New Phytologist

Research

Seed microbiota revealed by a large-scale meta-analysis including 50 plant species

Marie Simonin, Martial Briand, Guillaume Chesneau, Aude Rochefort, Coralie Marais, Alain Sarniguet and Matthieu Barret

Institut Agro, INRAE, IRHS, SFR QUASAV, Université d'Angers, F-49000 Angers, France

Author for correspondence:
Marie Simonin
Email: marie.simonin@inrae.fr

Received: *8 September 2021*
Accepted: *3 January 2022*

New Phytologist (2022)
doi: 10.1111/nph.18037

Key words: data synthesis, diversity, fungal community, metabarcoding, plant microbiome.

Summary

- Seed microbiota constitutes a primary inoculum for plants that is gaining attention owing to its role for plant health and productivity.
- Here, we performed a meta-analysis on 63 seed microbiota studies covering 50 plant species to synthesize knowledge on the diversity of this habitat.
- Seed microbiota are diverse and extremely variable, with taxa richness varying from one to thousands of taxa. Hence, seed microbiota presents a variable (i.e. flexible) microbial fraction but we also identified a stable (i.e. core) fraction across samples.
- Around 30 bacterial and fungal taxa are present in most plant species and in samples from all over the world. Core taxa, such as *Pantoea agglomerans*, *Pseudomonas viridiflava, P. fluorescens*, *Cladosporium perangustum* and *Alternaria* sp., are dominant seed taxa.
- The characterization of the core and flexible seed microbiota provided here will help uncover seed microbiota roles for plant health and design effective microbiome engineering.

Introduction

Seeds are key components of plant fitness and are central to the sustainability of the agri-food system. Seed consumption is the foundation of human food security with wheat, rice and maize seeds providing 42.5% of the world's food calorie supply (Reeves *et al.*, 2016). Moreover, the transition of seed to seedling represents a major bottleneck for both plant fitness and the assembly of plant microbiota. This transition has large implications in agricultural systems and for the maintenance of plant biodiversity in natural ecosystems (Leck *et al.*, 2008; Barret *et al.*, 2015; Paredes & Lebeis, 2016). Both seed quality for food consumption and seed vigour in agricultural settings can be influenced by the microorganisms living inside and on the surface of seeds (i.e. the seed microbiota) (Fels-Klerx *et al.*, 2012; Nelson, 2018). Knowledge regarding seed microbiota has long been lagging behind that of other plant compartments, like the rhizosphere, phyllosphere and endosphere (Shade *et al.*, 2017). However, recently renewed focus on reproductive tissues (seeds, flowers) and plant early life stages has emerged to better understand the dynamic and assembly processes of plant microbiota (Vandenkoornhuyse *et al.*, 2015; Ridout *et al.*, 2019; Fitzpatrick *et al.*, 2020).

Seed microbiota constitutes a primary microbial inoculum for the plant microbiota with potential long-term impacts on plant fitness. Seed-associated microorganisms can be acquired either horizontally from various environments (e.g. air, water, insects, seed processing) or vertically from the mother plant and transmitted across multiple generations (Shahzad *et al.*, 2018; Rodríguez *et al.*, 2019; Chesneau *et al.*, 2020). The prevalence of pathogens on seeds has been studied extensively but the occurrence and role of other commensal or mutualistic microorganisms constituting the majority of seed microbiota are mostly unknown (Darsonval *et al.*, 2009; Darrasse *et al.*, 2010; Bertolini *et al.*, 2015). Pioneer studies demonstrated that seed microbiota can influence plant fitness by modulating seed germination, seedling phenotypes or by promoting root symbiosis (Nelson, 2018; Ridout *et al.*, 2019; Rodríguez *et al.*, 2019). Still, limited attempts have been made to characterize the seed core microbiota of a specific plant species or shared across multiple plant species at a large scale (H. Chen *et al.*, 2018; Eyre *et al.*, 2019). Here, the definition of the core microbiota used corresponds to the 'common core' (*sensu* Risely, 2020) that represents the component of the microbiota that is found across a considerable proportion of hosts. The identification of the core and flexible microbiota of a plant habitat can help identify microbial taxa and functions that may be particularly important for host fitness (Shade & Handelsman, 2012). This identification can be achieved by large-scale data synthesis efforts (e.g. meta-analysis) which remain outstanding for seed microbiota.

A global analysis of seed microbial diversity appears even more timely as seeds appear as a key vector of solutions to promote sustainable agriculture. Seeds can play a double role: (1) as a source of innovations with seed-borne microorganisms representing key biotechnological resources (Berg & Raaijmakers, 2018); and (2) as carriers of microbial biostimulants or biocontrol solutions that greatly reduce the surface and volume of treatment applied to

fields, thus decreasing application costs and potential negative impacts on the environment (O'Callaghan, 2016; Ben-Jabeur *et al.*, 2019; Rocha *et al.*, 2019). A synthesized knowledge of seed microbiota will accelerate discoveries and help future practices promoting the presence of important seed microorganisms for plant health and productivity.

In this context, we performed a meta-analysis on available seed microbiota studies to synthesize current knowledge on the diversity of this habitat and to constitute an open database for the research community. This data synthesis effort enabled us to address the following questions:

(**1**) How diverse is the seed microbiota?

(**2**) Which taxonomic groups compose the seed microbiota?

(**3**) Can we find the evidence for a seed core microbiota shared across plant species?

(**4**) Do we detect specific patterns in seed microbiota composition and diversity by plant species?

These questions were addressed through a meta-analysis gathering a total of 63 seed microbiota studies yielding 3190 seed samples from 50 plant species collected in 28 countries. This study shows that the overall seed microbiota composition is highly variable from one seed sample to another, but most seeds share a core microbiota composed of few dominant bacterial and fungal taxa.

