identifying this initial complication led quickly to many others that could be formulated as research questions. For example: what are the sampling requirements necessary to appropriately represent the interactions of a microbial community; and, how do we detect change in the complex networks that might be constructed? It is here that the principal difficulty of this Ph.D. topic lies. It requires a multidisciplinary approach to answer all these questions. Ecology and Microbiology are the basic domains of research necessary to define and understand microbial interactions. Bioinformatics and Statistics provide the methods to process the metabarcoding data and reconstruct the interaction networks. To this we add the logic required to bridge the gap between our ecological knowledge and network reconstruction. This bridge was the Explainable Machine Learning (EML).

Explainable machine learning uses the understanding from a particular knowledge domain to process information in a human understandable way (Tamaddoni-Nezhad *et al.* 2021). EML made it possible to use the ecological definitions for

different interaction types, taken from the literature, as hypotheses to reconstruct microbial networks. These networks would, in principle, thereby exhibit links for each of the different interaction types hypothesised, giving a richer and fuller insight into the structure and function of microbial communities. This thesis is a first step towards an EML-based network reconstruction approach to biomonitoring and identifies the elements required for robust network based biomonitoring using metabarcode data.

1.1 Can metabarcoding data be used for biomonitoring?

The development of Next Generation Sequencing (NGS) approaches have lead to the establishment of standard pipelines for DNA sampled from the environment (eDNA), extraction of eDNA, sequencing and identification and clustering of the obtained sequences into OTUs. These pipelines also give count information for the abundance of sequences, and therefore species, in a sample. The consensus amongst scientists is that eDNA sequencing with an appropriate set of primers produces an accurate list of the taxa present in the environment (Burki *et al.* 2021). There is no agreement, however, for how accurate the counts are, as measures of abundance, following clustering and denoising of the sequence information. Some authors have stressed the importance of different biases that are introduced during metabarcoding, creating a distorted estimate of community abundances in the sequence counts of the OTUs (Fonseca 2018; Carr *et al.* 2019; Santi *et al.* 2021). Experiments using mock communities have shown that relatively accurate abundance measures can be obtained where appropriate sequencing and bioinformatic pipelines are followed (McGovern *et al.* 2018; Pauvert *et al.* 2019). This is consistent with the tests done with microbial communities as part of the NGB project (Dubart *et al.* 2021). The metabarcoding data used during this thesis followed the recommendations from these authors to obtain abundance information of the different ASVs found in grapevine leaves. In chapters II and VII, I detail how these metabarcoding datasets were obtained using the molecular biology pipelines, including DADA2, which is probably the most widely employed tool for ASV inference. Post-processing of ASVs to discard contaminants and low abundance ASVs that are more prone to be uninformative, used a specifcally designed pipeline (Cao *et al.* 2021). Similar pipelines could be developed to yield abundance information for reconstructing interaction networks, potentially allowing an automated sampling, sequencing and bioinformatic processing (Bohan *et al.* 2017). Obtaining the abundance data is therefore not an impediment, in principle, preventing network-based biomonitoring of ecosystems. Community diversity measures of ASV count data, analyzed in chapters II and VII, also show that microbial communities diverge when there is a change in the ecosystem, due either to a different cropping system or to grapevine disease status. As a consequence, we expect that reconstructed interaction networks should also reflect these changes, bringing a richer insight into the structure and function of microbial communities with ecosystem change.

1.2 Can explainable machine learning be used to infer interactions?

The networks reconstructed with correlation-based tools require interpretation by ecologists and microbiologists to obtain significant knowledge. This interpretation is subjective, being limited by the knowledge of the interpreter, the information available in databases and the literature, and the biases of each person doing the interpretation. Our goal of building explainable networks, useful for biomonitoring, demanded a network reconstruction procedure that avoids this subjective interpretation via a direct classification of the interaction mechanism between all interacting OTUs. The first step of this development, detailed in the chapter III.5, showed that it is possible to use explainable machine learning and a basic definition of interaction to accurately reconstruct networks from cross-sectional, eDNA sequence counts. The procedure used CProgol (Muggleton 1995), an explainable machine learning language for doing Abudctive/Inductive Logic Programming (A/ILP), to abduce the effects on the abundance of OTUs due to a hypothesised interaction. The abduction used the abundance of OTUs, from metabarcoding, together with a simple hypothesis of interaction describing the effect on OTU abundance, to obtain an estimate of Compression (Muggleton & Bryant 2000). The value of compression represents the amount of provided information supporting each effect on abundance. We formulated a framework to calculate a statistic, I, that quantifies the constancy of effect of an interaction on OTU abundance, using the compression values. This used the explicit assumption that to be important, an interaction should have a consistent effect direction on OTU abundance (up or down) over the whole data-set. Compression has previously been used to obtain probabilistic measures of abduced effects in different applications (Bryant *et al.* 2001; Tamaddoni-Nezhad *et al.* 2012). The potential of the A/ILP for reconstructing networks was evaluated by bootstrapping the OTU abundance changes to obtain significance estimates for the I values, thereby discriminating true from false interactions.

This framework was tested using an ecological-like model to generate different interaction types, as proposed by Weiss et al., (2016) (Weiss *et al.* 2016). Interaction datasets were generated by simulating the abundance of non-interacting OTUs and then modifying their abundance to mimic different interaction types, such as competition, commensalism, etc. The simulated interactions are modulated by a variable that defines the strength of the interaction on the abundance. The tests showed that it was possible to detect true interactions, with A/ILP, with an accuracy comparable to SparCC (Friedman & Alm 2012), one of the most widely used statistical network reconstruction tools. The SparCC results were broadly consistent with the findings of Weiss et al., (2016) and Tackmann et al., (2019). However, I found that the simple hypothesis of interaction used for the A/ILP learning was flawed. Detecting absent OTUs, due to the effects of strong negative interactions that led to exclusion, was difficult or even impossible because it was not considered in the hypothesis. The learning framework also illustrated some problems of implementation. The A/ILP abduction, when done in CProgol, took several hours to infer links for a relatively small dataset of 40 OTUs and 50 samples (Varghese *et al.* 2022). Given that a bootstrapping procedure for this size of

data-set also needed at least 100 re-samplings, the total execution time for a single OTU table was of the order of a few days. CProgol was written in C, some 20 years ago (Muggleton & Bryant 2000), and this code is no longer compliant with modern C compilers, significantly complicating the operational burden of running EML abduction on modern computers.

1.3 Can explainable machine learning be improved?

To address these issues, the EML framework to infer interaction networks was extended. In chapter IV, I detail how the improvement was developed. I started by evaluating different A/ILP machine learning languages for abduction, such as CProgol and Aleph (Srinivasan 2001). Our collaborators recently developed PyGol (Varghese *et al.* 2022), a much faster, python-based implementation of A/ILP and this simplified our choices. PyGol is able to obtain compression values from the same datasets at between 10 and 60 times faster than CProgol (Varghese *et al.* 2022), Figure A.6). The reduction of computation times from hours to minutes simplifies EML development by reducing the time required for testing, and thereby promoting a more interactive and experimental approach. The risks of a CProgol run made in error are much lower in PyGol. As a consequence of PyGol's efficiency, larger metabarcoding datasets can be learned on reasonable time scales, making the use of explainable machine learning for monitoring a realistic possibility. The method of discrimination of true interactions, from false, using the I statistic obtained from compression was also changed. Given that the previous bootstrapping procedure required a high number of re-samplings (Li *et al.* 2009) and did not take into account the stability of the different edges of the network, we adopted the StARS model selection procedure (Liu *et al.* 2010). The use of StARS reduced both the number of re-samplings that are necessary (Müller *et al.* 2016), reducing further the run times, and the number of false positives in the networks inferred from computer generated data by accounting for edge stability.

Together with these procedural improvements, the hypothesis of interaction was expanded to include the possibility of OTU exclusion and interdependence. The I statistic, obtained from this expanded theory, was found to be better at detecting computer-generated competition and amensalistic interactions. That a modified hypothesis of interaction can lead to different performance for detecting interactions illustrates the future potential of explainable machine learning for network reconstruction. Any ecological hypothesis, when expressed as a logical relation, could be used for abduction. Each obtained network would then be representative of the hypothesis used, depending the accuracy of detection and the utility of the hypothesis for describing the interactions in the ecosystem.

This improved framework for interaction network reconstruction was implemented as a tool within an R package (R Core Team 2022) called InfIntE (INFerence of INTeractions using Explainable machine learning) that calls PyGol as the engine for abduction. The tool has an automatic pipeline which: 1) transforms an OTU table into logic clauses; 2) abduces the effects on the abundance using a given set of hypotheses of interaction; 3) selects the edges of the network by calculating I values and running StARS; and 4) then directly classifies the significant interactions to their interaction type (Figure IV.1).

1.4 It is possible to perform a direct classification of interaction types using explainable machine learning?

InfIntE classifies the detected interactions by their type as a function of the effects on the abundance of the involved OTUs (Table I.1). The performance of the InfIntE interaction classification was evaluated in chapter IV using a combination of the ecological models to simululate interactions, from Weiss et al. (2016), and the OTU data of grapevine foliar communities, as described in chapter VII. The InfIntE detection of computer-generated amensalism and competition interactions detrimental to OTU abundance was significantly lower than the detection of beneficial interactions, such as commensalism and mutualism. This suggested that even where the incorporation of exclusion in the hypothesis increased the performance of InfIntE, the detection of detrimental interactions still posed considerable difficulties, caused by the lack of information provided by zero data in the OTU matrix. This issue is not unique to InfIntE, but is also shared with the statistical based inference (Connor *et al.* 2017). However, InfIntE correctly classified the majority of the computer generated interactions detected, indicating that the hypotheses of interaction and the classification mechanism were appropriate to identify the types of interactions based on the abundance changes. Consequently, we consider that where InfIntE detects an interaction, this interaction will be correctly classified. This posit was corroborated by reconstructing the microbial interactions on grapevine leaves suffering from downy mildew. Downy mildew is caused by the pathogen *Plasmopara viticola*, and due its impact in the grape production, pathogen interactions within the microbiome are an important subject of interest. Indeed the literature lists a number of fungal species described as *P. viticola* antagonists. All these previously discovered species, when present in our datasets, were shown to have detrimental interactions with the pathogen using InfIntE. This suggests, with some strength, that the interaction classification using real data works.

It is difficult to find work done to establish a link between an interaction type and its ecological meaning in the literature. Possibly the best examples of interaction type classification have been done using longitudinal (time series) metabarcoding data and adaptions of the Lotka-Volterra population dynamical functions (Tsai *et al.* 2015; Lo & Marculescu 2017a; Pinto *et al.* 2022). Proposals have also been made to simulate interactions types as a function of the metabolic pathways found in databases (Pacheco *et al.* 2019; Sambamoorthy & Raman 2022) and the literature (Sun *et al.* 2021). To date, however, there are no techniques to infer interaction types from interaction mechanisms/hypotheses in cross-sectional data, making InfIntE the first approach to deal with this subject.

Interaction types give insight into both how microbial communities are structured and how they might evolve (Pacheco *et al.* 2019). Interactions can have a great economic importance for ecological regulation and Ecosystem Services, such as for biocontrol (Musetti *et al.* 2006) and for the economic use of microbial communities in industry and industrial processes (Ghosh *et al.* 2016). The proportion of the different interactions types in a network provides information on microbial community assembly, dynamics and resilience and might therefore potentially be used for biomonitoring.

1.5 Considerations on InfIntE

The development of InfIntE required a continuous cycle of improvement, testing and validation. The performance of the different hypotheses of interaction and edge selection methods was tested using computer-generated datasets. Then, when I was satisfied that InfIntE offered sufficient performance, it was evaluated by reconstructing interaction networks from grapevine datasets, specifically focused on the pathobiome. The use of computer-generated data provided evidence that InfIntE can utilise OTUs tables with abundance changes caused by different types of interactions to reconstruct and classify those interactions at play. The use of 5-fold validation of grapevine community data demonstrated that InfIntE could also recover information from real datasets, reconstructing the pathobiome by detecting and classifying interactions based upon mechanisms expected from the literature. Overall, this suggests that InfIntE is a tool that has potential utility for reconstructing interaction networks, classified by their interaction types. It is, in addition, implemented in an easy-to-use R package that is available for ecologists wishing to use explainable machine learning to explore metabarcoding data.

1.6 Which network reconstruction tool should I choose for biomonitoring?

During my Ph.D., I have used two existing statistical network reconstruction tools, SparCC (Friedman & Alm 2012) and SPIEC-EASI (Kurtz *et al.* 2015), and developed an explainable machine learning based tool, InfIntE (chapters IV and V). The choice of the two statistical tools, among all the available options, was determined by their proven performance and wide use by microbial ecologists. This ensured that publications that benchmarked these tools, as methods of reconstructing networks from metabarcoding data, were also available. These statistical tools and InfIntE perform cross-sectional network inference using samples from an ecosystem taken at the same point in time. They have different strengths and weaknesses, offering different options to users for monitoring ecosystems and obtaining ecological knowledge. The value of each approach, and an any decision as to which one to use, varies as a function of the sampling size, the data and the need for information to support decision making.

Are interaction networks inferred from metabarcoding data replicable?

The principle of using interaction networks to monitor ecosystems is that we might detect changes in the interactions of networks that anticipate changes in the ecosystem. This principle presumes that interaction networks will change systematically when subjected to an ecosystem pressure or driver, whether biotic or abiotic. Consequently, it also assumes that networks will not change if they are not subject to ecosystem drivers. The first requirement for network reconstruction tools is that they deliver reproducible output, therefore. Networks obtained from the same ecosystem at the same time must be the similar, within the limits of sampling errors. The reproducibility of the three tools was evaluated using computer generated datasets that forced interactions in random OTU populations (Chapter IV).

No other factors influenced the OTU abundance data. For all three tools, the deviation of interaction detection accuracy between datasets with similar numbers of replicates and conditions was non significant. This indicates that, even where the tools use a level of randomness for re-sampling to assess interaction significance, they all detect the same network links when the conditions are the same.

Real metabarcoding data is riddled with many biases, from sampling through to the bioinformatic process, all of which might limit reproducibility. Such variability was evident even in samples of grapevine leaves from the same abiotic conditions. SparCC and SPIEC-EASI produced one to five consensus interactions, from the three networks reconstructed from conventional and organic samples. Only one consensus interaction was inferred by InfIntE between the interaction networks from nine vineyards. This divergence between computer-generated and real data suggests one possible explanation; that the lack of reproducibility is due to the high variance and stochasticity within and between microbial communities on the grapevine leaves (Ning *et al.* 2019; Peay & Bruns 2014). This would preclude reproducibility even though models simulating OTU abudances have previously been shown to accord well with real OTU abundances (Shoemaker *et al.* 2017). It should also be noted that conditions that are considered identical during sampling can include a myriad of unobserved biotic and abiotic differences. Our belief is that interaction networks are hard to construct from metabarcoding data, and that reproducibility depends on factors that affect the sample composition and abundance, as well as the number of sample replicates.

How many samples are necessary to reconstruct a network

There is no agreement in the literature on what is the minimum number of samples that are required to reconstruct a network. The recommended number ranges from 20 to several hundreds, depending on the source (Berry & Widder 2014; Hirano & Takemoto 2019). Sampling and sample number has important implications for any experimental or monitoring approach. The cost in time and money to sequence a sample has decreased markedly, thanks to next-generation sequencing. However, it is still expensive to obtain a sequence dataset from several hundreds of eDNA samples. Some upper limitation on data collection and the number of samples to be processed is necessary, simply for affordability. Insufficient sampling, by contrast, could result in non-informative interaction networks, invalidating the whole experiment.

Tests using different sized simulation datasets showed a detection accuracy that differed as a function of the sample number and the tool used. The accuracies of SparCC and SPIEC-EASI plateaued at around 20 samples, with more samples not improving the computed accuracy. InfIntE, by contrast, needed a number of samples closer to 50 for accuracy to plateau. In general, network reconstruction benchmarking exercises, including those that have considered SparCC and SPIEC-EASI, have used an excess of samples and therefore have not evaluated the sampling effort as a variable in the network reconstruction process (Weiss *et al.* 2016; Röttjers & Faust 2018). As we have seen, the sample size can have important effects on the interaction detection. Typically, 20 to 50 samples is an affordable number of samples for a metabarcoding essay. These sample sizes are not a point

of critical decision-making between the different tools, therefore. InfIntE can also use this order of the number of samples to classify the interaction types. However, for cases where only really limited number of samples are available and no interaction classification is required, statistical-based network inference tools would offer a valuable, first exploratory approach, possibly to justify further sampling and the use of InfIntE.

Can ecosystem change be assessed using interaction networks?

Interaction networks obtained from metabarcoding data are highly dependent on the condition of the samples from which they are inferred. Even sets of samples from same abiotic conditions, location and time points could produce different networks (Galiana *et al.* 2022). My work suggests that the networks produced by statistical and logical learning approaches are accurate representations of the interactions that caused the changes in the OTU abundance (chapters II & IV). Thus, network metrics can be used to evaluate change in interaction networks to assess ecosystem change (Tylianakis *et al.* 2010; Pellissier *et al.* 2018). Tests done using SparCC and SPIEC-EASI have shown that network α-diversity measures such as the number of edges or network diameter (Barabási *et al.* 2000), could not differentiate interaction networks from organic or conventional grapevine crops. Some network β-diversity measures, like the dissimilarity of associations (Poisot *et al.* 2012), could indicate differences between interaction networks from different cropping practices. These network metrics have not yet been used to evaluate the performance of InfIntE.

The networks obtained from different vineyards across France (chapter VII) are expected to be different, given the great geographical distance between them. We also expect that the networks obtained using InfIntE will reflect ecosystem change, since its accuracy is comparable to SparCC and SPIEC-EASI. It is important to remark that not all the differences found between networks were due to OTU change or presence. Networks also changed due to network plasticity (Gray *et al.* 2021), with links appearing or disappearing between networks. When combined with the few consensus interactions found across different networks, these findings point to a preliminary hypothesis that microbial interactions may not be conserved in an ecosystem and may be highly sensitive to the different factors and variables of the ecosystem. Even small variation in these factors and variables, whether abiotic (e.g. increase of temperature) or biotic (e.g. appearance of a new species) could exert a selective pressure across all possible interactions leading to a sweep of change across the network. It might be that the interaction networks then stabilize to a new stable state (Yuan *et al.* 2021). As a consequence, reconstructed interaction networks, irrespective of the used tool for reconstruction, are promising and sensitive ecological structure for the monitoring of ecosystems, where the metrics to compare networks are appropriate. This may be particularly relevant in a general context of global climate change were network modification might be the first indicator of the systemic risk of ecosystem change.

Do interaction networks provide information to manage ecosystems?

Detection of ecosystem change is important if we are to mitigate the change effects that climate change and other human-derived drivers cause to ecosystems (Cordier *et al.* 2021). When such change is detected, it is also necessary to take appropriate management decisions to mitigate the change, to conserve biodiversity and the delivery of ecosystem services. One of the main objectives of this thesis has been to provide ecologists with the necessary methodology to facilitate decision making at large scales using eDNA derived ecological networks. A key step in this is to explore the interaction networks. To this end we have designed an interactive tools to visualise and compare interaction networks (chapter V). The most important part of my work, nonetheless, is EML methods to obtain interaction networks that are explainable in terms of our ecological knowledge and hypotheses, and objectively classify directly the different microbial interaction types present. This work, encoded within the network reconstruction tool InfIntE, is a small, first step, towards the use of explainable machine learning for interaction network reconstruction. It establishes a baseline to obtain relevant ecological knowledge from metabarcoding data, directly and objectively, while reducing the subjective interpretation of the results often required for statistical approaches. This differentiates InfIntE from the statistical learning tools, such as SparCC and SPIEC-EASI. While InfIntE can detect interaction and reproduce networks with the same accuracy as SparCC and SPIEC-EASI, where sufficient samples are available, it does this with the benefit of directly classifying ecological interactions thereby retrieving significant ecological knowledge.

The classified interaction types proved to be useful for identifying all known antagonists of the pathogen *Plasmopara viticola* in metabarcoding samples from nine vineyards (chapter IV). Importantly, the methodology also predicted (discovered) three previously unknown antagonists as candidate biological control agents of *P. viticola* for future testing. Interaction types also have potential utility for biomonitoring. If a correlation association between two microorganisms stops, it might have many explanations. But, where ecologists also know that the association was due to a commensal interaction, for example, they might then conjecture that one of the microorganisms stopped producing the nutrient source for the other (Tshikantwa *et al.* 2018). Inferred classified interactions ultimately provide a better ecological explanation of change for ecosystem biomonitoring.

2 Future perspectives

2.1 InfIntE testing

InfIntE is the first explainable machine learning approach to be applied to the inference of microbial interactions. By the end of my Ph.D., InfIntE works as a standalone tool that can perform the reconstruction of interaction networks with sufficient precision to be compared against known interactions in the literature. The tool is encoded as an R package and stored and curated in a Github repository. My objective, in coding an R package, was to make InfIntE readily available to any ecologist with an interest in reconstructing networks using explainable machine

learning. The development of an R package is a constant and continuous process that does not end with the publication of the package. There is a necessary and ongoing process of testing and feedback from different end-users to identify and solve any possible bugs that might appear. To date, InfIntE has been tested by a few users, on a limited number of datasets, all of whom have been working in a Linux environment. The tool requires ongoing support and further testing, under different operating system environments and settings to ensure that there are no compatibility issues. The communication between R and PyGol's python scripts that are used to perform the abduction, are particularly sensitive. New versions of python have broken previous versions of the tool because the script no longer ran in the way expected during development. The network visualization tool also needs further development and feedback from the user community. This development should, in particular, address feedback that shapes the visualisation tool to user needs. The ease-of-use of this interactive tool should also be further improved to make sure that it can be used by any ecologist, even those without great computer skills, rendering the tool able to visualize relevant ecological knowledge from one or more interaction networks, and especially those ones reconstructed by InfIntE.

2.2 InfIntE improvement

The flexibility that explainable machine learning offers means that there is always room for further improvements in hypotheses, based upon the solid foundations established for InfIntE, as described here. Explainable machine learning uses knowledge from different scientific domains to discover new knowledge, following a human understandable process. The domain knowledge used in InfIntE to infer interactions is simple, logical hypotheses for how the abundances of different OTUs were affected by interactions with other OTUs. In chapter IV, I show that these initial hypotheses of interaction could be extended to include the possibility of exclusion of an OTU, caused by an interaction. In principle, it is possible to provide InfIntE with any logical hypothesis of interaction. The inclusion of new hypotheses of interaction, describing for example interactions involving more than two OTUs or interaction dynamics by including time, could help in the modelling and comprehension of more complex ecosystems. This would be useful in the case of interactions that change depending upon the life stage of the OTUs involved. Interactions might be hypothesised as a function of the ecology of the OTU (Chagnon *et al.* 2016), rather than abundance. In chapter IV, both metabarcode and pathogen qPCR data are employed to reconstruct pathobiome networks, illustrating how different sources of information about the OTU abundance can be used simultaneously in the inference. This could be expanded to include a great variety of abundance data, such as counts made using a microscope (Daley & Hobbie 1975) and spectrophotometric quantification (Zhou *et al.* 1998). What is required is data from a methodology that robustly quantifies the changes in abundance between samples. The use of InfIntE could also be extended to ecological other domains than the microbial, such as to learn interactions between invertebrates. This would *only* require appropriate descriptions of interactions, as logical hypotheses, and data for the change in abundance of the species concerned.

Domain knowledge is a broad, catch-all term for any information that might be

used to inform a process of learning. Any appropriate domain knowledge related to microorganisms could be included in the InfIntE learning where the knowledge can be related, logically, to the hypothesis of interaction used. The sources of microbial domain knowledge can include the expertise of ecologists, published papers and databases, such as for the traits of microorganisms for example. Including such knowledge in ecological network reconstruction would produce more accurate ecological networks, via more highly refined hypotheses of interaction. The link between this domain knowledge and the OTU abundance data is made using the taxonomic assignment of the OTU. As noted in chapter I, taxonomic assignment is performed by comparing the OTU sequence to curated sequences present in databases. This means that the assignment is highly dependant on the quality of the sequences and information in the database, and some of these can be very poor. Microbial taxonomy is also in constant flux, with the latin binomial nomenclature changing over time and some taxa having multiple names. This requires that the taxonomic assignment is regularly updated. If and when these issues are addressed, however, the introduction of different sources of domain knowledge into the hypotheses would result in an improvement of the network reconstruction. Microorganisms have many different ecological and functional characteristics, including body size, requirements for growth, metabolic pathways (Pacheco *et al.* 2019; Sambamoorthy & Raman 2022), ecological functions, and organisms with which they interact. Some of these characteristics or traits are stored in online databases, which could serve as valuable sources of domain knowledge. Information and the ecological function can be found in Fungaltraits (Põlme *et al.* 2020), Funguild (Nguyen *et al.* 2016b) and Faprotax (Louca *et al.* 2016). Metabolic information is curated in Metacyc (Caspi *et al.* 2014), Kegg (Kanehisa & Goto 2000) and Uniprot (The UniProt Consortium 2021). The NCBI (Sayers *et al.* 2022), FungiDB (Basenko *et al.* 2018) and MycoBank (Robert *et al.* 2013) have taxonomic and genetic data. Mechanisms can sometimes be postulated between different traits and characteristics. For example, a given metabolic pathways might produce a particular ecological function. It then becomes possible, for certain OTUs, to include the trait data as part of the search for ecological interactions, improving the network inference. Metabolic pathway information has already been used to infer interactions (Pacheco *et al.* 2019; Sambamoorthy & Raman 2022), but this was an isolated learning approach and it has never been used in metabarcoding based interaction inference.

OTU sequences are obtained by amplifying different parts of the eDNA coding for ribosomal RNA, depending on the taxonomic group of study. This part of the DNA has also been used to obtain evolutionary information. Ribosomal sequence alignment between two or more ASVs gives a distance measure related to the evolutionary distance between the ASVs (Van de Peer *et al.* 1993). ASVs that are closer, evolutionarily, are expected to share similar characteristics and traits. Thus, similarity between ASV sequences might also be used as domain knowledge to improve network inference. This information might also be included in the network reconstruction and provide insights into whether microorganisms with similar traits interact similarly, etc.. Where abiotic, environmental conditions at sampling can be recorded, these might be introduced into the inference process to control for sampling biases. This has been already done for some statistically based

interaction reconstruction tools (Ovaskainen *et al.* 2017; Chiquet *et al.* 2019).

EML is not only limited to using logical abduction for inference. There are other forms of EML that could be used to obtain relevant ecological knowledge from metabarcoding data. Ecologically, it would be especially interesting to examine the possibilities that the branch of EML called meta-interpretative learning (MIL) offers. MIL uses observations (the abundance change produced by interactions) and domain knowledge (e.g. functional information contained in databases) to learn (infer) new rules (Muggleton *et al.* 2014). For our microbial case, the rules learned would be new hypotheses of interaction. Comparing computer generated hypotheses of interaction with the hypotheses accepted by microbial ecologists could potentially revolutionise our understanding of how microorganisms interact. These rules would first and foremost have to be comprehensible to scientists; they would have to make sense. Tammaddoni-Nezhad et al., (2013) reconstructed predation rules in toy invertebrate networks, extracted from larger networks, learning that species with large body size eat smaller bodied species (big things eat small things). These rules were comprehensible as they directly reflected ecological theory for who eats whom in invertebrate systems (Bohan *et al.* 2011). MIL might therefore provide microbial ecology with novel interaction theory. MIL remains a nascent technology, however, and there are questions as to how well it can deal with the noise inherent in ecological data. Recent developments have started to provide evidence that once developed MIL has utility for the learning rules from images (Muggleton *et al.* 2018) and is also robust to noisy data (Patsantzis & Muggleton 2021).

2.3 Interaction testing

All the network reconstruction tools studied as part of this thesis use inference to produce a list of significant predictions of association between microorganisms. This list, obtained *in silico* from eDNA sequencing data, is grounded upon the effects on the abundance caused by an interaction. Irrespective of the degree of accuracy that can be ascribed to the different tools using various analytical calculations, the acceptance of an interaction as producing output that has value for science can only come with subsequent *in vivo* tests of the each interaction on the list. InfIntE was used to predict a list of *P.viticola* antagonists that had potential as biocontrol agents. Some of these antagonists have already been described in the literature, serving as a corroborative test. It is the as yet unknown antagonists that need to be tested for their biocontrol potential in small scale lab experiments to determine if they can have any detrimental effect on *P. viticola* abundance. If these tests show that the predicted antagonism is correct, then it validates InfIntE and also suggests that the tool could be used to identify biocontrol agents that might ultimately be used in large scale field applications. Culture experiments of the type necessary for validation are not straightforward, however. It is necessary to isolate the precise species, and possibly strains, that were detected using eDNA sequencing and therefore implicated in the networks. Some of these species may be very difficult or even impossible to grow in laboratory conditions (Pacheco *et al.* 2019). Ecological functions can also vary between different strains, depending on the strain. This is particularly marked in those taxa that employ horizontal gene

transfer (Thomas & Nielsen 2005).

Aside from the constraints of obtaining the correct taxa for the culture experiments, there are also difficulties in providing the interacting microorganisms the same conditions that provoked the interaction (Pacheco & Segrè 2019). Direct competition interactions that harm one or both species through the secretion of bioactive secondary metabolites (Derocles *et al.* 2018), produce readily observable results in a co-culture. Most other interactions are not so direct or observable, and many require appropriate culture conditions. Exploitative competition, as an example, will only happen if there is a scarcity of the resource needed by both interacting taxa. Mutualism will only be observed if there is a common benefit to be achieved by cooperation. Some interactions may be impossible to reproduce in culture media entirely. In such cases, direct inoculation of the taxa onto a controlled substrate (like a tissue) is the only option to evaluate interactions. For these cases, though, the absence of evidence for the interaction in the culture experiments does not provide evidence of absence of the interaction under natural conditions.

InfIntE appears to be an interesting tool to facilitate the development of culturomics validation. This interaction inference tool classifies interactions to their type, thereby making it much easier to define the experimental conditions necessary to validate the interactions than is the case with the correlational associations produced by statistical tools. As an example, where an interaction between two species is automatically classified as commensalism, the test experiment might focus on identifying and detecting the compound produced by one species that is fed upon by the other.

2.4 Final considerations

InfInte can be considered a strong basis for improving EML network reconstruction, as it is a functioning tool with demonstrable ability to reconstruct networks using eDNA. The parameters and requirements necessary for the data used by InfIntE to reconstruct interaction networks have been defined in this thesis. I also detail the bioinformatic pipelines to obtain abundance changes from metabarcoding data. Finally, the metrics necessary to compare interaction networks and evaluate InfIntE performance are identified and tested. I believe that everything is ready to start using InfIntE to monitor ecosystems.

This thesis begins with a reference to the book Anna Karenina and paraphrases the narrative for how *unhappy families are unhappy each in their own way*. Climate change is a source of "unhappiness" for many ecosystems. We need to start evaluating what is happening to the world's ecosystems, due to drivers like climate change, and taking appropriate corrective decisions. Over the chapters of this thesis, I have argued that it is possible to sample ecosystems for eDNA, reconstruct the microbial interaction networks of these ecosystems and obtain ecological knowledge on the changes that these systems undergo. Dubart et al., (2021) propose a number of different ecosystems that would be interesting to monitor from a biomonitoring theoretical point of view, and there are many ecosystems where monitoring would offer solutions to the many ecological problems those ecosystems face. At the end of the novel, Anna Karenina, overwhelmed by the impossibility

of being happy, throws herself under a train to end her life, only realising that this is a mistake when her death is inevitable. The impulse to give up is common, especially when the problems you face appear insurmountable. However, in most cases the alternative to not trying is worse. Climate change, and other global change problems, appear insurmountable and inevitable processes, but not trying to understand their effects and to stop or to mitigate those effects will lead to the loss of many ecosystems and cause great human suffering. I think that the work in this thesis is a small contribution towards the ecological understanding of microbial ecosystem change. I hope that it adds, in small degree, to better ecosystem monitoring, management and protection.

Chapter VII

Constructing ecological, microbial community data-sets from DNA data

Leaf microbiome data for European cultivated grapevine (*Vitis vinifera L.*) during downy mildew (*Plasmopara viticola*) epidemics in three wine-producing regions in France

Didac Barroso-Bergada, François Delmotte, Julie Faivre d'Arcier, Marie Massot, Emilie Chancerel, Isabelle Demeaux, Soizic Guimier, Erwan Guichoux. David A. Bohan, Corinne Vacher

Abstract:
Grapevine downy mildew (*Plasmopara viticola*) is a major disease of European cultivated grapevine (*Vitis vinifera L.*) against which a large amount of synthetic pesticides are used. Developing microbial biocontrol of *P. viticola* could reduce the use of pesticides in viticulture and preserve human and environmental health. To achieve this goal, it is necessary to better understand the interactions between *P. viticola* and the vine foliar microbiome. Here we present metabarcoding datasets describing the bacterial and fungal communities from more than 200 pairs of leaf samples collected during powdery mildew epidemics in three major wine-producing regions in France. The microbiome of both symptomatic and asymptomatic tissues was characterized. *P. viticola* abundance was quantified using qPCR. We provide the raw metabarcoding datasets, the Amplicon Sequence Variant (ASV) tables obtained after bioinformatic processing, the metadata describing sampling sites and tissue health condition and the code used for bioinformatic and descriptive statistical analysis.

1 Introduction

Downy mildew, caused by the oomycete *Plasmopara viticola*, is a major disease of European cultivated grapevine, *Vitis vinifera L.* (Fontaine *et al.* 2021). A current challenge in viticulture is to control downy mildew without using synthetic pesticides, in order to better preserve human and environmental health. Harnessing the plant microbiota is one of the possible avenues to reach this objective (Busby *et al.* 2017; Toju *et al.* 2018; D'Hondt *et al.* 2021). Experimental evidence suggests that microbial communities naturally associated with leaf tissues can contribute to grapevine resistance to downy mildew (Bruisson *et al.* 2019; Burruano *et al.* 2016), through different interaction mechanisms (interference competition, exploitative competition, hyperparasitism, immune priming). Leaf-associated microorganisms can interfere directly with *P. viticola*, as has been demonstrated for several members of the *Bacillus* genera (Bruisson *et al.* 2019) and the fungal species *Alternaria alternata* (Musetti *et al.* 2006). Some microorganisms can also compete with the pathogen for space and resources, or parasitize the pathogen (Zanzotto, Morroni, *et al.* 2016; Ghule *et al.* 2018). The plant microbiota can confer pathogen resistance by priming the plant immune response (Nishad *et al.*

2020; Hacquard *et al.* 2017). To foster the microbial biocontrol of downy mildew, it is necessary to have a more complete understanding of the microbiome associated with healthy and diseased tissues than currently exists. In this Resource Announcement we present novel metabarcoding datasets describing the bacterial and fungal communities associated with the leaves of *V. vinifera*, collected from nine vineyard plots in three major French wine-producing regions during downy mildew epidemics (Table VII.1). The datasets describe microbial communities associated with both downy mildew lesions and asymptomatic tissues. In addition to raw metabarcoding datasets, the Resource Announcement provides the amplicon sequence variant (ASV) tables obtained after bioinformatic processing, the metadata describing the sampling design and the code used for bioinformatic and descriptive statistical analysis.

2 Methods

2.1 Sampling

Samples were collected in June and July 2018 from three vineyard plots in each of three French wine-growing regions: Aquitaine (AQ), Champagne (CH) and Occitanie (OC) (Table VII.1; Figure VII.2A). Samples were collected during downy mildew (*P. viticola*) epidemics from 30 vines (*V. vinifera L.*) in each vineyard, on rows that were not treated against the pathogen. We collected one sporulating leaf from each vine using sterile gloves and placed it in an individual plastic bag. Leaves were processed on the day of collection with sterilized tools in the sterile field of a MICROBIO electric burner (MSEI, France). We collected four disks from each leaf: Two symptomatic foliar disks of 12mm diameter were taken from sporulating mildew lesions on each leaf (S for symptomatic) and placed together in a 2ml autoclaved collection tube stored in a box filled with silicagel. Two asymptomatic disks were also taken from each leaf (A for asymptomatic). Screw caps of the tubes were left loose to allow the disks to dry. All the samples were then freeze-dried.

2.2 DNA extraction

Total DNA was extracted with the DNeasy Plant Mini kit (Qiagen, France), with a slightly modified version of the protocol recommended by Kerdraon *et al.* 2019. Two autoclaved DNAase-free inox 420C beads were added to each tube and samples were ground at 1500 rpm with the Geno/Grinder® for 30 s, then 1 min and 1 min again, with manual shaking between each grinding step. Tubes were then centrifuged for 1 min at 6000 x g. Leaf powder and 200 µL of buffer AP1 preheated to 60°C were mixed by vortexing the tubes for 30 s twice at 1500 x g, and centrifuging them for 1 min at 3000 x g. 250 µL of preheated buffer AP1 and 4.5 µL of RNase A were added to each tube and mixed by vortexing the tubes for 30s twice at 1500 x g. After 5 min of rest, 130 µL of buffer P3 was added to each tube, which was then mixed by gentle inversion for 15 s, incubated at -20°C for 10 min and centrifuged for 1 min at 5000 x g. The supernatant (450 µL) was transferred to a spin column and centrifuged for 2 min at 20000 x g. The filtrate (200 µL) was

Table VII.1: Sampling design. Grapevine leaves were collected in three wine-growing regions in France, which are presented in the table from North to South: Champagne (CH), Aquitaine (AQ) and Occitanie (OC). Sampling took place in three vineyard plots (A to I) per region (Figure VII.2A). The grapevine variety and the sampling date are indicated for each plot. The GPS coordinates of each plot are given as Latitude (Lat) and Longitude (Lon). Samples were collected during downy mildew (*P. viticola*) epidemics from 30 vines (*V. vinifera L.*) in each vineyard at a date were the epidemics was at its peak of infection.

Vineyard	Region	Lat	Lon	Variety	Sampling Date
I	Champagne	49.063127	4.008855	Pinot noir	16/07/2018
H	Champagne	49.018104	3.980866	Meunier	11/07/2018
G	Champagne	49.017485	3.983795	Chardonnay	11/07/2018
D	Aquitaine	44.791315	-0.578224	Merlot	26/06/2018
E	Aquitaine	44.707248	0.244359	Cabernet Franc	27/07/2018
F	Aquitaine	44.628850	-0.263220	Merlot	27/07/2018
B	Occitanie	43.506132	4.754621	Chardonnay	14/06/2018
C	Occitanie	43.14223	3.13292	Gamay	15/06/2018
A	Occitanie	43.113225	2.095317	Chasan	13/06/2018

transferred to a new tube, to which sodium acetate (200 µL, 3 M, pH 5) and cold 2-propanol (600 µL) were added. DNA was precipitated by incubation at -20°C for a minimum of 1 hr and recovered by centrifugation (20 min, 13000 x g). The pellet was washed twice with cold ethanol (70%), dried at 50°C for approximately 30min and dissolved in 100 µL of AE buffer. The negative extraction controls were represented by extraction reagents in an autoclaved 2ml Eppendorf tube containing two autoclaved DNAase-free inox 420C beads.