Materials and Methods

Studies included in the Seed Microbiota Database

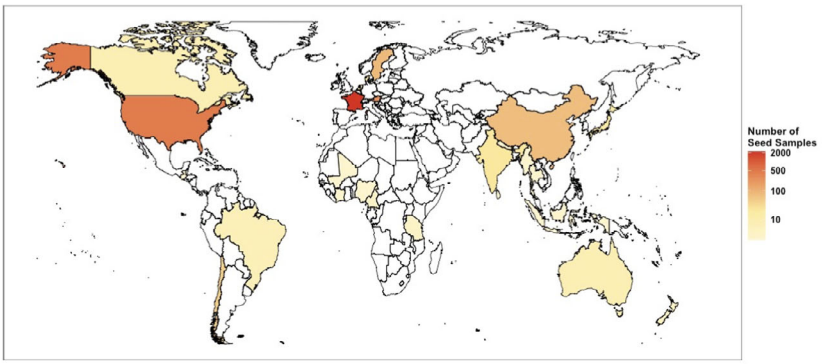
We identified seed microbiota studies from our team and from the literature by performing keyword searches in Google Scholar and by following references in reviews and seed microbiota papers (end of search 17 November 2020). Initially we identified 100 seed microbiota studies (59 for the 16S ribosomal (r)RNA gene, 14 for the *gyrB* gene and 27 for the internal transcribed spacer (ITS) region) that used amplicon sequencing (i.e. metabarcoding), but we decided to conserve in the meta-analysis only those studies using the Illumina sequencing technology that was the most common approach (90 datasets). This choice permitted us to use a similar bioinformatic pipeline for all studies and to not introduce biases in our interpretation as a consequence of differing sequencing technologies. Of these 90 studies, only 63 studies were finally included in this meta-analysis based on the following criteria: (1) availability of the raw sequence data on public repository (FASTQ files), (2) availability of metadata associated to each sample, (3) sequencing quality, and (4) the primer sets used. For the sequencing quality, studies with average PHRED quality scores < 20 were excluded and for the studies retained, each read presenting an instance of a quality score of ≤ 2 were excluded (DADA2 default setting). For the primers used, we focused on the primers or gene regions that were the most commonly studied in seed microbiota studies: for the 16S rRNA gene the V4 and V5–V6 regions and for ITS, the ITS1 and ITS2 regions. Most studies were downloaded from online repositories (e.g. ENA, NCBI-SRA) and some were obtained after personal communication with the authors. Detailed information on the 63 studies that constitute the Seed Microbiota Database can be found in Supporting Information Dataset S1.

Bioinformatic analysis and preparation of the datasets and subsets

Each study was independently re-processed with a standardized bioinformatic pipeline (code available at https://github.com/marie-simonin/Seed-Microbiota-Database) using QIIME 2 and DADA2 (Callahan *et al.*, 2016; Bolyen *et al.*, 2019). In brief, primer sequences were removed with CUTADAPT 2.7 and trimmed FASTQ files were processed with DADA2 v.1.10, using variable truncation parameters depending on the sequencing quality of the study. If necessary, because of small differences in the primers used, a trimming parameter was added in DADA2 to have all studies targeting the exact same region of the V4, V5–V6, ITS1 or ITS2 regions. Chimeric sequences were identified and removed with the removeBimeraDenovo function of DADA2. ITS1 and ITS2 reads were processed likewise, except that only the forward reads were included in the analysis, as recommended by Pauvert *et al.* (2019) to maximize fungal diversity detection. Amplicon sequence variant (ASV) taxonomic affiliations were performed with a scikit-learn multinomial naive Bayesian classifier implemented in QIIME 2 (qiime feature-classifier classify-sklearn; Bokulich *et al.*, 2018). The ASVs derived from 16S rRNA reads were classified with the SILVA 132 database (Quast *et al.*, 2013) trained for the specific gene region (V4 or V5–V6). *gyrB* reads were classified with an in-house *gyrB* database (train_set_gyrB_v4.fa.gz) available upon request. The ASVs derived from ITS1 and ITS2 reads were classified with the UNITE v.8 fungal database (Nilsson *et al.*, 2019). Unassigned ASVs or ASVs affiliated to mitochondria, chloroplasts or eukaryotes were removed from the 16S rRNA gene studies. Nonfungal ASVs were filtered from the ITS gene studies, and unassigned and *parE* ASVs (a *gyrB* paralogue) were filtered from the *gyrB* gene studies. After these filtering steps, samples with < 1000 reads were excluded from the meta-analysis. It should be noted that for some studies, several samples were lost at these filtering steps as a consequence of a high proportion of nontarget reads (plastid or nonfungal reads; see Dataset S1) and low sequencing depth, leading to a low sample replication for some plant species. For each study, two datasets were prepared: an unrarefied and a rarefied one. The rarefaction was performed at the lowest read number observed (min = 1000 reads/sample, max = 86 108 reads/sample; Dataset S1).

The next step was to merge all the studies targeting the same marker gene and region to form a final 'dataset'. This merging step between different studies was possible because all of the ASVs obtained targeted the exact same region of the marker gene considered (see above). A total of five datasets (16S rRNA-V4, 16S rRNA-V5-V6, *gyrB*, ITS1 and ITS2) were included in the meta-analysis (Fig. 1b). The unrarefied studies were merged into one dataset then further filtered to obtain a 'clean database' (e.g. Subset 1) by removing (1) ASVs present in only one sample and with < 20 reads in the entire dataset, (2) ASVs shorter than 200 bp, and (3) *gyrB* ASVs with a read length not compatible