2.3 Fungal ITS amplification

The ITS1 region of the fungal ITS rDNA gene (Schoch *et al.* 2012) was amplified using primers ITS1F-ITS2 (White *et al.* 1990, Gardes & Bruns 1993). To avoid a two-stage PCR protocol, each primer contained the Illumina adaptor sequence and a tag (ITS1F: 5′- CAAGCAGAAGACGGCATACGAGATGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTxxxxxxxxxxxxCTTGGTCATTTAGAG GAAGTAA-3′; ITS2: 5′- AATGATACGGCGACCACCGAGATCTACACTCT TTCCCTACACGACGCTCTTCCGATCTxxxxxxxxxxxxGCTGCGTTCTTCA TCGATGC-3′, where "x" is the 12 nucleotide tag). The PCR mixture (20 µL of final volume) consisted of 10 µL of 2X QIAGEN Multiplex PCR Master Mix (2X final), 2 µL each of the forward and reverse primers (0.1 µM final), 4 µL of water, 1 µL of 10 ng/µL BSA and 1 µL of DNA template. PCR cycling reactions were conducted on a Veriti 96-well Thermal Cycler (Applied Biosystems) using

the following conditions: initial denaturation at 95°C for 15 min followed by 35 cycles at 94°C for 30 s, 57°C for 90s, 72°C for 90 s with final extension of 72°C for 10 min. ITS1 amplification was confirmed by electrophoresis on a 2% agarose gel. Two marine fungal strains (*Candida oceani* and *Yamadazyma barbieri*) were used as positive controls as they were unlikely to be found in our samples. One positive control included 1 µL of 10 ng/µL DNA of *Candida oceani* only and the other included an equimolar mixture of both strains. The negative PCR controls were represented by PCR mix without any DNA template. Each PCR plate contained one negative extraction control, three negative PCR controls, one single-strain positive control and one two-strain positive control.

2.4 Bacterial 16S amplification

The V5-V6 region of the bacterial 16S rDNA gene was amplified using primers 799F-1115R (Redford *et al.* 2010, Chelius & Triplett 2001) 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 (799F: 5′- CAAGCAGAAGACGGCA TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTxxxxxxx xxxxxHS-AACMGGATTAGATACCCKG-3′; 1115R: 5′- AATGATACGGCGA CCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxx xxxxxxxxHS-AGGGTTGCGCTCGTTG-3′, where HS represents a 0-7-base-pair heterogeneity spacer and "x" a 12 nucleotide tag). The PCR mixture (20 µL of final volume) consisted of 10 µL of 2X QIAGEN Multiplex PCR Master Mix (2X final), 2 µL each of the forward and reverse primers (0.1 µM final), 4 µL of water, 1 µL of 10mg/ml BSA and 1 µL of DNA template. PCR cycling reactions were conducted on a Veriti 96-well Thermal Cycler (Applied Biosystems) using the following conditions: initial denaturation at 95°C for 15 min followed by 32 cycles at 94°C for 30 s, 53°C for 90 s, 72°C for 90 s with final extension of 72°C for 10 min. 16S amplification was confirmed by electrophoresis on a 2% agarose gel. Two marine bacterial strains (*Sulfitobacter pontiacus* and *Vibrio splendidus*) were used as positive controls as they were unlikely to be found in our samples. One positive control included 1 µL of 10 ng/µL DNA of *Vibrio splendidus* only and the other included an equimolar mixture of both strains. The negative PCR controls were represented by PCR mix without any DNA template. Each PCR plate contained one negative extraction control, three negative PCR controls, one single-strain positive control and one two-strain positive control.

2.5 MiSeq sequencing

PCR products were purified, quantified (Quant-it dsDNA assay kit; Invitrogen) and used to constitute equimolar pools (Hamilton Microlab STAR robot). Mean fragment size was determined with a Tapestation instrument (Agilent Technologies). ITS and 16S amplicons were sequenced on one and a half and two runs of the Miseq Instrument (Illumina), respectively, with the reagent kit v2 (500 cycles). Sequence demultiplexing (with an exact index search) was performed at the PGTB sequencing facility (Genome Transcriptome Facility of Bordeaux, Pierroton,

France) with DoubleTagDemultiplexer.

2.6 Quantification of *P. viticola*

The abundance of *P. viticola* was quantified with real-time quantitative PCR targeting the ITS1 region, which was amplified using the Giop primerset (Valsesia *et al.* 2005). Primers and probes sequences were synthesized by Integral DNA Technologies, the probe was labelled with 5′HEX, an InternalZEN™ and 3′IBFQ quenchers. The protocol was the following: qPCR was carried out in a 20 µL reaction volume with 2 µL of genomic template DNA, 900 nM of each primer, 500 nM of probe, 1X HOT FIREPol ®Probe qPCR Mix Plus no Rox (Solis BioDyne) and sterile MilliQ water. The PCR was performed with the following parameters: initial denaturing at 95°C for 12 min, followed by 40 cycles of 95°C, denaturing for 30 s, 60°C annealing for 1 min. Each PCR assay included duplicates of samples, negative controls without template DNA, and six points of validated and diluted *P. viticola*'s DNA. For the standard curve, *P. viticola* DNA (17 ng/µL) was serially diluted from $1:10$ to $1:10^7$ (concentration verified using a Qubit (Thermo Fisher Scientific)). The obtained standard curves were generated by plotting the DNA amounts against the CTs (threshold values at 2) measured with the LightCycler®480Software, Version 1.5 with Abs Quant/Fit Points analysis (Roche). The DNA quantification range was validated after multiple assays on triplicates giving a PCR efficiency between 90 and 100% with a limit of quantification at 17×10^{-5} ng/µL. Finally, the abundance of *P. viticola* in each sample was calculated as the ratio of the qPCR estimate over the total DNA amount, which was estimated using Quant iTTM PicoGreen™ ds DNA Assay Kit (Life Technologies).

2.7 Bioinformatics

The MiSeq sequences were processed using the DADA2 pipeline v1.22.0 (Callahan *et al.* 2016) implemented in R (R Core Team 2022). Primers were identified and removed using cutadapt 3.2 (Martin 2011) and the trimmed sequences were then parsed to the DADA2 algorithm to infer Amplicon Sequence Variants (ASVs). Sequencing quality scores were assessed using the plotQualityProfile function (Figure A.7). Chimeras were removed using the removeBimeraDenovo function of DADA2. ASV taxonomic assignment was performed using an implementation of the Naive Bayesian Classifier (Wang *et al.* 2007) included in the DADA2 pipeline. The databases used for taxonomic assignment were Silva v138.1 (Quast *et al.* 2012) and UNITE all eukaryotes v8.3 (Abarenkov *et al.* 2021) for 16S and ITS sequences, respectively. Three tables were obtained for both the 16S and ITS datasets: an ASV table with the sequence count in each sample; a table with the taxonomic assignment of each ASV sequence; and, a metadata table describing the collection conditions of each sample. The three tables were joined in a phyloseq object using the phyloseq bioconductor package v1.38.0 (McMurdie & Holmes 2013). To filter out possible contaminants, the combined method of the isContaminant function of the DECONTAM Bioconductor package v1.14.0 (Davis *et al.* 2018) was used, followed by the decontamination method described in Galan *et al.* 2016. Moreover, 16S ASVs identified as chloroplastic or mitochondrial with Metaxa2.2.3

(Bengtsson-Palme *et al.* 2015), or according to their taxonomic assignment in the Silva database, were removed. The remaining ASVs were clustered using the Lulu algorithm (Frøslev *et al.* 2017) with default parameters. ASVs that could not be assigned to a bacterial or fungal phylum were removed. ASVs present in less than 1% of the samples were removed to make sure that the data were free of sequencing artifacts and low abundant contaminants (Cao *et al.* 2021). Finally, pairs of samples where, at least, one of the samples had less than 1000 ASV counts were removed.

2.8 Descriptive statistics

Variations in community alpha- and beta-diversity were analyzed using the statistic environment R v4.1.2 (R Core Team 2022). ASV counts were rarefied to the minimum sample sequencing depth (Figure A.8) and alpha-diversity measures (richness, diversity and evenness) were computed using the phyloseq package. The symptomatology effect (symptomatic vs. asymptomatic) on this measures was assessed using a Wilcoxon rank sum paired test. The abundance of *P. viticola* estimated by qPCR was also assessed using a Wilcoxon rank sum paired test. Compositional dissimilarities among samples were estimated using the Aitchison distance (Aitchison 1982). ASV counts were first transformed using a centered log ratio (clr) transformation to obtain scale-invariant values (Gloor *et al.* 2017) and then the euclidean distance between clr-transformed counts was used as a dissimilarity index. Symptomatic (S) and Asymptomatic (A) samples were represented in a Principal Coordinate Analysis (PCoA) and rotated using Procrustes function from vegan package v2.6-2 (Oksanen *et al.* 2022. Plots were done using ggplot2 package v3.3.5 (Wickham 2016). Significance of Procrustes statistic was assessed using a permutational analysis.

3 Results

This Resource Announcement provides the raw ITS sequence dataset, the raw 16S sequence dataset, and the corresponding ASV tables in R phyloseq format (McMurdie & Holmes 2013). Three ASV tables are provided for each sequence dataset: the raw ASV table which includes the positive and negative control samples; the filtered ASV table; the filtered ASV table after aggregation of highly-similar ASVs using LULU (Frøslev *et al.* 2017). Each phyloseq object includes the ASV table, a table with the taxonomic assignment of all ASVs and a metadata table. The metadata include, for each sample, the sampling site (Table VII.1 and Figure VII.2A), the tissue health condition (A or S), the *P. viticola* DNA concentration estimated by qPCR and the DNA concentration before equimolar pooling. The code used for bioinformatic and descriptive statistical analysis is also provided, as well as the number of reads retained at each step of the bioinformatic process for each sample.

3.1 Abundance of *P. viticola*

The visual assessments of disease symptoms (A or S) were congruent with qPCR data for *P. viticola*. The pathogen was significantly more abundant in sporulating

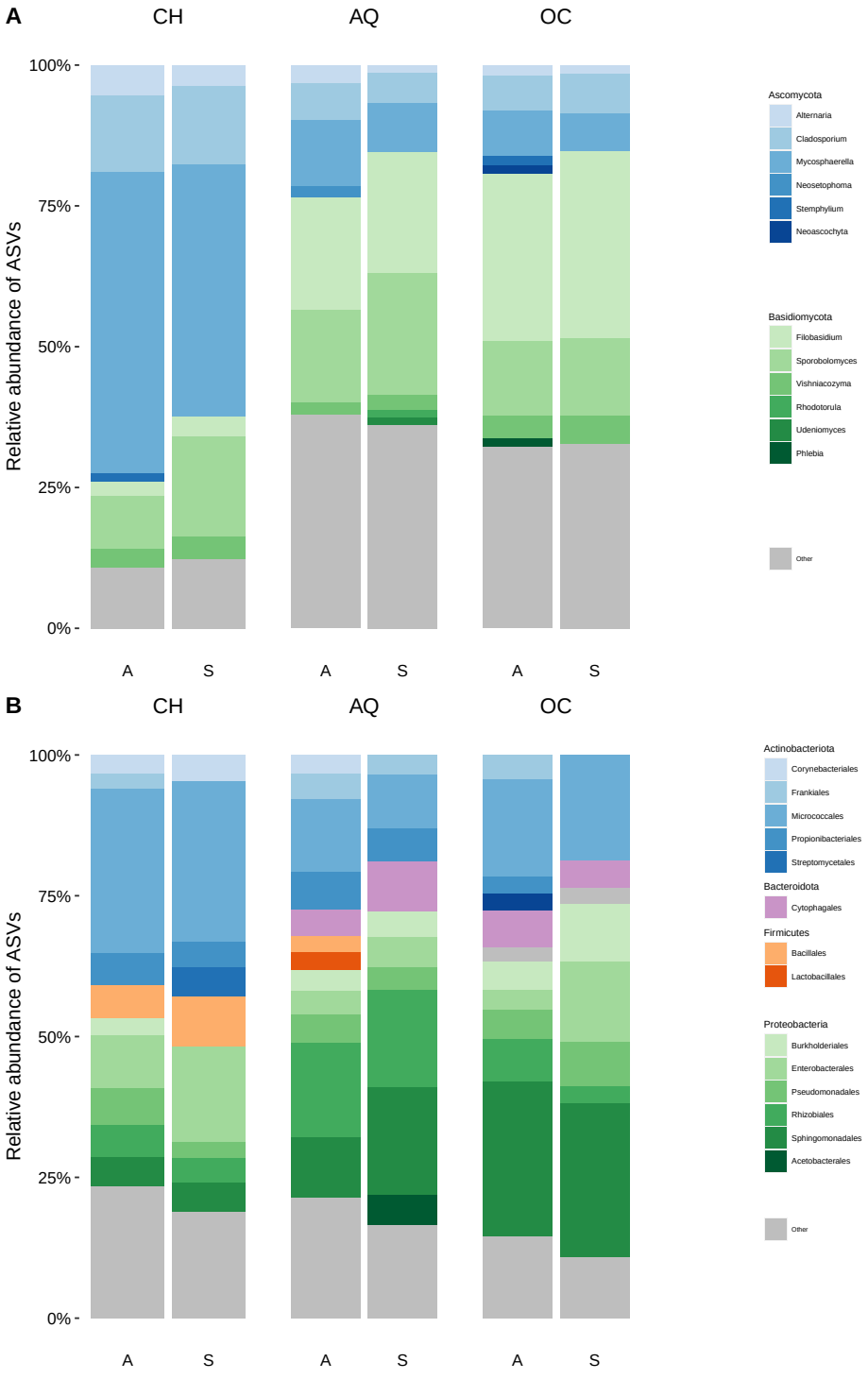


Figure VII.1: Taxonomic barplots showing the relative abundance of A fungal genera and B bacterial orders in grapevine (*Vitis vinifera L.*) leaves depending on the geographic region (Champagne (CH), Aquitaine (AQ) or Occitanie (OC)) and tissue health condition (asymptomatic (A) or sporulating downy mildew lesion (S)).

lesions than in visually healthy samples (Figure VII.2B).

3.2 Fungal community

267 pairs of leaf DNA samples were amplified with fungal ITS primers. The sequencing gave a total of 19125131 raw sequences, with an average of 35814 sequences per sample (SD: 18449; min: 21; max: 295191). The DADA2 pipeline (Callahan *et al.* 2016) retained 14513974 quality sequences representing 5559 ASVs distributed in 531 samples. 3 samples did not generate enough quality sequences to perform ASV inference. The final table obtained after the filtering process and the aggregation of highly-similar ASVs using LULU (Frøslev *et al.* 2017) was composed of 13633258 sequences distributed among 648 ASVs and 251 pairs of samples. The average number of reads per sample was 27158 (SD: 14789; min: 3565; max: 228423).

Taxonomic composition The foliar fungal community was dominated by Ascomycota, which represented 55% of the total number of sequences. Ascomycota pathogens were a small part of the fungal community. *Erysiphe necator* was 0.4% of the total sequence counts while *Botrytis cinerea* was only the 0.002%. Basidiomycota represented 45% of the total number of sequences. Fungal communities in Aquitaine and Occitanie were the most similar according to the taxonomic barplots, while Champagne was distinguished by a greater abundance of the *Mycosphaerella* genus (Figure VII.1A).

Alpha and Beta diversity Alpha diversity measures were differently affected by the tissue health conditions. Fungal community richness differed significantly between symptomatic and asymptomatic samples (Figure VII.3). On the other hand, diversity and evenness were not affected by the health condition. Fungal community composition differed between geographic regions and tissue health conditions (Figure VII.2 C & E). Procrustes test was significant for reordering the asymptomatic and symptomatic PCoA plots.

3.3 Bacterial community

267 pairs leaf DNA samples were amplified with bacterial 16S primers. The sequencing gave a total of 24049900 raw sequences, with an average of 45037 sequences per sample (SD: 25578, min: 8; max: 135255). The DADA2 pipeline (Callahan *et al.* 2016) retained 20309264 quality sequences representing 12669 ASVs. 6 samples did not generate enough sequences to perform the ASV inference. The final table obtained after the filtering process and the aggregation of highly-similar ASVs using LULU (Frøslev *et al.* 2017) was composed of 6512073 sequences distributed among 986 ASVs and 195 pairs of samples. The average number of reads per sample was 16698 (SD: 16554; min: 1105; max: 86598).

Taxonomic composition The most abundant bacterial phylum was Proteobacteria (59.5% of sequence counts), followed by Actinobacteria (29.4%), Firmicutes (5.2%) and Bacteroidetes (4.2%) (Figure VII.1B). Proteobacteria dominated the

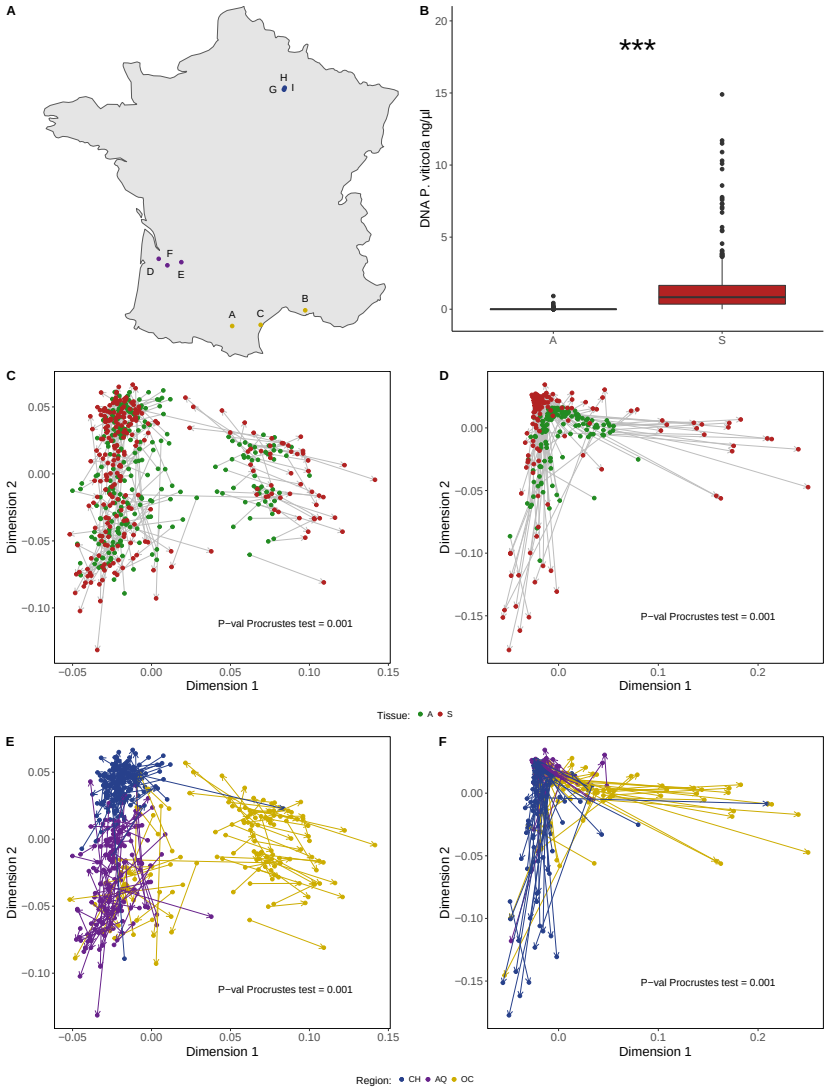


Figure VII.2: Effects of geography and disease on the leaf microbiome of European cultivated grapevine (*Vitis vinifera L.*). **A**: Map of France showing the location of the sampling sites. Vineyard plots A, B and C are located in Aquitaine (AQ); D, F and E are located in Occitanie (OC); G, H and I are located in Champagne (CH). **B**: Concentration of *Plasmopara viticola* DNA, estimated by qPCR, in sporulating mildew lesions (S) and visually asymptomatic leaf samples (A). **C, D, F & F** Procrustes rotation of the Principal Coordinates Analysis (PCoA) plots of symptomatic and asymptomatic samples showing the similarity of fungal communities across leaf samples for fungal and bacterial datasets. Plots are colored by the geographic region (CH, AQ or OC) and tissue health condition (A or S). For both datasets, the permutation test to assess the significance of the Procrustes statistic was significant.

community in Aquitaine and Occitanie, while Actinobacteria was the most abundant phylum in Champagne (Figure VII.1B).

Alpha and Beta diversity Tissue health condition was an structuring factor of the bacterial alpha diversity measures. Community richness, evenness and diversity differed significantly between symptomatic and asymptomatic samples (Figure VII.3). All measures tended to be higher in asymptomatic tissues than in downy mildew lesions. As for the fungi, the bacterial community composition differed between geographic regions and tissue health conditions (Figure VII.2 D & F). However the bacterial communities were less spatially structured than the fungal communities. Procrustes test was significant for reordering the asymptomatic and symptomatic PCoA plots.

4 Future directions

This Resource Announcement provides two metabarcoding datasets describing the changes undergone by fungal and bacterial communities of leaf tissues infected by downy mildew (*P. viticola*) in untreated grapevines (*V. vinifera L.*) across three geographic regions in France. With more than 500 samples sequenced for both the bacterial 16S and fungal ITS regions, they represent a considerable sequencing effort and provide valuable knowledge on the grapevine foliar microbiome. These datasets could be combined with other microbiome datasets (Fort *et al.* 2016; Barroso-Bergadà *et al.* 2021; Zarraonaindia *et al.* 2015; Singh *et al.* 2018) to define the grapevine 'core microbiome', a prerequisite in the development of sustainable viticulture (Toju *et al.* 2018). They also provide useful information for the inference of microbial interaction networks (Röttjers & Faust 2018; Barroso-Bergadà *et al.* 2021) necessary to understand the role of microbial communities in downy mildew development and to discover candidate biocontrol agents.

5 Availability of Data and Materials

The sequence datasets have been deposited in NCBI SRA in bioproject PRJNA797225 (http://www.ncbi.nlm.nih.gov/bioproject/797225) and bioproject PRJNA797948 (http://www.ncbi.nlm.nih.gov/bioproject/797948). The biosample accession numbers are SAMN24973302 to SAMN24973835; Bioinformatic scripts and raw and filtered ASV tables in R phyloseq format have been deposited in Dataverse (https://doi.org/10.15454/2YDSBL). The tables showing variation in sequence counts during the bioinformatic process and the scripts used for data processing and statistical analysis were included in the Dataverse deposit (https://doi.org/10.15454/2YDSBL).

6 Author contributions

Author contributions were as follows: (i) C.V. and D.A.B. conceived and coordinated the study; (ii) F.D., I.D. and S.G. designed the sampling and collected

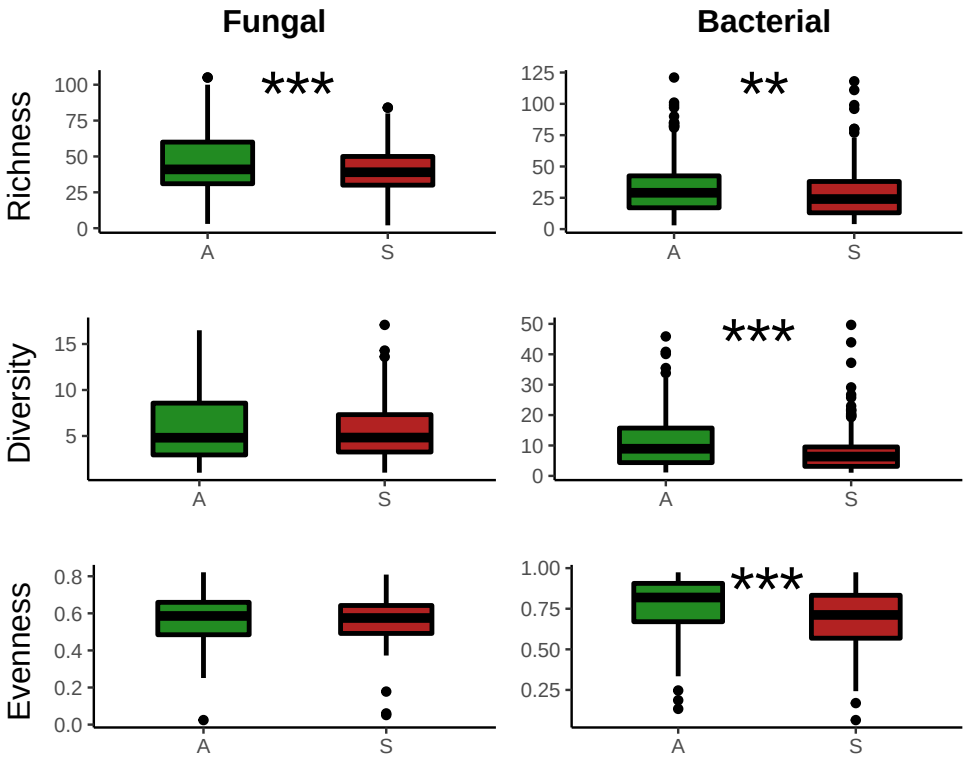


Figure VII.3: Variations in microbial community alpha-diversity (observed richness, Shannon diversity and inverse Simpson evenness) in grapevine leaves, among tissue health conditions (asymptomatic (A) and sporulating downy mildew lesion (S)).

samples and metadata; (iii) M.M., J.F.A., E.C. and E.G. developed protocols and performed the molecular biological work; (iv) D.B.B. performed data analysis and wrote the manuscript; (v) all authors edited the manuscript.

7 Acknowledgments

We thank all members of the *Consortium Biocontrôle* for their support for this work. We thank Pascale Pienne for her contribution to the sampling, Céline Lalanne, Adline Delcamp, Christophe Boury and all other members of the PGTB sequencing facility for their support in molecular biology, and Gaétan Burgaud and Frédéric Garabetian for kindly providing marine strains used as positive controls. We are very grateful to Valérie Laval and Carole Couture for valuable advice when developing protocols. We also thank Frédéric Barraquand, Stéphane Robin, Charlie Pauvert and Andreas Makiola for helpful discussions at the beginning of the project.

8 Funding

This work was supported by the *Consortium Biocontrôle* within the 'BCMicrobiome' project. Additional funding was received from the Aquitaine Region (Athene project, n°2016-1R20301-00007218) and the ANR (NGB project, ANR-17-CE32-0011; Labex COTE, ANR-10-LABX-45). The Genome Transcriptome Facility of Bordeaux (PGTB) received grants from *Investissements d'avenir* and *convention attributive d'aide EquipEx Xyloforest* (ANR-10-EQPX-16-01).

Bibliography

Bibliography

1. Abarenkov, K. *et al. UNITE general FASTA release for eukaryotes* (ed Community, U.) version 10.05.2021. 2021.
2. Abdelfattah, A., Malacrinò, A., Wisniewski, M., Cacciola, S. O. & Schena, L. Metabarcoding: A powerful tool to investigate microbial communities and shape future plant protection strategies. *Biol. Control* **120,** 1–10. DOI: 10.1016/j.biocontrol.2017.07.009 (2018).
3. Agler, M. T. *et al.* Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLoS Biol.* **14,** e1002352. DOI: 10.1371/journal.pbio.1002352 (2016).
4. Ai, L., Muggleton, S. H., Hocquette, C., Gromowski, M. & Schmid, U. Beneficial and harmful explanatory machine learning. *Mach. Learn.* **110,** 695–721. DOI: 10.1007/s10994-020-05941-0 (2021).
5. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–160. DOI: 10.1111/j.2517-6161.1982.tb01195.x (1982).
6. Alemao, C. A. *et al.* Impact of diet and the bacterial microbiome on the mucous barrier and immune disorders. *Allergy* **76,** 714–734. DOI: 10.1111/all.14548 (2021).
7. Alonso-Alemany, D., Clemente, J. C., Jansson, J. & Valiente, G. Taxonomic Assignment in Metagenomics with TANGO. *EMBnet.journal* **17,** 16–20. DOI: 10.14806/ej.17.2.237 (2011).
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410. DOI: 10.1016/S0022-2836(05)80360-2 (1990).
9. Amini, A., Muggleton, S. H., Lodhi, H. & Sternberg, M. J. E. A Novel Logic-Based Approach for Quantitative Toxicology Prediction. *J. Chem. Inf. Model.* **47,** 998–1006. DOI: 10.1021/ci600223d (2007).
10. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2.** DOI: 10.1128/mSystems.00191-16 (2017).
11. Anderson, M. J. Permutation tests for univariate or multivariate analysis of variance and regression. *Can. J. Fish. Aquat. Sci.* DOI: 10.1139/f01-004 (2011).
12. Aragona, M. *et al.* New-Generation Sequencing Technology in Diagnosis of Fungal Plant Pathogens: A Dream Comes True? *J. Fungi* **8,** 737. DOI: 10.3390/jof8070737 (2022).

13. Araújo, M. B., Pearson, R. G., Thuiller, W. & Erhard, M. Validation of species–climate impact models under climate change. *Global Change Biol.* **11,** 1504–1513. DOI: 10.1111/j.1365-2486.2005.01000.x (2005).

14. Armijo, G. *et al.* Grapevine Pathogenic Microorganisms: Understanding Infection Strategies and Host Response Scenarios. *Front. Plant Sci.* **0.** DOI: 10.3389/fpls.2016.00382 (2016).

15. Arnold, A. E. *et al.* Fungal endophytes limit pathogen damage in a tropical tree. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 15649–15654. DOI: 10.1073/pnas.2533483100 (2003).

16. Baird, D. J. & Hajibabaei, M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* **21,** 2039–2044. DOI: 10.1111/j.1365-294X.2012.05519.x (2012).

17. Bakshi, S., Sztejnberg, A. & Yarden, O. Isolation and Characterization of a Cold-Tolerant Strain of *Fusarium proliferatum*, a Biocontrol Agent of Grape Downy Mildew. *Phytopathology* **91,** 1062–1068. DOI: 10.1094/PHYTO.2001.91.11.1062 (2001).

18. Baksi, K. D., Kuntal, B. K. & Mande, S. S. 'TIME': A Web Application for Obtaining Insights into Microbial Ecology Using Longitudinal Microbiome Data. *Front. Microbiol.* **9,** 36. DOI: 10.3389/fmicb.2018.00036 (2018).

19. Bálint, M. *et al.* Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiol. Rev.* **6,** 189–96. DOI: 10.1093/femsre/fuw017 (2016).

20. Bálint, M. *et al.* Relocation, high-latitude warming and host genetic identity shape the foliar fungal microbiome of poplars. *Mol. Ecol.* **24,** 235–248. DOI: 10.1111/mec.13018 (2015).

21. Bang, C. *et al.* Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? *Zoology* **127,** 1–19. DOI: 10.1016/j.zool.2018.02.004 (2018).

22. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* **115,** 6506–6511. DOI: 10.1073/pnas.1711842115 (2018).

23. Barabási, A.-L., Albert, R. & Jeong, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A* **281,** 69–77. DOI: 10.1016/S0378-4371(00)00018-2 (2000).

24. Barroso-Bergada, D. *et al.* in *Inductive Logic Programming* 26–40 (Springer International Publishing, Cham, Switzerland, 2022). DOI: 10.1007/978-3-030-97454-1_3.

25. Barroso-Bergadà, D. *et al.* Metagenomic Next-Generation Sequencing (mNGS) Data Reveal the Phyllosphere Microbiome of Wheat Plants Infected by the Fungal Pathogen *Zymoseptoria tritici*. *Phytobiomes Journal.* DOI: 10.1094/PBIOMES-02-22-0008-FI (2022).

26. Barroso-Bergadà, D. *et al.* Microbial networks inferred from environmental DNA data for biomonitoring ecosystem change: Strengths and pitfalls. *Mol. Ecol. Resour.* **21,** 762–780. DOI: 10.1111/1755-0998.13302 (2021).

27. Basenko, E. Y. *et al.* FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *Journal of Fungi* **4.** DOI: 10.3390/jof4010039 (2018).

28. Bastian, M., Heymann, S. & Jacomy, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks* in (2009). DOI: 10.13140/2.1.1341.1520.

29. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Soft.* **67,** 1–48. DOI: 10.18637/jss.v067.i01 (2015).

30. Beckers, B. *et al.* Performance of 16s rDNA Primer Pairs in the Study of Rhizosphere and Endosphere Bacterial Microbiomes in Metabarcoding Studies. *Frontiers in Microbiology* **7.** DOI: 10.3389/fmicb.2016.00650 (2016).

31. Beckh, K. *et al.* Explainable Machine Learning with Prior Knowledge: An Overview. *arXiv.* DOI: 10.48550/arXiv.2105.10172 (2021).

32. Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W. & Courchamp, F. Impacts of climate change on the future of biodiversity. *Ecol. Lett.* **15,** 365–377. DOI: 10.1111/j.1461-0248.2011.01736.x (2012).

33. Belle, V. & Papantonis, I. Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data* **4.** DOI: 10.3389/fdata.2021.688969 (2021).

34. Bengtsson-Palme, J. *et al.* metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* **15,** 1403–1414. DOI: 10.1111/1755-0998.12399 (2015).

35. Bensch, K. *et al.* Common but different: The expanding realm of *Cladosporium*. *Stud. Mycol.* **82,** 23–74. DOI: 10.1016/j.simyco.2015.10.001 (2015).

36. Berendsen, R. L., Pieterse, C. M. J. & Bakker, P. A. H. M. The rhizosphere microbiome and plant health. *Trends Plant Sci.* **17,** 478–486. DOI: 10.1016/j.tplants.2012.04.001 (2012).

37. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* **5,** 219. DOI: 10.3389/fmicb.2014.00219 (2014).

38. Bettenfeld, P. *et al.* The microbiota of the grapevine holobiont: A key component of plant health. *J. Adv. Res.* DOI: 10.1016/j.jare.2021.12.008 (2021).

39. Blanchard, J. L. A rewired food web. *Nature* **527,** 173–174. DOI: 10.1038/nature16311 (2015).

40. Błaszczyk, L., Salamon, S. & Mikołajczak, K. Fungi Inhabiting the Wheat Endosphere. *Pathogens* **10.** DOI: 10.3390/pathogens10101288 (2021).

41. Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A. & Tamaddoni-Nezhad, A. Automated Discovery of Food Webs from Ecological Data Using Logic-Based Machine Learning. *PLOS ONE* **6,** 1–9. DOI: 10.1371/journal.pone.0029028 (2011).

42. Bohan, D. A., Richter, A., Bane, M., Therond, O. & Pocock, M. J. O. Farmer-led agroecology for biodiversity with climate change. *Trends Ecol. Evol.* **0.** DOI: 10.1016/j.tree.2022.07.006 (2022).

43. Bohan, D. A. *et al.* Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. *Trends in Ecology & Evolution* **32,** 477–487. DOI: https://doi.org/10.1016/j.tree.2017.03.001 (2017).

44. Bolker, B. M. *et al.* Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* **24,** 127–135. DOI: 10.1016/j.tree.2008.10.008 (2009).

45. Brader, G. *et al.* Ecology and Genomic Insights into Plant-Pathogenic and Plant-Nonpathogenic Endophytes. *Annu. Rev. Phytopathol.* **55,** 61–83. DOI: 10.1146/annurev-phyto-080516-035641 (2017).

46. Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J. & Cristescu, M. E. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecol. Evol.* **5,** 2234–2251. DOI: 10.1002/ece3.1485 (2015).

47. Bruisson, S. *et al.* Endophytes and Epiphytes From the Grapevine Leaf Microbiome as Potential Biocontrol Agents Against Phytopathogens. *Frontiers in Microbiology* **10.** DOI: 10.3389/fmicb.2019.02726 (2019).

48. Bryant, C. *et al.* Combining Inductive Logic Programming, Active Learning and Robotics to Discover the Function of Genes. *Electron. Trans. Artif. Intell.* **5,** 1–36 (2001).

49. Burki, F., Sandin, M. M. & Jamy, M. Diversity and ecology of protists revealed by metabarcoding. *Curr. Biol.* **31,** R1267–R1280. DOI: 10.1016/j.cub.2021.07.066 (2021).

50. Burruano, S., Mondello, V. & Conigliaro, G. Endophytic fungi in asymptomatic *Vitis vinifera L.* and their effects on *Plasmopara viticola. Biocontrol of major grapevine diseases: leading research,* 98–112. DOI: 10.1079/9781780647128.0098 (2016).

51. Busby, P. E. *et al.* Research priorities for harnessing plant microbiomes in sustainable agriculture. *PLoS Biol.* **15,** e2001793. DOI: 10.1371/journal.pbio.2001793 (2017).

52. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11,** 2639–2643. DOI: 10.1038/ismej.2017.119 (2017).

53. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data - Nature Methods. *Nat. Methods* **13,** 581–583. DOI: 10.1038/nmeth.3869 (2016).

54. Cambon, D. M. C. *et al.* Drought tolerance traits in Neotropical trees correlate with the composition of phyllosphere fungal communities. *Phytobiomes Journal.* DOI: 10.1094/PBIOMES-04-22-0023-R (2022).

55. Cao, Q. *et al.* Effects of Rare Microbiome Taxa Filtering on Statistical Analysis. *Frontiers in Microbiology* **11.** DOI: 10.3389/fmicb.2020.607325 (2021).

56. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data - Nature Methods. *Nat. Methods* **7,** 335–336. DOI: 10.1038/nmeth.f.303 (2010).

57. Caron, D. A. *et al.* Defining DNA-Based Operational Taxonomic Units for Microbial-Eukaryote Ecology. *Applied and Environmental Microbiology* **75,** 5797–5808. DOI: 10.1128/AEM.00298-09 (2009).

58. Carr, A., Diener, C., Baliga, N. S. & Gibbons, S. M. Use and abuse of correlation analyses in microbial ecology. *ISME J.* **13,** 2647–2655. DOI: 10.1038/s41396-019-0459-z (2019).

59. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42,** D459–D471. DOI: 10.1093/nar/gkt1103 (2014).

60. Castañeda, L. E., Miura, T., Sánchez, R. & Barbosa, O. Effects of agricultural management on phyllosphere fungal diversity in vineyards and the association with adjacent native forests. *PeerJ* **6,** e5715. DOI: 10.7717/peerj.5715 (2018).

61. Castelvecchi, D. Can we open the black box of AI? *Nature News* **538,** 20. DOI: 10.1038/538020a (2016).

62. Chagnon, P.-L., U'Ren, J. M., Miadlikowska, J., Lutzoni, F. & Elizabeth Arnold, A. Interaction type influences ecological network structure more than local abiotic conditions: evidence from endophytic and endolichenic fungi at a continental scale. *Oecologia* **180,** 181–191. DOI: 10.1007/s00442-015-3457-5 (2016).

63. Chang, W. *et al. shiny: Web Application Framework for R* R package version 1.7.1 (2021).

64. Chao, A., Chazdon, R. L., Colwell, R. K. & Shen, T.-J. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol. Lett.* **8,** 148–159. DOI: 10.1111/j.1461-0248.2004.00707.x (2005).

65. Chaudhry, V. *et al.* Shaping the leaf microbiota: plant–microbe–microbe interactions. *Journal of Experimental Botany* **72,** 36–56. DOI: 10.1093/jxb/eraa417 (2020).

66. Chelius, M. K. & Triplett, E. W. The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. *Microb. Ecol.* **41,** 252–263. DOI: 10.1007/s002480000087 (2001).

67. Chen, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots* R package version 1.7.3 (2022).

68. Chiquet, J., Robin, S. & Mariadassou, M. *Variational Inference for sparse network reconstruction from count data* in *Proceedings of the 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) **97** (PMLR, 2019), 1162–1171.

69. Cline, L. C., Song, Z., Al-Ghalith, G. A., Knights, D. & Kennedy, P. G. Moving beyond de novo clustering in fungal community ecology. *The New Phytologist* **216,** 629–634 (2017).

70. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42,** issue.
DOI: 10.1093/nar/gkt1244 (2014).

71. Collins, W. *Collins Online Dictionary | Definitions, Thesaurus and Translations* [Online; accessed 18. Jul. 2022]. 2022.

72. Compson, Z. G. *et al.* Network-Based Biomonitoring: Exploring Freshwater Food Webs With Stable Isotope Analysis and DNA Metabarcoding. *Frontiers in Ecology and Evolution* **7.** DOI: 10.3389/fevo.2019.00395 (2019).

73. Connor, N., Barberán, A. & Clauset, A. Using null models to infer microbial co-occurrence networks. *PLoS One* **12,** e0176751.
DOI: 10.1371/journal.pone.0176751 (2017).

74. Cordier, T. *et al.* Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Mol. Ecol.* **30,** 2937–2958. DOI: 10.1111/mec.15472 (2021).

75. Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M. & Relman, D. A. The Application of Ecological Theory Toward an Understanding of the Human Microbiome. *Science* **336,** 1255–1262.
DOI: 10.1126/science.1224203 (2012).

76. Council of the European Union. *Council Regulation (EC) No834/2007 of 28 June 2007. On organic production and labelling of organic products and repealing Regulation (EEC) No 2092/91.* 2007.

77. Crhanova, M. *et al.* Systematic Culturomics Shows that Half of Chicken Caecal Microbiota Members can be Grown in Vitro Except for Two Lineages of *Clostridiales* and a Single Lineage of *Bacteroidetes*. *Microorganisms* **7,** 496. DOI: 10.3390/microorganisms7110496 (2019).

78. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695,** 1–9 (2006).

79. D'Hondt, K. *et al.* Microbiome innovations for a sustainable future - Nature Microbiology. *Nat. Microbiol.* **6,** 138–142.
DOI: 10.1038/s41564-020-00857-w (2021).

80. Dai, W.-Z., Xu, Q., Yu, Y. & Zhou, Z.-H. *Bridging Machine Learning and Logical Reasoning by Abductive Learning* in *Advances in Neural Information Processing Systems* (eds Wallach, H. *et al.*) **32** (Curran Associates, Inc., 2019).

81. Daley, R. J. & Hobbie, J. E. Direct counts of aquatic bacteria by a modified epifluorescence technique1. *Limnol. Oceanogr.* **20,** 875–882. DOI: 10.4319/lo.1975.20.5.0875 (1975).

82. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6,** 1–14. DOI: 10.1186/s40168-018-0605-2 (2018).

83. De Queiroz, K. Species Concepts and Species Delimitation. *Syst. Biol.* **56,** 879–886. DOI: 10.1080/10635150701701083 (2007).

84. Deliere, L. *et al.* Le réseau Ecoviti Bordeaux expérimente des systémes de cultures viticoles "bas intrants". *Phytoma* **1,** 51–55 (2014).

85. Delmas, E. *et al.* Analysing ecological networks of species interactions. *Biol. Rev. Camb. Philos. Soc.* DOI: 10.1111/brv.12433 (2018).

86. Derocles, S. A. P. *et al.* Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis. *Advances in Ecological Research* **58,** 1–62. DOI: 10.1016/bs.aecr.2017.12.001 (2018).

87. De Vries, F. T. *et al.* Abiotic drivers and plant traits explain landscape-scale patterns in soil microbial communities. *Ecology Letters* **15,** 1230–1239. DOI: https://doi.org/10.1111/j.1461-0248.2012.01844.x (2012).

88. De Vries, F. T. *et al.* Soil bacterial networks are less stable under drought than fungal networks. *Nat. Commun.* **9,** 1–12. DOI: 10.1038/s41467-018-05516-7 (2018).

89. Dickie, I. A. *et al.* Towards robust and repeatable sampling methods in eDNA-based studies. *Mol. Ecol. Resour.* **18,** 940–952. DOI: 10.1111/1755-0998.12907 (2018).

90. Directorate-General for Agriculture and Rural Development. *The Commission publishes previsions for EU wine production 2021/22* 2021.

91. Dissanayake, A. J., Purahong, W., Wubet, T., Hyde, K. D. & Yan, J. Direct comparison of culture-dependent and culture-independent molecular approaches reveal the diversity of fungal endophytic communities in stems of grapevine (*Vitis vinifera*). *Fungal Diversity* **90.** DOI: 10.1007/s13225-018-0399-3 (2018).

92. Djemiel, C. *et al.* Inferring microbiota functions from taxonomic genes: a review. *GigaScience* **11,** giab090. DOI: 10.1093/gigascience/giab090 (2022).

93. Dohlman, A. B. & Shen, X. Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference. *Experimental Biology and Medicine* **244.** PMID: 30880449, 445–458. DOI: 10.1177/1535370219836771 (2019).

94. Dollive, S. *et al.* A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol.* **13,** R60. DOI: 10.1186/gb-2012-13-7-r60 (2012).

95. Dubart, M. *et al.* Coupling ecological network analysis with high-throughput sequencing-based surveys: Lessons from the next-generation biomonitoring project. *Advances in Ecological Research* **65,** 367–430. DOI: 10.1016/bs.aecr.2021.10.007 (2021).

96. Duhan, D. *et al.* Functional Characterization of the Nep1-Like Protein Effectors of the Necrotrophic Pathogen – *Alternaria brassicae*. *Frontiers in Microbiology* **12.** DOI: 10.3389/fmicb.2021.738617 (2021).

97. Durán, P. *et al.* Microbial Community Composition in Take-All Suppressive Soils. *Frontiers in Microbiology* **9.** DOI: 10.3389/fmicb.2018.02198 (2018).

98. Duvivier, M., Dedeurwaerder, G., De Proft, M., Moreau, J.-M. & Legrève, A. Real-time PCR quantification and spatio-temporal distribution of airborne inoculum of *Mycosphaerella graminicola* in Belgium. *Eur. J. Plant Pathol.* **137,** 325–341. DOI: 10.1007/s10658-013-0245-0 (2013).

99. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*. DOI: https://doi.org/10.1101/081257 (2016).

100. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6,** e4652. DOI: 10.7717/peerj.4652 (2018).

101. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10,** 996–998. DOI: 10.1038/nmeth.2604 (2013).

102. Edgar, R. C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31,** 3476–3482. DOI: 10.1093/bioinformatics/btv401 (2015).

103. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap Monographs on Statistics and Applied Probability* **57** (Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993).

104. European Commission. *DIRECTIVE 2009/128/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 October 2009. Establishing a framework for Community action to achieve the sustainable use of pesticides.* 2009.

105. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–1039. DOI: 10.1126/science.1153213 (2008).

106. Fan, J., Upadhye, S. & Worster, A. Understanding receiver operating characteristic (ROC) curves. *CJEM* **8,** 19–20. DOI: 10.1017/s1481803500013336 (2006).

107. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31,** 3172–3180. DOI: 10.1093/bioinformatics/btv349 (2015).

108. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25,** 56–66. DOI: 10.1016/j.mib.2015.04.004 (2015).

109. Faust, K. & Raes, J. CoNet app: inference of biological association networks using Cytoscape. *F1000Research* **5.**
DOI: 10.12688/f1000research.9050.2 (2016).

110. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10,** 538–550. DOI: 10.1038/nrmicro2832 (2012).

111. Faust, K. *et al.* Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput. Biol.* **8,** e1002606.
DOI: 10.1371/journal.pcbi.1002606 (2012).

112. Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS One* **9,** e102451. DOI: 10.1371/journal.pone.0102451 (2014).

113. Fleming, A. On the Antibacterial Action of Cultures of a *Penicillium*, with Special Reference to their Use in the Isolation of B. influenzæ. *Br. J. Exp. Pathol.* **10,** 226 (1929).

114. Fones, H. & Gurr, S. The impact of *Septoria tritici* Blotch disease on wheat: An EU perspective. *Fungal Genet. Biol.* **79,** 3–7.
DOI: 10.1016/j.fgb.2015.04.004 (2015).

115. Fonseca, V. G. "Pitfalls in relative abundance estimation using eDNA metabarcoding". *Mol. Ecol. Resour.* **18,** 923–926. DOI: 10.1111/1755-0998.12902 (2018).

116. Fontaine, M. C. *et al.* Europe as a bridgehead in the worldwide invasion history of grapevine downy mildew, *Plasmopara viticola*. *Curr. Biol.* **31,** 2155–2166.e4. DOI: 10.1016/j.cub.2021.03.009 (2021).

117. Fort, T., Robin, C., Capdevielle, X., Delière, L. & Vacher, C. Foliar fungal communities strongly differ between habitat patches in a landscape mosaic. *PeerJ* **4,** e2656. DOI: 10.7717/peerj.2656 (2016).

118. Francioli, D., Lentendu, G., Lewin, S. & Kolb, S. DNA Metabarcoding for the Characterization of Terrestrial Microbiota—Pitfalls and Solutions. *Microorganisms* **9,** 361. DOI: 10.3390/microorganisms9020361 (2021).

119. Franco, N. R. *et al.* Bacterial composition and diversity in deep-sea sediments from the Southern Colombian Caribbean Sea. *Diversity* **13,** 10 (2020).

120. Fredrickson, A. G. & Stephanopoulos, G. Microbial Competition. *Science* **213,** 972–979. DOI: 10.1126/science.7268409 (1981).

121. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology* **8,** 1–11.
DOI: 10.1371/journal.pcbi.1002687 (2012).

122. Friesen, T. L. *et al.* Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat. Genet.* **38,** 953–956. DOI: 10.1038/ng1839 (2006).

123. Frøslev, T. G. *et al.* Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates - Nature Communications. *Nat. Commun.* **8,** 1–11. DOI: 10.1038/s41467-017-01312-x (2017).

124. Fruchterman, T. M. J. & Reingold, E. M. Graph drawing by force-directed placement. *Softw.: Pract. Exper.* **21,** 1129–1164. DOI: 10.1002/spe.4380211102 (1991).

125. Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459,** 193–199. DOI: 10.1038/nature08058 (2009).

126. Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* **13,** 133–146. DOI: 10.1038/nrmicro3417 (2015).

127. Galan, M. *et al.* 16S rRNA Amplicon Sequencing for Epidemiological Surveys of Bacteria in Wildlife. *mSystems* **1.** DOI: 10.1128/mSystems.00032-16 (2016).

128. Galiana, N. *et al.* Ecological network complexity scales with area. *Nat. Ecol. Evol.* **6,** 307–314. DOI: 10.1038/s41559-021-01644-4 (2022).

129. Gao, Z., Kang, Y., Yu, J. & Ren, L. Human Pharyngeal Microbiome May Play A Protective Role in Respiratory Tract Infections. *Genomics Proteomics Bioinformatics* **12,** 144–150. DOI: 10.1016/j.gpb.2014.06.001 (2014).

130. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for basidiomycetes–application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2,** 113–118. DOI: 10.1111/j.1365-294x.1993.tb00005.x (1993).

131. Gerber, G. K. The dynamic microbiome. *FEBS Lett.* **588,** 4131–4139. DOI: 10.1016/j.febslet.2014.02.037 (2014).

132. Getoor, L. & Taskar, B. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)* (The MIT Press, 2007).

133. Ghosh, S., Chowdhury, R. & Bhattacharya, P. Mixed consortia in bioprocesses: role of microbial interactions. *Appl. Microbiol. Biotechnol.* **100,** 4283–4295. DOI: 10.1007/s00253-016-7448-1 (2016).

134. Ghule, M. R., Sawant, I. S., Sawant, S. D., Sharma, R. & Shouche, Y. S. Identification of *Fusarium* species as putative mycoparasites of *Plasmopara viticola* causing downy mildew in grapevines. *Australasian Plant Dis. Notes* **13,** 1–6. DOI: 10.1007/s13314-018-0297-2 (2018).

135. Gilbert, J. A. & Neufeld, J. D. Life in a World without Microbes. *PLoS Biol.* **12,** e1002020. DOI: 10.1371/journal.pbio.1002020 (2014).

136. Gilpin, L. H. *et al.* in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* 80–89 (IEEE, 2018). DOI: 10.1109/DSAA.2018.00018.

137. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8,** 2224. DOI: 10.3389/fmicb.2017.02224 (2017).

138. Goel, N. S., Maitra, S. C. & Montroll, E. W. On the Volterra and Other Nonlinear Models of Interacting Populations. *Rev. Mod. Phys.* **43,** 231–276. DOI: 10.1103/RevModPhys.43.231 (1971).

139. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* **3,** 679–687. DOI: 10.1038/nrmicro1204 (2005).

140. Golubev, W. *Antagonistic Interactions Among Yeasts* 197–219. DOI: 10.1007/3-540-30985-3_10 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).

141. Gotelli, N. J. NULL MODEL ANALYSIS OF SPECIES CO-OCCURRENCE PATTERNS. *Ecology* **81,** 2606–2621. DOI: 10.1890/0012-9658(2000)081[2606:NMAOSC]2.0.CO;2 (2000).

142. Gray, C. *et al.* Ecological plasticity governs ecosystem services in multilayer networks. *Commun. Biol.* **4,** 1–7. DOI: 10.1038/s42003-020-01547-3 (2021).

143. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41,** D597–D604. DOI: 10.1093/nar/gks1160 (2013).

144. Hacquard, S., Spaepen, S., Garrido-Oter, R. & Schulze-Lefert, P. Interplay Between Innate Immunity and the Plant Microbiota. *Annu. Rev. Phytopathol.* **55,** 565–589. DOI: 10.1146/annurev-phyto-080516-035623 (2017).

145. Han, M., Sun, L., Gan, D., Fu, L. & Zhu, B. Root functional traits are key determinants of the rhizosphere effect on soil organic matter decomposition across 14 temperate hardwood species. *Soil Biol. Biochem.* **151,** 108019. DOI: 10.1016/j.soilbio.2020.108019 (2020).

146. Handelsman, J. Metagenomics and Microbial Communities. *eLS.* DOI: 10.1002/9780470015902.a0020367 (2007).

147. Harm, A., Kassemeyer, H.-H., Seibicke, T. & Regner, F. Evaluation of Chemical and Natural Resistance Inducers against Downy Mildew (*Plasmopara viticola*) in Grapevine. *Am. J. Enol. Vitic.* **62,** 184–192. DOI: 10.5344/ajev.2011.09054 (2011).

148. Hassani, M. A., Durán, P. & Hacquard, S. Microbial interactions within the plant holobiont. *Microbiome* **6,** 1–17. DOI: 10.1186/s40168-018-0445-0 (2018).

149. He, Y. *et al.* Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **3,** 1–10. DOI: 10.1186/s40168-015-0081-x (2015).

150. Heeger, F. *et al.* Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Mol. Ecol. Resour.* **18,** 1500–1514. DOI: 10.1111/1755-0998.12937 (2018).

151. Hilber-Bodmer, M., Schmid, M., Ahrens, C. H. & Freimoser, F. M. Competition assays and physiological experiments of soil and phyllosphere yeasts identify *Candida subhashii* as a novel antagonist of filamentous fungi. *BMC Microbiol.* **17,** 1–15. DOI: 10.1186/s12866-016-0908-z (2017).

152. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinf.* **20,** 329. DOI: 10.1186/s12859-019-2915-1 (2019).

153. Hooke, R. *Micrographia or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries Thereupon* (John Martyn, & James Allestry, London, England, UK, 1664).

154. Hromada, S. *et al.* Negative interactions determine Clostridioides difficile growth in synthetic human gut communities. *Mol. Syst. Biol.* **17,** e10355. DOI: 10.15252/msb.202110355 (2021).

155. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **82,** 801–811. DOI: 10.1016/j.humimm.2021.02.012 (2021).

156. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1,** 1–6. DOI: 10.1038/nmicrobiol.2016.48 (2016).

157. Ings, T. C. *et al.* Ecological networks–beyond food webs. *J. Anim. Ecol.* **78,** 253–269. DOI: 10.1111/j.1365-2656.2008.01460.x (2009).

158. Ishaq, S. L. Plant-microbial interactions in agriculture and the use of farming systems to improve diversity and productivity. *AIMS Microbiol.* **3,** 335. DOI: 10.3934/microbiol.2017.2.335 (2017).

159. Jakuschkin, B. *et al.* Deciphering the Pathobiome: Intra- and Interkingdom Interactions Involving the Pathogen *Erysiphe alphitoides*. *Microb. Ecol.* **72,** 870–880. DOI: 10.1007/s00248-016-0777-x (2016).

160. Joos, L. *et al.* Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. *BMC Genomics* **21,** 1–17. DOI: 10.1186/s12864-020-07126-4 (2020).

161. Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237,** 82–88. DOI: 10.1038/237082a0 (1972).

162. Kakas, A. C., Kowalski, R. A. & Toni, F. Abductive Logic Programming. *J. Logic Comput.* **2,** 719–770. DOI: 10.1093/logcom/2.6.719 (1992).

163. Kakas, A. & Papadopoulos, G. Parallel Abduction in Logic Programming. *First International Symposium on Parallel Symbolic Computation* (1996).

164. Kamada, N., Chen, G. Y., Inohara, N. & Núñez, G. Control of pathogens and pathobionts by the gut microbiota. *Nat. Immunol.* **14,** 685–690. DOI: 10.1038/ni.2608 (2013).

165. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31,** 7–15. DOI: 10.1016/0020-0190(89)90102-6 (1989).

166. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30. DOI: 10.1093/nar/28.1.27 (2000).

167. Karimi, B. *et al.* Microbial diversity and ecological networks as indicators of environmental quality. *Environ. Chem. Lett.* **15,** 265–281. DOI: 10.1007/s10311-017-0614-6 (2017).

168. Kauffman, J., Kittas, A., Bennett, L. & Tsoka, S. DyCoNet: A Gephi Plugin for Community Detection in Dynamic Complex Networks. *PLoS One* **9,** e101357. DOI: 10.1371/journal.pone.0101357 (2014).

169. Kemen, E. Microbe–microbe interactions determine oomycete and fungal host colonization. *Curr. Opin. Plant Biol.* **20,** 75–81. DOI: 10.1016/j.pbi.2014.04.005 (2014).

170. Kemler, M., Witfeld, F., Begerow, D. & Yurkov, A. Phylloplane Yeasts in Temperate Climates. *Yeasts in Natural Ecosystems: Diversity,* 171–197. DOI: 10.1007/978-3-319-62683-3_6 (2017).

171. Kerdraon, L., Barret, M., Laval, V. & Suffert, F. Differential dynamics of microbial community networks help identify microorganisms interacting with residue-borne pathogens: the case of *Zymoseptoria tritici* in wheat. *Microbiome* **7,** 1–17. DOI: 10.1186/s40168-019-0736-0 (2019).

172. Kernaghan, G., Mayerhofer, M. & Griffin, A. Fungal endophytes of wild and hybrid *Vitis* leaves and their potential for vineyard biocontrol. *Can. J. Microbiol.* **63,** 583–595. DOI: 10.1139/cjm-2016-0740 (2017).

173. Klassen, R., Schaffrath, R., Buzzini, P. & Ganter, P. F. Antagonistic Interactions and Killer Yeasts. *Yeasts in Natural Ecosystems: Ecology,* 229–275. DOI: 10.1007/978-3-319-61575-2_9 (2017).

174. Ko, D. K. & Brandizzi, F. Network-based approaches for understanding gene regulation and function in plants. *Plant J.* **104,** 302–317. DOI: 10.1111/tpj.14940 (2020).

175. Koch, H. & Schmid-Hempel, P. Socially transmitted gut microbiota protect bumble bees against an intestinal parasite. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 19288–19292. DOI: 10.1073/pnas.1110474108 (2011).

176. Koch, R. An Address on *Cholera* and its *Bacillus*. *Br. Med. J.* **2,** 453. DOI: 10.1136/bmj.2.1236.453 (1884).

177. Kodikara, S., Ellul, S. & Lê Cao, K.-A. Statistical challenges in longitudinal microbiome data analysis. *Briefings Bioinf.* **23,** bbac273. DOI: 10.1093/bib/bbac273 (2022).

178. Konopka, A. What is microbial community ecology? - The ISME Journal. *ISME J.* **3,** 1223–1230. DOI: 10.1038/ismej.2009.88 (2009).

179. Kortsch, S., Primicerio, R., Fossheim, M., Dolgov, A. V. & Aschan, M. Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists. *Proc. R. Soc. B.* **282,** 20151546. DOI: 10.1098/rspb.2015.1546 (2015).

180. Kumar, A. & Verma, J. P. Does plant—Microbe interaction confer stress tolerance in plants: A review? *Microbiol. Res.* **207,** 41–52. DOI: 10.1016/j.micres.2017.11.004 (2018).

181. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology* **11,** 1–25. DOI: 10.1371/journal.pcbi.1004226 (2015).

182. Kuzyakov, Y. & Cheng, W. Photosynthesis controls of rhizosphere respiration and organic matter decomposition. *Soil Biol. Biochem.* **33,** 1915–1925. DOI: 10.1016/S0038-0717(01)00117-1 (2001).

183. Lafferty, K. D., Dobson, A. P. & Kuris, A. M. Parasites dominate food web links. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 11211–11216. DOI: 10.1073/pnas.0604755103 (2006).

184. Laforest-Lapointe, I., Paquette, A., Messier, C. & Kembel, S. W. Leaf bacterial diversity mediates plant diversity and ecosystem function relationships. *Nature* **546,** 145–147. DOI: 10.1038/nature22399 (2017).

185. Lagier, J.-C. *et al.* Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16,** 540–550. DOI: 10.1038/s41579-018-0041-0 (2018).

186. Lakshmanan, V., Ray, P. & Craven, K. D. in *Plant Stress Tolerance* 69–84 (Springer, New York, NY, USA, 2017). DOI: 10.1007/978-1-4939-7136-7_4.

187. Lamb, P. D. *et al.* How quantitative is metabarcoding: A meta-analytical approach. *Mol. Ecol.* **28,** 420–430. DOI: 10.1111/mec.14920 (2019).

188. Lane, N. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Phil. Trans. R. Soc. B* **370,** 20140344. DOI: 10.1098/rstb.2014.0344 (2015).

189. Latz, M. A. C. *et al.* Short- and long-read metabarcoding of the eukaryotic rRNA operon: Evaluation of primers and comparison to shotgun metagenomics sequencing. *Mol. Ecol. Resour.* **22,** 2304–2318. DOI: 10.1111/1755-0998.13623 (2022).

190. Laur, J. *et al.* Effectors involved in fungal-fungal interaction lead to a rare phenomenon of hyperbiotrophy in the tritrophic system biocontrol agent-powdery mildew-plant. *New Phytol.* **217,** 713–725. DOI: 10.1111/nph.14851 (2018).

191. Leander, B. S. Predatory protists. *Curr. Biol.* **30,** 510–516. DOI: 10.1016/j.cub.2020.03.052 (2020).

192. Lennard, K. *et al.* Microbial Composition Predicts Genital Tract Inflammation and Persistent Bacterial Vaginosis in South African Adolescent Females. *Infection and Immunity* **86,** e00410–17. DOI: 10.1128/IAI.00410-17 (2018).

193. Li, C., Lim, K. M. K., Chng, K. R. & Nagarajan, N. Predicting microbial interactions through computational approaches. *Methods* **102,** 12–9. DOI: 10.1016/j.ymeth.2016.02.019 (2016).

194. Li, J., Tai, B. C. & Nott, D. J. Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test. *Journal of Nonparametric Statistics* **21,** 649–661. DOI: 10.1080/10485250902770035 (2009).

195. Lim, K. M. K., Li, C., Chng, K. R. & Nagarajan, N. @MInter: automated text-mining of microbial interactions. *Bioinformatics* **32,** 2981–2987. DOI: 10.1093/bioinformatics/btw357 (2016).

196. Links, M. G., Chaban, B., Hemmingsen, S. M., Muirhead, K. & Hill, J. E. mPUMA: a computational approach to microbiota analysis by de novo assembly of operational taxonomic units based on protein-coding barcode sequences. *Microbiome* **1,** 1–7. DOI: 10.1186/2049-2618-1-23 (2013).

197. Liu, H., Roeder, K. & Wasserman, L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Neural Information Processing Systems* **24** (2010).

198. Liu, W. *et al.* Interspecific Bacterial Interactions are Reflected in Multispecies Biofilm Spatial Organization. *Front. Microbiol.* **7.** DOI: 10.3389/fmicb.2016.01366 (2016).

199. Lloyd, D. P. & Allen, R. J. Competition for space during bacterial colonization of a surface. *J. R. Soc. Interface* **12,** 20150608. DOI: 10.1098/rsif.2015.0608 (2015).

200. Lo, C. & Marculescu, R. Inferring Microbial Interactions from Metagenomic Time-series Using Prior Biological Knowledge. *ACM-BCB '17: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics,* 168–177. DOI: 10.1145/3107411.3107435 (2017).

201. Lo, C. & Marculescu, R. MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLOS Computational Biology* **13,** 1–20. DOI: 10.1371/journal.pcbi.1005915 (2017).

202. Lofgren, L. A. *et al.* Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Mol. Ecol.* **28,** 721–730. DOI: 10.1111/mec.14995 (2019).

203. Lopezaraiza-Mikel, M., Hayes, R., Whalley, M. & Memmott, J. The impact of an alien plant on a native plant-pollinator network: An experimental approach. *Ecology letters* **10,** 539–50. DOI: 10.1111/j.1461-0248.2007.01055.x (2007).

204. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353,** 1272–1277. DOI: 10.1126/science.aaf4507 (2016).

205. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 1–21. DOI: 10.1186/s13059-014-0550-8 (2014).

206. Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G. & Bar-Joseph, Z. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome* **7,** 1–14. DOI: 10.1186/s40168-019-0660-3 (2019).

207. Lutz, M. C., Lopes, C. A., Rodriguez, M. E., Sosa, M. C. & Sangorrın, M. P. Efficacy and putative mode of action of native and commercial antagonistic yeasts against postharvest pathogens of pear. *Int. J. Food Microbiol.* **164,** 166–172. DOI: 10.1016/j.ijfoodmicro.2013.04.005 (2013).

208. Ma, A. *et al.* Ecological networks reveal resilience of agro-ecosystems to changes in farming management. *Nat. Ecol. Evol.* **3,** 260–264. DOI: 10.1038/s41559-018-0757-2 (2019).

209. Mace, G. M., Norris, K. & Fitter, A. H. Biodiversity and ecosystem services: a multilayered relationship. *Trends in Ecology & Evolution* **27,** 19–26. DOI: https://doi.org/10.1016/j.tree.2011.08.006 (2012).

210. Macfadyen, S., Gibson, R., Raso, L., Sint, D. & Memmott, J. Parasitoid control of aphids in organic and conventional farming systems. *Agric. Ecosyst. Environ.* **133,** 14–18. DOI: 10.1016/j.agee.2009.04.012 (2009).

211. Magana, M. *et al.* Options and Limitations in Clinical Investigation of Bacterial Biofilms. *Clin. Microbiol. Rev.* **31.** DOI: 10.1128/CMR.00084-16 (2018).

212. Makiola, A. *et al.* Key Questions for Next-Generation Biomonitoring. *Frontiers in Environmental Science* **7.** DOI: 10.3389/fenvs.2019.00197 (2020).

213. Marchesi, J. R. *et al.* Design and Evaluation of Useful Bacterium-Specific PCR Primers That Amplify Genes Coding for Bacterial 16S rRNA. *Applied and Environmental Microbiology* **64,** 795–799. DOI: 10.1128/AEM.64.2.795-799.1998 (1998).

214. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences* **104,** 11889–11894. DOI: 10.1073/pnas.0704662104 (2007).

215. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif).* **6.** DOI: 10.1146/annurev-anchem-062012-092628 (2013).

216. Margulis, L. *et al.* Symbiogenesis and symbionticism. *Symbiosis as a source of evolutionary innovation: Speciation and morphogenesis* **10** (1991).

217. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12. DOI: 10.14806/ej.17.1.200 (2011).

218. Martini, M. *et al.* DNA-Dependent Detection of the Grapevine Fungal Endophytes *Aureobasidium pullulans* and *Epicoccum nigrum*. *Plant Dis.* **93,** 993–998. DOI: 10.1094/PDIS-93-10-0993 (2009).

219. Matchado, M. S. *et al.* Network analysis methods for studying microbial communities: A mini review. *Comput. Struct. Biotechnol. J.* **19,** 2687–2698. DOI: 10.1016/j.csbj.2021.05.001 (2021).

220. Mathon, L. *et al.* Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Mol. Ecol. Resour.* **21,** 2565–2579. DOI: 10.1111/1755-0998.13430 (2021).

221. Mayr, E. *The growth of biological thought: Diversity, evolution, and inheritance* (Harvard University Press, 1982).

222. McCauley, D. J. *et al.* From wing to wing: the persistence of long ecological interaction chains in less-disturbed ecosystems. *Sci. Rep.* **2,** 1–5. DOI: 10.1038/srep00409 (2012).

223. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6,** 610–618. DOI: 10.1038/ismej.2011.139 (2012).

224. McGovern, E., Waters, S. M., Blackshields, G. & McCabe, M. S. Evaluating Established Methods for Rumen 16S rRNA Amplicon Sequencing With Mock Microbial Populations. *Frontiers in Microbiology* **9.** DOI: 10.3389/fmicb.2018.01365 (2018).

225. McGovern, P. E., Glusker, D. L., Exner, L. J. & Voigt, M. M. Neolithic resinated wine. *Nature* **381,** 480–481. DOI: 10.1038/381480a0 (1996).

226. McIntosh, R. P. Raunkiaer's "Law of Frequency". *Ecology* **43,** 533–535. DOI: 10.2307/1933384 (1962).

227. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8,** 1–11. DOI: 10.1371/journal.pone.0061217 (2013).

228. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* **10,** e1003531. DOI: 10.1371/journal.pcbi.1003531 (2014).

229. Mencher, A., Morales, P., Tronchoni, J. & Gonzalez, R. Mechanisms Involved in Interspecific Communication between Wine Yeasts. *Foods* **10,** 1734. DOI: 10.3390/foods10081734 (2021).

230. Merkel, A. Y., Tarnovetskii, I. Y., Podosokorskaya, O. A. & Toshchakov, S. V. Analysis of 16S rRNA Primer Systems for Profiling of Thermophilic Microbial Communities. *Microbiology* **88,** 671–680. DOI: 10.1134/S0026261719060110 (2019).

231. Min, E. *et al. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media* in *Proceedings of the ACM Web Conference 2022* (Association for Computing Machinery, Virtual Event, Lyon, France, 2022), 1148–1158. DOI: 10.1145/3485447.3512163.

232. Mohan, S., Thirumalai, C. & Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **7,** 81542–81554. DOI: 10.1109/ACCESS.2019.2923707 (2019).

233. Molitor, D. & Beyer, M. Epidemiology, identification and disease management of grape black rot and potentially useful metabolites of black rot pathogens for industrial applications – a review. *Ann. Appl. Biol.* **165,** 305–317. DOI: 10.1111/aab.12155 (2014).

234. Montoya, D., Rogers, L. & Memmott, J. Emerging perspectives in the restoration of biodiversity-based ecosystem services. *Trends Ecol. Evol.* **27,** 666–672. DOI: 10.1016/j.tree.2012.07.004 (2012).

235. Moore, D. R. The Anna Karenina Principle Applied to Ecological Risk Assessments of Multiple Stressors. *Human and Ecological Risk Assessment: An International Journal* **7,** 231–237. DOI: 10.1080/20018091094349 (2001).

236. Morais, D., Gélisse, S., Laval, V., Sache, I. & Suffert, F. Inferring the origin of primary inoculum of *Zymoseptoria tritici* from differential adaptation of resident and immigrant populations to wheat cultivars. *Eur. J. Plant Pathol.* **145,** 393–404. DOI: 10.1007/s10658-015-0853-y (2016).

237. Moreau, D., Bardgett, R. D., Finlay, R. D., Jones, D. L. & Philippot, L. A plant perspective on nitrogen cycling in the rhizosphere. *Funct. Ecol.* **33,** 540–552. DOI: 10.1111/1365-2435.13303 (2019).

238. Morlon, H., Kefi, S. & Martinez, N. D. Effects of trophic similarity on community composition. *Ecol. Lett.* **17,** 1495–1506. DOI: 10.1111/ele.12356 (2014).

239. Morriën, W. E. *et al.* Soil networks become more connected and take up more carbon as nature restoration progresses. *Nat. Commun.* **8,** 14349. DOI: 10.1038/ncomms14349 (2017).

240. Morris, A., Meyer, K. & Bohannan, B. Linking microbial communities to ecosystem functions: what we can learn from genotype–phenotype mapping in organisms. *Phil. Trans. R. Soc. B* **375,** 20190244. DOI: 10.1098/rstb.2019.0244 (2020).

241. Mortimer, P. Koch's colonies and the culinary contribution of Fanny Hesse. *Microbiology Today* **28,** 136–137 (2001).

242. Mounier, J. *et al.* Microbial Interactions within a Cheese Microbial Community. *Applied and Environmental Microbiology* **74,** 172–181. DOI: 10.1128/AEM.01338-07 (2008).

243. Muggleton, S. Inverse entailment and progol. *NGCO* **13,** 245–286. DOI: 10.1007/BF03037227 (1995).

244. Muggleton, S. *et al.* Meta-Interpretive Learning from noisy images. *Mach. Learn.* **107,** 1097–1118. DOI: 10.1007/s10994-018-5710-8 (2018).

245. Muggleton, S. H. & Bryant, C. H. *Theory Completion Using Inverse Entailment* 130–146 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2000).

246. Muggleton, S. H., Lin, D., Pahlavi, N. & Tamaddoni-Nezhad, A. Meta-interpretive learning: application to grammatical inference. *Mach. Learn.* **94,** 25–49. DOI: 10.1007/s10994-013-5358-3 (2014).

247. Müller, C. L., Bonneau, R. & Kurtz, Z. Generalized Stability Approach for Regularized Graphical Models. *arXiv.* DOI: 10.48550/arXiv.1605.07072 (2016).

248. Mullis, K. *et al.* Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symp. Quant. Biol.* **51,** 263–273. DOI: 10.1101/SQB.1986.051.01.032 (1986).

249. Murall, C. L. *et al.* Invasions of Host-Associated Microbiome Networks. *Advances in Ecological Research* **57,** 201–281. DOI: 10.1016/bs.aecr.2016.11.002 (2017).

250. Musetti, R. *et al.* Antifungal activity of diketopiperazines extracted from *Alternaria alternata* against *Plasmopara viticola*: An ultrastructural study. *Micron* **38,** 643–650. DOI: 10.1016/j.micron.2006.09.001 (2007).

251. Musetti, R. *et al.* Inhibition of Sporulation and Ultrastructural Alterations of Grapevine Downy Mildew by the Endophytic Fungus *Alternaria alternata*. *Phytopathology* **96,** 689–698. DOI: 10.1094/PHYTO-96-0689 (2006).

252. Newman, M., Barabási, A.-L. & Watts, D. J. *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, NJ, USA, 2006).

253. Nguyen, N.-P., Warnow, T., Pop, M. & White, B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npj Biofilms Microbiomes* **2,** 1–8. DOI: 10.1038/npjbiofilms.2016.4 (2016).

254. Nguyen, N. H. *et al.* FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecol.* **20,** 241–248. DOI: 10.1016/j.funeco.2015.06.006 (2016).

255. Nicolaisen, M., Justesen, A., Knorr, K., Wang, J. & Pinnschmidt, H. Fungal communities in wheat grain show significant co-existence patterns among species. *Fungal Ecology* **11,** 145–153. DOI: https://doi.org/10.1016/j.funeco.2014.06.002 (2014).

256. Nieuwdorp, M., Gilijamse, P. W., Pai, N. & Kaplan, L. M. Role of the Microbiome in Energy Regulation and Metabolism. *Gastroenterology* **146,** 1525–1533. DOI: 10.1053/j.gastro.2014.02.008 (2014).

257. Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K.-H. Intraspecific ITS Variability in the Kingdom Fungi as Expressed in the International Sequence Databases and Its Implications for Molecular Species Identification. *Evolutionary Bioinformatics Online* **4,** 193. DOI: 10.4137/ebo.s653 (2008).

258. Nilsson, R. H. *et al.* The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* **47,** 259–264. DOI: 10.1093/nar/gky1022 (2019).

259. Ning, D., Deng, Y., Tiedje, J. M. & Zhou, J. A general framework for quantitatively assessing ecological stochasticity. *Proc. Natl. Acad. Sci. U.S.A.* **116,** 16892–16898. DOI: 10.1073/pnas.1904623116 (2019).

260. Nishad, R., Ahmed, T., Rahman, V. J. & Kareem, A. Modulation of Plant Defense System in Response to Microbial Interactions. *Frontiers in Microbiology* **11.** DOI: 10.3389/fmicb.2020.01298 (2020).

261. Ohlmann, M. *et al.* Mapping the imprint of biotic interactions on β-diversity. *Ecol. Lett.* **21,** 1660–1669. DOI: 10.1111/ele.13143 (2018).

262. Oksanen, J. *et al. vegan: Community Ecology Package* R package version 2.6-2 (2022).

263. Orellana-Torrejon, C. *et al.* Annual dynamics of *Zymoseptoria tritici* populations in wheat cultivar mixtures: A compromise between the efficacy and durability of a recently broken-down resistance gene? *Plant Pathol.* **71,** 289–303. DOI: 10.1111/ppa.13458 (2021).

264. Otu, H. H. & Sayood, K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* **19,** 2122–2130. DOI: 10.1093/bioinformatics/btg295 (2003).

265. Ovaskainen, O. *et al.* How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20,** 561–576. DOI: 10.1111/ele.12757 (2017).

266. Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat. Commun.* **10,** 1–12. DOI: 10.1038/s41467-018-07946-9 (2019).

267. Pacheco, A. R. & Segrè, D. A multidimensional perspective on microbial interactions. *FEMS Microbiol. Lett.* **366,** fnz125. DOI: 10.1093/femsle/fnz125 (2019).

268. Pacifico, D. *et al.* The Role of the Endophytic Microbiome in the Grapevine Response to Environmental Triggers. *Frontiers in Plant Science* **10.** DOI: 10.3389/fpls.2019.01256 (2019).