(a) Origin of seed samples



(b) Plant species and studies included in the meta-analysis

Plant Family	Plant Species	Number of Seed Samples: 16S / gyrB / ITS	Number of Studies: 16S / gyrB / ITS	References included in Meta-analysis
Apiaceae	Daucus carota (Carrot)	4 / 4 / 4	1 / 1 / 1	Barret et al. 2015
	Astrantia major (Great Masterwort)	4 / / 4	1 / / 1	Wassermann et al. 2019
Asteraceae	Helianthus annuus (Sunflower)	18 / / 22	1 / / 1	Leff et al 2017
Brassicaceae	Alliaria petiolata	8 / 7 / 8	1 / 1 / 1	Lecorff et al. 2018 unp
	Arabidopsis thaliana	10 / 9 / 9	2 / 2 / 2	Barret et al. 2015; Lecorff et al. 2018 unp
	Barbarea vulgaris	2 / 2 / 2	1 / 1 / 1	Lecorff et al. 2018 unp
	Berteroa incana	2 / 2 / 2	1 / 1 / 1	Lecorff et al. 2018 unp
	Brassica nigra	2 / 2 / 2	1 / 1 / 1	Lecorff et al. 2018 unp
	Brassica oleracea var. italica (Broccoli)	2 / 2 / 2	1 / 1 / 1	Barret et al. 2015
	Brassica oleracea var. capitata (Cabbage)	28 / 28 / 18	2 / 2 / 1	Barret et al. 2015; Goertz et al. 2020 unp
	Capsella bursa-pastoris	4 / 2 / 4	1 / 1 / 1	Lecorff et al. 2018 unp
	Cardamine hirsuta	3 / 1 / 4	1 / 1 / 1	Lecorff et al. 2018 unp
	Brassica oleracea var. botrytis (Cauliflower)	2 / 2 / 2	1 / 1 / 1	Lecorff et al. 2018 unp
	Erophila verna	4 / 3 / 4	1 / 1 / 1	Lecorff et al. 2018 unp
	Eruca vesicaria (Garden rocket)	2 / 2 / 2	1 / 1 / 1	Barret et al. 2015
	Raphanus sativus (Radish)	203 / 295 / 187	4 / 7 / 3	Barret et al. 2015,2017 unp; Chesneau et al. 2020; Lecorff et al. 2018 unp; Rezki et al. 2016, 2018; Torres-Cortes et al. 2019
	Brassica napus (Rapeseed)	350 / 302 / 238	7 / 4 / 5	Barret et al. 2015; Huet et al. 2020; Mougel et al. 2017 unp; Prado et al. 2020; Rochefort et al. 2019; Rybakova et al. 2017; Sarniguet et al. 2016 unp
	Rorippa sylvestris	2 / / 2	1 / / 1	Lecorff et al. 2018 unp
	Sinapis arvensis	3 / 3 / 3	1 / 1 / 1	Lecorff et al. 2018 unp
	Brassica rapa var. rapa (Turnip)	8 / 8 / 8	1 / 1 / 1	Barret et al. 2015
Caprifoliaceae	Pincushion Flower	4 / / 5	1 / / 1	Wassermann et al. 2019
Caryophyllaceae	Heliosperma alpestre	4 / / 4	1 / / 1	Wassermann et al. 2019
Celastraceae	Grass of Parnassus	4 / / 4	1 / / 1	Wassermann et al. 2019
Cucurbitaceae	Melon	10 / 10 /	1 / 1 /	Goertz et al. 2020 unp
Fabaceae	Alfalfa	1 / / 3	1 / / 1	Dai et al. 2020
	Phaseolus vulgaris (Bean)	176 / 50 / 186	5 / 4 / 5	Barret et al. 2015; Barret et al. 2015-2 unp; Barret et al. 2017 unp; Barret et al. 2020 unp; Bintarti et al. 2020; Chesneau et al. 2020; Klaedtke et al. 2016
	Vicia villosa (Hairy vetch)	2 / / 3	1 / / 1	Dai et al. 2020
	Medicago truncatula	1 / 2 / 2	1 / 1 / 1	Barret et al. 2015
	Pisum sativum (Pea)	5 / / 17	1 / / 1	Chartrel et al. 2021
	Trifolium pratense (Red Clover)	/ / 3	/ / 1	Dai et al. 2020
	Trifolium repens (White Clover)	/ / 6	/ / 2	Dai et al. 2020; Idbella et al. 2020
Fagaceae	Quercus robur (Oak)	30 / / 46	1 / / 1	Abdelfattah et al. 2020
Gentianaceae	Gentiana germanica (Chiltern Gentian)	4 / / 4	1 / / 1	Wassermann et al. 2019
	Gentiana asclepiadea (Willow Gentian)	4 / / 4	1 / / 1	Wassermann et al. 2019
Orobanchaceae	Euphrasia rostkoviana (Eyebright)	4 / / 4	1 / / 1	Wassermann et al. 2019
	Phelipanche ramosa	76 / / 76	1 / / 1	Huet et al. 2020
	Rhinanthus glacialis	4 / / 3	1 / / 1	Wassermann et al. 2019
Poaceae	Agrostis stolonifera (Creeping Bentgrass)	/ / 7	/ / 1	Doherty et al. 2020
	Elymus dahuricus (Dahurian wildrye)	3 / / 3	1 / / 1	Dai et al. 2020
	Festuca rubra	6 / /	1 / /	Chen Q et al. 2020
	Lolium arundinacea	2 / /	1 / /	Chen Q et al. 2020
	Lolium perenne	16 / /	1 / /	Chen Q et al. 2020
	Avena sativa (Oat)	3 / / 3	1 / / 1	Dai et al. 2020
	Oryza sp. (Rice)	120 / / 43	4 / / 1	Eyre et al. 2019; Kim et al. 2020; Liu et al. 2019; Raj et al. 2019
	Setaria pumila	33 / /	1 / /	Escobar-Rodriguez et al. 2018
	Setaria viridis	123 / /	2 / /	Escobar-Rodriguez et al. 2018, 2019
	Elymus sibiricus (Siberian wildrye)	3 / / 3	1 / / 1	Dai et al. 2020
	Triticum aestivum (Wheat)	184 / / 165	2 / / 1	Bakker & McCormick 2019; Mitter et al. 2017
Solanaceae	Nicotiana tabacum (Tobacco)	20 / /	1 / /	Chen X et al. 2020
	Solanum lycopersicum (Tomato)	28 / 18 / 4	3 / 2 / 1	Barret et al. 2015; Bergna et al. 2018; Goertz et al. 2020 unp
Total: 14 Families	**50 Species**	**1531 / 745 / 1125**	**31 / 12 / 20**	

Fig. 1 Overview of the seed microbiota meta-analysis database. (a) Map presenting the countries of origin of seed samples ($n = 28$ countries). The countries are coloured based on the number of seed samples that originated from this area. (b) Table presenting each plant species included in the meta-analysis ($n = 50$ plant species): the number of samples and studies for the three marker genes and the references of the studies. unp, unpublished dataset originating from our research group.

with codon triplets (multiple of 3). The rarefied studies were merged into one dataset (Subset 2) to look at study-specific patterns in seed microbiota diversity. To investigate community structure and composition patterns across all studies, an additional subset was prepared by performing a global rarefaction across all samples (Subset 3, rarefaction at 1000 to 1056 reads/sample depending on the dataset; Table S1).

Community and taxa-level analyses

Diversity and community structure analyses were performed in R 3.6.2 using the PHYLOSEQ (v.1.28.0), VEGAN (v.2.5-7) and MICROBIOME (v.1.7.21) packages (Oksanen *et al.*, 2007; McMurdie & Holmes, 2013; Lahti *et al.*, 2017). The code and files to reproduce the analyses and figures are available at https://github.com/marie-simonin/Seed-Microbiota-Database. Seed microbiota diversity across all studies was explored using ASV richness, Shannon and Pielou's evenness indexes. Changes in seed microbial community composition were assessed using Bray–Curtis dissimilarity and principal coordinate analysis (PCoA) was used to plot the ordination.

Seed core and flexible microbiota across all plant species or specific to some plant species were identified. For the analyses across all plant species, we used the arbitrary criteria that ASVs detected in ≥ 20 plant species would be considered as core taxa, as they are prevalent in seed microbiota from diverse plants and countries.

For the analyses by plant species, we focused on plants for which a sufficient number of independent studies (≥ 3 studies) and samples (≥ 50 samples) were available for a robust analysis. Four plant species responded to these criteria for the 16S rRNA gene-V4 dataset (bean, radish, rapeseed and rice) and three plant species for the *gyrB* and ITS1 datasets (bean, radish, rapeseed). To identify the core ASVs specific to these plant species, we used the arbitrary criteria that an ASV should be detected in a minimum of two independent studies and be present in at least half of the samples (minimum prevalence 50%). On the same subset of four (16S rRNA gene-V4 dataset) or three plant species (ITS1 dataset), we investigated the statistical effects of plant species on seed microbiota diversity and structure. We also tested the statistical influence of plant botanical families on a subset of plant families (Brassicaceae, Fabaceae, Orobanchaceae, Poaceae for the 16S rRNA gene-V4 dataset, and Brassicaceae, Fabaceae, Orobanchaceae for ITS1) for which we had multiple species characterized (> 3 species) and a high seed sample number (> 83 samples). We assessed statistical effects of plant species and families on the ASV richness and relative abundance of bacterial phyla or fungal classes using Kruskall–Wallis nonparametric tests in R. The effects on seed community structure and predicted metagenomes (see below) were analyzed using a PERMANOVA with the adonis function in VEGAN.

For the 16S-V4 and *gyrB* datasets (using Subset 1), we used PICRUST2 (Douglas *et al.*, 2020) to predict metagenome functions of seed microbiota from metabarcoding data.

All figures were prepared using the GGPLOT2 (v.3.3.3) (Wickham, 2016) and GGPUBR (v.0.4.0) packages, and the data management was done using the DPLYR (v.1.0.4) and TIDYVERSE (v.1.3.0) packages in R and in DB Browser for SQLITE (v3.12.2).

Results

Creation of the Seed Microbiota Database to perform the meta-analysis

In order to synthesize knowledge on the diversity and composition of seed microbiota, we identified, re-processed and re-analyzed raw data from published microbiome studies, and also unpublished data from our laboratory. We initially identified 100 seed microbiota studies (59 for 16S rRNA gene; 14 for *gyrB* and 27 for fungal ITS regions) that used amplicon sequencing (i.e. metabarcoding) to characterize community structure. To enable the comparison of different studies and the use of a common bioinformatic pipeline, only studies using the Illumina sequencing technology were considered ($n = 90$). A total of 63 studies were finally included in the publicly available database that we called Seed Microbiota Database, based on the (1) availability of FASTQ files and associated metadata, (2) sequencing quality and (3) primer sets employed (Fig. 1). The Seed Microbiota Database is available on the Data INRAE portal (https://doi.org/10.15454/2ANNJM). Detailed information on the 63 studies and references can be found in Dataset S1 and Fig. 1(b). Each study was independently reprocessed with a standardized bioinformatic pipeline (https://github.com/marie-simonin/Seed-Microbiota-Database) using QIIME2 and DADA2 before being merged with other studies targeting the same marker gene and region to form a final 'dataset'. A total of five final datasets were included in the meta-analysis, which corresponds to the most common molecular markers (16S rRNA-V4, 16S rRNA-V5-V6, *gyrB*, ITS1 and ITS2) used to characterize seed microbiomes (Fig. S1). Detailed information on the description of the samples (e.g. origin, plant species, seed preparation, seed fraction, seed number) for the different molecular markers employed can be found in Figs S2–S4.

In the end, this meta-analysis was conducted on a total of 3190 samples from 50 plant species collected in 28 countries. The community profiles of these samples were estimated via three bacterial markers (*gyrB*, V4 and V5–V6 regions of the 16S rRNA gene) and two fungal markers (ITS1 and ITS2) with 105 million paired-end reads and thousands of ASVs identified (Fig. 1). The 16S-V4, *gyrB* and ITS1 datasets presented the highest number of samples and plant diversity (Fig. 1b) and thus have been used primarily for community profiling of the seed microbiota in this meta-analysis. Each dataset was prepared in three different subsets adapted to different types of downstream analyses (Subset 1: no rarefaction; Subset 2: rarefaction at the study-level; Subset 3: rarefaction across all studies; Fig. S1; Table S1).

The seed samples included in the meta-analysis cover an important diversity of plant families ($n = 14$). Still, it should be noted that 73% of all samples came from four species from Brassicaceae (*Raphanus sativus* (radish), *Brassica napus* (rapeseed)), Fabaceae (*Phaseolus vulgaris* (bean)) and Poaceae (*Triticum aestivum* (wheat)) (Figs 1b, S2–S4). Most of the other plant species

investigated were covered by < 40 seed samples ($n = 37$), with the exception of *Oryza* sp. (rice), *Setaria viridis*, *Phelipanche ramosa*, *Solanum lycopersicum* (tomato), *Brassica oleracea* var. *capitata* (cabbage) and *Helianthus annuus* (sunflower) (Figs S2–S4). Seed samples were harvested across all continents with a significant number of samples from France ($n = 2159$), USA ($n = 457$), Austria ($n = 261$), Sweden ($n = 79$) and China ($n = 71$), and were analyzed by 15 research institutes/universities. Different procedures were employed for characterizing seed microbiota structure including raw seed samples (i.e. total fraction), surface-disinfected, surface-washed and dissected seeds. Additionally, seeds were prepared following different approaches before DNA extraction (e.g. soaking, grinding, washing; Figs S2–S4). All studies except one (Abdelfattah *et al.*, 2021, *Quercus robur* L. (oak)) characterized seed microbiota structure on seed lots (> 3 seeds) because the microbial DNA present on a single seed often is insufficient to perform metabarcoding (Caruso *et al.*, 2019). Hence, seed microbiota were characterized on samples with a variable number of seeds depending on the study, but the majority of samples were composed of 1000 seeds providing a view of the seed 'metacommunity' of a given site.

Part 1: Patterns across all plant species

Highly variable seed microbiota diversity The alpha diversity patterns of seed microbiota were explored using observed ASV richness, Pielou's evenness and Shannon diversity index. A large variability in ASV richness per seed sample was observed within and across plant species (16S rRNA gene markers, 1–2224 ASVs; *gyrB*, 3–718 ASVs; ITS markers, 1–423 ASVs; Fig. 2). The median ASV richness across all samples were 48 ± 8 for the 16S rRNA gene markers, 63 ± 8.6 for *gyrB* and 52 ± 2.7 for ITS markers indicating similar levels of prokaryotic (Archaea + Bacteria) and fungal richness in seed microbiota. When considering only the plant species that had a high number of samples (> 120) and several independent studies (> 4), we observed that rapeseed presented a significantly higher prokaryotic diversity (median 57 ASVs), whereas rice (33 ASVs), radish (23 ASVs) and bean seeds (23 ASVs) were on the lower end (Fig. S5D). At the plant family level, Orobanchaceae had a significantly higher prokaryotic richness (median 91 ASVs) than Poaceae (56 ASVs), Brassicaceae (41 ASVs) and Fabaceae (23 ASVs; Fig. S6D).