269. Pancher, M. *et al.* Fungal Endophytic Communities in Grapevines (*Vitis vinifera L.*) Respond to Crop Management. *Appl. Environ. Microbiol.* **78,** 4308. DOI: 10.1128/AEM.07655-11 (2012).

270. Pasteur, L., Chamberland, C., Joubert, J., *et al.* Théorie des germes et ses applications à la médecine et à la chirurgie. *Libraire de l'Académie de Médecine* (1878).

271. Patsantzis, S. & Muggleton, S. H. Top program construction and reduction for polynomial time Meta-Interpretive learning. *Mach. Learn.* **110,** 755–778. DOI: 10.1007/s10994-020-05945-w (2021).

272. Pauvert, C. *et al.* Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology* **41,** 23–33. DOI: https://doi.org/10.1016/j.funeco.2019.03.005 (2019).

273. Pauvert, C. *et al.* Microbial association networks give relevant insights into plant pathobiomes. *bioRxiv.* DOI: 10.1101/2020.02.21.958033 (2020).

274. Pearman, J. K. *et al.* Comparing sediment DNA extraction methods for assessing organic enrichment associated with marine aquaculture. *PeerJ* **8.** DOI: 10.7717/peerj.10231 (2020).

275. Pearson, K. *Correlation coefficient* in *Royal Society Proceedings* **58** (1895), 214.

276. Peay, K. G. & Bruns, T. D. Spore dispersal of basidiomycete fungi at the landscape scale is driven by stochastic and deterministic processes and generates variability in plant-fungal interactions. *New Phytol.* **204,** 180–191. DOI: 10.1111/nph.12906 (2014).

277. Pedersen, T. L. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks* R package version 2.0.6 (2022).

278. Pellissier, L. *et al.* Comparing species interaction networks along environmental gradients. *Biol. Rev.* **93,** 785–800. DOI: 10.1111/brv.12366 (2018).

279. Perazzolli, M. *et al.* Downy mildew resistance induced by *Trichoderma harzianum* T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. *BMC Genomics* **13,** 1–19. DOI: 10.1186/1471-2164-13-660 (2012).

280. Pielou, E. C. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13,** 131–144. DOI: 10.1016/0022-5193(66)90013-0 (1966).

281. Pierce, E. C. & Dutton, R. J. Putting microbial interactions back into community contexts. *Curr. Opin. Microbiol.* **65,** 56–63. DOI: 10.1016/j.mib.2021.10.008 (2022).

282. Pilosof, S., Porter, M. A., Pascual, M. & Kéfi, S. The multilayer nature of ecological networks. *Nat. Ecol. Evol.* **1,** 1–9. DOI: 10.1038/s41559-017-0101 (2017).

283. Pinto, C. *et al.* Unravelling the Diversity of Grapevine Microbiome. *PLoS One* **9,** e85622. DOI: 10.1371/journal.pone.0085622 (2014).

284. Pinto, S., Benincà, E., van Nes, E. H., Scheffer, M. & Bogaards, J. A. Species abundance correlations carry limited information about microbial network interactions. *PLoS Comput. Biol.* **18,** e1010491. DOI: 10.1371/journal.pcbi.1010491 (2022).

285. Pitt, D. & Aubin, J.-M. Joseph Lister: father of modern surgery. *Can. J. Surg.* **55,** E8. DOI: 10.1503/cjs.007112 (2012).

286. Pocock, M. J. O., Evans, D. M. & Memmott, J. The Robustness and Restoration of a Network of Ecological Networks. *Science* **335,** 973–977. DOI: 10.1126/science.1214915 (2012).

287. Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. The dissimilarity of species interaction networks. *Ecol. Lett.* **15,** 1353–1361. DOI: 10.1111/ele.12002 (2012).

288. Põlme, S. *et al.* FungalTraits: a user-friendly traits database of fungi and fungus-like stramenopiles. *Fungal Diversity* **105,** 1–16. DOI: 10.1007/s13225-020-00466-2 (2020).

289. Poudel, R. *et al.* Microbiome Networks: A Systems Framework for Identifying Candidate Microbial Assemblages for Disease Management. *Phytopathology* **106,** 1083–1096. DOI: 10.1094/PHYTO-02-16-0058-FI (2016).

290. Poveda, J., Roeschlin, R. A., Marano, M. R. & Favaro, M. A. Microorganisms as biocontrol agents against bacterial citrus diseases. *Biological Control* **158,** 104602. DOI: https://doi.org/10.1016/j.biocontrol.2021.104602 (2021).

291. Prodan, A. *et al.* Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One* **15,** e0227434. DOI: 10.1371/journal.pone.0227434 (2020).

292. Qu, K., Guo, F., Liu, X., Lin, Y. & Zou, Q. Application of Machine Learning in Microbiology. *Frontiers in Microbiology* **10.** DOI: 10.3389/fmicb.2019.00827 (2019).

293. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41,** D590–D596. DOI: 10.1093/nar/gks1219 (2012).

294. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35,** 833–844. DOI: 10.1038/nbt.3935 (2017).

295. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022).

296. Raimundo, R. L. G., Guimarães Jr., P. R. & Evans, D. M. Adaptive Networks for Restoration Ecology. *Trends Ecol. Evol.* **33,** 664–675. DOI: 10.1016/j.tree.2018.06.002 (2018).

297. Redford, A. J., Bowers, R. M., Knight, R., Linhart, Y. & Fierer, N. The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environmental Microbiology* **12,** 2885–2893. DOI: https://doi.org/10.1111/j.1462-2920.2010.02258.x (2010).

298. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334,** 1518–1524. DOI: 10.1126/science.1205438 (2011).

299. Ritpitakphong, U. *et al.* The microbiome of the leaf surface of Arabidopsis protects against a fungal pathogen. *New Phytol.* **210,** 1033–1043. DOI: 10.1111/nph.13808 (2016).

300. Riva, V. *et al.* Exploitation of Rhizosphere Microbiome Services. *Methods in Rhizosphere Biology Research,* 105–132. DOI: 10.1007/978-981-13-5767-1_7 (2019).

301. Robert, V. *et al.* MycoBank gearing up for new horizons. *IMA Fungus* **4,** 371. DOI: 10.5598/imafungus.2013.04.02.16 (2013).

302. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).

303. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4,** e2584. DOI: 10.7717/peerj.2584 (2016).

304. Rohwer, R. R., Hamilton, J. J., Newton, R. J. & McMahon, K. D. TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* **3,** e00327–18. DOI: 10.1128/mSphere.00327-18 (2018).

305. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* **8,** 42200–42216. DOI: 10.1109/ACCESS.2020.2976199 (2020).

306. Röttjers, L. & Faust, K. From hairballs to hypotheses-biological insights from microbial networks. *FEMS Microbiol. Rev.* **42,** 761–780. DOI: 10.1093/femsre/fuy030 (2018).

307. Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecol. Conserv.* **17,** e00547. DOI: 10.1016/j.gecco.2019.e00547 (2019).

308. Saak, C. C., Dinh, C. B. & Dutton, R. J. Experimental approaches to tracking mobile genetic elements in microbial communities. *FEMS Microbiology Reviews* **44,** 606–630. DOI: 10.1093/femsre/fuaa025 (2020).

309. Saccá, M. L., Barra Caracciolo, A., Di Lenola, M. & Grenni, P. Ecosystem Services Provided By Soil Microorganisms. *Soil Biological Communities and Ecosystem Resilience,* 9–24. DOI: 10.1007/978-3-319-63336-7_2 (2017).

310. Saito, R. *et al.* A travel guide to Cytoscape plugins - Nature Methods. *Nat. Methods* **9,** 1069–1076. DOI: 10.1038/nmeth.2212 (2012).

311. Salter-Townshend, M., White, A., Gollini, I. & Murphy, T. B. Review of Statistical Network Analysis: Models, Algorithms, and Software. *Stat. Anal. Data Min.* **5,** 243–264. DOI: 10.1002/sam.11146 (2012).

312. Sambamoorthy, G. & Raman, K. Deciphering the evolution of microbial interactions: in silico studies of two-member microbial communities. *bioRxiv.* DOI: 10.1101/2022.01.14.476316 (2022).

313. Samek, W., Wiegand, T. & Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv.* DOI: 10.48550/arXiv.1708.08296 (2017).

314. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 5463–5467. DOI: 10.1073/pnas.74.12.5463 (1977).

315. Santi, I., Kasapidis, P., Karakassis, I. & Pitta, P. A Comparison of DNA Metabarcoding and Microscopy Methodologies for the Study of Aquatic Microbial Eukaryotes. *Diversity* **13,** 180. DOI: 10.3390/d13050180 (2021).

316. Sapkota, R., Knorr, K., Jørgensen, L. N., O'Hanlon, K. A. & Nicolaisen, M. Host genotype is an important determinant of the cereal phyllosphere mycobiome. *New Phytol.* **207,** 1134–1144. DOI: 10.1111/nph.13418 (2015).

317. Sasaki, S., Hatano, R., Ohwada, H. & Nishiyama, H. *Estimating Productivity of Dairy Cows by Inductive Logic Programming* in *The 29th ILP conference* (2019).

318. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50,** 20–26. DOI: 10.1093/nar/gkab1112 (2022).

319. Schieber, T. A. *et al.* Quantification of network structural dissimilarities - Nature Communications. *Nat. Commun.* **8,** 1–10. DOI: 10.1038/ncomms13928 (2017).

320. Schielzeth, H. & Nakagawa, S. Nested by design: model fitting and interpretation in a mixed model era. *Methods Ecol. Evol.* **4,** 14–24. DOI: 10.1111/j.2041-210x.2012.00251.x (2013).

321. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541. DOI: 10.1128/AEM.01541-09 (2009).

322. Schmid, F., Moser, G., Müller, H. & Berg, G. Functional and Structural Microbial Diversity in Organic and Conventional Viticulture: Organic Farming Benefits Natural Biocontrol Agents. *Appl. Environ. Microbiol.* **77,** 2188. DOI: 10.1128/AEM.02187-10 (2011).

323. Schmidt, R., Cordovez, V., de Boer, W., Raaijmakers, J. & Garbeva, P. Volatile affairs in microbial interactions. *ISME J.* **9,** 2329–2335. DOI: 10.1038/ismej.2015.42 (2015).

324. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 6241–6246. DOI: 10.1073/pnas.1117018109 (2012).

325. Setati, M. E., Jacobson, D. & Bauer, F. F. Sequence-based Analysis of the *Vitis vinifera L.* cv Cabernet Sauvignon Grape Must Mycobiome in Three South African Vineyards Employing Distinct Agronomic Systems. *Frontiers in Microbiology* **6.** DOI: 10.3389/fmicb.2015.01358 (2015).

326. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13,** 2498. DOI: 10.1101/gr.1239303 (2003).

327. Shelton, A. O. *et al.* Toward quantitative metabarcoding. *bioRxiv.* DOI: 10.1101/2022.04.26.489602 (2022).

328. Shoemaker, W. R., Locey, K. J. & Lennon, J. T. A macroecological theory of microbial biodiversity. *Nat. Ecol. Evol.* **1,** 1–6. DOI: 10.1038/s41559-017-0107 (2017).

329. Shrestha, A. *et al.* Enhancement of Nitrogen-fixing Activity of *Enterobacteriaceae* Strains Isolated from Sago Palm (Metroxylon sagu) by Microbial Interaction with Non-nitrogen Fixers. *Microbes Environ.* **22,** 59–70. DOI: 10.1264/jsme2.22.59 (2007).

330. Simpson, E. H. Measurement of Diversity. *Nature* **163,** 688. DOI: 10.1038/163688a0 (1949).

331. Simpson, G. L. *permute: Functions for Generating Restricted Permutations of Data* R package version 0.9-7 (2022).

332. Singh, P., Santoni, S., This, P. & Péros, J.-P. Genotype-Environment Interaction Shapes the Microbial Assemblage in Grapevine's Phyllosphere and Carposphere: An NGS Approach. *Microorganisms* **6.** DOI: 10.3390/microorganisms6040096 (2018).

333. Skillings, D. Holobionts and the ecology of organisms: Multi-species communities or integrated individuals? *Biol. Philos.* **31,** 875–892. DOI: 10.1007/s10539-016-9544-0 (2016).

334. Smart, A. S. *et al.* Assessing the cost-efficiency of environmental DNA sampling. *Methods Ecol. Evol.* **7,** 1291–1298. DOI: 10.1111/2041-210X.12598 (2016).

335. Spearman, C. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology* **15,** 88–103 (1904).

336. Srinivasan, A. *A Learning Engine for Proposing Hypotheses (Aleph)* 2001.

337. Stewart, E. J. Growing Unculturable Bacteria. *Journal of Bacteriology* **194,** 4151–4160. DOI: 10.1128/JB.00345-12 (2012).

338. Stulberg, E. *et al.* An assessment of US microbiome research. *Nat. Microbiol.* **1,** 1–7. DOI: 10.1038/nmicrobiol.2015.15 (2016).

339. Sun, X. *et al. Multi-type Microbial Relation Extraction by Transfer Learning* in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2021), 266–269. DOI: 10.1109/BIBM52615.2021.9669738.

340. Swett, C. L., Bourret, T. & Gubler, W. D. Characterizing the Brown Spot Pathosystem in Late-Harvest Table Grapes (*Vitis vinifera L.*) in the California Central Valley. *Plant Disease* **100.** PMID: 30682913, 2204–2210. DOI: 10.1094/PDIS-11-15-1343-RE (2016).

341. Tackmann, J., Matias Rodrigues, J. F. & von Mering, C. Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Systems* **9,** 286–296.e8. DOI: 10.1016/j.cels.2019.08.002 (2019).

342. Takai, K. & Horikoshi, K. Rapid Detection and Quantification of Members of the Archaeal Community by Quantitative PCR Using Fluorogenic Probes. *Applied and Environmental Microbiology* **66,** 5066–5072. DOI: 10.1128/AEM.66.11.5066-5072.2000 (2000).

343. Tamaddoni-Nezhad, A., Bohan, D., Milani, G. A., Raybould, A. & Muggleton, S. Human–Machine Scientific Discovery. *Human-Like Machine Intelligence,* 297–315. DOI: 10.1093/oso/9780198862536.003.0015 (2021).

344. Tamaddoni-Nezhad, A., Bohan, D., Raybould, A. & Muggleton, S. H. Machine Learning a Probabilistic Network of Ecological Interactions. *Inductive Logic Programming,* 332–346. DOI: 10.1007/978-3-642-31951-8_28 (2012).

345. Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A. & Muggleton, S. Application of abductive ILP to learning metabolic network inhibition from temporal data. *Mach. Learn.* **64,** 209–230. DOI: 10.1007/s10994-006-8988-x (2006).

346. Tamaddoni-Nezhad, A., Kakas, A., Muggleton, S. & Pazos, F. in *Inductive Logic Programming* 305–322 (Springer, Berlin, Germany, 2004). DOI: 10.1007/978-3-540-30109-7_23.

347. Tamaddoni-Nezhad, A., Milani, G. A., Raybould, A., Muggleton, S. & Bohan, D. A. Construction and Validation of Food Webs Using Logic-Based Machine Learning and Text Mining. *Advances in Ecological Research* **49,** 225–289. DOI: 10.1016/B978-0-12-420002-9.00004-4 (2013).

348. Tedersoo, L., Albertsen, M., Anslan, S. & Callahan, B. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Applied and Environmental Microbiology* **87,** e00626–21. DOI: 10.1128/AEM.00626-21 (2021).

349. Teng, F. *et al.* Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Sci. Rep.* **8,** 1–12. DOI: 10.1038/s41598-018-34294-x (2018).

350. Thambugala, K. M., Daranagama, D. A., Phillips, A. J. L., Kannangara, S. D. & Promputtha, I. Fungi vs. Fungi in Biocontrol: An Overview of Fungal Antagonists Applied Against Fungal Plant Pathogens. *Frontiers in Cellular and Infection Microbiology* **10.** DOI: 10.3389/fcimb.2020.604923 (2020).

351. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49,** D480–D489. DOI: 10.1093/nar/gkaa1100 (2021).

352. This, P., Lacombe, T. & Thomas, M. R. Historical origins and genetic diversity of wine grapes. *Trends Genet.* **22,** 511–519. DOI: 10.1016/j.tig.2006.07.008 (2006).

353. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat. Rev. Microbiol.* **3,** 711–721. DOI: 10.1038/nrmicro1234 (2005).

354. Thomsen, P. F. & Willerslev, E. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* **183,** 4–18. DOI: 10.1016/j.biocon.2014.11.019 (2015).

355. Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. Agricultural sustainability and intensive production practices. *Nature* **418,** 671–677. DOI: 10.1038/nature01014 (2002).

356. Toju, H. *et al.* Core microbiomes for sustainable agroecosystems - Nature Plants. *Nat. Plants* **4,** 247–257. DOI: 10.1038/s41477-018-0139-4 (2018).

357. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. *What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use* in *Proceedings of the 4th Machine Learning for Healthcare Conference* (eds Doshi-Velez, F. *et al.*) **106** (PMLR, 2019), 359–380.

358. Tsai, K.-N., Lin, S.-H., Liu, W.-C. & Wang, D. Inferring microbial interaction network from microbiome data using RMN algorithm. *BMC Syst. Biol.* **9,** 1–10. DOI: 10.1186/s12918-015-0199-2 (2015).

359. Tshikantwa, T. S., Ullah, M. W., He, F. & Yang, G. Current Trends and Potential Applications of Microbial Interactions for Human Welfare. *Frontiers in Microbiology* **9,** 1156. DOI: 10.3389/fmicb.2018.01156 (2018).

360. Tuck, S. L. *et al.* Land-use intensity and the effects of organic farming on biodiversity: a hierarchical meta-analysis. *J. Appl. Ecol.* **51,** 746–755. DOI: 10.1111/1365-2664.12219 (2014).

361. Tylianakis, J. M., Laliberté, E., Nielsen, A. & Bascompte, J. Conservation of species interaction networks. *Biol. Conserv.* **143,** 2270–2279. DOI: 10.1016/j.biocon.2009.12.004 (2010).

362. Tylianakis, J. M. & Morris, R. J. Ecological Networks Across Environmental Gradients. *Annu. Rev. Ecol. Evol. Syst.* **48,** 25–48. DOI: 10.1146/annurev-ecolsys-110316-022821 (2017).

363. Tylianakis, J. M., Tscharntke, T. & Lewis, O. T. Habitat modification alters the structure of tropical host-parasitoid food webs. *Nature* **445,** 202–205. DOI: 10.1038/nature05429 (2007).

364. Ushey, K., Allaire, J. & Tang, Y. *reticulate: Interface to 'Python'* R package version 1.24 (2022).

365. Ushio, M. *et al.* Fluctuating interaction network and time-varying stability of a natural fish community. *Nature* **554.** DOI: 10.1038/nature25504 (2018).

366. Vacher, C. *et al. Chapter One - Learning Ecological Networks from Next-Generation Sequencing Data* 1–39. DOI: 10.1016/bs.aecr.2015.10.004 (Academic Press, 2016).

367. Valsesia, G. *et al.* Development of a High-Throughput Method for Quantification of *Plasmopara viticola* DNA in Grapevine Leaves by Means of Quantitative Real-Time Polymerase Chain Reaction. *Phytopathology®* **95.** PMID: 18943784, 672–678. DOI: 10.1094/PHYTO-95-0672 (2005).

368. Van de Peer, Y., Neefs, J.-M., De Rijk, P. & De Wachter, R. Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: Calibration of the molecular clock. *J. Mol. Evol.* **37,** 221–232. DOI: 10.1007/BF02407359 (1993).

369. Van der Heijden, M. G. A., Bardgett, R. D. & van Straalen, N. M. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol. Lett.* **11,** 296–310. DOI: 10.1111/j.1461-0248.2007.01139.x (2008).

370. Varanda, C. M. R. *et al.* Fungal endophytic communities associated to the phyllosphere of grapevine cultivars under different types of management. *Fungal Biol.* **120,** 1525–1536. DOI: 10.1016/j.funbio.2016.08.002 (2016).

371. Varghese, D., Barroso-Bergada, D., Bohan, D. A. & Tamaddoni-Nezhad, A. Efficient Abductive Learning of Microbial Interactions using Meta Inverse Entailment. *In Proceedings of the 31st International Conference on ILP* (2022).

372. Varghese, D. & Tamaddoni-Nezhad, A. *PyGol* https://github.com/danyvarghese/PyGol/. 2022.

373. Vayssier-Taussat, M. *et al.* Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Frontiers in Cellular and Infection Microbiology* **4.** DOI: 10.3389/fcimb.2014.00029 (2014).

374. Větrovský, T. *et al.* GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. *Sci. Data* **7,** 1–14. DOI: 10.1038/s41597-020-0567-7 (2020).

375. Vionnet, L. *et al.* Microbial life in the grapevine: what can we expect from the leaf microbiome? *OENO One* **52,** 219–224. DOI: 10.20870/oeno-one.2018.52.3.2120 (2018).

376. Vogel, C., Bodenhausen, N., Gruissem, W. & Vorholt, J. A. The Arabidopsis leaf transcriptome reveals distinct but also overlapping responses to colonization by phyllosphere commensals and pathogen infection with impact on plant health. *New Phytol.* **212,** 192–207. DOI: 10.1111/nph.14036 (2016).

377. Volterra, V. Variations and fluctuations of the number of individuals in animal species living together. *Animal Ecology,* 409–448 (1926).

378. Vorholt, J. A. Microbial life in the phyllosphere. *Nat. Rev. Microbiol.* **10,** 828–840. DOI: 10.1038/nrmicro2910 (2012).

379. Wagg, C., Schlaeppi, K., Banerjee, S., Kuramae, E. E. & van der Heijden, M. G. A. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. *Nat. Commun.* **10,** 1–10. DOI: 10.1038/s41467-019-12798-y (2019).

380. Wang, G. C. & Wang, Y. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl. Environ. Microbiol.* **63,** 4645–4650. DOI: 10.1128/aem.63.12.4645-4650.1997 (1997).

381. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267. DOI: 10.1128/AEM.00062-07 (2007).

382. Ward, D. M. *et al.* Genomics, environmental genomics and the issue of microbial species. *Heredity* **100,** 207–219. DOI: 10.1038/sj.hdy.6801011 (2008).

383. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738. DOI: 10.1038/171737a0 (1953).

384. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35,** 1064–1066. DOI: 10.1093/bioinformatics/bty734 (2018).

385. Weiland-Bräuer, N. Friends or Foes—Microbial Interactions in Nature. *Biology* **10,** 496. DOI: 10.3390/biology10060496 (2021).

386. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10,** 1669–1681. DOI: 10.1038/ismej.2015.235 (2016).

387. White, T. J., Bruns, T. D., Lee, S. B., Taylor, J. W. & Sninsky, J. J. Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics. *Pcr Protocols: a Guide to Methods and Applications,* **31,** 315–322 (1990).

388. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

389. Wielemaker, J., Schrijvers, T., Triska, M. & Lager, T. SWI-Prolog. *Theory Pract. Logic Program.* **12,** 67–96. DOI: 10.1017/S1471068411000494 (2012).

390. Wilke, C. O. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'* R package version 1.1.1 (2020).

391. World Health Organization. *Ecosystems and human well-being: health synthesis: a report of the Millennium Ecosystem Assessment* (World Health Organization, 2005).

392. Wu, B. *et al.* Current insights into fungal species diversity and perspective on naming the environmental DNA sequences of fungi. *Mycology* **10,** 127–140. DOI: 10.1080/21501203.2019.1614106 (2019).

393. Xiong, Q., Hu, J., Wei, H., Zhang, H. & Zhu, J. Relationship between Plant Roots, Rhizosphere Microorganisms, and Nitrogen and Its Special Focus on Rice. *Agriculture* **11,** 234. DOI: 10.3390/agriculture11030234 (2021).

394. Yuan, M. M. *et al.* Climate warming enhances microbial network complexity and stability. *Nat. Clim. Change* **11,** 343–348. DOI: 10.1038/s41558-021-00989-9 (2021).

395. Zanzotto, A., Morroni, M., *et al.* Major biocontrol studies and measures against fungal and oomycete pathogens of grapevine. *Biocontrol of Major Grapevine Diseases: Leading Research. CABI, Oxford, UK,* 1–34. DOI: https://doi.org/10.1079/9781780647128.0001 (2016).

396. Zarraonaindia, I. & Gilbert, J. A. Understanding grapevine-microbiome interactions: implications for viticulture industry. *Microbial Cell* **2,** 171. DOI: 10.15698/mic2015.05.204 (2015).

397. Zarraonaindia, I. *et al.* The Soil Microbiome Influences Grapevine-Associated Microbiota. *mBio* **6,** e02527–14. DOI: 10.1128/mBio.02527-14 (2015).

398. Zhang, X., Johnston, E. R., Liu, W., Li, L. & Han, X. Environmental changes affect the assembly of soil bacterial community primarily by mediating stochastic processes. *Global Change Biol.* **22,** 198–207. DOI: 10.1111/gcb.13080 (2016).

399. Zhou, Y., Martins, E., Groboillot, A., Champagne, C. P. & Neufeld, R. J. Spectrophotometric quantification of lactic bacteria in alginate and control of cell release with chitosan coating. *J. Appl. Microbiol.* **84,** 342–348. DOI: 10.1046/j.1365-2672.1998.00348.x (1998).

400. Zhu, Y.-G. & Penuelas, J. Changes in the environmental microbiome in the Anthropocene. *Global Change Biol.* **26,** 3175–3177. DOI: 10.1111/gcb.15086 (2020).

401. Zinger, L. *et al.* DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.* **28,** 1857–1862. DOI: 10.1111/mec.15060 (2019).

402. Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. & Smith, G. M. in *Mixed effects models and extensions in ecology with R* 101–142 (Springer, New York, NY, New York, NY, USA, 2009). DOI: 10.1007/978-0-387-87458-6_5.

Appendix A

Supplementary Figures and Tables

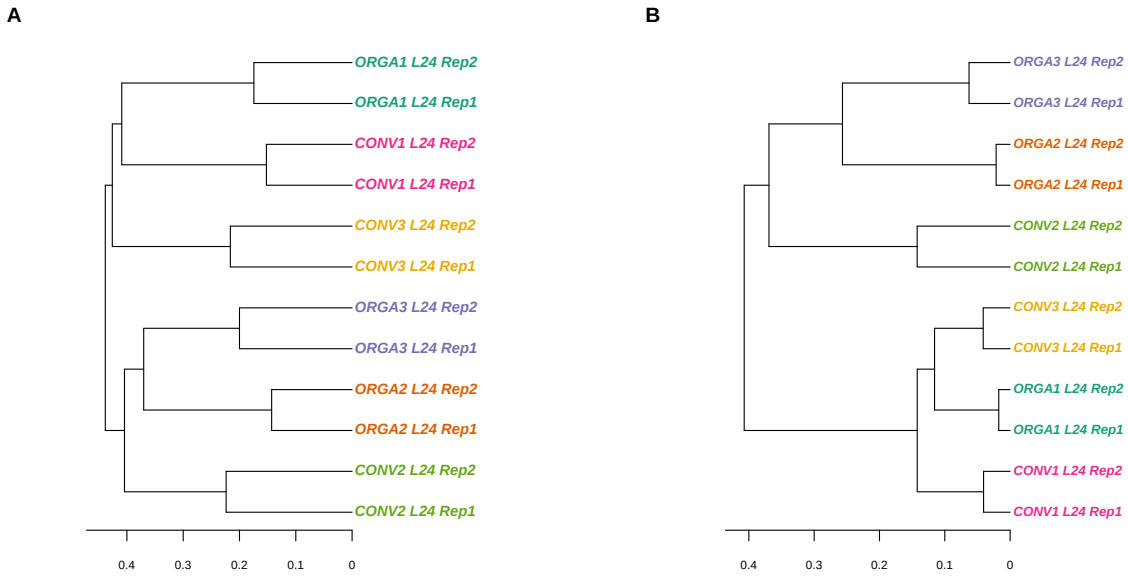


Figure A.1: Dendrogram plot of compositional dissimilarities between technical replicates for sequencing. Technical replicates were created by splitting six lots of PCR products in half and sequencing the two halves independently. The PCR products used were those corresponding to the leaf collected on vine number 24 (L24) in each of the six plots studied (ORGA1, ORGA2, ORGA3, CONV1, CONV2, CONV3; see Figure 1). Compositional dissimilarities between samples were computed with **A** the binary Jaccard index and **B** the quantitative Jaccard index. The dendrogram was built using a hierarchical clustering algorithm (complete linkage method). Compositional dissimilarities between the two technical replicates of the same sample were significantly smaller than the dissimilarities among samples (PERMANOVA: $F = 39.98$; $R2 = 0.97$; $p = 0.001$).

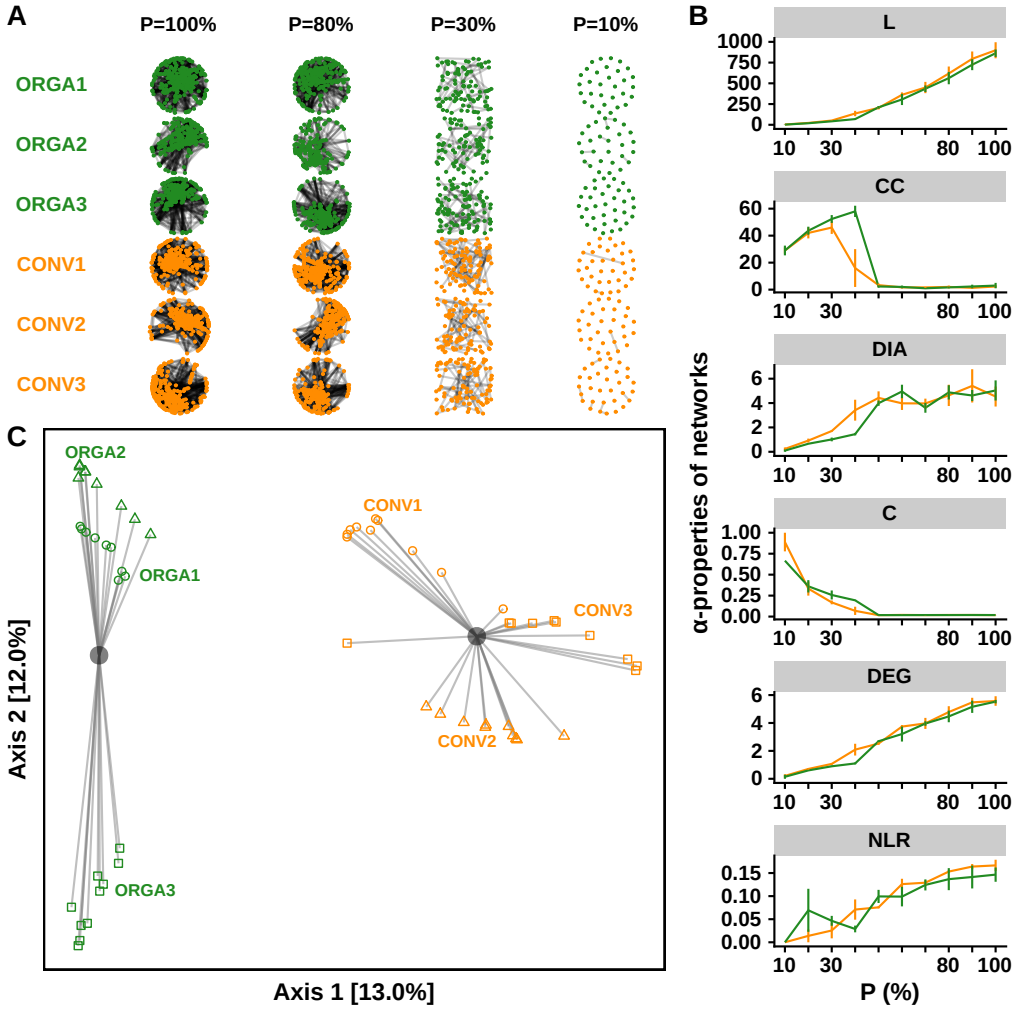


Figure A.2: Effect of cropping system — conventional (CONV) versus organic (ORGA) — on the α-properties and β-propertie of grapevine foliar fungal networks. **A** Association networks inferred from fungal metabarcoding data with SPIEC-EASI (Kurtz *et al.* 2015). A total of 60 networks were inferred, corresponding to 2 cropping systems × 3 replicates (blocks) × 10 P values, with P the percentage of most abundant ASVs used for network inference. Only four values of P are shown on the Figure. **B** Variations in network α-properties. The following properties (Table II.1) were calculated for each network: the number of links (L) and connected components (CC), the network diameter (DIA) and connectance (C) and the mean degree (DEG) and negative link ratio (NLR). The percentage P of ASVs used for network reconstruction had a significant influence on all properties (Table A.10), whereas the cropping system did not (Table A.9). **C** Principal coordinate analysis (PCoA) represents dissimilarities between networks, measured with the βOS index (Poisot *et al.* 2012) calculated with the binary Jaccard index. βOS measures the dissimilarity between two networks in terms of the presence-absence of associations between shared ASVs. The centroids for each cropping system are represented by gray circles. The effect of the cropping system on βOS was significant, in interaction with the percentage P of most abundant ASVs used for network inference (Table A.11). Networks were inferred with SPIEC-EASI (Kurtz *et al.* 2015).

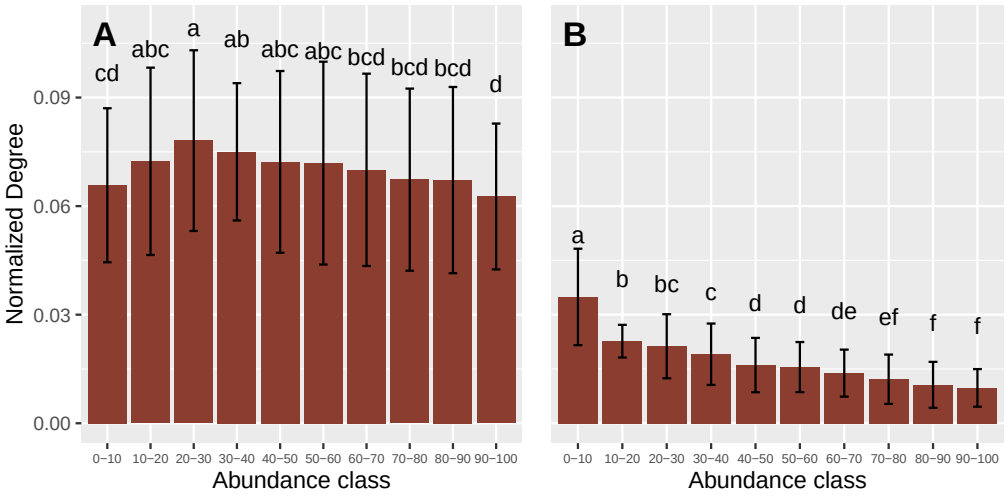


Figure A.3: Normalized degree of nodes in networks inferred with A SparCC or B SPIEC-EASI. Nodes were classified according to the relative abundance of their corresponding ASVs. Abundance class 0-10 corresponds to the 10% most abundant nodes, while abundance class 90-100 corresponds to the 10% less abundant nodes. Normalized degree was obtained by dividing the node degree by n-1, where n is the total number of nodes in the network. The effect of abundance class on the normalized node degree was analyzed with ANOVA followed by post-hoc Tukey's test. Effect of abundance class was significant in both cases (SparCC: $F = 6.797$, $p < 0.001$; SPIEC-EASI: $F = 173.8$, $p < 0.001$).

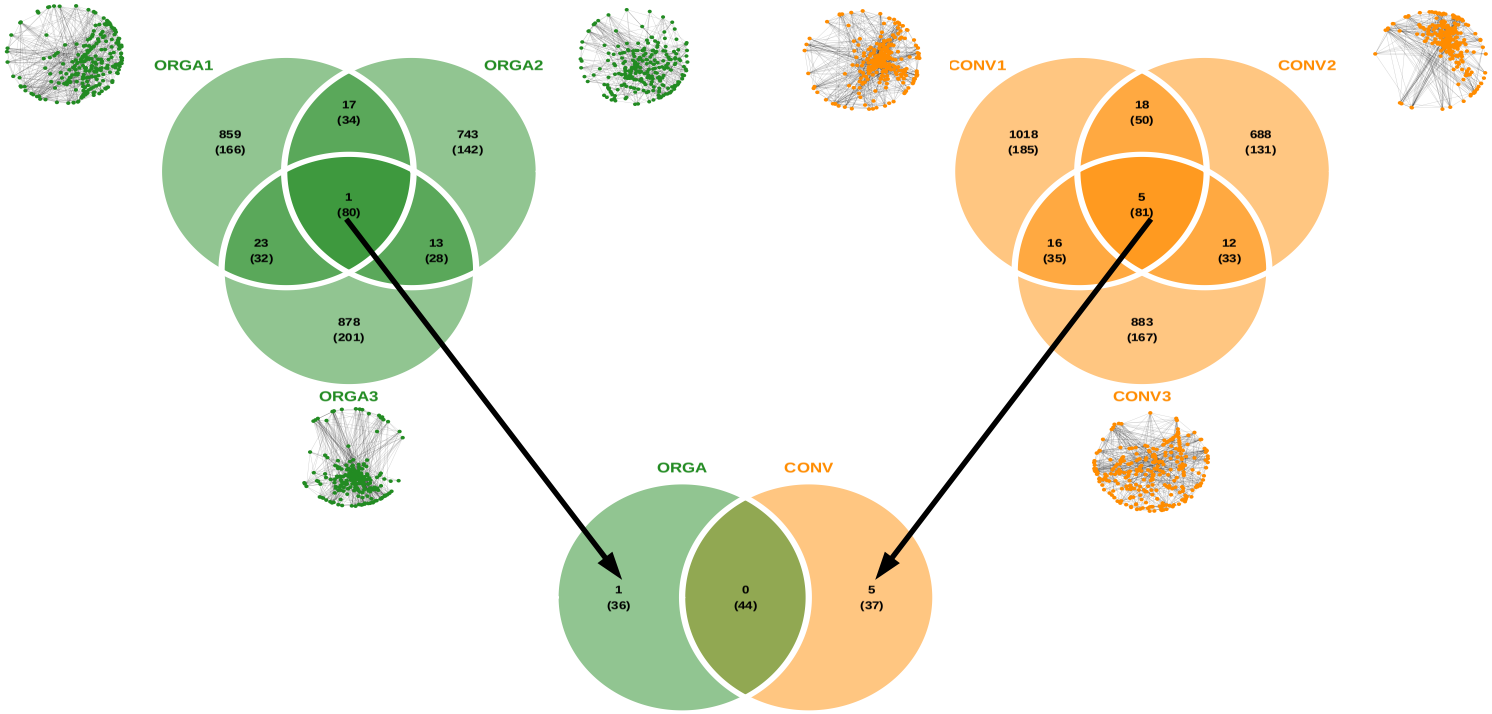


Figure A.4: Venn diagrams showing the number of fungal associations common to network replicates. A Associations common to the three network replicates inferred for the organic cropping system (ORGA1, ORGA2, ORGA3) and **B** the three network replicates inferred for the conventional cropping system (CONV1, CONV2, CONV3), regardless of the sign of the association, in the situation in which all ASVs were used for network construction (P=100%). **C** Associations common to the six networks. Networks were inferred with SPIEC-EASI (Kurtz *et al.* 2015). The number of nodes shared by the network replicates is indicated into brackets.