For the fungal community, rapeseed (median: 67 ASVs) and bean (67 ASVs) had significantly higher richness than radish seeds (51 ASVs; Fig. S5E). At the plant family level, Fabaceae (median: 64 ASVs) and Orobanchaceae (63 ASVs) had an overall similar fungal richness to Brassicaceae (60 ASVs; Fig. S6E).

A high variability between seed samples also was observed in Pielou's evenness values (Fig. S7) and Shannon index (Fig. S8). Pielou's index provides a measurement of the evenness of the taxa in seed microbiota ranging from 0 (no evenness) to 1 (complete evenness). Between samples from the same study, seed bacterial and fungal communities can range from being extremely uneven (< 0.25) to highly even (> 0.75) for several plant species, such as bean, radish, rapeseed and wheat (Fig. S7), again showing the high variability in seed microbiota composition.

We also considered the influence of the seed fraction studied, as many studies performed a surface-disinfection of the seeds to enrich the endophytic fraction of the microbiota. The median ASV richness of surface-disinfected seeds was only slightly lower (40 ± 6, $n = 415$) than nondisinfected seeds (Total fraction, 50 ± 10, $n = 1039$) for the 16S rRNA gene samples but the loss of ASVs with surface-disinfection was more pronounced for the ITS samples (16 ± 2, $n = 193$ vs 47 ± 3, $n = 930$; Fig. S9). Regarding the seed preparation technique used before DNA extraction, similar levels of prokaryotic ASV richness were observed between the different techniques except for the fungal ASV richness that was higher with seed soaking (61 ± 3, $n = 802$) compared to grinding (21 ± 2, $n = 322$, Fig. S10). However, this latter result could be driven mainly by the fact that different plant species were studied with different techniques. We also explored if the variability observed in seed microbiota diversity was associated with the number of seeds included in the seed sample (from 1 to 1000 seeds). We found that ASV richness generally increased with the number of seeds in the sample for both prokaryotic and fungal communities, but a high variability still existed for each given seed sample size (e.g. 1000 seeds; Fig. S11).

Altogether, these results show that seed microbiota are diverse with a median of 100 ASVs ± 11 by sample (sum of prokaryotic and fungal diversity) and also are extremely variable from one sample to another even within the same study or for a given plant species. These analyses also suggest that seed surface-disinfection and preparation have larger effects on the fungal community than the prokaryotic community.

Next, we explored the relationships between the prokaryotic and fungal richness to determine the group of dominance and their diversity correlation. This analysis was performed only on samples for which both 16S rRNA gene and ITS metabarcoding were performed ($n = 884$). The prokaryotic and fungal ASV richness were weakly positively correlated ($R^2 = 0.11$, $P < 0.001$; Fig. 3a). Seed microbiota presented a variable proportion of bacterial and fungal taxa but, interestingly, communities were not skewed towards one particular microbial group (Fig. 3a). Values were distributed close to the 1 : 1 line (similar diversity in both groups) or on both sides of the lines for most species and plant families (Fig. 3a,b). These results indicate that seed microbiota can be dominated either by fungal or prokaryotic taxa in terms of diversity. Only Poaceae (three studies) and Orobanchaceae (two studies) seeds appeared to present a higher proportion of prokaryotic richness (Fig. 3a,b). Still, large differences in the proportion of prokaryotic and fungal richness were observed between and among plant species (Fig. 3b). For instance, Brassicaceae species ranged from being dominated by fungal taxa (e.g. radish) to being highly dominated by prokaryotic taxa (e.g. cauliflower, *Arabidopsis thaliana*). In summary, these results show a similar contribution of the prokaryotic and fungal diversity to the structure of seed microbiota, with the exception of a few plant species that appeared dominated by prokaryotes.

Additionally, we explored the distribution of ASVs' relative abundance by sample using ranked abundance curves and compared the patterns between prokaryotic and fungal communities

New Phytologist

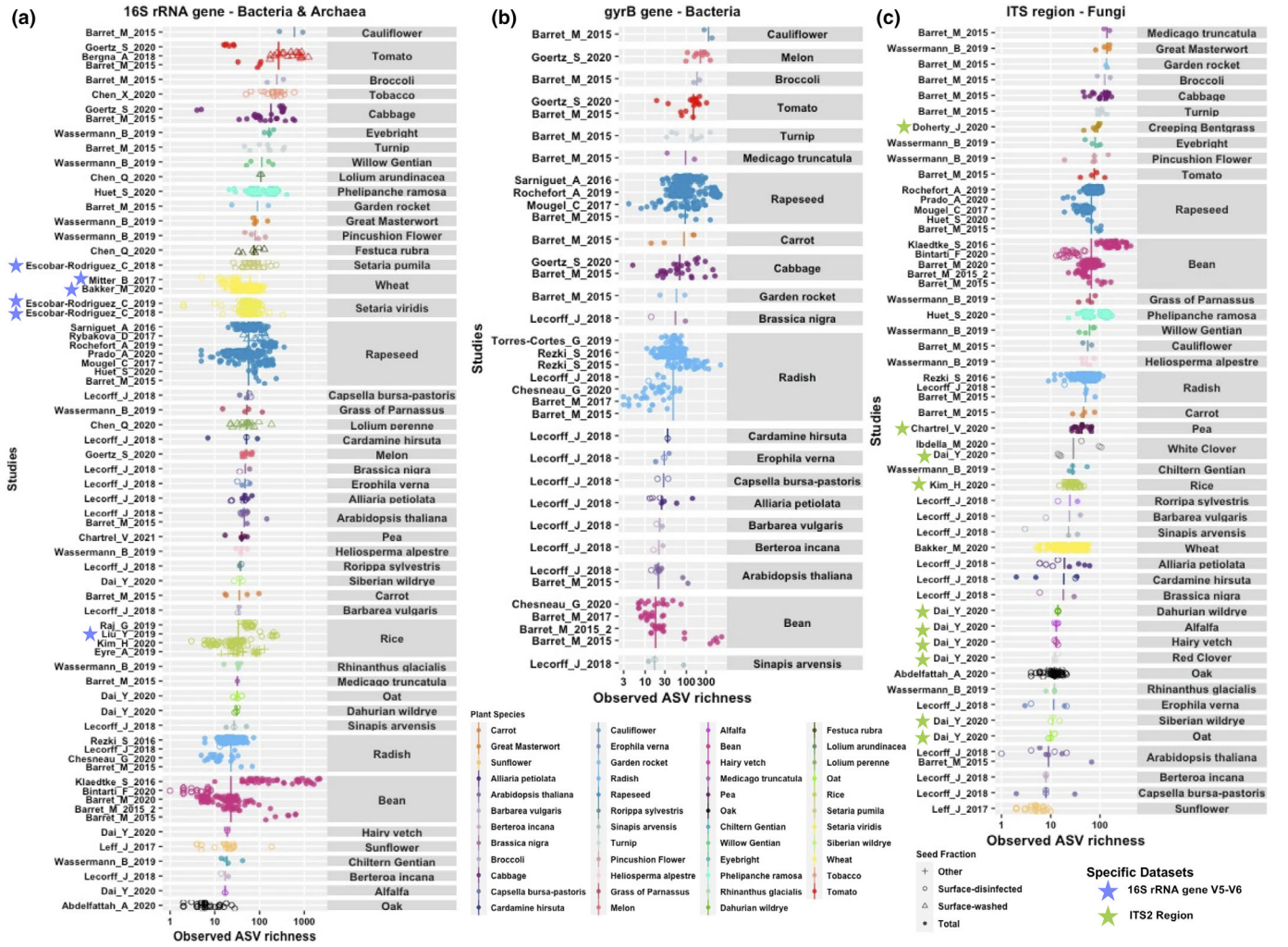


Fig. 2 Seed microbiota diversity patterns within and across plant species. Observed amplicon sequence variant (ASV) richness of all the seed samples included in the meta-analysis for the 16S ribosomal (r)RNA gene markers (a; $n = 1531$, V4 and V5–V6 datasets combined), *gyrB* (b; $n = 754$) and internal transcribed spacer (ITS) markers (c; $n = 1125$, ITS1 and ITS2 datasets combined). Blue star, the few 16S-V5–V6 studies; green star, ITS2 studies. Each point represents a seed sample that can be composed of multiple seeds (up to a thousand seeds). The shape of the points represents the seed fraction considered (Total, no pre-treatment of the seeds; Surface-washed, seeds were rinsed with sterile water; Surface-disinfected, seeds surface-sterilized with chemical products; Other, seeds dissected or received specific treatments). For each plant species, the vertical line corresponds to the median value of the samples. Note that the *x*-axis is presented on a log-scale and that the range differs on each panel. On each panel, species are ordered based on the median richness value of the plant species (lowest richness at the bottom and highest at the top).

(Fig. 3c,d). Both communities appeared to be dominated by few ASVs that collectively represented > 50% or 75% of the reads in seed microbiota. In particular, we observed that seed fungal communities were generally highly dominated by one ASV (median rank 1 ASV: 57.9% relative abundance), whereas the rank 1 ASV of the prokaryotic community generally had a lower abundance (median 33.1%). It should be noted that the identity of the ranked ASVs generally is different between samples. These results indicate that seed communities often are highly uneven with few dominant prokaryotic and fungal ASVs.

Identification of seven microbial clades dominating seed microbiota Abundance–occupancy curves were plotted for each phylum to identify the ASVs' distribution across all samples of the meta-analysis (Figs 4, S12, S13). The abundance–occupancy curves by phylum or class presented expected distributions with positive relationships between ASV abundance and prevalence, indicating that the most prevalent ASVs were also dominant in terms of relative abundance across the entire dataset. For Prokaryotes, ASVs were distributed across 43 phyla (two archaeal and 41 bacterial phyla; Fig. S12) but 12 phyla were dominant in terms of relative abundance (Fig. 4a) and collectively represented 99.79% of the reads in the dataset and 97% of the ASV richness (7948 ASVs). In particular, the four bacterial phyla that dominate the seed microbiota both in terms of diversity and abundance were Proteobacteria, Actinobacteria, Firmicutes and Bacteroidetes (Figs 4a, S12). A low detection of Archaea on seeds was observed with 51 ASVs that represented only 0.1% of the reads (16S rRNA gene-V4 dataset), but it should be noted that the primers targeting the 16S rRNA gene V4 region are not