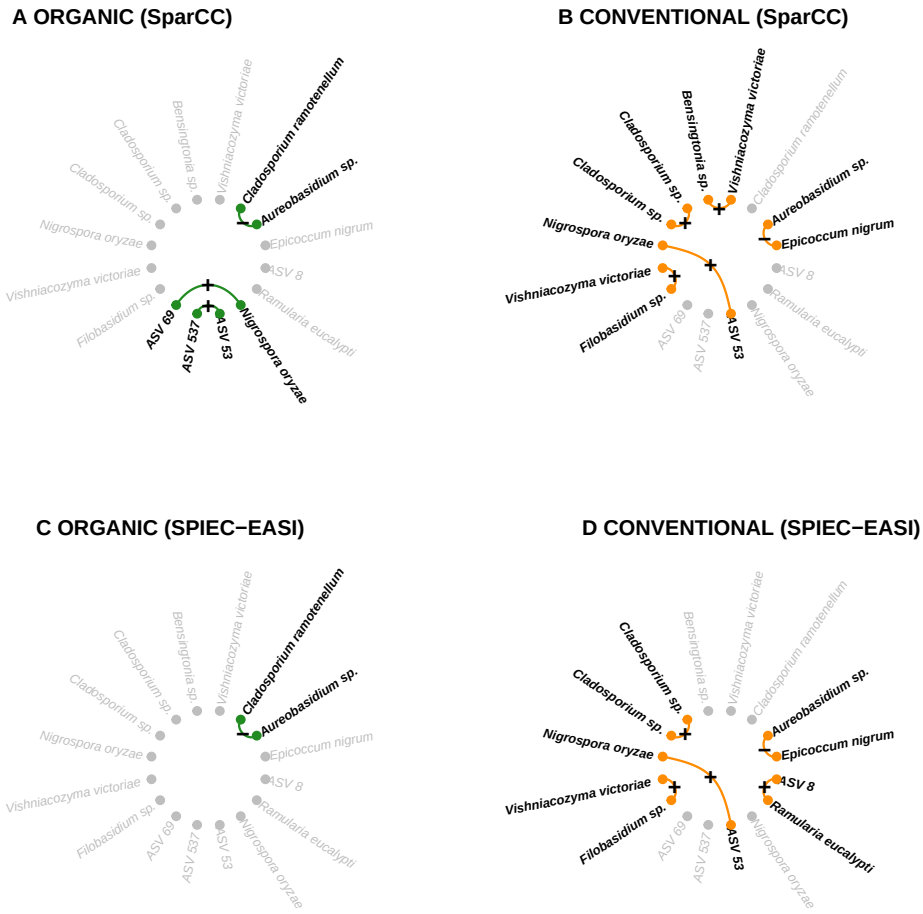


Figure A.5: Consensus networks between the three network replicates for the organic (ORGA) and the conventional (CONV) cropping systems depending on the method for network inference. Network nodes represent fungal ASVs and links represent significant positive (+) or negative (-) associations common to the three network replicates (Fig. 6 and S4). The fungal ASVs absent from a network are indicated in gray. Networks were inferred with SparCC (Friedman & Alm 2012) or SPIEC-EASI (Kurtz *et al.* 2015).

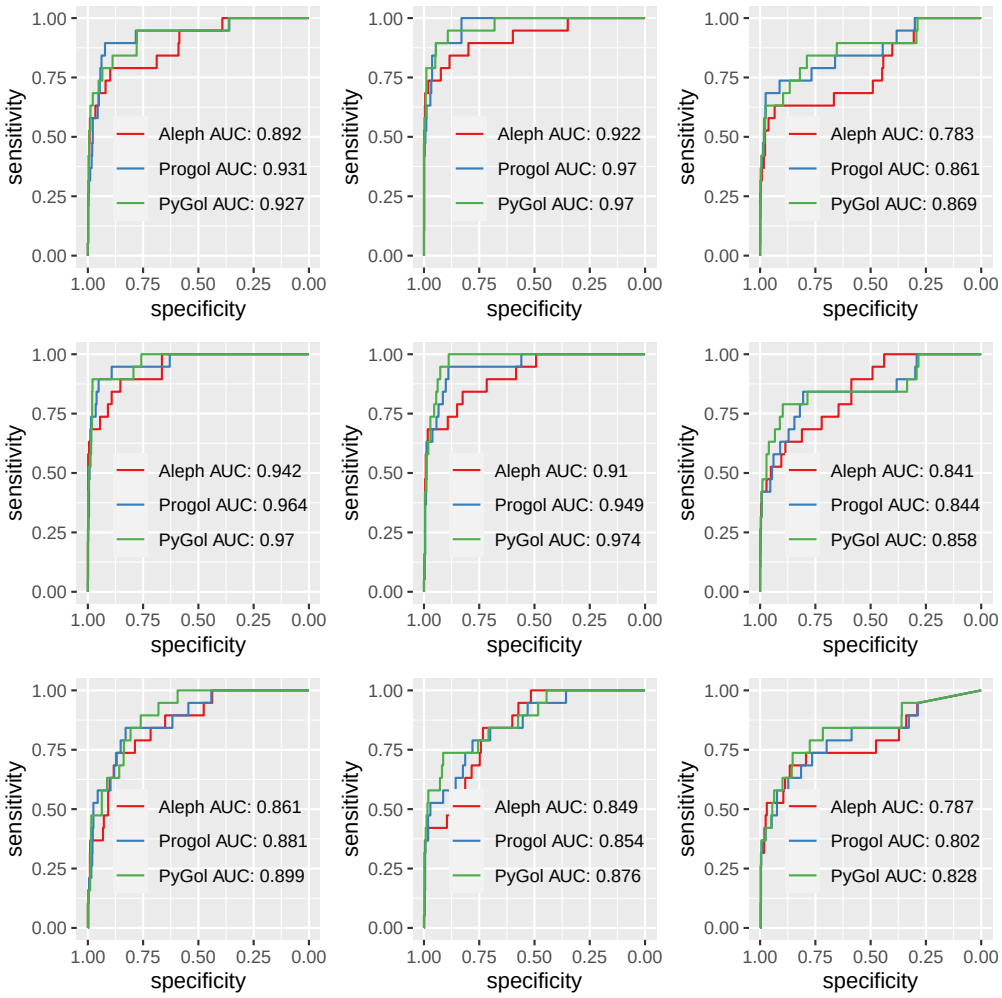


Figure A.6: Comparison of the performance of Progol, Pygol and Aleph. ROC curves showing the performance of I statistic computed from the compression values obtained from abducing 9 computer generated datasets. Data-sets are randomly generated following the ecological models of Weiss et al. (2016). Higher area under the curve (AUC) indicates better predictive power. Line color show the A/ILP programm used.

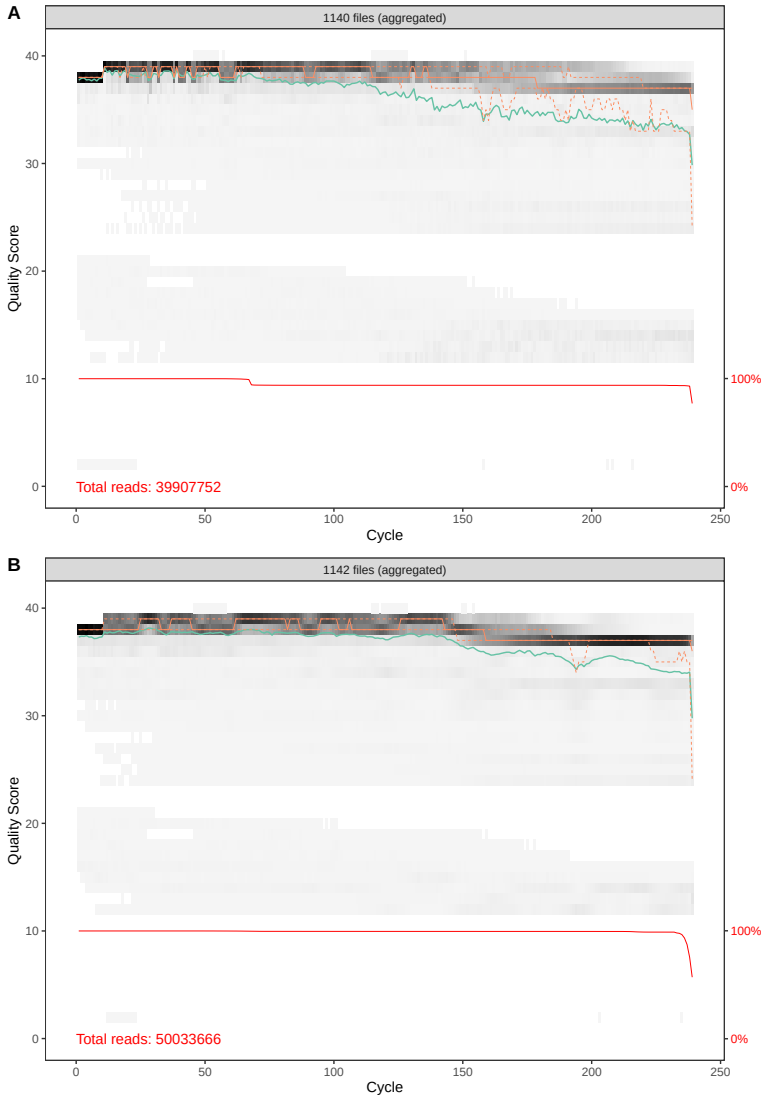


Figure A.7: Average quality scores distribution in function of sequence position for the sequenced ITS A and 16S B input fastq files. Fastq files were obtained sequencing the eDNA present in grapevine leaves. The plot was done using the plotQualityProfile function of the DADA2 pipeline.

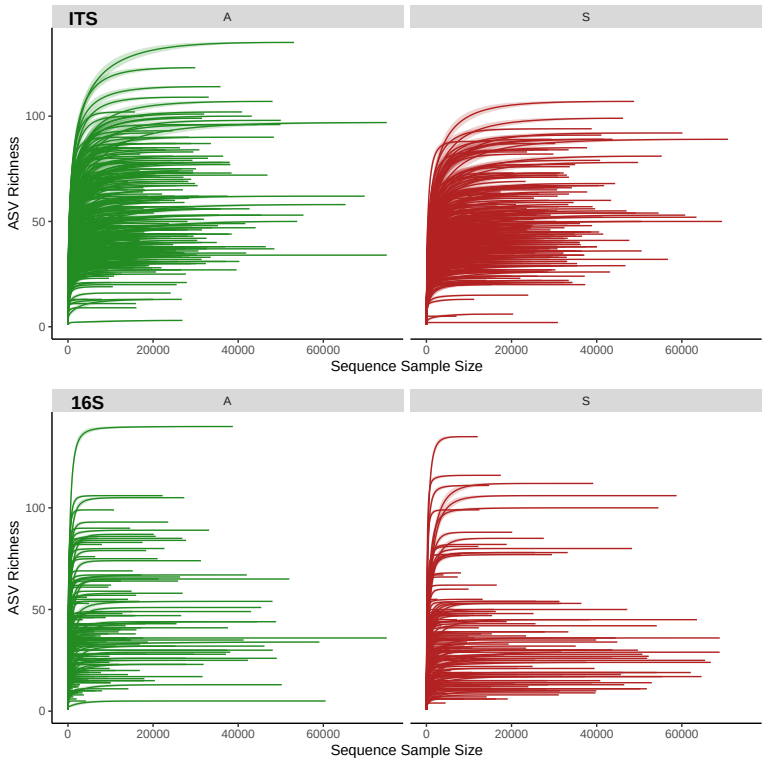


Figure A.8: Rarefaction curves showing the variation of ASV richness in function of the rarefied sequencing depth for the ITS and 16S metabarcoding datasets. Each plot shows the rarefaction curve for asymptomatic (A) and sporulating downy mildew lesion (S) samples.

Table A.1: List of phytosanitary products and active ingredients applied in the year of the sampling campaign, together with their normalized dose (also referred to as the treatment frequency index). PM = powdery mildew, caused by the fungal pathogen *Erysiphe necator* and DM = downy mildew, caused by the oomycete pathogen *Plasmopara viticola*. Leaf sampling was performed on September 10, 2015, more than one month after the last phytosanitary treatment and a couple of hours before grape harvest. The treatment frequency index did not differ between cropping systems (ANOVA: df = 21; F = 0.436; p = 0.516).

Date	**Cropping System**	**Fungicides**	**Active ingredients**	**Target disease**	
				PM	**DM**
2015-04-30	ORGA	Heliocuivre©	Copper		0.145
2015-04-30	ORGA	Citrothiol DG©	Micronized sulfur	0.371	
2015-05-07	CONV	Chaoline©	Fosetyl aluminum + metirame		0.292
2015-05-07	CONV	Dynali©	Cyflufenamid + difenoconazole	0.289	
2015-05-13	ORGA	Heliocuivre©	Copper		0.167
2015-05-13	ORGA	Citrothiol DG©	Micronized sulfur	0.4	
2015-05-19	CONV	Cabrio Top©	Metirame-zinc + pyraclostrobin	0.5	
2015-05-28	ORGA	Citrothiol DG©	Micronized sulfur	0.8	
2015-05-28	ORGA	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.533
2015-06-04	CONV	Vivando©	Metrafenone	0.833	
2015-06-04	CONV	Chaoline©	Fosetyl aluminum + metirame		0.708
2015-06-09	ORGA	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.533
2015-06-09	ORGA	Citrothiol DG©	Micronized sulfur	0.6	
2015-06-25	ORGA	Citrothiol DG©	Micronized sulfur	0.6	
2015-06-25	CONV	Citrothiol DG©	Micronized sulfur	0.6	
2015-07-01	ORGA	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.533
2015-07-01	CONV	Cabrio Top©	Metirame-zinc + pyraclostrobin	0.75	
2015-07-17	ORGA	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.4
2015-07-17	ORGA	Heliocuivre©	Copper		0.083
2015-07-17	CONV	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.4
2015-07-17	CONV	Heliocuivre©	Copper		0.083
2015-08-03	ORGA	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.533
2015-08-03	CONV	Bouillie Bordelaise RSR® Dispersss® NC	Copper		0.533

Table A.2: Effect of cropping system — conventional (CONV) versus organic (ORGA) — on the incidence and severity of foliar disease symptoms at harvest time (2015-09-07). Disease incidence is defined as the percentage of leaves displaying symptoms, whereas disease severity is defined as the percentage leaf damage. Symptom incidence and severity were estimated visually on 40 grapevines for each plot (40×3 per cropping system). The mean values are reported for each cropping system as a percentage. Wald $\chi 2$ tests were used for comparisons after linear mixed model analysis with cropping system as a fixed effect and block as a random effect.

Disease		**ORGA (%)**	**CONV (%)**	χ**2**	**p-value**
Downy Mildew	Incidence	0.749	0.688	0.57	0.45
	Severity	0.037	0.03	1.93	0.164
Powdery Mildew	Incidence	0.113	1.346	12.49	**<0.001**
	Severity	0.003	0.102	7.97	**0.005**
Black rot	Incidence	0.188	0.354	19.02	**<0.001**
	Severity	0.007	0.014	5.49	**0.019**

Table A.3: Primer pairs used to amplify the fungal ITS1 region.

1st PCR with regular primers (bold)

Forward ITS1F: 5'-CTTGGTCATTTAGAGGAAGTAA-3'

Reverse ITS2: 5'-GCTGCGTTCTTCATCGATGC-3'

2nd nested PCR with pre-tagged primers (italics)

Forward ITS1F-pre-tag: 5'-*CTTTCCCTACACGACGCTCTTCCGATCT* **CTTGGTCATTTAGAGGAAGTAA**-3'

Reverse ITS2-pre-tag: 5'-*GGAGTTCAGACGTGTGCTCTTCCGATCT* **GCTGCGTTCTTCATCGATGC**-3'

Table A.4: Effect of cropping system — conventional (CONV) versus organic (ORGA) — on community α-properties. Generalized linear mixed models included the cropping system as a fixed treatment effect and the sampling depth as an offset. For every community α-property (as defined in Table II.1), we compared the likelihood of a full model including the block and its interaction with the cropping system as random effects and a simplified model including only the block factor as random effect. Only the results of the best model are shown. The ORGA system was taken as the reference.

Fixed effects	**Estimate**	**SE**	**z**	**P(>|z|)**
Richness				
(Intercept)	-6.9569	0.0466	-149.4	<2e-16
Cropping_System (CONV)	-0.1206	0.0554	-2.2	**0.029**
Diversity				
(Intercept)	-9.5533	0.0655	-145.7	<2e-16
Cropping_System (CONV)	-0.3079	0.107	-2.9	**0.004**
Evenness				
(Intercept)	-11.6042	0.0675	-171.9	<2e-16
Cropping_System (CONV)	-0.281	0.0787	-3.6	**<0.001**
Random effects	**Variance**	**SD**		
Richness				
Block	0.0019	0.0433		
Residual				
Diversity				
Block	0	0		
Residual	2.126	1.458		
Evenness				
Block	0.0025	0.0504		
Residual	0.0194	0.1393		

Table A.5: Effect of cropping system — conventional (CONV) versus organic (ORGA) — on community α-properties. Generalized linear mixed models included the cropping system and the proportion of reads assigned to the *Erysiphe* genus as fixed effects and the sampling depth as an offset. For every community α-property (as defined in Table II.1), we compared the likelihood of a full model including the block and its interaction with the cropping system as random effects and a simplified model including only the block factor as random effect. Only the results of the best model are shown. The ORGA system was taken as the reference.

Fixed effects	**Estimate**	**SE**	**z**	**P(>|z|)**
Richness				
(Intercept)	-6.9571	0.0469	-148.4	<2e-16
Cropping_System (CONV)	-0.1255	0.0567	-2.2	**0.027**
Erysiphe reads %	0.2352	0.6048	0.4	0.697
Diversity				
(Intercept)	-9.5541	0.0652	-146.5	<2e-16
Cropping_System (CONV)	-0.3347	0.1110	-3.0	**0.003**
Erysiphe reads %	1.3255	1.0417	1.3	0.203
Evenness				
(Intercept)	-11.6060	0.0694	-167.2	<2e-16
Cropping_System (CONV)	-0.3043	0.0802	-3.8	**<0.001**
Erysiphe reads %	1.2734	0.7344	1.7	0.083
Random effects	**Variance**	**SD**		
Richness				
Block	0.00197	0.04439		
Residual				
Diversity				
Block	0	0		
Residual	2.10467	1.4507		
Evenness				
Block	0.00281	0.05303		
Residual	0.01905	0.13801		

Table A.6: Effect of cropping system —conventional (CONV) versus organic (ORGA) — on the level of stochasticity in community assembly. The relative contribution of deterministic and stochastic processes to community assembly was assessed for each cropping system with the Normalized Stochasticity Ratio (NST) defined by Ning et al. (2019), that ranges from 0 to 100, where 0 means a completely deterministic assembly process and 100 a completely stochastic assembly process. NST was calculated using the tNST function with the quantitative and binary Jaccard dissimilarity indices, the FE null model, and other parameters by default values. Differences in NST values between both cropping systems were tested using permutational analysis of variance.

ASV	**Dissimilarity index**	**NST value (%)**		**F**	**p-value**
		ORGA	**CONV**		
all	Quantitative Jaccard	29.28	33.62	11.6	0.416
all	Binary Jaccard	78.38	94.8	404.9	**0.001**

Table A.7: Effect of cropping system on the α-properties of fungal association networks inferred with SparCC. Properties (as defined in Table II.1) were compared between cropping systems for every value of the percentage P of the most abundant ASVs used for network inference. The U and p-values of Wilcoxon rank-sum tests are reported. The p-value is not available (NA) for situations in which property values were equal for all networks. The p-values after Benjamini-Hochberg adjustment are not reported because all were equal to one.

P (%)	L	CC	DIA	C	DEG	NLR
10	U = 1; p = 0.19	U = 9; p = 0.077	U = 3; p = 0.505	U = 8; p = 0.19	U = 1; p = 0.19	U = 5; p = 1
20	U = 1; p = 0.19	U = 6.5; p = 0.48	U = 7; p = 0.354	U = 2; p = 0.383	U = 2; p = 0.383	U = 9; p = 0.081
30	U = 1; p = 0.19	U = 6; p = 0.505	U = 6.5; p = 0.48	U = 0; p = 0.081	U = 1; p = 0.19	U = 7; p = 0.383
40	U = 1; p = 0.19	U = 6; p = 0.505	U = 5.5; p = 0.814	U = 2; p = 0.383	U = 1; p = 0.19	U = 5; p = 1
50	U = 1; p = 0.19	U = 4.5; p = NA	U = 3; p = 0.505	U = 2; p = 0.383	U = 2; p = 0.383	U = 6; p = 0.663
60	U = 1; p = 0.19	U = 4.5; p = NA	U = 6; p = 0.505	U = 2; p = 0.383	U = 1; p = 0.19	U = 5; p = 1
70	U = 3; p = 0.663	U = 4.5; p = NA	U = 3; p = 0.619	U = 4; p = 1	U = 2; p = 0.383	U = 4; p = 1
80	U = 3; p = 0.663	U = 4.5; p = NA	U = 6; p = 0.505	U = 2; p = 0.383	U = 2; p = 0.383	U = 4; p = 1
90	U = 3; p = 0.663	U = 4.5; p = NA	U = 6; p = 0.505	U = 2; p = 0.383	U = 2; p = 0.383	U = 2; p = 0.383
100	U = 3; p = 0.663	U = 4.5; p = NA	U = 6; p = 0.505	U = 3; p = 0.663	U = 4; p = 1	U = 4; p = 1

Table A.8: Effect of the percentage P of the most abundant ASVs used for network inference on the α-properties of fungal association networks inferred with SparCC. Spearman's correlation coefficient and the results of Spearman's rank correlation tests are reported for each network property (as defined in Table II.1). The p-values are reported after Benjamini-Hochberg adjustment.

Property	**Correlation (ρ)**	**S**	**p-value**
L	0.98	839	$<$**0.001**
CC	-0.63	58685	$<$**0.001**
DIA	-0.84	66296	$<$**0.001**
C	-0.71	61374	$<$**0.001**
DEG	0.95	1647	$<$**0.001**
NLR	-0.57	56550	$<$**0.001**

Table A.9: Effect of cropping system on the α-properties of fungal association networks inferred with SPIEC-EASI. Properties (as defined in Table II.1) were compared between cropping systems for every value of the percentage P of the most abundant ASVs used for network inference. The U and p-values of Wilcoxon rank-sum tests are reported. The p-value is not available (NA) for situations in which property values were equal for all networks. The p-values after Benjamini-Hochberg adjustment are not reported because all were equal to one.

P(%)	L	CC	DIA	C	DEG	NLR
10	U = 2; p = 0.376	U = 5; p = 1	U = 2; p = 0.354	U = 0.5; p = 0.617	U = 2; p = 0.376	U = 1.5; p = NA
20	U = 0; p = 0.081	U = 5; p = 1	U = 3; p = 0.663	U = 5.5; p = 0.814	U = 1; p = 0.19	U = 7; p = 0.354
30	U = 1; p = 0.19	U = 6.5; p = 0.507	U = 0; p = 0.081	U = 7; p = 0.383	U = 0; p = 0.081	U = 7; p = 0.383
40	U = 1; p = 0.19	U = 9; p = 0.081	U = 0; p = 0.081	U = 9; p = 0.077	U = 0; p = 0.081	U = 2; p = 0.383
50	U = 5; p = 1	U = 3; p = 0.653	U = 3; p = 0.663	U = 7; p = 0.383	U = 8; p = 0.19	U = 6.5; p = 0.507
60	U = 3; p = 0.663	U = 5; p = 1	U = 8; p = 0.19	U = 3; p = 0.663	U = 3; p = 0.663	U = 2; p = 0.383
70	U = 4; p = 1	U = 1.5; p = 0.188	U = 2; p = 0.383	U = 5; p = 1	U = 6; p = 0.663	U = 4; p = 1
80	U = 4; p = 1	U = 4; p = 1	U = 4; p = 1	U = 3; p = 0.663	U = 4; p = 1	U = 3; p = 0.663
90	U = 3; p = 0.663	U = 5; p = 1	U = 4; p = 1	U = 4; p = 1	U = 2; p = 0.383	U = 3; p = 0.663
100	U = 3; p = 0.663	U = 4; p = 1	U = 4; p = 1	U = 6; p = 0.663	U = 3; p = 0.663	U = 2; p = 0.383

Table A.10: Effect of the percentage P of the most abundant ASVs used for network inference on the α-properties of fungal association networks inferred with SPIEC-EASI. Spearman's correlation coefficient and the results of Spearman's rank correlation tests are reported for each network property (as defined in Table II.1). The p-values are reported after Benjamini-Hochberg adjustment.

Property	**Correlation (ρ)**	**S**	**p-value**
L	0.98	621	$<$**0.001**
CC	-0.69	60722	$<$**0.001**
DIA	0.79	7723	$<$**0.001**
C	-0.69	55067	$<$**0.001**
DEG	0.97	971	$<$**0.001**
NLR	0.84	5158	$<$**0.001**

Table A.11: Effect of cropping system — conventional versus organic — on the β-properties of grapevine foliar fungal networks inferred with SPIEC-EASI. The D index quantifies the topological dissimilarity between networks (Schieber *et al.* 2017) whereas the other three metrics (βWN, βOS and βST), which were calculated with the binary Jaccard index, quantify differences in associations between networks (Poisot *et al.* 2012). The effect of the percentage P of the most abundant ASVs used for network inference, and the effect of cropping system on the dissimilarities between networks were evaluated in permutational analysis of variance (PERMANOVA). The number of permutations was set to 999 and permutations were constrained by block.

Dissimilarity index	**PERMANOVA**				
Topological dissimilarity (Schieber's D)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	100.89	0.65	**<0.01**
	Cropping_System (CS)	1	0.99	0.01	0.31
	P × CS	1	0.31	0	0.68
	Residuals	54		0.35	
	Total	57		1	
Overall dissimilarity of associations (βWN)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	2.689	0.04	**<0.01**
	Cropping_System (CS)	1	5.06	0.08	**<0.01**
	P × CS	1	2.547	0.04	**<0.01**
	Residuals	54		0.84	
	Total	57		1	
Dissimilarity of associations between shared ASVs (βOS)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	3.863	0.06	**<0.01**
	Cropping_System (CS)	1	8.799	0.13	**<0.01**
	P × CS	1	3.145	0.05	**<0.01**
	Residuals	54		0.77	
	Total	57		1	
Dissimilarity of associations due to ASV turnover (βST)	**Variable**	**Df**	***F***	**R2**	**Pr(>F)**
	Percent_ASV (P)	1	0.279	0.01	1
	Cropping_System (CS)	1	0.2259	0.01	1
	P × CS	1	0.2948	0.01	1
	Residuals	54		0.97	
	Total	57		1	

Appendix B

Metagenomic next-generation sequencing (mNGS) data reveals the phyllosphere microbiome of wheat plants infected by the fungal pathogen *Zymoseptoria tritici*

Metagenomic next-generation sequencing (mNGS) data reveals the phyllosphere microbiome of wheat plants infected by the fungal pathogen *Zymoseptoria tritici*

Didac Barroso-Bergada, Marie Massot, Noémie Vignolles, Julie Faivre d'Arcier, Emilie Chancerel, Erwan Guichoux, Anne-Sophie Walker, Corinne Vacher, David A. Bohan, Valérie Laval, Frédéric Suffert

Abstract:
The fungal pathogen *Zymoseptoria tritici* is the causal agent of *Septoria tritici* blotch (STB), a major wheat disease in Western Europe. Microorganisms inhabiting wheat leaves might act as beneficial, biocontrol or facilitating agents that could limit or stimulate the development of *Z. tritici*. Improving our understanding of microbial communities in the wheat phyllosphere would lead to new insights into STB management. This resource announcement provides fungal and bacterial metabarcoding datasets obtained by sampling wheat leaves with and without symptoms caused by *Z. tritici*. Tissues were sampled from three commercial wheat varieties on three sampling dates during a cropping season. Weeds around wheat fields were sampled as well. In total, more than 450 leaf samples were collected. The pathogen *Z. tritici* was quantified using qPCR. We provide the raw metabarcoding datasets, the Amplicon Sequence Variant (ASV) tables obtained after bioinformatic processing, the metadata associated to each sample (sampling date, wheat variety and tissue health condition), a preliminary descriptive analysis of the data, and the code used for bioinformatic and descriptive statistical analysis.

Keywords *community succession, fungal pathogen, microbial diversity, microbiome, leaf microbiota, phyllosphere, wheat*

1 Introduction

Wheat crops are exposed to many fungal plant pathogens, including *Zymoseptoria tritici*, the causal agent of *Septoria tritici* blotch (STB), a major disease in Western Europe (Fones & Gurr 2015). In field conditions, wheat leaves host a multitude of other microorganisms — endophytic, epiphytic, pathogenic and saprophytic (Błaszczyk *et al.* 2021) — some of which interact directly or indirectly with *Z. tritici* (Kerdraon *et al.* 2019). Several taxa may also have antagonistic or synergistic activity while interacting with other taxa, and could be considered potential biocontrol agents or facilitating agents that can limit or stimulate STB development (Chaudhry *et al.* 2020). Maximizing the chance of highlighting important interactions, for instance within co-occurrence network analysis, requires a thorough description of communities under different biotic and abiotic conditions (Röttjers & Faust 2018). In this resource announcement, we present fungal and bacterial community datasets collected on wheat leaves over the course of a wheat

cropping season, taking into account: (i) the physiological stage of wheat; (ii) the dynamics of STB development; and (iii) different wheat cultivars. We collected leaf samples in monovarietal plots of three wheat cultivars, at three dates over a growing season. One of the wheat varieties carried resistance genes to STB that also potentially impact the development of other taxa of the microbial community. Three types of leaf sections were collected, which differed in the presence of symptoms caused by *Z. tritici*: (i) sections with no STB lesion from a visually healthy leaf; (ii) sections with no STB lesion from a visually symptomatic leaf; and (iii) sections with STB lesions. This dataset could be used to explore the co-occurrence of microbial species and thereby improve our understanding of the community dynamics associated with the development of *Z. tritici* on wheat leaves. Weeds in the margins and edges of cultivated wheat fields can act as alternative hosts for microbial species present in the crop. For this reason, a second dataset composed of weed leaf samples was collected to get an insight into the host range of the microorganisms associate to wheat leaves and potentially interacting with *Z. tritici* in the agroecosystem.

2 Methods

2.1 Sampling

Wheat

Samples were collected in 2018 at the Grignon experimental station (Yvelines, France; X:48.842, Y:1.943) from three varieties of winter-sown bread wheat (*Triticum aestivum*). Two varieties, Soissons (SOI) and Apache (APA), were considered susceptible to *Z. tritici* (both rated 5 on the ARVALIS-Institut du Végétal/CTPS scale, from 1 to 9, with 9 being to the most resistant cultivar), while the variety Cellule (CEL), carrying the gene Stb16q, was considered to be more resistant (rated 7). Leaf samples of each variety were collected in three plots of 30 m2. The three APA and CEL plots were independent experimental plots described in Orellana-Torrejon et al. (202) (Orellana-Torrejon *et al.* 2021 while the three SOI plots were delineated within a larger (1 ha) wheat field described in Morais et al. (2016) (Morais *et al.* 2016) and Kerdraon et al. (2019) (Kerdraon *et al.* 2019). Within each plot, five samples were taken at locations spaced 1 m apart along a transect. For each sample, three pieces of leaf were collected: an asymptomatic leaf piece taken from a leaf without any STB lesions (G); an asymptomatic leaf piece from a leaf with STB lesions (GS); and a symptomatic leaf piece including a portion of sporulating lesion (S), i.e. bearing pycnidia (*Z. tritici* asexual fruiting bodies). Three sampling campaigns were performed: the first on March 14th (SOI) and 15th (APA and CEL); the second on May 3rd; and the third on June 13th. In March and May, leaf pieces measured 5 cm long. G samples were taken from the central part of the second leaf (F2) of a plant located as close as possible to the sampling point. S and GS samples were collected from the third leaf (F3) of another plant. S samples were taken from the distal part of the leaf and GS samples were cut from the basal part (closer to the stem insertion) of the same leaf. In June, leaf pieces measured 3 cm long because leaves were broader and

our goal was to collect an approximately similar amount of tissue on all sampling dates. All leaves were found to be symptomatic in June so we only collected S and GS samples from the third leaf of different plants.

Weeds

On July 16th, samples were collected on eight species of weeds to produce a complementary dataset. Some of these weeds presented symptoms, caused by undetermined fungal pathogens that were not *Z. tritici* that is specific to wheat. Five GS and five S samples were collected on *Lolium perenne* (LOLPE) individuals growing within the SOI field and on *Arrhenatherum elatius* (ARREL) individuals growing on a slope 5m away from the SOI field. These two weed species were dominant weeds at the time of sampling. Five G samples were also collected on *Senecio vulgaris* (SENVU) individuals growing within the SOI field, *Poa annua* (POAAN) individuals growing on the path along the SOI field, *Hordeum murinum* (HORMU) and *Plantago lanceolata* (PLALA) individuals growing between the field and the path, and *Urtica dioica* (URTDI) and *Geranium molle* (GERMO) individuals growing on the slope 2m away from the SOI field. All leaf samples were cut with scissors and placed in 2 ml autoclaved collection tubes. They were then brought back to the laboratory and stored at -20°C prior to freeze-drying.

2.2 DNA extraction

Total DNA was extracted with the DNeasy Plant Mini kit (Qiagen, France), using a protocol slightly modified from that recommended by Kerdraon *et al.* 2019. Two autoclaved DNAase-free inox 420C beads were added to each tube and samples were ground at 1500 rpm with the Geno/Grinder® for 30 s, then 1 min and 1 min again, with manual shaking between each grinding step. Tubes were then centrifuged for 1 min at 6000 g. Leaf powder and 200 µL of buffer AP1 preheated to 60°C were mixed by vortexing the tubes for 30 s twice at 1500 g, and centrifuging them for 1 min at 3000 g. 250 µL of preheated buffer AP1 and 4.5 µL of RNase A were added to each tube and mixed by vortexing the tubes for 30s twice at 1500 g. After 5 min of rest, 130 µL of buffer P3 was added to each tube, which was then mixed by gentle inversion for 15 s, incubated at -20°C for 10 min and centrifuged for 1 min at 5000 g. The supernatant (450 µL) was transferred to a spin column and centrifuged for 2 min at 20000 g. The filtrate (200 µL) was transferred to a new tube, to which sodium acetate (200 µL, 3 M, pH 5) and cold 2-propanol (600 µL) were added. DNA was precipitated by incubation at -20°C for a minimum of 1 hr and recovered by centrifugation (20 min, 13000 g). The pellet was washed with cold ethanol (70%), dried at 50°C for about 30min, and dissolved in 100 µL of AE buffer.

2.3 Bacterial 16S amplification

The V5-V6 region of the bacterial 16S rDNA gene was amplified using primers 799F-1115R (Redford *et al.* 2010, Chelius & Triplett 2001) 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 (799F: 5′- CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTxxxxxxxxxxxxHS-AACMGGATTAGATACCCKG-3′; 1115R: 5′- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxxxxxxxHS-AGGGTTGCGCTCGTTG-3′, where HS represents a 0-7-base-pair heterogeneity spacer and "x" a 12 nucleotide tag). The PCR mixture (20 µL of final volume) consisted of 4 µL of buffer Phusion High-Fidelity 5X (ThermoFisher) (1X final), 2 µL each of the forward and reverse primers (0.2 µM final), 2 µL of 2 mM dNTPs (200 µM final), 8.2 µL of water, 0.6 µL of SO, DM0, 2 µL of Phusion Hot Start II Polymerase (ThermoFisher) and 1 µL of DNA template. PCR cycling reactions were conducted on a Veriti 96-well Thermal Cycler (Applied Biosystems) using the following conditions: initial denaturation at 98°C for 30s followed by 30 cycles at 98°C for 15 s, 60°C for 30 s, 72°C for 30 s with final extension of 72°C for 10 min. Two marine fungal strains (*Candida oceani* and *Yamadazyma barbieri*) were used as positive controls as they were unlikely to be found in our samples. One positive control included 1 µL of 10 ng/µL DNA of *Candida oceani* only and the other included an equimolar mixture of both strains. The negative PCR controls were represented by PCR mix without any DNA template. Each PCR plate contained one negative extraction control, three negative PCR controls, one single-strain positive control and one two-strain positive control.

2.4 Fungal ITS amplification

The ITS1 region of the fungal ITS rDNA gene (Schoch *et al.* 2012) was amplified using primers ITS1F-ITS2 (White *et al.* 1990, Gardes & 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 nucleotide tag). The PCR mixture (20 µL of final volume) consisted of 10 µL of 2X QIAGEN Multiplex PCR Master Mix (2X final), 2 µL each of the forward and reverse primers (0.1 µM final), 4 µL of water, 1 µL of 10 ng/µL BSA and 1 µL of DNA template. PCR cycling reactions were conducted on a Veriti 96-well Thermal Cycler (Applied Biosystems) using the following conditions: initial denaturation at 95°C for 15 min followed by 35 cycles at 94°C for 30 s, 57°C for 90 s, 72°C for 90 s with final extension of 72°C for 10 min. ITS1 amplification was confirmed by electrophoresis on a 2% agarose gel. Two marine fungal strains (*Candida oceani* and *Yamadazyma barbieri*) were used as positive controls as they were unlikely to be found in our samples. One positive control included 1 µL of 10 ng/µL DNA of *Candida oceani* only and the other included an equimolar mixture of both strains. The negative PCR controls were represented by PCR mix without any DNA template. Each PCR plate contained one negative extraction control, three negative PCR controls, one single-strain positive control and one two-strain positive control.

2.5 Sequencing

MiSeq sequencing, PCR products purification (CleanPCR, MokaScience), library sequencing on an Illumina MiSeq platform (v2 chemistry, 2 × 250 bp) and sequence demultiplexing (with exact index search) were performed at the PGTB sequencing facility (Genome Transcriptome Platform of Bordeaux, Pierroton, France). Fungal ITS1 amplicons were sequenced on three runs and bacterial 16S amplicons were sequenced on four runs.

2.6 Bioinformatic treatment

The MiSeq sequences produced were processed using the DADA2 pipeline version 1.22.0 (Callahan *et al.* 2016) implemented in R. Primers were identified and removed using cutadapt 3.2 (Martin 2011) and the trimmed sequences were then parsed to the DADA2 algorithm. Chimeras were removed using the removeBimeraDenovo functionality of DADA2. Taxonomic assignment of amplicon sequence variants (ASVs) taxonomic assignment was performed using an implementation of the Naive Bayesian Classifier (Wang *et al.* 2007) included in the DADA2 pipeline. The databases used for taxonomic assignment were the Silva v138.1 (Quast *et al.* 2012) and the UNITE all eukaryotes v8.3 (Abarenkov *et al.* 2021) for 16S and ITS sequences, respectively. Three tables were obtained at the end of this process: an ASV table with the sequence count in each sample; a table with the taxonomic assignment of each ASV sequence; and a metadata table describing the collection conditions of each sample. The three tables were joined in a phyloseq object using the phyloseq bioconductor package v1.38.0 (McMurdie & Holmes 2013). To filter out possible contaminants, the combined method of the *isContaminant* function of the DECONTAM Bioconductor package v1.14.0 (Davis *et al.* 2018) was used, followed by the decontamination method described in Galan et al. (2016). Moreover, 16S ASVs identified as chloroplastic or mitochondrial with Metaxa2.2.3 (Bengtsson-Palme *et al.* 2015), or according to their taxonomic assignment in the Silva database, were removed. The remaining ASVs were clustered using the Lulu algorithm (Frøslev *et al.* 2017) with default parameters. ASVs that could not be assigned to a bacterial or fungal phylum were removed. Finally, ASVs present in less than 1% of the samples were removed to make sure that the data were free of sequencing artifacts and low abundant contaminants (Cao *et al.* 2021).

2.7 Quantification of *Z. tritici* by qPCR

The abundance of *Z. tritici* in wheat tissues was estimated using the quantitative PCR assay developed by Duvivier *et al.* 2013. The specific set of primers included a forward primer (50-ATTGGCGAGAGGGATGAAGG-30), a reverse primer (50-TTCGTGTCCCAGTGCGTGTA-30), both leading to an amplification product of 101 pb, and a Taqman fluorogenic probe (50-ACGACTCGC GGCTTTCACCCAACG-30). The probe was labelled with a FAM fluorescent reporter dye and a BHQ-1 quencher. The quantification reaction was performed with the CFX96 Real time System C1000 Thermal Cycler (BIORAD, USA), using hard shell PCR 96-well WHT/CLR plates. The mix reaction was composed of reverse and forward primers at 500 nM per reaction, the probe at 500 nM per reaction in a final volume of 25

µL, with 5 µL of DNA introduced per well. All samples (standard DNA, eDNA to be analysed, and negative controls) were analyzed with three replicates. The PCR program was 95°C for 10 min and (95°C for 15 s, 60°C for 20 s, 72°C for 40 s) repeated for 40 cycles. The concentration of DNA in the unknown samples was calculated by comparing cycle threshold (Ct) values of the samples with known standard quantities of *Z. tritici* genomic DNA, using a tenfold serial dilution from 0.5 ng to 5.10-5 ng per well. Ct values were plotted against the log of the initial concentration of *Z. tritici* genomic DNA to produce the standard curve used for sample quantity determination.

2.8 Analysis

Data contained in the phyloseq object were analyzed using the statistical environment R v4.1.2 (R Core Team 2022) to characterize the fungal and bacterial community composition and to assess the effect of the different experimental factors on these communities. The analysis was performed using only the samples obtained from wheat plants at March and May. Samples obtained in June were not included in the analysis because there were no healthy (G) samples available. DNA sequencers can only read a maximum number of sequences. As a result, the reads obtained by the sequencer is a random sample of the total number of DNA sequences, and thus, compositional (Gloor *et al.* 2017). ASV counts were transformed using a clr transformation (Aitchison 1982) to obtain scale-invariant values, avoiding the compositional effect. Then, the phyloseq R package was used to obtain the euclidean distance between samples and to perform a Principal Coordinate Analysis (PCoA). The PCoA was plotted using ggplot2 package v3.3.5 (Wickham 2016). A permutational multivariate analysis of variance (PERMANOVA) was performed to assess the effect of the experimental design on the communities, using the adonis2 function of the vegan R package v2.5.7 (Oksanen *et al.* 2022) following the experimental formula "tissue × date × variety/plot". Alpha diversity measures were obtained using the phyloseq package and fitted in a generalised mixed model using the lme4 R package v1.1-27.1 (Bates *et al.* 2015). *Z. tritici* qPCR analysis was also fitted in a generalised mixed model using lme4.

3 Results

This resource announcement provides two sets of raw sequence files, one set obtained using primers for the fungal ITS region and another obtained using primers for the bacterial 16S region. The sequences are available in the Dataverse files (see section Availability of Data and Materials). The raw and filtered ASV tables obtained during the dereplication and filtering process are provided in the form of phyloseq objects (McMurdie & Holmes 2013). Each phyloseq object includes the ASV table, a table with the ASV taxonomic assignment and a metadata table. The raw ASV tables also include the positive and negative control samples used for the filtering. The samples obtained in June, as well as samples obtained from weeds growing in the vicinity of the wheat crop, are included in the phyloseq objects but were not analyzed in the present study. The metadata table includes, for each sample, the wheat variety or weed species sampled, the sampling date,

the plot, the visual assessment of symptoms and the *Z. tritici* DNA concentration obtained by qPCR. The impacts of the visual assessment of symptoms, the date and the variety on the *Z. tritici* DNA concentration were significant (Figure B.1A). While the varieties Apache and Soissons are more susceptible to STB than Cellule, the difference in the concentration of *Z. tritici* DNA between them were limited, which is not surprising since sampling was based on similar leaf symptom criteria (G, GS, S). The tables showing the change in number of reads in each sample during the bioinformatic process are also provided in the Dataverse files.

3.1 Fungal communities

All leaf samples (n = 360) were sequenced using ITS primers and gave an average of 31,586 raw fungal sequences per sample with a minimum of 45 reads and a maximum of 246,434 reads per sample. The ASV inference process identified an average of 28,178 high quality sequences per sample distributed in 2,821 unique ASVs in the 360 samples. The ASV table obtained after the filtering process, which deleted contaminants and low abundant ASVs, was made up of an average of 27,609 sequences per sample distributed between 391 ASVs and 357 samples. Three samples did not generate enough sequences to assign ASVs. The minimum number of reads in a sample was 20 and the maximum was 223,756. Several samples of weeds (n = 101) growing close to the field were also sequenced using ITS primers. The bioinformatic process yielded an average of 45,631 sequences per sample and a total of 337 ASVs from these weed samples. The minimum number of reads in a sample was 1,331 and the maximum was 365,035. The number of reads in each sample at each step of the bioinformatic process is supplied in the Dataverse files. Taxonomic composition — For the wheat dataset, sequences assigned to *Ascomycota* represented 74% of the total counts, while sequences assigned to *Basidiomycota* represented 25 (Figure B.2A). As expected, ASVs assigned to the genus *Zymoseptoria* were the most abundant (60% of the sequences). *Zymoseptoria* was also more abundant in symptomatic than in asymptomatic leaf samples. It was also slightly more abundant in the Soissons (SOI) and the Apache (APA) varieties, than in the Cellule (CEL) cultivar, which is less susceptible because it carries the Stb16q resistance gene. Alpha diversity — Fungal community richness (number of species), diversity (Shannon diversity index) and evenness (inverse Simpson index) differed significantly between dates and tissue health conditions. Wheat variety had a minor effect in all alpha diversity measures, being significant only for richness and diversity. Beta diversity — The composition of wheat foliar fungal communities (Figure B.1B) differed significantly among dates, varieties and tissue health conditions, (Table B.1). Tissue was the most important factor, explaining 10% of the variance in the permutational analysis of variance. Samples collected in June were not included in the analysis to avoid a potential bias caused by the absence of healthy leaves (G samples) at that time.

3.2 Bacterial communities

The 360 samples used for ITS sequencing were also sequenced using 16S primers, obtaining an average of 40,724 raw bacterial sequences per sample with a mini-

mum of 0 reads and a maximum of 92,520 reads per sample. The ASV inference process identified a mean of 31,969 high quality sequences shared between 12,349 unique ASVs in 350 samples. Ten samples did not generate enough sequences to assign ASVs inference. The filtering process deleted contaminants and low abundant ASVs. The ASV table obtained had an average of 13,964 sequences per sample distributed between 1,495 ASVs and 340 samples. The minimum number of reads in a sample was 2 and the maximum was 71,051. The samples of weeds (n = 102) growing surrounding the wheat plots were also sequenced using 16S primers. The bioinformatic process produced an average of 29,991 sequences per sample and 1,068 unique ASVs from the weed samples. The minimum number of reads in a sample was 30 and the maximum was 70,999. The number of reads of each sample at each step of the bioinformatic process is supplied in the Dataverse files. Taxonomic composition — The most abundant bacterial phyla were *Proteobacteria* (39% of ASVs), followed by *Actinobacteria* (35%), *Bacteroidetes* (12%) and *Firmicutes* (11%) (Figure B.2B). *Proteobacteria* were more evident in later sampling dates while *Actinobacteria* were more present at the March sampling. Tissue condition and wheat variety did not seem to have an important effect on the community composition of the phyla. Alpha diversity — Bacterial community richness, evenness and diversity differed significantly between tissue health conditions. Sampling date affected richness and evenness, while wheat variety only affected diversity. Beta diversity — The composition of the bacterial communities of wheat leaves (Figure B.1C) differed significantly among dates, varieties and tissue conditions, (Table B.1). As with the fungal component of the community, date was the most important structuring factor, explaining some 8% of the variance in the permutational analysis of variance. Samples obtained in June were not included in the analysis to avoid a potential for bias caused by the absence of healthy leaf samples on that sample date.

4 Conclusions

Preliminary statistical analyses revealed that sampling date, wheat variety and STB symptoms had significant effects on fungal and bacterial communities of the wheat phyllosphere. While the three factors tested affected the community structure, the date of sampling exhibited the strongest effect. As expected, we found a relationship between the presence of *Z. tritici* assessed by eye (STB symptoms on the leaves) and by qPCR (concentration of *Z. tritici* DNA within the leaf tissue). Moreover, the large community overlap between samples from the three wheat varieties (Figures B.1 B and C) suggests that microbiomes are similar despite their variation in susceptibility to STB. These findings have implications for the concept that variation in the microbiome could be relevant to define and optimize biological control of STB taking into account the impact of the wheat varieties. The preliminary analysis of this wheat dataset confirms the effectiveness of the sampling strategy and the need for further in-depth investigations. For instance, co-occurrence network analyses could be used to characterize the dynamics of the community associated with *Z. tritici* and might help identify individual taxa of interest as potential biocontrol or beneficial agents to improve wheat health. The

weed dataset could also be examined for *Z. tritici* interactions with different communities present on non-crop plants within and in the margins of the field.

5 Availability of Data and Materials

The sequence datasets were deposited in NCBI SRA in bioproject PRJNA803042 (https://www.ncbi.nlm.nih.gov/bioproject/803042). The biosample accession numbers are SAMN25610777 to SAMN25611238. Bioinformatic scripts and raw and filtered ASV tables in R phyloseq format were deposited in Dataverse

(https://doi.org/10.15454/QTXFP9). The tables showing variation in sequence counts during the bioinformatic process and the scripts used for data processing and statistical analysis were included in the Dataverse deposit.

6 Author contributions

Author contributions were as follows: (i) D.A.B., C.V., F.S. and V.L coordinated the study; (ii) F.S., V.L., A.-S.W. and M.M. designed the sampling, collected samples and metadata; (iii) N.V., J.F.A., E.C. and E.G. developed protocols and performed the molecular biology work; (iv) D.B.B. performed data analysis and wrote the manuscript; (v) all authors edited the manuscript.

7 Acknowledgments

We thank all members of the Consortium Biocontrôle for their support for the project. We thank Céline Lalanne, Adline Delcamp, Christophe Boury and all other members of PGTB (Genome Transcriptome Platform of Bordeaux) for their support in molecular biology, and Gaétan Burgaud and Frédéric Garabetian for kindly providing marine strains used as positive controls. We also thank Frédéric Barraquand, Stéphane Robin, Charlie Pauvert and Andreas Makiola for helpful discussions at the beginning of the project. We acknowledge support from the ANR NGB project (ANR-17-CE32-0011). PTGB and BIOGER received support from Investissements d'avenir and convention attributive d'aide EquipEx Xyloforest (ANR-10-EQPX-16-01) and from Saclay Plant Sciences-SPS (ANR-17-EUR-0007), respectively. References

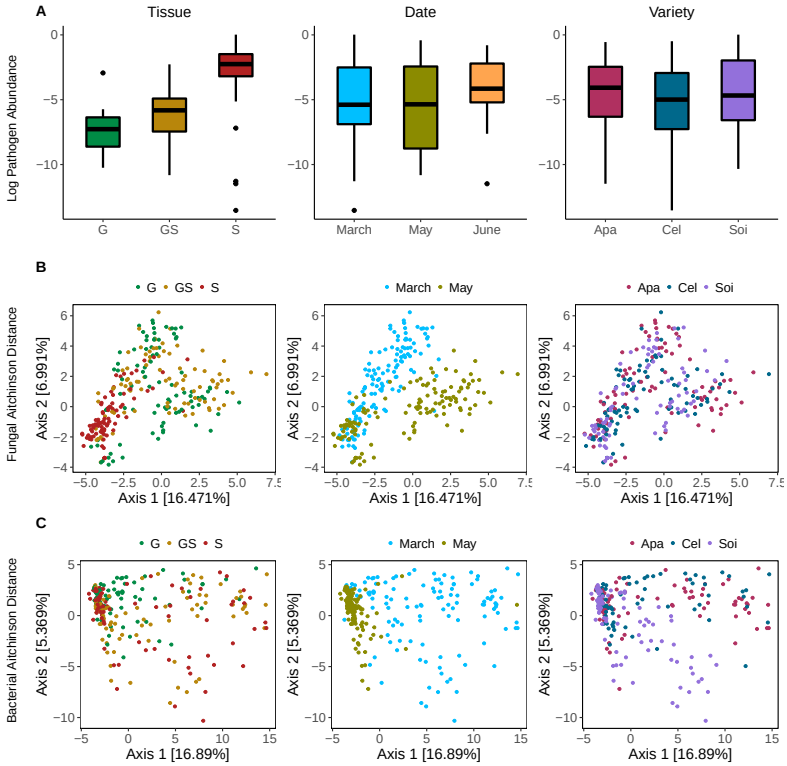


Figure B.1: Community composition of wheat leaves. **A**, Abundance of the pathogen (*Zymoseptoria tritici*) on wheat leaves measured using quantitative PCR. Principal coordinates analysis plots showing the similarity of **B**, fungal and **C**, bacterial communities from different samples. Ordination was performed using the Aitchison distance (Aitchison 1982). Plots were colored by the *Septoria tritici* blotch symptoms: leaf samples collected on asymptomatic leaves (G), green parts of a symptomatic leaf (GS), and symptomatic parts of a leaf (S); the sampling seasons of March, May, and June; and the wheat varieties sampled: Apache (Apa), Cellule (Cel), and Soissons (Soi).

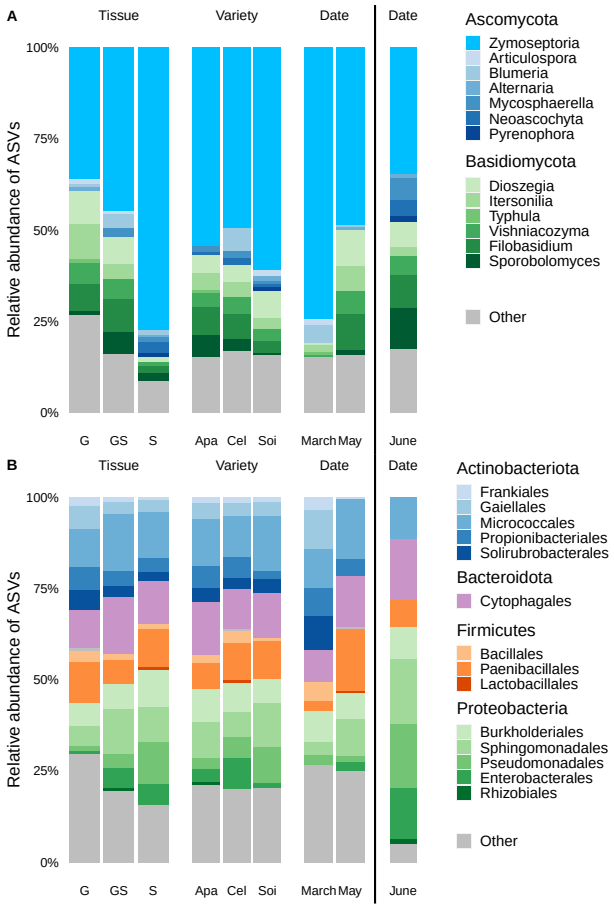


Figure B.2: Relative abundance of different taxa over different sampling dates, wheat varieties, and leaf tissue conditions. **A**, Relative abundance of different fungal genera and **B**, relative abundance of different bacterial orders. ASV = amplicon sequence variant. Bar plots are separated by Septoria leaf blotch symptoms: leaf samples collected on asymptomatic leaves (G), green parts of a symptomatic leaf (GS), and symptomatic parts of a leaf (S); the wheat varieties sampled: Apache (Apa), Cellule (Cel), and Soissons (Soi); and the sampling dates in March, May, and June (June is separated by a bar because no asymptomatic (G) samples were found).

Table B.1: Permutational multivariate analysis of variance between samples. The Aitchinson distance (Aitchison 1982) between samples in the abundance matrix is used as a distance metric. The analysis were performed for the fungal and bacterial ASV tables separately using the R vegan package (Oksanen *et al.* 2022). Factors are: Tissue, correspondiong to septoria leaf blotch symptoms; Date, corresponding to the sampling dates; and Variety, corresponding to the wheat varieties sampled. The plots were considered nested to the wheat variety. Values of p = * < .05; ** < .01; *** < .001

Fungi	**Df**	**Sum of squares**	**R2**	**F**	**P value**
Tissue	2	1369.149	0.051	12.319	***
Date	2	2856.615	0.107	25.702	***
Variety	2	615.13	0.023	5.534	***
Tissue:Date	3	968.617	0.036	5.81	***
Tissue:Variety	4	370.541	0.014	1.667	***
Date:Variety	4	803.509	0.03	3.615	***
Tissue:Date:Variety	6	488.472	0.018	1.465	***
Tissue:Date:Variety:Plot	48	3510.502	0.131	1.316	***
Residual	285	15838.138	0.591		
Total	356	26820.673	1		
Bacteria	**Df**	**Sum of squares**	**R2**	**F**	**P value**
Tissue	2	822.805	0.017	3.699	***
Date	2	6153.514	0.126	27.664	***
Variety	2	1189.58	0.024	5.348	***
Tissue:Date	3	856.926	0.017	2.568	***
Tissue:Variety	4	574.873	0.012	1.292	*
Date:Variety	4	1918.814	0.039	4.313	***
Tissue:Date:Variety	6	909.533	0.019	1.363	**
Tissue:Date:Variety:Plot	48	6775.259	0.138	1.269	***
Residual	268	29806.437	0.608		
Total	339	49007.741	1		

Appendix C

List of Figures and Tables

List of Figures

I.1 **Change in the abundances of a taxon following the establishment of a second taxon in the media.** **A**: Exploitative competition interaction where both taxa arrive at an abundance equilibrium. **B**: Competitive exclusion process where one taxa is able to exclude the other from the media. **C**: Cyclical dynamics of abundance in predator-prey interaction. 21

I.2 **Steps to reconstruct an interaction network using eDNA:** 1. Ecological samples are taken from the environment; 2. eDNA is extracted from the ecological sample; 3. More copies of the DNA are obtained by amplification using primers specific for a taxonomic group; 4. DNA is sequenced; 5. A bioinformatic pipeline is followed to cluster the sequences in OTUs and obtain a measure of OTU abundance; 6. Interaction networks are reconstructed using OTU abundance data. 28

I.3 **Cross-sectional and longitudinal abundance measures of four OTUs.** Cross-sectional network reconstruction uses the change in abundance of samples at the same point in time, but different point in space for each OTU. Longitudinal network reconstruction uses samples from the same point in space, but different points in time for each OTU. 29

I.4 **Logic inference processes as a function of the input information.** Deduction uses rules and causes to determine the exact effect. Induction uses causes and effects to determine the rules that drive them. Abduction proposes plausible causes for effects based on rules. 34

II.1 **Experimental design**. Foliar fungal communities were characterized in three conventional (CONV) and three organic (ORGA) vineyard plots by a metabarcoding approach. We analyzed 20 foliar samples per plot. For each plot, we obtained 20 community profiles (described in terms of amplicon sequence variants (ASV)) and one association network (inferred either with the SparCC software developed by Friedman & Alm, 2012 or with the SPIEC-EASI software developed by Kurtz et al, 2015). More networks were then obtained by varying network reconstruction parameters (Figure A.3). The effects of cropping system (CONV versus ORGA) on the grapevine foliar microbiota were assessed with both community and network α- and β-properties. 44

II.2 **Effect of cropping system —conventional (CONV) versus organic (ORGA) — on the α-diversity and β-diversity metrics of grapevine foliar fungal communities.** **A** Community richness, defined as the number of ASVs. **B** Community diversity, measured with the inverse Simpson index. **C** Community evenness, measured with Pielou's index. Differences in α-diversity metrics between cropping systems were significant (Table S4; * $p<0.05$; **$p<0.01$; ***$p<0.001$). **D** Principal coordinate analysis (PCoA) was used to represent dissimilarities in composition between samples, as assessed with the quantitative and **E** binary Jaccard indices. The effect of the cropping system on both β-diversity metrics was significant, as a single effect for the quantitative Jaccard index and in interaction with block for the binary index (Table II.2). Green circles, squares and triangles correspond to samples collected in the ORGA1, ORGA2 and ORGA3 plots, respectively. Orange circles, squares and triangles correspond to the CONV1, CONV2 and CONV3 plots, respectively (Figure II.1). **F** Log-transformed ratio of ASV relative abundance in CONV plots over that in ORGA plots, for 14 ASVs identified as differentially abundant between cropping systems by DESeq2 analysis followed by Benjamini-Hochberg adjustment (Love *et al.* 2014). 52

II.3 **Effect of cropping system — conventional (CONV) versus organic (ORGA) — on the α-properties and β-properties of grapevine foliar fungal networks.** **A** Association networks inferred from fungal metabarcoding data with SparCC (Friedman & Alm 2012). A total of 60 networks were inferred, corresponding to 2 cropping systems $\times$ 3 replicates (blocks) $\times$ 10 P values, with P the percentage of most abundant ASVs used for network inference. Only four values of P are shown on the Figure. **B** Variations in network α-properties. The following properties (Table II.1) were calculated for each network: the number of links (L) and connected components (CC), the network diameter (DIA) and connectance (C) and the mean degree (DEG) and negative link ratio (NLR). The percentage P of ASVs used for network reconstruction had a significant influence on all properties (Table A.8), whereas the cropping system did not (Table A.7). **C** Principal coordinate analysis (PCoA) represents dissimilarities between networks, measured with the βOS index (Poisot *et al.* 2012) calculated with the binary Jaccard index. βOS measures the dissimilarity between two networks in terms of the presence-absence of associations between shared ASVs. The centroids for each cropping system are represented by gray circles. The effect of the cropping system on βOS was significant (Table II.4). Networks were inferred with SparCC (Friedman & Alm 2012). 55

II.4 **Venn diagrams showing the number of fungal associations common to network replicates. A** Associations common to the three network replicates inferred for the organic cropping system (ORGA1, ORGA2, ORGA3) and **B** the three network replicates inferred for the conventional cropping system (CONV1, CONV2, CONV3), regardless of the sign of the association, in the situation in which all ASVs were used for network construction (P=100%). **C** Associations common to the six networks. Networks were inferred with SparCC (Friedman & Alm 2012). The number of nodes shared by the network replicates is indicated into brackets. 59

III.1 **Description of the interaction inference process.** Microbial communities are shaped by the interaction between their members. DNA sequencing together with bioinformatic processes allow estimation of the abundance of the different microbes present in the communities. Using the abundance information from different communities as training examples, and the rules of interaction as background knowledge, it is possible to infer an interaction network that generalizes the interactions between microbes. 69

III.2 **Summary of the inference process of microbial interactions using A/ILP.** Observations are obtained assessing the abundance change between OTUs. The ecological theory describes how the presence of an OTU can affect the abundance of a second OTU. Abduction is performed using the observations and theory. Significant interactions are assessed by bootstrapping the compression value from different permutations of observations 74

III.3 **Area under the ROC curve values (AUC) obtained using different number of permutations.** Each plot shows the AUCs obtained for interactions of different strengths. Each line represents a method used to obtain the estimators. Error bars show the standard deviation of the means. 76

IV.1 **Schematic diagram representing the InfIntE pipeline.** The pipeline performs: conversion of abundance data contained in the OTU table to logical clauses, based upon our ecological knowledge; abduction of interaction effects from the logical clauses using PyGol; selection of important edges using StARS; and, the direct classification of interaction types depicted by the different edge colors. Arrows show the direction of effects or interactions. Edges without arrow in the network represent non-directed interactions. 90

IV.2 **Relationship between number of samples and interaction inference performance for different strengths of interaction**. Datasets were computer-generated simulating four different interaction types: amensalism, commensalism, competition and mutualism. **A:** Area under the roc curve values (AUC)(Fan *et al.* 2006) obtained by I statistic with and without exclusion. Larger AUC values represent better specificity and sensitivity in interaction detection. I statistic is used by InfIntE as a numeric measure of interaction. **B:** Area under the roc curve values (AUC) obtained by InfIntE's I statistic and SparCC and SPIEC-EASI correlation like measures. InfIntE used the hypothesis of interactions including exclusion. SparCC and SPIEC-EASI were executed with default settings. **C:** Accuracy of interaction detection computed as described in section 2.2. InfIntE used the hypothesis of interactions including exclusion. SparCC and SPIEC-EASI were executed with default settings. 94

IV.3 **Nightingale rose charts comparing the percentage of correct interaction classification by types.** OTU tables were synthetically generated simulating groups of 60 replicated samples mixing four different types of interactions: amensalism, commensalism, competition and mutualism. Inference of interactions was performed using InfIntE, SparCC and SPIEC-EASI. Charts in the top row show the percentage of each interaction type correctly detected by each tool. Bottom row charts show the percentage of false positives proposed by each tool over the total possible false positives. Each petal of a rose chart is colored as a function of the classification of the detected interaction type given for each inference tool. InfIntE automatically classifies interactions as amensalism, commensalism, competition and mutualism while SparCC and SPIEC-EASI return positive (+) or negative (-) associations. InfIntE correctly detects and classifies most mutualism and commensalism as well as around 20% of competition interactions with few false positives. SparCC detects most interactions at the expense of a large amount of false positives. SPIEC-EASI has a similar performance to InfIntE, but without interaction type classification. . . 96

IV.4 **Accuracy of abundance change prediction in grapevine metabarcoding data** as a function of the number of observations used for the inference. Each dataset consisted of 60 samples. Each point presents the mean accuracy of prediction for different fold combinations of the 9 vineyards 97

IV.5 **Interaction networks predicted for each of the 9 different vineyards** in the dataset, inferred using InfIntE. Each vineyard dataset was composed by 60 samples. The edge colours follow the interaction typology. The pie chart associated with each network indicates the relative percentage of each interaction type in the network 98

V.1 **Screenshot image showing the network visualisation tool.** Interactive visualisation options are available in the left-hand panel. The image of the network is displayed at the right-hand panel. The network displayed was predicted in the vignette example. The network image shows the interactions between bacterial OTUs inhabiting wheat leaves 115

VII.1 **Taxonomic barplots showing the relative abundance of A fungal genera and B bacterial orders in grapevine (*Vitis vinifera L.*)** leaves depending on the geographic region (Champagne (CH), Aquitaine (AQ) or Occitanie (OC)) and tissue health condition (asymptomatic (A) or sporulating downy mildew lesion (S)). 138

VII.2 **Effects of geography and disease on the leaf microbiome of European cultivated grapevine (*Vitis vinifera L.*).** **A**: Map of France showing the location of the sampling sites. Vineyard plots A, B and C are located in Aquitaine (AQ); D, F and E are located in Occitanie (OC); G, H and I are located in Champagne (CH). **B**: Concentration of *Plasmopara viticola* DNA, estimated by qPCR, in sporulating mildew lesions (S) and visually asymptomatic leaf samples (A). **C, D, F & F** Procrustes rotation of the Principal Coordinates Analysis (PCoA) plots of symptomatic and asymptomatic samples showing the similarity of fungal communities across leaf samples for fungal and bacterial datasets. Plots are colored by the geographic region (CH, AQ or OC) and tissue health condition (A or S). For both datasets, the permutation test to assess the significance of the Procrustes statistic was significant. 140

VII.3 **Variations in microbial community alpha-diversity** (observed richness, Shannon diversity and inverse Simpson evenness) in grapevine leaves, among tissue health conditions (asymptomatic (A) and sporulating downy mildew lesion (S)). 142

A.1 **Dendrogram plot of compositional dissimilarities between technical replicates for sequencing.** Technical replicates were created by splitting six lots of PCR products in half and sequencing the two halves independently. The PCR products used were those corresponding to the leaf collected on vine number 24 (L24) in each of the six plots studied (ORGA1, ORGA2, ORGA3, CONV1, CONV2, CONV3; see Figure 1). Compositional dissimilarities between samples were computed with **A** the binary Jaccard index and **B** the quantitative Jaccard index. The dendrogram was built using a hierarchical clustering algorithm (complete linkage method). Compositional dissimilarities between the two technical replicates of the same sample were significantly smaller than the dissimilarities among samples (PERMANOVA: F = 39.98; R2 = 0.97; p = 0.001). 176

A.2 **Effect of cropping system — conventional (CONV) versus organic (ORGA) — on the α-properties and β-propertie of grapevine foliar fungal networks. A** Association networks inferred from fungal metabarcoding data with SPIEC-EASI (Kurtz *et al.* 2015). A total of 60 networks were inferred, corresponding to 2 cropping systems × 3 replicates (blocks) × 10 P values, with P the percentage of most abundant ASVs used for network inference. Only four values of P are shown on the Figure. **B** Variations in network α-properties. The following properties (Table II.1) were calculated for each network: the number of links (L) and connected components (CC), the network diameter (DIA) and connectance (C) and the mean degree (DEG) and negative link ratio (NLR). The percentage P of ASVs used for network reconstruction had a significant influence on all properties (Table A.10), whereas the cropping system did not (Table A.9). **C** Principal coordinate analysis (PCoA) represents dissimilarities between networks, measured with the βOS index (Poisot *et al.* 2012) calculated with the binary Jaccard index. βOS measures the dissimilarity between two networks in terms of the presence-absence of associations between shared ASVs. The centroids for each cropping system are represented by gray circles. The effect of the cropping system on βOS was significant, in interaction with the percentage P of most abundant ASVs used for network inference (Table A.11). Networks were inferred with SPIEC-EASI (Kurtz *et al.* 2015). 177

A.3 **Normalized degree of nodes in networks inferred with A SparCC or B SPIEC-EASI.** Nodes were classified according to the relative abundance of their corresponding ASVs. Abundance class 0-10 corresponds to the 10% most abundant nodes, while abundance class 90-100 corresponds to the 10% less abundant nodes. Normalized degree was obtained by dividing the node degree by n-1, where n is the total number of nodes in the network. The effect of abundance class on the normalized node degree was analyzed with ANOVA followed by post-hoc Tukey's test. Effect of abundance class was significant in both cases (SparCC: $F = 6.797$, $p < 0.001$; SPIEC-EASI: $F = 173.8$, $p < 0.001$). 178

A.4 **Venn diagrams showing the number of fungal associations common to network replicates. A** Associations common to the three network replicates inferred for the organic cropping system (ORGA1, ORGA2, ORGA3) and **B** the three network replicates inferred for the conventional cropping system (CONV1, CONV2, CONV3), regardless of the sign of the association, in the situation in which all ASVs were used for network construction (P=100%). **C** Associations common to the six networks. Networks were inferred with SPIEC-EASI (Kurtz *et al.* 2015). The number of nodes shared by the network replicates is indicated into brackets. 179

A.5 **Consensus networks between the three network replicates for the organic (ORGA) and the conventional (CONV) cropping systems depending on the method for network inference.** Network nodes represent fungal ASVs and links represent significant positive (+) or negative (-) associations common to the three network replicates (Fig. 6 and S4). The fungal ASVs absent from a network are indicated in gray. Networks were inferred with SparCC (Friedman & Alm 2012) or SPIEC-EASI (Kurtz *et al.* 2015). 180

A.6 **Comparison of the performance of Progol, Pygol and Aleph.** ROC curves showing the performance of I statistic computed from the compression values obtained from abducing 9 computer generated datasets. Data-sets are randomly generated following the ecological models of Weiss et al. (2016). Higher area under the curve (AUC) indicates better predictive power. Line color show the A/ILP programm used. 181

A.7 **Average quality scores distribution in function of sequence position for the sequenced ITS A and 16S B input fastq files.** Fastq files were obtained sequencing the eDNA present in grapevine leaves. The plot was done using the plotQualityProfile function of the DADA2 pipeline. 182

A.8 **Rarefaction curves showing the variation of ASV richness in function of the rarefied sequencing depth for the ITS and 16S metabarcoding datasets.** Each plot shows the rarefaction curve for asymptomatic (A) and sporulating downy mildew lesion (S) samples. 183

B.1 **Community composition of wheat leaves. A**, Abundance of the pathogen (*Zymoseptoria tritici*) on wheat leaves measured using quantitative PCR. Principal coordinates analysis plots showing the similarity of **B**, fungal and **C**, bacterial communities from different samples. Ordination was performed using the Aitchison distance (Aitchison 1982). Plots were colored by the *Septoria tritici* blotch symptoms: leaf samples collected on asymptomatic leaves (G), green parts of a symptomatic leaf (GS), and symptomatic parts of a leaf (S); the sampling seasons of March, May, and June; and the wheat varieties sampled: Apache (Apa), Cellule (Cel), and Soissons (Soi). 204

B.2 **Relative abundance of different taxa over different sampling dates, wheat varieties, and leaf tissue conditions. A**, Relative abundance of different fungal genera and **B**, relative abundance of different bacterial orders. ASV = amplicon sequence variant. Bar plots are separated by Septoria leaf blotch symptoms: leaf samples collected on asymptomatic leaves (G), green parts of a symptomatic leaf (GS), and symptomatic parts of a leaf (S); the wheat varieties sampled: Apache (Apa), Cellule (Cel), and Soissons (Soi); and the sampling dates in March, May, and June (June is separated by a bar because no asymptomatic (G) samples were found). 205

List of Tables

I.1 **Interaction types as a function of the effect on the involved taxa.** Inspired by Derocles *et al.* 2018; Faust & Raes 2012. 19
I.2 **DNA region amplified to metabarcode the different taxonomic groups present in microbial communities.** 24

II.1 **List of community-level and network-level α- and β-properties analyzed in the study.** The number of independent observations (N) and the size of corresponding dissimilarity matrices (S) are indicated. The last column indicates if the property varied significantly (Yes/No) with change in the cropping system (CS). 45
II.2 **Effect of cropping system — conventional versus organic — on the β-diversity metrics of grapevine foliar fungal communities.** Dissimilarities in community composition between samples were assessed with both the quantitative and binary Jaccard indices. The effects of sequencing depth (SD, log-transformed), cropping system (CS) and block (B) on compositional dissimilarities between communities were evaluated using permutational analysis of variance (PERMANOVA), with the number of permutations set to 999. 54
II.3 **Most abundant amplicon sequence variants (ASVs) in grapevine foliar fungal communities according to the cropping system.** The relative abundances (RA, in %) and ranks of ASVs were calculated for all leaf samples (TOTAL; n = 112) and for samples collected from organic (ORGA; n = 55) and conventional plots (CONV; n = 57). 56
II.4 **Effect of cropping system — conventional versus organic — on the β-properties of grapevine foliar fungal networks inferred with SparCC.** The D index quantifies the topological dissimilarity between networks (Schieber *et al.* 2017) whereas the other three metrics (βWN, βOS and βST), which were calculated with the binary Jaccard index, quantify differences in associations between networks (Poisot *et al.* 2012). The effect of the percentage P of the most abundant ASVs used for network inference, and the effect of cropping system (CS) on the dissimilarities between networks were evaluated in permutational analysis of variance (PERMANOVA). The number of permutations was set to 999 and permutations were constrained by block. 57

II.5 **Number of associations shared between network replicates within each cropping system — conventional (CONV) and organic (ORGA) — depending on the method of network inference.** Networks were inferred with SparCC (Friedman & Alm 2012) or SPIEC-EASI (Kurtz *et al.* 2015), by aggregating or not the ASVs at the genus level, and by including various percentages P of the most abundant ASVs or genera in the network. The number of shared ASVs or genera between the three network replicates is given into brackets. For every combination of parameters, three random networks having the same number of nodes and links than the three inferred networks were simulated. The pseudo p-value is the probability, estimated with 999 simulations, that the three random networks shared more associations than the three inferred networks (* p<0.05; ** p<0.01; *** p<0.001). 61

III.1 **Type of interactions in function of the changes in abundance following Derocles *et al.* 2018.** 69

III.2 **Performance of the bootstrapping estimator compared with optimal threshold obtained from the ROC curve and SparCC.** The three datsets used for the interaction inference have 16 real interactions over 496 possible interactions. 80

IV.1 **Interactions types as described in Derocles et al. 2018.** (Derocles *et al.* 2018) 84

IV.2 **Potential *Plasmopara viticola* antagonists found by InfIntE.** The table shows the fungal species found to have a potential interaction able to reduce the abundance of *P. viticola*. A bibliographical search in Google Scholar, Pubmed and Science Direct was conducted to identify whether potential antagonists have previously been described as biocontrol agents of *P. viticola* or other pathogens in the literature. The keywords used for the search were the name of the potential antagonist, "*Plasmopara viticola*", "biocontrol" and "antagonist". Those taxa identified with an asterisk were not automatically assigned to a taxonomic grouping in UNITE and required manual curation and assignment using BLAST. When there was more than one OTU assigned to the same species having the same interaction it is noted with xn, where n is the number of OTUs . . . 101

VII.1 **Sampling design.** Grapevine leaves were collected in three winegrowing regions in France, which are presented in the table from North to South: Champagne (CH), Aquitaine (AQ) and Occitanie (OC). Sampling took place in three vineyard plots (A to I) per region (Figure VII.2A). The grapevine variety and the sampling date are indicated for each plot. The GPS coordinates of each plot are given as Latitude (Lat) and Longitude (Lon). Samples were collected during downy mildew (*P. viticola*) epidemics from 30 vines (*V. vinifera L.*) in each vineyard at a date were the epidemics was at its peak of infection. 134