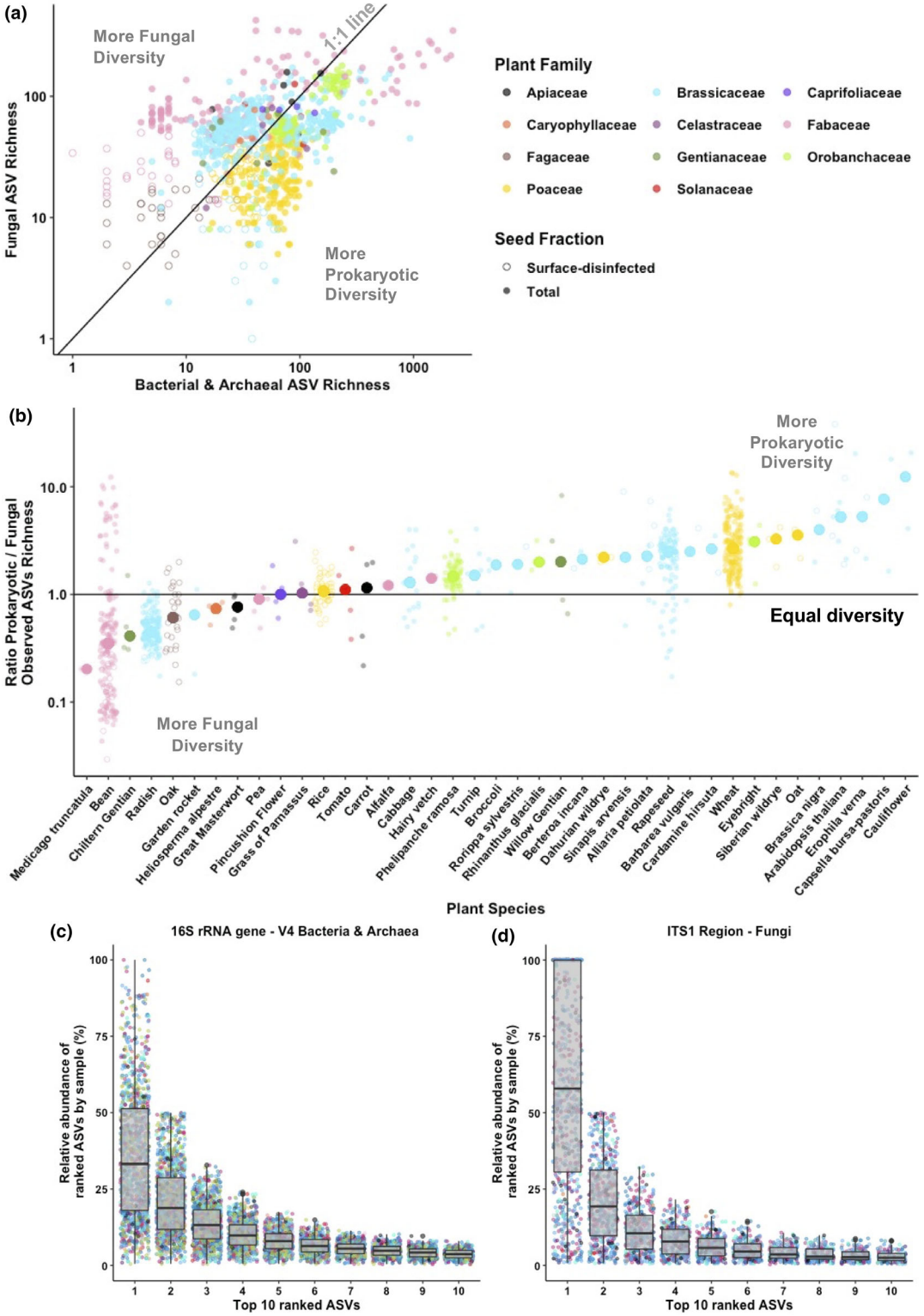
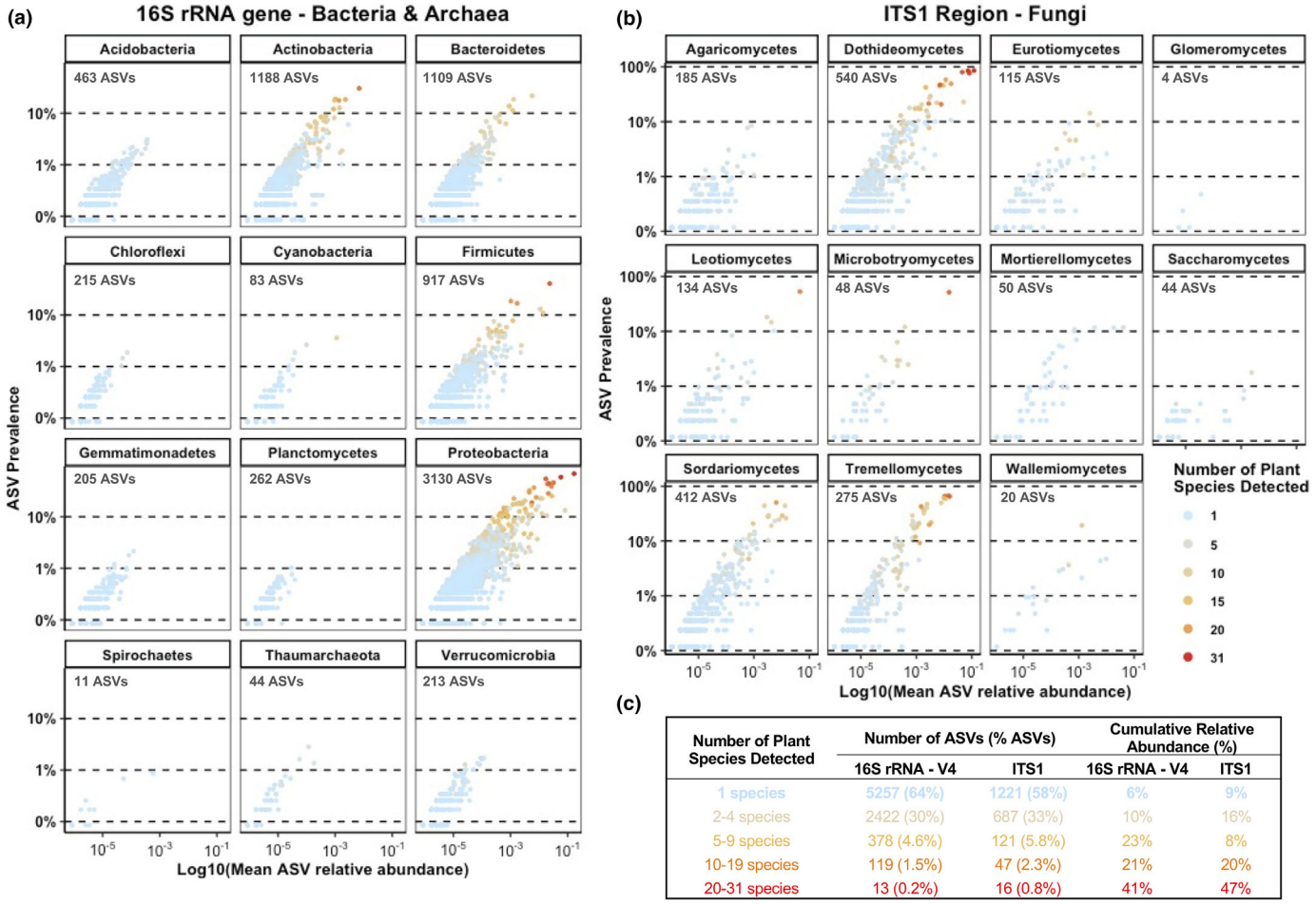


Fig. 3 Proportion of prokaryotic (bacteria and archaea) and fungal richness in seed microbiota for samples where both communities were characterized simultaneously ($n = 884$). (a) Relationship between amplicon sequence variant (ASV) richness values of the prokaryotic community (16S ribosomal (r)RNA gene-V4 dataset) and of the fungal community (internal transcribed spacer (ITS)1 region dataset). The 1 : 1 line facilitates the visualization of those plant families presenting an enrichment of one of the taxonomic groups over another. Note that on (a) both axes are on a log-scale. (b) Ratios of prokaryotic and fungal ASV richness for each plant species. The vertical line represents a ratio of 1 when communities present an equal diversity. Note that the *y*-axis is presented on a log-scale. Relative abundance of the top 10 ranked ASVs by sample for the prokaryotic community (c; $n = 1182$ samples) and fungal community (d; $n = 850$ samples). Each point represents a sample and is coloured by the plant species (see legend in Fig. 2). The boxplots represent the median (middle of the box) and top and bottom of the box corresponds to the first and third quartiles.