A.1 **List of phytosanitary products and active ingredients applied in the year of the sampling campaign, together with their normalized dose (also referred to as the treatment frequency index).** PM = powdery mildew, caused by the fungal pathogen *Erysiphe necator* and DM = downy mildew, caused by the oomycete pathogen *Plasmopara viticola*. Leaf sampling was performed on September 10, 2015, more than one month after the last phytosanitary treatment and a couple of hours before grape harvest. The treatment frequency index did not differ between cropping systems (ANOVA: df = 21; F = 0.436; p = 0.516). 184
A.2 **Effect of cropping system — conventional (CONV) versus organic (ORGA) — on the incidence and severity of foliar disease symptoms at harvest time (2015-09-07).** Disease incidence is defined as the percentage of leaves displaying symptoms, whereas disease severity is defined as the percentage leaf damage. Symptom incidence and severity were estimated visually on 40 grapevines for each plot (40 × 3 per cropping system). The mean values are reported for each cropping system as a percentage. Wald $\chi 2$ tests were used for comparisons after linear mixed model analysis with cropping system as a fixed effect and block as a random effect. 185
A.3 **Primer pairs used to amplify the fungal ITS1 region.** 185
A.4 **Effect of cropping system — conventional (CONV) versus organic (ORGA) — on community α-properties.** Generalized linear mixed models included the cropping system as a fixed treatment effect and the sampling depth as an offset. For every community α-property (as defined in Table II.1), we compared the likelihood of a full model including the block and its interaction with the cropping system as random effects and a simplified model including only the block factor as random effect. Only the results of the best model are shown. The ORGA system was taken as the reference. 186
A.5 **Effect of cropping system — conventional (CONV) versus organic (ORGA) — on community α-properties.** Generalized linear mixed models included the cropping system and the proportion of reads assigned to the *Erysiphe* genus as fixed effects and the sampling depth as an offset. For every community α-property (as defined in Table II.1), we compared the likelihood of a full model including the block and its interaction with the cropping system as random effects and a simplified model including only the block factor as random effect. Only the results of the best model are shown. The ORGA system was taken as the reference. 187

A.6 **Effect of cropping system —conventional (CONV) versus organic (ORGA) — on the level of stochasticity in community assembly.** The relative contribution of deterministic and stochastic processes to community assembly was assessed for each cropping system with the Normalized Stochasticity Ratio (NST) defined by Ning et al. (2019), that ranges from 0 to 100, where 0 means a completely deterministic assembly process and 100 a completely stochastic assembly process. NST was calculated using the tNST function with the quantitative and binary Jaccard dissimilarity indices, the FE null model, and other parameters by default values. Differences in NST values between both cropping systems were tested using permutational analysis of variance. 188

A.7 **Effect of cropping system on the α-properties of fungal association networks inferred with SparCC.** Properties (as defined in Table II.1) were compared between cropping systems for every value of the percentage P of the most abundant ASVs used for network inference. The U and p-values of Wilcoxon rank-sum tests are reported. The p-value is not available (NA) for situations in which property values were equal for all networks. The p-values after Benjamini-Hochberg adjustment are not reported because all were equal to one. 189

A.8 **Effect of the percentage P of the most abundant ASVs used for network inference on the α-properties of fungal association networks inferred with SparCC.** Spearman's correlation coefficient and the results of Spearman's rank correlation tests are reported for each network property (as defined in Table II.1). The p-values are reported after Benjamini-Hochberg adjustment. 190

A.9 **Effect of cropping system on the α-properties of fungal association networks inferred with SPIEC-EASI.** Properties (as defined in Table II.1) were compared between cropping systems for every value of the percentage P of the most abundant ASVs used for network inference. The U and p-values of Wilcoxon rank-sum tests are reported. The p-value is not available (NA) for situations in which property values were equal for all networks. The p-values after Benjamini-Hochberg adjustment are not reported because all were equal to one. 191

A.10 **Effect of the percentage P of the most abundant ASVs used for network inference on the α-properties of fungal association networks inferred with SPIEC-EASI.** Spearman's correlation coefficient and the results of Spearman's rank correlation tests are reported for each network property (as defined in Table II.1). The p-values are reported after Benjamini-Hochberg adjustment. . . 192

A.11 **Effect of cropping system — conventional versus organic — on the β-properties of grapevine foliar fungal networks inferred with SPIEC-EASI.** The D index quantifies the topological dissimilarity between networks (Schieber *et al.* 2017) whereas the other three metrics (βWN, βOS and βST), which were calculated with the binary Jaccard index, quantify differences in associations between networks (Poisot *et al.* 2012). The effect of the percentage P of the most abundant ASVs used for network inference, and the effect of cropping system on the dissimilarities between networks were evaluated in permutational analysis of variance (PERMANOVA). The number of permutations was set to 999 and permutations were constrained by block. 193

B.1 **Permutational multivariate analysis of variance between samples.** The Aitchinson distance (Aitchison 1982) between samples in the abundance matrix is used as a distance metric. The analysis were performed for the fungal and bacterial ASV tables separately using the R vegan package (Oksanen *et al.* 2022). Factors are: Tissue, correspondiong to septoria leaf blotch symptoms; Date, corresponding to the sampling dates; and Variety, corresponding to the wheat varieties sampled. The plots were considered nested to the wheat variety. Values of p = * $<$ .05; ** $<$ .01; *** $<$ .001 . . . 206

Appendix D

Glossary

Term	Description
Abductive/Inductive Logic Programming (A/ILP)	High-level knowledge-representation framework that can be used to solve problems based on abductive reasoning.
Amplicon Sequence Variant (ASV)	Group of DNA amplicons that are identical considering the probability of sequencing errors, it can be considered as an OTU synonym
Bio-monitoring	Control the changes and evolution of a biological context
Ecosystem	Complex of living organisms, their physical environment, and all their interrelationships in a particular unit of space.
Environment	Delimited physical space where living organisms habit.
Environmnetal DNA (eDNA)	DNA material obtained by sampling and environment
Explainable Machine Learning (EML)	Machine learning branch focused in human-understandable computation
Interaction	Action or influence of one taxa on another that changes the abundance of one or both taxa
Meta-Interpretative Learning (MIL)	ILP technique which uses higher-order meta-rules to support predicate invention and learning of recursive definitions
Metabarcoding	Process of identifying the taxa present in eDNA in base to their sequence
Microbial Community	multi-species assemblages within which microorganisms live and interact, in a contiguous environment.
Microorganism	Living being not observable at naked eye due its reduced size
Next Generation Sequencing (NGS)	Parallel high-throughput methodology to obtain the nucleotide sequencing of multiple DNA fragments
Next-Generation Biomonitorin prject (NGB)	Project to monitor ecosystems using automated eDNA sampling and sequencing
Operational Taxonomic Unit (OTU)	Group of DNA amplicons clustered around a sequence
Sequencing Depth	Total number of sequences in a sample processed by the sequences

Appendix E

Résumé Français

Introduction

1 Projet Next Generation Biomonitoring

L'activité humaine a un impact important sur l'environnement. L'homme modifie l'environnement au niveau local (déforestation ou rejets polluants dans une rivière) et au niveau mondial (changement climatique). Évaluer la façon dont nous modifions notre environnement peut nous aider à mieux comprendre pourquoi ces changements se sont produits et à trouver des moyens de les atténuer ou de les empêcher de se produire. L'évaluation environnementale repose en grande partie sur la bio-surveillance des écosystèmes. Le projet Next Generation Biomonitoring (NGB) (Bohan et al. (2017)) propose d'utiliser le séquençage de nouvelle génération (NGS) et la déduction de la structure écologique pour améliorer la biosurveillance à l'aide de l'ADN environnemental (ADNe). L'ADNe est l'ADN obtenu directement à partir d'échantillons environnementaux (sol, sédiments, eau, etc.) sans aucun signe évident de matériel biologique source (Thomsen and Willerslev (2015)).

L'ADNe offre la possibilité d'obtenir des informations sur des contextes écologiques qui ne peuvent pas être facilement étudiées à l'aide d'autres techniques. L'une des plus importantes de ces contextes est celle des microorganismes. Les microorganismes vivent dans des communautés complexes, composées d'individus présentant de grandes différences taxonomiques, morphologiques et fonctionnelles. La diversité des communautés microbiennes est à l'origine d'un large éventail de processus écologiques impliquant des microbes, de la pathogénicité à la fourniture de services écosystémiques (Ishaq (2017)). Par conséquent, L'étude de la dynamique des communautés microbiennes peut conduire à une meilleure compréhension des processus écologiques, et de leurs liens avec LES activités humaines.

Les communautés microbiennes sont façonnées par leur environnement et les interactions entre leurs membres. Ainsi, une surveillance efficace nécessite la connaissance des différentes interactions qui affectent l'abondance des différents micro-organismes. Pour étudier les interactions, il devient important d'obtenir des informations sur l'abondance microbienne à partir des données d'ADNe, en regroupant ces informations en unités taxonomiques opérationnelles.

Les unités taxonomiques opérationnelles (OTU) peuvent être identifiées à partir de l'ADNe séquencé en utilisant différentes techniques bioinformatiques (Caporaso et al. (2010) ; Schloss et al. (2009)). Une OTU peut être définie comme le regroupement de différentes séquences d'ADNe qui partagent une homologie supérieure à un seuil donné. Callahan et al. (2016), a développé un outil permettant d'obtenir des variantes de séquences d'amplicons (ASV), qui sont une version à plus haute résolution des OTU. Les comptes d'ASV/OTU sont normalement organisés dans une matrice où chaque échantillon d'ADNe est une colonne, chaque ASV/OTU est une ligne et chaque cellule contient le compte du nombre de séquences de chaque échantillon appartenant à chaque ASV/OTU (Dohlman and Shen (2019)). Le nombre de séquences appartenant à la même ASV/OTU peut être utilisé comme une mesure d'abondance (Schloss et al. (2009)). Dans le section suivante, il est décrit le processus suivi pour obtenir les tableaux ASV qui seront utilisés dans les autres parties.

L'ASV/OTU peut être attribué à un taxon en utilisant des bases de données de référence par un processus appelé métabarcoding. Il est donc possible de déterminer les espèces qui font partie de la communauté présente dans un échantillon d'ADNe et leur abondance. Les communautés évoluent cependant en permanence, façonnées par les interactions de leurs membres avec l'environnement et les milieux (interactions abiotiques), et les autres membres de la communauté (interactions biotiques) (Konopka (2009)). Par conséquent, pour suivre les communautés à l'aide du métabarcoding, il est important de prendre en compte non seulement tous les facteurs abiotiques qui affectent la structure de la communauté, comme la température, l'humidité, la composition du milieu, mais aussi les différentes interactions entre les organismes présents.

2 Inférence de réseaux d'interactions

Étant donné que tous les membres d'une communauté partagent le même espace, leurs interactions ne peuvent être comprises de manière isolée. Des interactions différentes ont des conséquences différentes pour un organisme donné et il est important de connaître toutes les interactions entre les membres de la communauté pour prédire les effets à l'échelle de la communauté. Les réseaux d'interaction ont été utilisés pour visualiser toutes les interactions dans un contexte écologique donné. Ces réseaux présentent tous les membres des communautés comme les nœuds du réseau et les différentes interactions comme les liens reliant ces nœuds. Les réseaux ont été utilisés pour comprendre les relations plantes-pollinisateurs (Lopezaraiza-Mikel et al. (2007)), les communautés de poissons (Ushio et al. (2018)) et les micro-organismes (Nicolaisen et al. (2014)).

Les réseaux d'interaction peuvent être déduits à partir des informations contenues dans une matrice d'OTU. Ce processus n'est cependant pas simple. L'un des principaux problèmes que nous devons résoudre pour l'inférence de réseaux à partir d'une matrice d'OTU est que les données de séquençage sont compositionnelles (Gloor et al. (2017)). Les séquenceurs ne sont capables de traiter qu'un nombre donné de séquences d'ADN et, par conséquent, le nombre d'OTU est relatif au nombre maximal de séquences traitées. Un deuxième problème est celui de la rareté des données. Les OTUs ne sont pas toujours présents dans tous les échantillons, ce qui conduit à des ensembles de données gonflés à zéro. De nombreux outils statistiques, comme SparCC (Friedman and Alm (2012)), SPIEC-EASI (Kurtz et al. (2015)), CCLasso (Fang et al. (2015)) et PLN (Chiquet et al. (2019)), ont été développés pour inférer des réseaux à partir d'une matrice d'OTU. Ces outils sont capables de traiter des données éparses et compositionnelles en transformant les données, et la plupart d'entre eux ont upeuvent être utilisés sur R (Dohlman and Shen (2019)). Cependant, il n'y a pas d'accord sur le nombre d'échantillons requis pour construire un seul réseau à l'aide d'outils d'inférence de réseau (Berry and Widder (2014); Hirano and Takemoto (2019)). La variation de la quantité d'informations nécessaires à la construction d'un réseau entraîne des difficultés pour obtenir des réseaux reproductibles à partir d'un même contexte écologique. De plus, l'importante variabilité entre les réseaux d'interaction augmente augmente le nombre d'échantillons nécessaires pour évaluer les différences entre les réseaux. Il faut donc évaluer la reproductibilité des outils actuels d'inférence de réseau et mettre en œuvre des méthodes statistiques pour obtenir des mesures de réseau fiables. Nous avons abordé ces questions en utilisant différentes mesures de réseau pour étudier les propriétés des réseaux d'interaction produits par deux outils largement utilisés, SparCC et SPIEC EASI.

3 Explainable Machine Learning (EML)

Les réseaux basés sur la corrélation présentent également des problèmes non résolus pour relier les liens du réseau aux interactions écologiques (Röttjers and Faust (2018)). Pour faire face à ces problèmes, une alternative aux méthodes basées sur la corrélation est l'apprentissage automatique explicable (EML). Nous avons d'abord choisi d'utiliser Progol (Muggleton (1995)), une implémentation du langage de programmation logique abductive/inductive (A/ILP) pour développer un outil d'inférence de réseau basé sur l'EML capable d'identifier les interactions écologiques à partir de données de séquençage. A/ILP nous permet de décrire les interactions écologiques en utilisant des règles basées sur la logique symbolique (Tshikantwa et al. (2018)). Nous pouvons alors prédire, de manière explicite, les réseaux d'interactions. Au cours du développement de la thèse, une nouvelle implémentation de A/ILP appelée PyGol a permis de simplifier l'utilisation de EML pour réaliser de nombreux processus, réduisant ainsi le temps d'exécution (Varghese et al. (2022)).

L'ILP a été utilisé dans de nombreux domaines de connaissance, notamment l'inférence de réseaux métaboliques (Tamaddoni-Nezhad et al. (2006)), l'élaboration des processus impliqués dans la production de lait de vache (Sasaki et al. (2019)), et l'inférence de relations trophiques à partir de données d'observation d'arthropodes (Bohan et al. (2011)). Dans le cas de la reconstruction de réseaux d'interactions microbiennes, nous proposons d'utiliser une application de l'ILP appelée Abductive ILP (A/ILP). Le processus d'apprentissage abductif consiste à utiliser les informations contenues dans un ensemble d'observations pour étendre une théorie, via des hypothèses inférées, afin d'enrichir

les connaissances qui existent dans un domaine scientifique. L'abduction génère des hypothèses qui compriment un ensemble d'observations expérimentales. La quantité d'informations comprimées peut être comprise comme une mesure de la probabilité d'une hypothèse. La programmation logique abductive (Kakas and Papadopoulos (1996)) est généralement appliquée à des problèmes qui peuvent être séparés en deux ensembles disjoints de prédicats : les prédicats observables et les prédicats abductibles. En pratique, les prédicats observables décrivent les observations empiriques que nous essayons de modéliser, telles que les informations sur l'abondance des espèces. Les prédicats abductibles - ici les interactions que nous déduisons - décrivent les relations sous-jacentes de notre modèle qui ne sont pas observables directement mais qui peuvent, grâce à la théorie, apporter des informations observables. Nous pouvons également avoir des prédicats d'arrière-plan (connaissances préalables), qui sont des relations auxiliaires nous aidant à relier les informations observables et abductibles (Tamaddoni-Nezhad et al. (2021)). Il n'existe pas de norme d'or pour tester les outils d'inférence de réseaux d'interactions Röttjers and Faust (2018). Habituellement, la validation est effectuée en combinant des ensembles de données générées par ordinateur où les interactions sont simulées Weiss et al. (2016) avec l'évaluation des propriétés générales des ensembles de données NGS réels où il n'est pas possible de connaître les interactions existantes Chiquet et al. (2019). Par conséquent, il est nécessaire de valider l'inférence de réseau basée sur A/ILP en utilisant à la fois des données synthétiques et des données NGS. Ensuite, les données écologiques NGS produites par le métabarcodage peuvent être utilisées pour obtenir des informations écologiques pertinentes.

Construire des ensembles de données sur les communautés microbiennes à partir de données ADN

4 Introduction

L'inférence des réseaux d'interaction nécessite des informations précises sur la composition des communautés microbiennes. Pour obtenir ces informations, il est d'abord nécessaire d'identifier le contexte écologique des communautés microbiennes étudiées et de concevoir un échantillonnage approprié pour l'ADNe. Pour éviter d'introduire des biais dans l'inférence du réseau, il est important d'identifier les conditions abiotiques des échantillons et d'obtenir suffisamment d'échantillons partageant ces conditions. Une fois l'échantillonnage effectué, l'ADNe doit être extrait du matériel échantillonné, amplifié à l'aide d'amorces appropriées au groupe taxonomique étudié et séquencé pour obtenir les séquences d'ADN de tous les organismes présents. Ensuite, un processus bioinformatique est mis en œuvre pour obtenir les tableaux ASV mentionnés dans le introduction générale. Dans cette partie est détaillé le processus d'obtention d'informations sur les communautés microbiennes situées dans les feuilles de vignes provenant de vignobles de différentes régions françaises. Nous émettons l'hypothèse que la structure de la communauté microbienne peut conférer une résistance aux microbes pathogènes. Pour l'étude de cas de la vigne, des tissus sains et des tissus infectés par le pathogène *Plasmopara viticola* ont été collectés pour évaluer le rôle des membres du réseau microbien dans la pathogénèse.

5 Méthodes

5.1 Échantillonnage

Des échantillons ont été collectés en juin et juillet 2018 dans trois vignobles de trois régions viticoles françaises : Aquitaine (AQ), Champagne (CH) et Occitanie (OC). Les échantillons ont été prélevés pendant les épidémies de mildiou (*Plasmopara viticola*) sur 30 vignes dans chaque vignoble, sur des rangs non traités contre le pathogène. Une feuille sporulée a été prélevée sur chaque vigne à l'aide de gants stériles et placée dans un sac plastique individuel. Les feuilles ont été traitées le jour de la collecte avec des outils stérilisés dans le champ stérile d'un brûleur électrique MICROBIO (MSEI, France). Deux disques foliaires symptomatiques de 12 mm de diamètre ont été prélevés sur des lésions de mildiou sporulant sur chaque feuille. Deux disques, considérés comme asymptomatiques, ont également été prélevés sur chaque feuille. Les disques ont été placés individuellement dans des tubes de collecte de 2ml autoclavés et conservés dans une boîte remplie de silicagel. Les bouchons à vis des tubes ont été laissés desserrés pour permettre aux disques de sécher. Tous les échantillons ont ensuite été lyophilisés.

5.2 Amplification ITS fongique

L'ADN total a été extrait avec le kit DNeasy Plant Mini (Qiagen, France), avec une version légèrement modifiée du protocole recommandé par Kerdraon et al. (2019). La région ITS1 du gène ITS rDNA fongique (Schoch et al. (2012)) a été amplifiée à l'aide des amorces ITS1F-ITS2 (White et al. (1990), Gardes and Bruns (1993)). Pour éviter un protocole PCR en deux étapes, chaque amorce contenait la séquence adaptatrice d'Illumina et une étiquette (ITS1F : 5′ - CAAGCAGAAGACG GCATACGAGATGTGACTGGAGTTCAGGTGCTTCCGATxxxxxxxxCTTGGTCATTTAGAG GAAGTAA - 3′ ; ITS2 : 5′ - AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCTxxxxxxxxGCTGCGTTCTTCATCGATGC - 3′, où "x" est l'étiquette de 12 nucléotides). Deux espèces de champignons marins (*Candida oceani* et *Yamadazyma barbieri*) ont été utilisées comme témoins positifs car il était peu probable de les trouver dans nos échantillons.

Un contrôle positif comprenait 1 µL de 10 ng/µL d'ADN de *Candida oceani*, seul, et l'autre comprenait un mélange équimolaire des deux souches. Les contrôles PCR négatifs étaient représentés par un mélange PCR sans matrice d'ADN. Chaque plaque PCR contenait un contrôle d'extraction négatif, trois contrôles PCR négatifs, un contrôle positif à une seule souche et un contrôle positif à deux souches.

5.3 Sequencing

Séquençage MiSeq - La purification des produits PCR (CleanPCR, MokaScience), le séquençage des librairies sur une plateforme Illumina MiSeq (chimie v2, 2x250 bp) et le démultiplexage des séquences (avec recherche d'index exact) ont été réalisés au centre de séquençage PGTB (Pierroton, France). Les amplicons ITS1 fongiques ont été séquencés sur trois passages et les amplicons 16S bactériens ont été séquencés sur quatre passages.

5.4 Traitement bioinformatique

Les séquences MiSeq produites ont été traitées à l'aide du pipeline DADA2 version 1.16.0 (Callahan et al. (2016)), implémenté sur R. Les amorces ont été identifiées et supprimées à l'aide de cutadapt 2.8 (Martin (2011)) et les séquences découpées ont ensuite été analysées par l'algorithme DADA2. DADA2 produit un tableau des variantes de séquences d'amplicons (ASV), un analogue à plus haute résolution du tableau des OTU couramment utilisé, qui enregistre le nombre de fois où chaque variante de séquence d'amplicon a été observée dans un échantillon, dans tous les échantillons. Les chimères ont été supprimées à l'aide de la fonctionnalité removeBimeraDenovo de DADA2. Les séquences identiques qui avaient des longueurs différentes ont été jointes. Pour filtrer les éventuels contaminants (séquences d'ADN qui sont ajoutées à l'échantillon pendant le processus d'extraction et de séquençage) et les séquences (ASV) de faible abondance, un pipeline personnalisé a été développé. Tout d'abord, la fonction isContaminant du paquet Bioconductor DECONTAM v 1.8.0(Davis et al. (2018)) a été utilisée pour éliminer les séquences contaminantes. Ensuite, tous les contaminants restants ont été supprimés en utilisant les contrôles positifs et négatifs, comme décrit dans Galan et al. (2016). Les séquences identifiées comme chloroplastiques ou mitochondriales ont été supprimées à l'aide de Metaxa2.2 (Bengtsson-Palme et al. (2015)). Les séquences ASV avec une affectation taxonomique " mitochondrie " ont également été supprimées. Une fois que toutes les ASV contaminantes ont été supprimées, les séquences restantes ont été regroupées au niveau d'identité de séquence de 99 %. Dans une dernière étape, les ASV présents dans moins de 1% des échantillons ont été supprimés.

5.5 Attribution taxonomique

Une assignation taxonomique a été effectuée en utilisant une implémentation d'un classificateur bayésien naïf inclus dans le pipeline DADA2. Les bases de données utilisées pour l'affectation taxonomique étaient les bases de données Silva v132 (Quast et al. (2012)) et UNITE v8.1 (Nilsson et al. (2019)) pour les séquences 16S et ITS, respectivement. Les ASV qui n'ont pas pu être assignés à un phylum ont été supprimés. Trois tableaux ont été obtenus à la fin de ce processus : un tableau ASV avec le nombre de séquences dans chaque échantillon ; un tableau avec l'affectation taxonomique de chaque séquence ASV ; et, un tableau de métadonnées décrivant le plan expérimental de chaque échantillon. Les trois tableaux ont été réunis dans un objet phyloseq à l'aide du package bioconducteur phyloseq v1.32.0 (McMurdie and Holmes (2013)).

5.6 Analyse bioinformatique

Les données contenues dans l'objet phyloseq ont été analysées dans R v4.0.3 (R Core Team (2022)) pour évaluer la composition des communautés fongiques et bactériennes et l'effet du plan expérimental sur ces communautés. Les comptages ASV obtenus à l'aide du pipeline dada2 ont été transformés à l'aide d'une transformation clr (Aitchison (1982)). Le paquet phyloseq a été utilisé pour obtenir

la distance euclidienne entre les échantillons et effectuer une analyse des coordonnées principales (PCoA) dans R. La PCoA a été tracée à l'aide du paquet ggplot v3.3.2 (Wickham (2016)). La composition relative de la communauté a également été représentée à l'aide de ggplot. Une analyse de variance multivariée permutationnelle (permanova), réalisée pour évaluer l'effet du plan expérimental sur les communautés, a été évaluée à l'aide de la fonction adonis2 du paquet R vegan v2.5.6 (Oksanen et al. (2022)).

6 Analyse de la communauté ITS

La composition des communautés fongiques foliaires de la vigne différait significativementen fonction des régions et des conditions tissulaires (Figure 1C). La région géographique était le facteur expérimental le plus significatif, expliquant environ 11 % de la variance de la communauté fongique de la vigne. Au sein de chaque région, le vignoble a également un impact important sur la composition de la communauté, expliquant près de 10 % de la variance de la communauté. Les différentes conditions climatiques (par exemple, la température, l'humidité) propres à chaque vignoble, semblent être des facteurs important dans la structure de la communauté fongique. L'importance de l'état des tissus, bien que significative, était relativement mineure. Les communautés d'Aquitaine et d'Occitanie étaient les plus similaires, tandis que la présence plus importante du genre Mycosphaerella a modifié la forme des communautés microbiennes de Champagne. L'abondance des genres fongiques n'était pas significativement affectée par l'état des tissus symptomatiques ou asymptomatiques (Figure 1A).

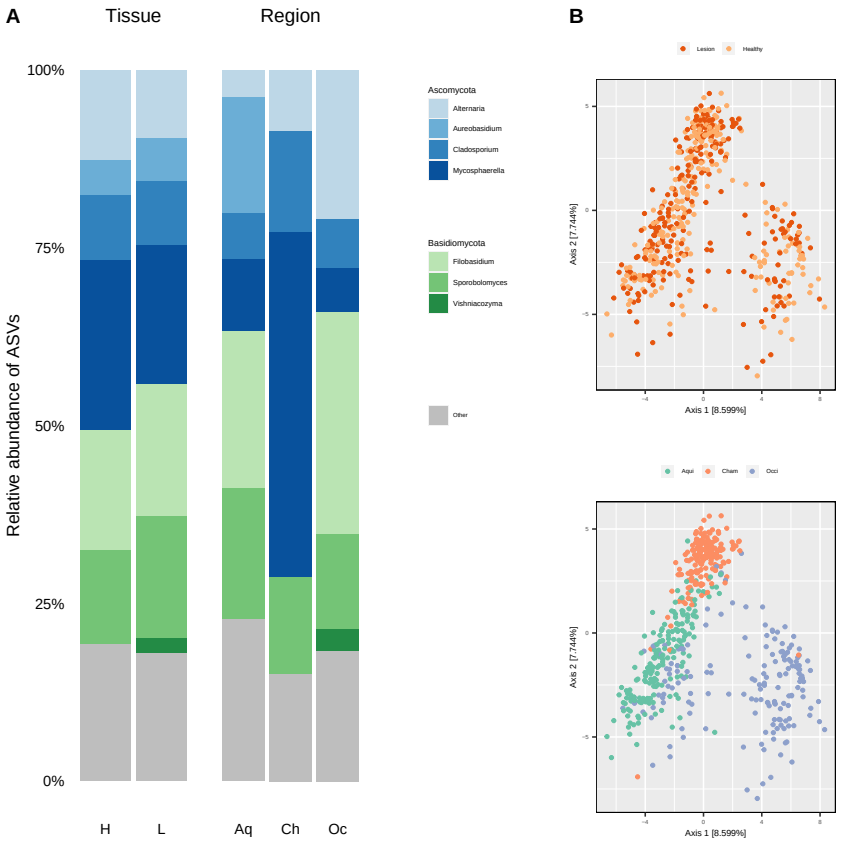


Figure 1: **A:** Composition de la communauté fongique au niveau du genre. **B:** Ordination des échantillons en utilisant la distance d'Aitchinson. Les échantillons sont colorés en fonction de l'état du tissu et de la région.

Approches fondées sur la corrélation pour la biosurveillance à l'aide de l'ADN

7 Introduction

L'un des objectifs du projet Next-Generation Biomonitoring est d'utiliser les données de séquençage pour reconstruire les réseaux d'interaction écologique en temps réel, puis de calculer les propriétés au niveau du réseau pour évaluer les changements dans l'écosystème. Pour mettre en œuvre l'utilisation des réseaux d'interaction écologique, il est nécessaire de compter sur des méthodes fiables pour déduire ces réseaux. Dans ce partie, deux méthodes d'inférence de réseaux largement utilisées, SparCC (Friedman and Alm (2012)) et SPIEC-EASI (Kurtz et al. (2015)) sont utilisées pour inférer des réseaux à partir de données écologiques. Nous évaluons ensuite : (1) la reproductibilité des réseaux basés sur l'ADN en l'absence de changement dans l'écosystème ; et (2) les avantages et les inconvénients des propriétés au niveau de la communauté et du réseau pour le suivi du changement. Nous avons choisi un réseau microbien associé aux cultures comme étude de cas. Les réseaux sont supposés soutenir les services de régulation des maladies dans les agroécosystèmes. Nous avons analysé leur réponse au changement de pratique agricole entre les systèmes biologiques et conventionnels.

8 Matériels et méthodes

8.1 Échantillonnage et séquençage

Des échantillons de feuilles de vigne ont été prélevés le 10 septembre 2015 dans un vignoble expérimental. Le vignoble expérimental a été planté en 2011 pour comparer deux systèmes de culture : l'agriculture conventionnelle durable (CONV) et l'agriculture biologique (ORGA) (Delière et al. (2015)). Les systèmes de culture différaient par les types de pesticides appliqués et la période d'application. Des feuilles de vigne ont été collectées dans 3 parcelles CONV et 3 parcelles ORGA avec 19 vignes par parcelle. L'extraction de l'ADN a été effectuée selon le protocole CTAB chloroforme/alcool isoamylique (24:1). La région de l'espaceur transcrit interne (ITS) du ribosome nucléaire, qui est considérée comme la région universelle du code-barres des champignons, a été amplifiée à l'aide d'amorces spécifiques. Le séquençage a eu lieu sur une plateforme Illumina Miseq.

8.2 Analyse bioinformatique

Les séquences brutes ont été traitées à l'aide du pipeline DADA2 (Callahan et al. (2016) comme recommandé dans (Pauvert et al. (2019)). Le pipeline DADA2 identifie les différents Amplicon Sequence Variants (ASV), une version à plus haute résolution des OTU. Le tableau des ASV obtenu à l'aide du pipeline DADA2 a été filtré selon la description de (Galan et al. (2016)), en supprimant les ASV inférieurs à la limite de contamination détectée avec les contrôles d'extraction. Avant l'inférence des réseaux, les différentes propriétés des communautés ont été évaluées à l'aide du package R vegan et testées à l'aide du paquet R lme4. Les réseaux ont été construits à l'aide de l'algorithme SparCC (Friedman and Alm (2012)) mis en œuvre dans FastSpar (Watts et al. (2018)) avec les valeurs par défaut de SparCC, et de la méthode SPIEC-EASI (Kurtz et al. (2015)) exécutée dans l'environnement R en utilisant la procédure MB pour la sélection des bords. Pour chaque méthode d'inférence de réseau, dix réseaux par parcelle ont été construits en faisant varier le pourcentage P des ASV inclus dans le réseau (avec P allant de 10% à 100% des ASV les plus abondants dans la parcelle). Nous avons fait varier P car nous nous attendions à ce qu'il influence la reproductibilité des réseaux. Nous nous attendions, en particulier, à ce que les réseaux construits uniquement à partir des ASV les plus abondants présentent une meilleure reproductibilité. 60 réseaux ont été obtenus, 3 parcelles x 2 traitements x 10 valeurs de P. Les propriétés α et β au niveau du réseau ont été analysées en utilisant Permanova pour évaluer les différences des métriques entre les systèmes de culture.

9 Résultats

La figure 2A montre l'effet de la valeur P dans les réseaux inférés. La valeur de P a un effet significatif sur toutes les propriétés α. Le système de culture n'a pas montré d'effet sur les réseaux α-propriétés obtenus avec SparCC, comme on peut le voir sur la figure 2B. En revanche, le système de culture a eu un effet significatif sur le nombre d'associations entre les ASV partagés (βOS) pour les réseaux SparCC. Ces résultats suggèrent que la variation du réseau entre les systèmes de culture est due à la rotation des associations (saisie par βOS), plutôt qu'à la rotation des ASV, et montrent que les propriétés β du réseau définies dans le cadre de l'écologie théorique de Poisot et al. (2012) peuvent détecter les différences entre les systèmes de culture. Des résultats similaires ont été obtenus pour les réseaux inférés avec SPIEC-EASI. Aucune association n'était partagée entre les deux systèmes de culture indépendamment de l'outil d'inférence de réseau. De plus, seuls quelques liens étaient partagés entre des parcelles appartenant au même système de culture. Bien que ces parcelles partagent plus d'associations que les réseaux aléatoires de même taille, le nombre réduit d'associations partagées démontre les difficultés d'obtenir un réseau reproductible à partir des mêmes conditions abiotiques.

10 Conclusions

1. Les mesures de β-properties au niveau du réseau peuvent détecter les changements produits par le système de culture.

2. α-properties ne détectent pas les changements produits par les systèmes de culture. Ainsi, les différences entre les systèmes de culture sont basées sur le renouvellement des ASV et la réassociation des liens, et non sur les changements des caractéristiques du réseau.

3. Les réseaux microbiens obtenus à l'aide des outils existants d'inférence statistique de réseaux varient de façon marquée entre les répliques d'un même ensemble de conditions environnementales.

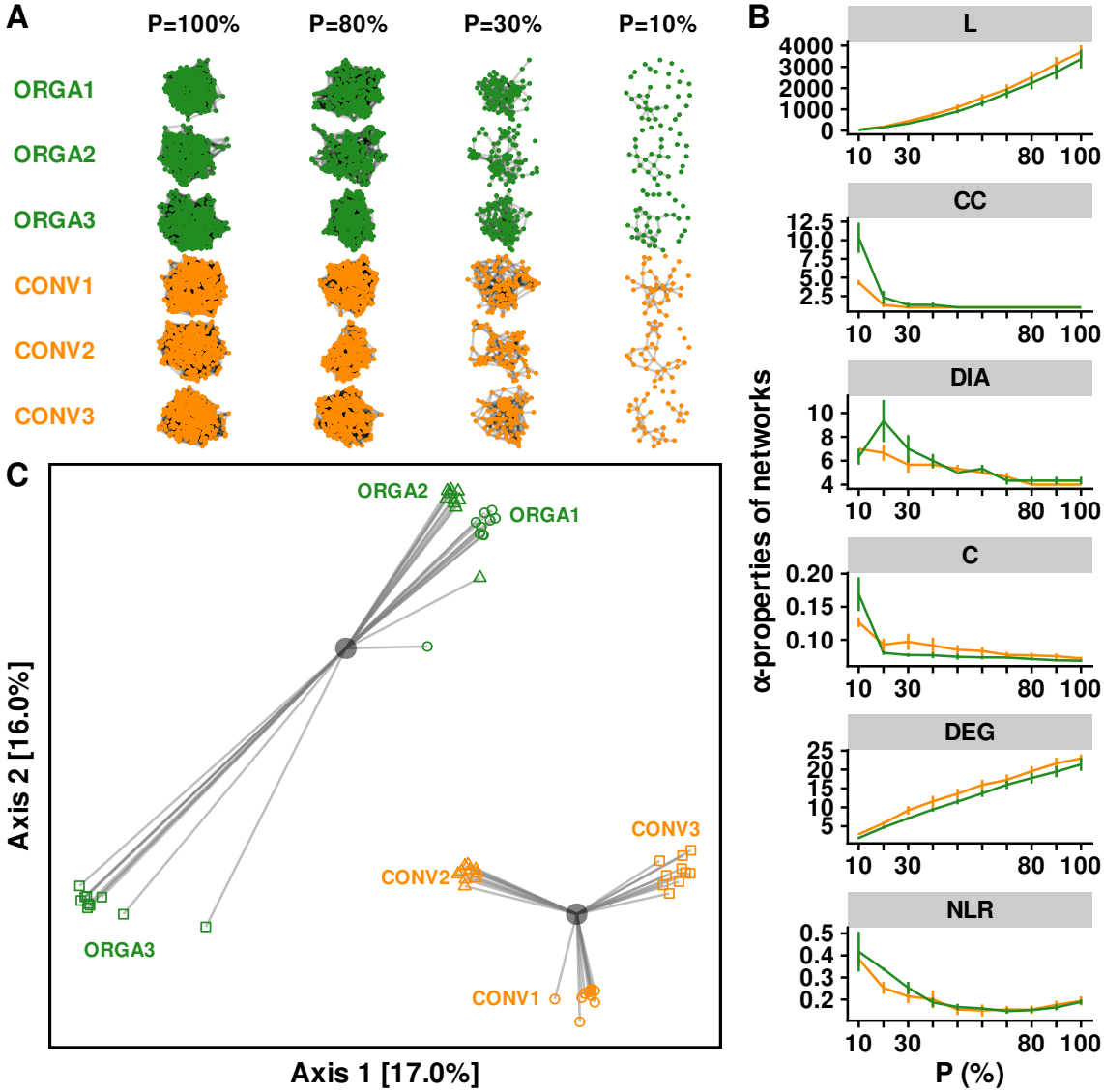


Figure 2: Effet du système de culture - conventionnel (CONV) versus biologique (ORGA) - sur les propriétés α et les propriétés β des réseaux fongiques foliaires de la vigne. (**A**) Réseaux d'association déduits des données de métabarcodage fongique avec SparCC (Friedman and Alm (2012)). Au total, 60 réseaux ont été inférés, correspondant à 2 systèmes de culture $\times$ 3 réplicats (blocs) $\times$ 10 valeurs de P, P étant le pourcentage des ASV les plus abondants utilisés pour l'inférence des réseaux. Seules quatre valeurs de P sont représentées sur la figure. (**B**) Variations des propriétés *alpha* du réseau. Les propriétés suivantes ont été calculées pour chaque réseau : le nombre de liens (L) et de composants connectés (CC), le diamètre du réseau (DIA) et la connectivité (C) ainsi que le degré moyen (DEG) et le ratio de liens négatifs (NLR). (**C**) L'analyse des coordonnées principales (PCoA) représente les dissimilitudes entre les réseaux, mesurées avec l'indice βOS (Poisot et al. (2012)) calculé avec l'indice binaire de Jaccard.

Utilisation d'une approche basée sur la logique pour déduire des interactions à partir de données d'ADN simulées

11 Introduction

Comme nous l'avons montré dans le partie précédente, les méthodes statistiques existantes pour déduire les réseaux d'interactions microbiennes peuvent être utilisées pour explorer les interactions possibles qui peuvent être déduites d'un tableau ASV. Cependant, il est difficile d'établir une relation entre les associations suggérées à partir de la corrélation et les véritables interactions microbiennes. Il est donc difficile de fournir une interprétation écologique correcte à partir des réseaux d'interaction déduits. A/ILP pourrait potentiellement résoudre ce problème. Les connaissances écologiques existantes peuvent être utilisées par A/ILP pour déduire, directement, les interactions écologiques, contournant ainsi la nécessité de l'étape d'interprétation. Dans ce partie, nous décrivons la mise en œuvre de l'inférence d'interaction basée sur le PIILA et la méthodologie utilisée pour l'évaluer.

Tester la performance des réseaux d'inférence écologique en utilisant des données réelles de communautés microbiennes est difficile car il y a peu ou pas d'informations sur la majorité des interactions. Par conséquent, les performances de l'inférence A/ILP sont évaluées dans cette partie à l'aide d'une méthodologie *in silico* proposée par Weiss et al. (2016). La méthodologie Weiss et al. (2016) simule des tableaux ASV répliqués de type écologique en utilisant différents mécanismes écologiques d'interaction. Ensuite, A/ILP est utilisé pour évaluer la présence des interactions simulées, comme un ensemble connu d'attentes. Ce calcul de test produit une statistique de test qui est utilisée pour déterminer la méthodologie la plus sensible ; et, évaluer la signification probabiliste de la statistique pour la discrimination des interactions.

12 Méthodes

12.1 Inférence d'interaction A/ILP

Le processus d'inférence de l'interaction A/ILP est divisé en cinq étapes différentes :

1. **Transformation des données:** Les informations d'abondance contenues dans la matrice ASV sont transformées en clauses logiques. Les clauses logiques sont basées sur une variation de l'abondance entre les échantillons, calculée à l'aide d'un test du Khi-deux.

2. **Description logique de l'effet d'abondance:** La description des interactions, comprise comme un effet sur l'abondance de l'espèce en interaction, est exprimée à l'aide de clauses logiques. Une espèce peut provoquer deux effets sur l'abondance d'autres espèces, une augmentation (vers le haut) ou une diminution (vers le bas).

3. **Abduction des interactions:** Les descriptions logiques des effets dans l'abondance et les clauses logiques décrivant les changements d'abondance sont ensuite utilisées dans le programme A/ILP, Progol, pour déduire les effets que les espèces ont sur les autres espèces. Ce processus, appelé abduction, renvoie une valeur de compression. La compression est une mesure numérique de la quantité d'information qui soutient chaque effet.

4. **Calcul de l'estimateur I:** Le processus d'abduction dépend fortement de l'ordre dans lequel les clauses logiques sont fournies. Par conséquent, il est nécessaire de répéter le processus d'abduction plusieurs fois en permutant l'ordre des clauses logiques. Ensuite, toutes les valeurs de compression permutées sont mises en commun, et différentes méthodologies sont employées pour obtenir une valeur d'estimateur final pour une interaction I. Les performances de la statistique I calculée à l'aide des différentes méthodologies sont testées dans une évaluation expérimentale.

5. **Bootstrapping de l'estimateur de I:** Les valeurs de compression obtenues à partir de différentes permutations sont rééchantillonnées dans un processus de bootstrapping pour obtenir la signification statistique des différents estimateurs de I.

12.2 Ensembles de données simulées

Weiss et al. (2016) a proposé un modèle de simulation pour créer des tableaux générés par ordinateur incluant les effets des interactions linéaires de type écologique. Le modèle utilise la distribution log-normale pour simuler l'abondance des espèces sans interaction dans un ensemble de communautés microbiennes ou d'échantillons. Les interactions sont ensuite introduites en modifiant l'abondance des espèces selon les différents mécanismes d'interaction écologique suivant Faust and Raes (2012). En utilisant la méthode proposée dans Weiss et al. (2016), nous avons généré trois tableaux contenant les abondances de 16 paires d'espèces en interaction dans 50 communautés. Les tableaux ont été simulés en utilisant des interactions de différentes valeurs de force (2, 3 et 5), et quatre mécanismes d'interaction écologique différents : amensalisme, commensalisme, compétition et mutualisme Derocles et al. (2018). Comme un séquenceur ne peut lire qu'un nombre limité de séquences dans un échantillon, et que celles-ci sont partagées entre les espèces, l'imposition d'un biais de composition dans les données (Gloor et al. (2017)). Pour générer des tableaux microbiens de type écologique, il est donc nécessaire de réintroduire la compositionnalité dans les ensembles de données simulées. Pour ce faire, nous avons normalisé la profondeur de séquençage sous forme de probabilités dans une distribution multinomiale, puis nous avons rééchantillonné la distribution pour obtenir les comptes simulés à travers une profondeur de séquençage commune.

13 Évaluation expérimentale

Les performances de l'inférence microbienne basée sur A/ILP sont évaluées à l'aide des ensembles de données générés par ordinateur. Tout d'abord, nous testons un certain nombre d'échantillons de l'espace des hypothèses et différentes fonctions pour obtenir la statistique I. Ensuite, le meilleur paramètre est utilisé pour évaluer les performances de la procédure de bootstrapping par rapport à un seuil pour I et SparCC.

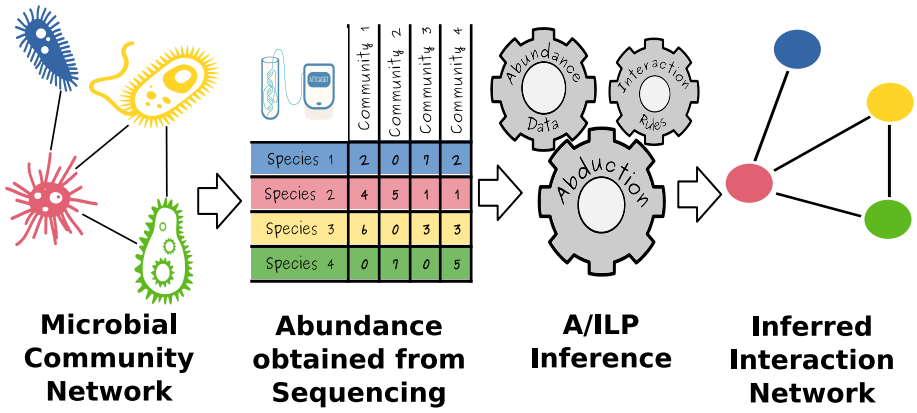


Figure 3: Description du processus d'inférence des interactions. Les communautés microbiennes sont façonnées par l'interaction entre leurs membres. Le séquençage de l'ADN et les processus bioinformatiques permettent d'estimer l'abondance des différents microbes présents dans les communautés. En utilisant les informations sur l'abondance des différentes communautés comme exemples d'apprentissage, et les règles d'interaction comme connaissances de base, il est possible d'inférer un réseau d'interaction qui généralise les interactions entre les microbes.

13.1 Expérience 1

Hypothèse nulle 1:

L'utilisation de l'estimateur I tel que défini dans (12.1) à l'aide de différentes fonctions n'entraîne pas une précision supérieure à celle de l'approche basée sur la fréquence Hypothesis Frequency Estimator (Tamaddoni-Nezhad et al. (2013)) pour la prédiction des interactions microbiennes.

Matériels et méthodes:

Tableaux générés par ordinateur, comme décrit dans la section 2.1, calculés selon la méthodologie de Weiss et al. (2016), 100 abductions d'effets possibles sont effectuées pour chaque tableau. Les observations produites à partir des tableaux sont permutées aléatoirement à chaque exécution. La description logique de l'effet est utilisée comme connaissance de base. Ensuite, les estimateurs sont obtenus en utilisant les différentes fonctions décrites précédemment.

Puisque les interactions qui déterminent les abondances des tableaux générés par l'ordinateur sont connues, il est possible de traiter l'inférence des interactions comme un problème de classification. Les interactions peuvent être classées comme existantes ou non existantes et les valeurs des estimateurs obtenues en utilisant les différentes fonctions indiquent la précision de la classification. Ainsi, l'aire sous la courbe (AUC) du taux de vrais positifs par rapport au taux de faux positifs (courbe ROC) peut être utilisée comme mesure de la performance. L'AUC est calculée pour toutes les fonctions aux permutations = 1, 5, 10, 25 et 50 enlèvements. Un test ANOVA est effectué, ainsi qu'un test d'amplitude de Tuckey, pour évaluer la significativité des différences de valeurs de l'AUC entre toutes les fonctions.

Résultats et discussion:

Les valeurs de l'AUC pour les différentes méthodologies d'obtention des estimateurs et le nombre de permutations sont présentées dans la figure 4. Comme prévu, les valeurs de l'AUC augmentent à mesure que le nombre de permutations utilisées pour l'inférence augmente. Elles se stabilisent à environ n = 50 permutations. Les valeurs de l'AUC sont similaires lorsque la force d'interaction simulée est faible, étant significativement plus faibles à la force d'interaction simulée la plus élevée. Cela peut s'expliquer par la faible performance du modèle logique pour le cas spécifique d'une interaction réduisant à 0 l'abondance d'une espèce donnée, dont la probabilité augmente avec la force des interactions. Ce processus écologique, appelé exclusion, réduit fortement la cooccurrence entre espèces et, par conséquent, l'information disponible pour inférer une interaction. La compression maximale utilisée pour obtenir I est la métrique qui donne les valeurs les plus élevées de l'AUC, pour un nombre donné de permutations et de forces d'interaction. Les permutations HFE, Sum et indépendantes ont des valeurs d'AUC similaires aux forces 2 et 5, tandis que la méthode de permutation indépendante est plus performante à la force 3. L'ANOVA montre que toutes les fonctions ont des valeurs d'AUC significativement différentes, à l'exception des permutations indépendantes et des permutations indépendantes.

13.2 Expérience 2

Matériel et méthodes:

La procédure de bootstrapping est réalisée à l'aide des trois tableaux générés par ordinateur utilisés dans l'expérience précédente. Comme expliqué dans la section 3.2, la procédure utilise la compression maximale pour obtenir l'estimateur I. Deux techniques différentes de bootstrapping sont évaluées : ordinaire et par strates. Le bootstrapping ordinaire effectue le bootstrapping indépendamment sur toutes les valeurs de compression tandis que la méthode strata contraint le bootstrapping aux valeurs de compression par effet. Les interactions avec une valeur $p < 0.05$ sont considérées comme existantes. La précision du bootstrapping est comparée à la précision de la prédiction en utilisant un seuil optimal pour l'estimateur I. La métrique du seuil optimal est obtenue automatiquement à partir des courbes ROC du test précédent en utilisant la méthode du meilleur seuil du package R

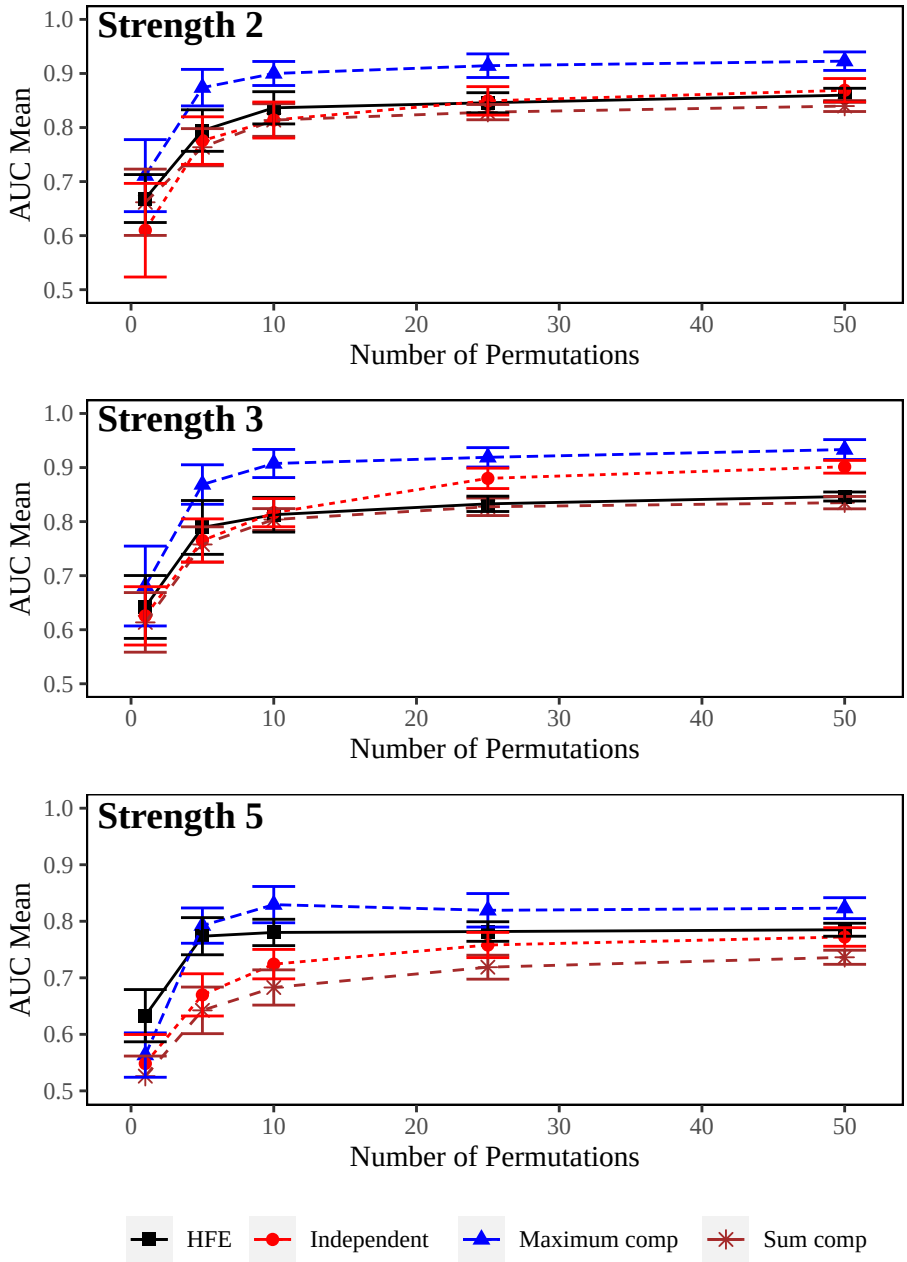


Figure 4: Aire sous la courbe ROC (AUC) obtenue en utilisant différents nombres de permutations. Chaque graphique montre les AUC obtenues en déduisant des interactions de différentes forces. Chaque ligne représente une méthode utilisée pour obtenir les estimateurs. Les barres d'erreur indiquent l'écart type des moyennes.

pROC, selon Robin et al. (2011). Afin de disposer d'une référence pour comparer la performance de l'inférence des interactions par A/ILP, l'inférence des interactions a également été réalisée à l'aide de SparCC Friedman and Alm (2012), un outil d'inférence statistique largement utilisé. Le processus a été réalisé en utilisant la mise en œuvre FastSpar 1.0 (Watts et al. (2018)) avec les paramètres par défaut.

Résultats et discussion:

Les mesures de précision sont présentées dans le tableau 1. Le bootstrap ordinaire présente une meilleure précision que le bootstrap par strates à la force d'interaction = 2 et 3, tandis que les strates sont plus performantes à la force = 5. Cependant, le bootstrap ordinaire permet la détection d'un plus grand nombre de vrais positifs contrairement au strata. Nous pensons donc que c'est la meilleure option à utiliser pour la détection des interactions. Dans tous les cas, la précision du bootstrap est supérieure à la précision du seuil optimal. Les valeurs de sensibilité "bootstrapped" A/ILP sont significativement inférieures à celles de SparCC à toutes les forces. Cependant, les valeurs de spécificité sont légèrement supérieures. Ainsi, SparCC présente un grand nombre de faux positifs, tandis que A/ILP génère un plus grand nombre de faux négatifs. Cela produit des mesures de précision similaires pour SparCC et A/ILP bootstrapped, indépendamment de la force de l'interaction. Nous rejetons donc l'hypothèse nulle.

14 Conclusions

1. Notre travail montre que A/ILP peut être utilisé pour déduire avec précision des interactions de type écologique à partir de jeux de données générés par ordinateur.
2. La compression maximale est la meilleure méthodologie pour calculer les statistiques I à partir des valeurs de compression.
3. La procédure de bootstrapping permet de discriminer les liens avec une précision acceptable.

Table 1: Performance de l'estimateur bootstrapping comparé au seuil optimal obtenu à partir de la courbe ROC et de SparCC. Les trois ensembles de données utilisés pour l'inférence d'interaction ont 16 interactions réelles sur 496 interactions possibles.

Strength 2			
	Optimal threshold	Ordinary Bootstrap	SparCC
Total	40	13	26
TP	13	9	12
FP	27	4	14
TN	453	476	466
FN	3	7	4
Sensitivity	0.812	0.5620.75	
Specificity	0.944	0.992	0.971
Accuracy	0.94	0.978	0.964
Strength 3			
	Optimal threshold	Ordinary Bootstrap	SparCC
Total	69	7	31
TP	14	6	11
FP	55	10	20
TN	425	479	460
FN	2	10	5
Sensitivity	0.875	0.375	0.688
Specificity	0.885	0.998	0.958
Accuracy	0.885	0.978	0.95
Strength 5			
	Optimal threshold	Ordinary Bootstrap	SparCC
Total	50	27	40
TP	12	10	13
FP	38	17	27
TN	442	463	453
FN	4	6	3
Sensitivity	0.75	0.625	0.812
Specificity	0.921	0.965	0.944
Accuracy	0.915	0.954	0.94

Inférence logique des interactions à partir de données environnementales sur l'ADN

14.1 La toile noire des communautés microbiennes

La diversité fonctionnelle des communautés microbiennes résulte de la combinaison du grand nombre d'espèces et des nombreux types d'interaction, tels que la compétition, le mutualisme, la prédation ou le parasitisme, dans les réseaux écologiques microbiens. Comprendre la relation entre les réseaux microbiens et les services et fonctions fournis par les communautés microbiennes est un défi majeur pour l'écologie microbienne, d'autant plus que nombre de ces interactions sont difficiles à observer et à caractériser. Nous pensons que cette "toile noire" d'interactions pourrait être démêlée en utilisant une approche d'apprentissage automatique explicable, appelée programmation logique abductive/inductive (A/ILP) dans le package R InfIntE, qui utilise des règles mécanistes (hypothèses d'interaction) pour déduire directement la structure du réseau et les types d'interaction. Nous tentons ici de démêler la toile noire du microbiome des plantes en utilisant des données de métabarcodage échantillonnées à partir du microbiome foliaire de la vigne. En utilisant des données synthétiques simulées, nous montrons d'abord qu'il est possible de reconstruire de manière satisfaisante les réseaux microbiens en utilisant l'apprentissage automatique explicable. Nous confirmons ensuite que la toile noire du microbiome de la vigne est diversifiée, étant composée d'une gamme de types d'interactions conformes à la littérature. Cette première tentative d'utilisation de l'apprentissage automatique explicable pour déduire les réseaux d'interaction microbienne fait progresser notre compréhension des processus écologiques qui se produisent dans les communautés microbiennes et nous a permis de déduire des types spécifiques d'interaction au sein du microbiome de la vigne qui pourraient être validés par l'expérimentation. Ce travail aura des applications potentiellement précieuses, comme la découverte d'interactions antagonistes qui pourraient être utilisées pour identifier des agents de contrôle biologique potentiels au sein du microbiome.

15 Matériaux et méthodes

15.1 Cadre d'hypothèse pour l'apprentissage des interactions écologiques microbiennes par logique abductive

Les approches explicatives pour déduire les interactions écologiques commencent par une déclaration claire des règles d'une interaction écologique. Les mécanismes des interactions écologiques entre deux OTU, voire plus, peuvent être décrits d'une multitude de façons différentes (Faust and Raes, 2012; Tshikantwa et al., 2018). Nous posons que les faits communs minimums pour toutes les interactions hypothétiques sont les suivants : les deux OTUs subissant une interaction doivent être présentes ensemble dans au moins un échantillon ; et, au moins une des OTUs impliquées dans une interaction subit un changement d'abondance. Ici, l'abondance est comprise comme une mesure de la taille d'une population d'OTUs dans un échantillon et est dérivée du nombre de lectures de séquences d'OTUs trouvées dans chaque échantillon. Ainsi, pour évaluer le changement d'abondance de toutes les OTU, dans tous les échantillons, le nombre de séquences d'une OTU et la profondeur totale des séquences dans deux échantillons prélevés dans les mêmes conditions biotiques et abiotiques, sont utilisés pour construire un tableau de contingence, l'importance du changement d'abondance des OTU entre les échantillons étant évaluée par un test d'indépendance de χ^2. Les changements significatifs sont alors classés soit comme une augmentation, *up*, soit comme une diminution, *down*, par rapport à l'abondance relative de l'OTU, dans les deux échantillons. Symboliquement, cela peut être exprimé par la clause logique $abundance(s1, x, y, up/down)$. Ici, $s1$ est une OTU donnée, (x, y), sont deux échantillons donnés partageant les mêmes conditions et $up/down$ décrit la direction du changement d'abondance significatif. Les changements d'abondance sont calculés de cette manière pour tous les OTUs dans tous les échantillons. Cela permet d'éviter de nombreux biais de composition inhérents au traitement des données de séquences d'ADN sous forme de comptages, puisque seuls les comptages

d'une même OTU sont comparés et que la profondeur totale de la séquence est prise en compte par le test d'indépendance χ^2. La présence, *yes*, ou l'absence, *no*, d'une OTU dans un échantillon, x, peut être exprimée par la clause : $presence(s1, x, yes/no)$.

Le processus de logique abductive utilise ces clauses pour trouver des explications possibles (effets) pour les changements observés dans l'abondance et la présence en utilisant des hypothèses *a priori* pour les interactions écologiques qui reflètent l'état actuel des connaissances écologiques, présentes dans la littérature. Dans ce cas, nous émettons l'hypothèse qu'une interaction aura eu lieu lorsque la présence d'au moins une OTU ($s1$) a produit un effet cohérent sur l'abondance d'une autre OTU ($s2$) dans les échantillons. Les relations logiques pour l'effet d'une telle interaction peuvent être décrites à partir des clauses d'abondance et de présence comme suit :

$$\begin{aligned} effect\_up(s1, s2) \quad \text{if:} \quad & \begin{cases} abundance(x, y, s2, up) \\ presence(s1, x, no) \\ presence(s1, y, yes) \end{cases} \\ effect\_down(s1, s2) \quad \text{if:} \quad & \begin{cases} abundance(x, y, s2, down) \\ presence(s1, x, no) \\ presence(s1, y, yes) \end{cases} \end{aligned} \tag{1}$$

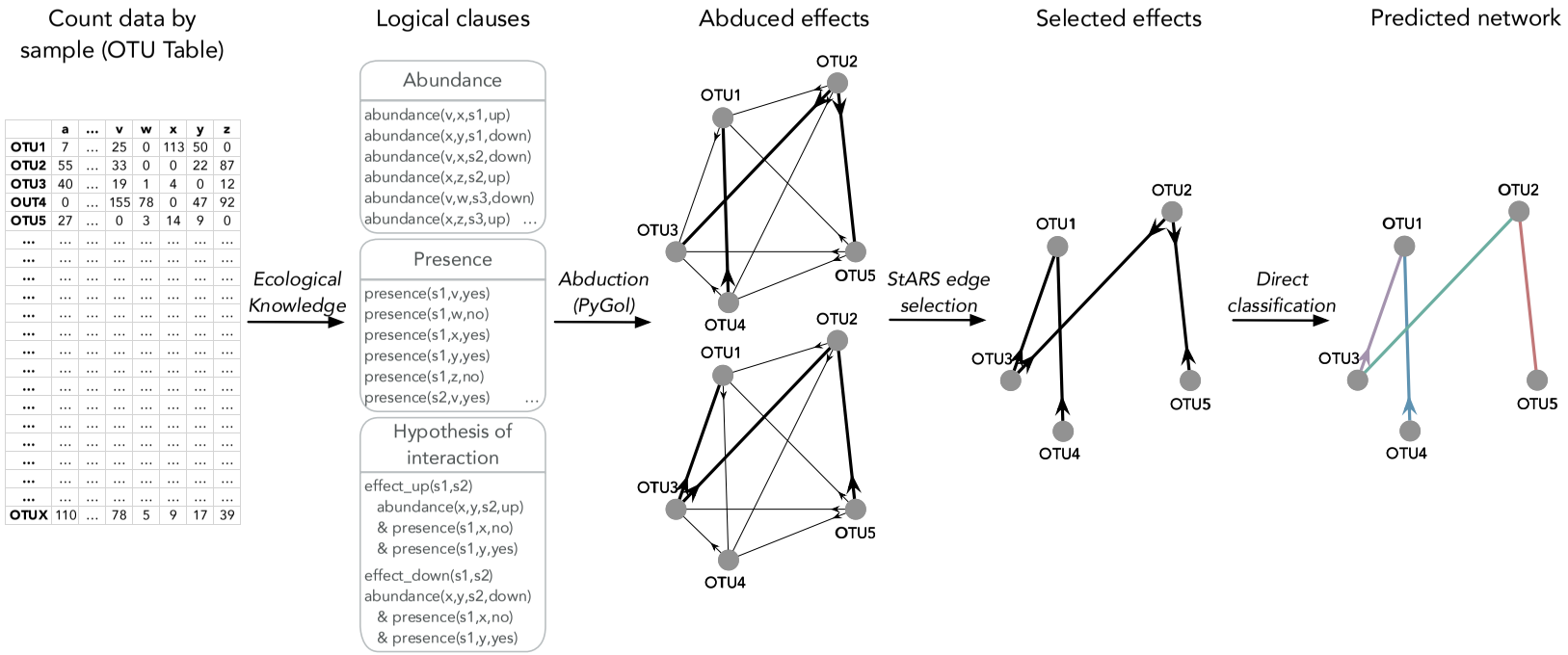
15.2 Détection d'interaction

Pour chaque paire d'OTUs considérée, la valeur de I est traitée comme le poids d'une arête dirigée dans un réseau d'interaction écologique. En fixant un seuil, λ, pour la valeur absolue, I, on sélectionne une liste d'arêtes inférées pour un réseau. Un seuil de $\lambda = 0$ sélectionnerait toutes les arêtes possibles, tandis qu'un seuil de $\lambda = max(I)$ donnerait un réseau vide, sans arêtes sélectionnées. Cependant, $max(I) = nobservations$ et il dépend du nombre d'observations dans un ensemble de données, et il n'est pas possible d'établir une valeur commune de λ pour la reconstruction d'un réseau quelconque. Nous sélectionnons donc les arêtes d'interaction significatives de manière empirique en utilisant une méthodologie de sous-échantillonnage appelée StARS (Liu et al., 2010). La procédure StARS sous-échantillonne 80% des échantillons, plusieurs fois, et effectue l'abduction des bords du réseau. Le réseau d'interactions le plus stable est ensuite identifié en utilisant la fréquence d'apparition des bords à différentes valeurs de λ. Nous utilisons ici 50 rééchantillonnages des données et 50 valeurs de λ augmentant linéairement de 0 à $max(I)$. Le nombre de sous-échantillons et la longueur du chemin lambda sont choisis en suivant les recommandations de Muller et al., 2016 (Müller et al., 2016), avec un seuil de stabilité restrictif de 0,01 de manière à minimiser le nombre d'interactions faussement positives. Tout le processus suivi pour inférer et classer les interactions est implémenté dans un package R appelé Inference of Interactions using Explainable Machine Learning. Il est détaillé dans la figure 5.

16 Résultats

16.1 Expérience 1 : évaluation d'InfIntE avec des données générées par ordinateur

. InfIntE a été utilisé pour inférer des réseaux d'interaction pour chaque tableau d'OTU simulé (comme décrit dans (Weiss et al., 2016)) pour les hypothèses d'interaction. L'aire sous la courbe caractéristique d'exploitation du récepteur (AUC)(Fan et al., 2006) a ensuite été évaluée. L'AUC a été traitée comme une mesure de l'efficacité avec laquelle l'outil a détecté des interactions dans les ensembles de données simulées que nous savions être réelles, c'est-à-dire présentes dans l'ensemble de données, ou fausses, c'est-à-dire non présentes dans l'ensemble de données. L'inférence des interactions a également été réalisée à l'aide des outils d'inférence statistique SparCC (Friedman2012) et SPIEC-EASI glasso (Kurtz2015), afin de comparer les performances de détection des interactions



	a	...	v	w	x	y	z
OTU1	7	...	25	0	113	50	0
OTU2	55	...	33	0	0	22	87
OTU3	40	...	19	1	4	0	12
OUT4	0	...	155	78	0	47	92
OTU5	27	...	0	3	14	9	0
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
OTUX	110	...	78	5	9	17	39

Figure 5: Schéma représentant le processus d'InfIntE. Le processus effectue : la conversion des données d'abondance contenues dans le tableau des OTU en clauses logiques, sur la base de nos connaissances écologiques ; l'abduction des effets d'interaction à partir des clauses logiques à l'aide de PyGol ; la sélection des bords importants à l'aide de StARS ; et, la classification directe des types d'interaction représentés par les différentes couleurs des liens. Les flèches indiquent la direction des effets ou des interactions. Les arêtes sans flèche dans le réseau représentent des interactions non dirigées.

entre notre approche d'inférence logique et les outils d'inférence de réseaux statistiques actuels, largement adoptés en écologie microbienne. L'inférence SparCC a été réalisée dans FastSpar v1.0.0 (Watts et al., 2018) et l'inférence glasso SPIEC-EASI a été exécutée dans le paquet R SpiecEasi v1.1.2, tous deux avec leurs paramètres par défaut respectifs. La statistique I a été calculée pour les hypothèses avec et sans exclusion dans InfIntE. Les corrélations SparCC ont été obtenues directement et les corrélations SPIEC-EASI ont été obtenues à partir de la matrice de covariance inverse à $\lambda = 0$. Étant donné que ces trois outils différents produisent différents types de réseaux d'interaction, soit des réseaux d'interaction classés, soit des réseaux corrélationnels, la plus grande valeur de I ou de corrélation, obtenue pour chaque paire d'OTU, a été utilisée pour comparer les performances des statistiques utilisées par les différents outils (figure 6).

16.2 Expérience 2 : Identification d'agents potentiels de lutte biologique par inférence InfIntE

Les réseaux d'interaction issus de données réelles sur la vigne ont été inférés à l'aide de InfIntE. Un total de 20 ASV fongiques, correspondant à 12 espèces et 2 genres différents, ont été identifiés comme antagonistes potentiels de *P. viticola* (Tableau 2). Parmi ceux-ci, cinq ASV fongiques des genres *Alternaria* et *Fusarium genus* sont déjà apparus dans la littérature comme des antagonistes de *P. viticola* (Musetti et al., 2006; Ghule et al., 2018). Ces ASV ont été classés pour avoir des interactions de compétition avec *P. viticola* dans au moins un des réseaux de vignobles. L'espèce fongique, *Aureobasidium pullulans*, a également été inférée comme ayant une interaction de compétition avec le pathogène, et a été décrite comme antagoniste de *P. viticola*(Harm et al., 2011). Les ASV correspondant aux genres *Cladosporium*, *Phlebia*, *Sporobolomyces* et *Vishniacozyma* n'ont pas été décrits précédemment comme antagonistes de *P. viticola*, mais ont été identifiés comme antagonistes de pathogènes dans d'autres systèmes foliaires. Les ASV attribuées à *Mycosphaerella tassiana* et deux espèces du genre *Filobasidium* ont également été classées comme antagonistes de *P. viticola* par inférence, mais elles ne sont pas décrites comme antagonistes dans la littérature et représentent de nouveaux agents de biocontrôle potentiels.

17 Conclusions

1. L'EML peut être utilisé pour détecter les interactions dans la même plage de précision que les autres outils statistiques.
2. L'EML application, Infinte, peut classer les interactions par type, en détectant les antagonistes potentiels des agents pathogènes.
3. InfInte est codé dans un package R et sera disponible pour tout utilisateur.

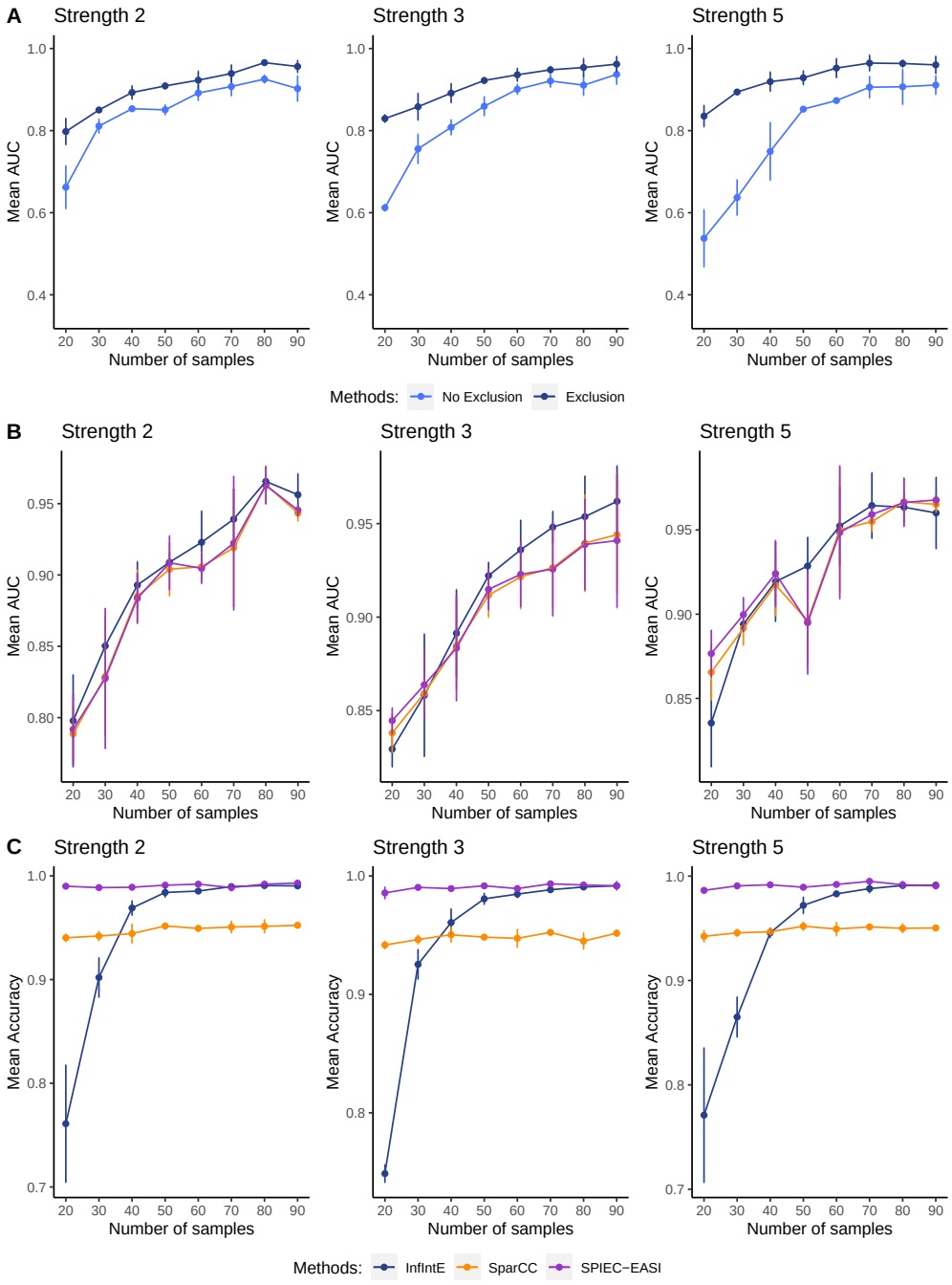


Figure 6: Relation entre le nombre d'échantillons et les performances d'inférence d'interaction pour différentes forces d'interaction. Les ensembles de données ont été générés par ordinateur en simulant quatre types d'interaction différents : amensalisme, commensalisme, compétition et mutualisme. **A:** Valeurs de l'aire sous la courbe (AUC)(Fan et al., 2006) obtenues par la statistique I avec et sans exclusion. Des valeurs AUC plus élevées représentent une meilleure spécificité et sensibilité dans la détection des interactions. La statistique I est utilisée par InfIntE comme une mesure numérique de l'interaction. **B:** Valeurs de l'aire sous la courbe (AUC) obtenues par la statistique I d'InfIntE et les mesures similaires à la corrélation SparCC et SPIEC-EASI. InfIntE a utilisé l'hypothèse d'interactions incluant l'exclusion. SparCC et SPIEC-EASI ont été exécutés avec les paramètres par défaut. **C:** Précision de la détection des interactions calculée en fonction de toutes les interactions possibles. InfIntE a utilisé l'hypothèse d'interactions incluant l'exclusion. SparCC et SPIEC-EASI ont été exécutés avec les paramètres par défaut.

Table 2: Les antagonistes potentiels de *Plasmopara viticola* trouvés par InfIntE. Le tableau montre les espèces fongiques trouvées pour avoir une interaction potentielle capable de réduire l'abondance de *P. viticola*. Une recherche bibliographique dans Google Scholar, Pubmed et Science Direct a été effectuée afin d'identifier si les antagonistes potentiels avaient déjà été décrits comme agents de biocontrôle du *P. viticola* ou d'autres pathogènes dans la littérature. Les mots-clés utilisés pour la recherche étaient le nom de l'antagoniste potentiel, "*Plasmopara viticola*", "biocontrôle" et "antagoniste". Les taxons identifiés par un astérisque n'ont pas été automatiquement assignés à un groupe taxonomique dans UNITE et ont dû être traités manuellement et assignés à l'aide de BLAST. Lorsque plus d'une OTU a été attribuée à la même espèce ayant la même interaction, cela est indiqué par xn, où n est le nombre d'OTU

.

Name	**Vineyard**	**Interaction**	**Bibliography against plasmopara**	**Bibliography biocontrol**
Cladosporium delicatulum	I	competition		Kohl et al. 2019; Baharvandi et al. 2015; Becker et al. 2020
Mycosphaerella tassiana	I	competition		
*Alternaria sp.**	A	amensalism	Mussetti et al. 2006	
*Alternaria alternata**	I	competition	Mussetti et al. 2006, 2007	
Alternaria brasicae	B	competition	Duhan et al. 2021	
*Aureobasidium pullulans**	I	competition	Harm et al. 2011	
Filobasidium chernovii	Ix2	competition		
*Filobasidium magnu*m*	D	competition		
*Fusarium sp.**	A, B, E	competition	Ghule et al. 2018; Bakshi et al. 2001	
Phlebia rufa	E	amensalism		White and Boddy 1992
Sporobolomyces roseus	Ix3	competition		Janisiewicz et al. 1994; Filonow et al. 1996;
*Sporobolomyces pararoseus**	A,G	competition		Li et al. 2017 (in grapes)
Vishniacozyma victoriae	B,C	amensalism, competition		Gramisci et al. 2018; Lutz et al. 2020
Vishniacozyma carnescens	D	amensalism		Becker et al. 2020