(c)

Number of Plant Species Detected	Number of ASVs (% ASVs)		Cumulative Relative Abundance (%)	
	16S rRNA - V4	ITS1	16S rRNA - V4	ITS1
1 species	5257 (64%)	1221 (58%)	6%	9%
2-4 species	2422 (30%)	687 (33%)	10%	16%
5-9 species	378 (4.6%)	121 (5.8%)	23%	8%
10-19 species	119 (1.5%)	47 (2.3%)	21%	20%
20-31 species	13 (0.2%)	16 (0.8%)	41%	47%

Fig. 4 Abundance–occupancy curves for (a) the most abundant prokaryotic phyla (Bacteria and Archaea, 16S ribosomal (r)RNA gene-V4 dataset) representing > 0.5% of the reads in the dataset and (b) most abundant fungal classes (internal transcribed spacer (ITS)1 region dataset). Each point represents an amplicon sequence variant (ASV) and the points are coloured based on the number of plant species in which they were detected. In the different panels are indicated the ASV richness of each prokaryotic phylum or fungal class. (c) Summary table presenting the number and percentage of ASVs detected in one to 20–31 plant species. Note that on (a) and (b), both axes are presented on a log-scale.

adapted to accurately capture archaeal diversity (Taffner *et al.*, 2020).

For Fungi, ASVs were distributed across 32 classes but 11 classes were dominant in terms of relative abundance (Fig. 4b) and collectively represented 98% of the reads in the dataset and 87% of the ASV richness (1827 ASVs). In particular, three classes were highly dominant: Dothideomycetes (53% of reads, 26% of ASVs), Sordariomycetes (12% of reads, 20% of ASVs) and Tremellomycetes (11% of reads, 13% of ASVs), that collectively represented 76% of the reads in the dataset and 59% of the ASV richness (1227 ASVs; Fig. 4b).

For each ASV, we calculated the number of plant species in which they were detected to identify highly prevalent taxa that could be considered as seed core taxa (Fig. 4). The vast majority of ASVs were detected in only one plant species (64% for prokaryotic ASVs and 58% for fungal ASVs), generally were detected in a few samples and not abundant (i.e. rare taxa; Fig. 4a–c). On the contrary, 13 prokaryotic ASVs and 16 fungal ASVs were detected in > 20 different plant species, and they were all extremely abundant in seed microbiota, presenting collectively a relative abundance of 41% and 47% (Fig. 4a–c). Based on the arbitrary criteria of a detection in ≥ 20 plant species, these taxa could be considered as core taxa as they are both prevalent and abundant in seed microbiota from diverse plants and countries.

In summary, seven microbial clades (Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria, Dothideomycetes, Sordariomycetes and Tremellomycetes) dominate seed microbiota and these communities appear to be structured in two main fractions: a core microbiota composed of few ASVs that are abundant and present in most samples (i.e. dominant taxa), and a flexible microbiota composed of a high diversity of ASVs but with an extremely variable composition between samples and a low abundance (i.e. rare taxa).

Identification of a seed core microbiota across all plant species Next, we characterized the taxonomy, prevalence and relative abundance of each seed core ASV across all samples (Fig. 5). In the 16S-V4 dataset, the 13 core bacterial ASVs were predominantly Proteobacteria ($n = 11$; Fig. 5a). One core bacterial ASV affiliated to *Pantoea* was highly dominant (16.9% of all

Fig. 5 Identification of bacterial and fungal core taxa of seed microbiota across all plant species. Distribution and relative abundance of amplicon sequence variants (ASVs) present in > 20 plant species (> 12 species for the *gyrB* dataset) across all seed samples of the meta-analysis. (a) For the V4 region of the 16S ribosomal (r)RNA gene dataset, 13 bacterial ASVs were identified as core taxa. (b) For the *gyrB* dataset, 18 bacterial ASVs were identified. (c) For the internal transcribed spacer (ITS)1 region dataset, 16 fungal ASVs were identified. For each core ASV identified, the prevalence (% of samples), relative abundance in the dataset (% of reads) and number of plant species (no. of plant species) in which the ASV is detected are presented in a table. The ridges represent the relative abundance of each ASV in the different samples of the dataset. The ASVs are ordered based on their prevalence (most prevalent ASV on top).

reads) and was detected in 68.5% of samples collected from 27 plant species (Fig. 5a). The other core bacterial ASVs were affiliated to *Pseudomonas* ($n = 5$), *Rhizobium* ($n = 2$), *Sphingomonas*, *Methylobacterium* and *Paenibacillus*. The 16S rRNA gene markers facilitated a taxonomic affiliation only at the genus level but the *gyrB* marker enabled the refinement of the taxonomy of core taxa at the species level. For the *gyrB* dataset, we identified 18 bacterial ASVs present in > 12 plant species (lower plant diversity available than in the 16S rRNA gene dataset; Figs 1b, 5b).

Therefore, the dominant *Pantoea* ASV was affiliated to the *P. agglomerans* species and was composed of three distinct ASVs, representing different populations. Interestingly, one of the three *P. agglomerans* ASVs was extremely dominant with a prevalence of 88% and a relative abundance of 24% (Fig. 5b). The *gyrB* dataset also indicated that *Pseudomonas* core ASVs were affiliated to four distinct species: *P. viridiflava* (three ASVs) and three species belonging to the *P. fluorescens* subgroup (Hesse *et al.*, 2018): *P. fluorescens* (three ASVs), *P. poae* and *P. antarctica*. Additional core taxa were identified in the *gyrB* dataset that were not detected in the 16S rRNA gene dataset, such as ASVs affiliated to *Cutibacterium acnes*, *Erwinia persicina* and *Stenotrophomonas* sp. (Fig. 5a,b). In both 16S and *gyrB* datasets, we observed that core ASVs affiliated to *Rhizobium* were present in a high number of samples but at a low abundance (except in rice).

In the ITS1 dataset, we identified 16 core fungal ASVs affiliated to the *Dothideomycetes* (nine ASVs), *Tremellomycetes* (four ASVs), *Microbotryomycetes* (one ASV), *Sordariomycetes* (one ASV) and *Leotiomycetes* (one ASV) classes (Fig. 5c). The core fungal community was dominated by four extremely prevalent ASVs (> 77% of samples, presence in > 28 plant species) with a *Cladosporium perangustum* ASV being the most abundant (12% of reads), followed by an *Alternaria metachromatica* ASV (8.2%), an unclassified *Capnodiales* ASV (7.4%) and an *Alternaria* sp. (4.5%). The other core fungal ASVs were affiliated to the diverse genera *Filobasidium*, *Vishniacozyma*, *Sporobolomycetes*, *Epicoccum*, *Aureobasidium* and *Gibberella* (Fig. 5c). The full list of core and flexible taxa across all plant species is available in Dataset S2.

Our results suggest that the seed core microbiota is less variable on the fungal side than on the bacterial side. Based on the threshold of a detection in 20 plant species, the core fungal ASVs presented a highest prevalence (range 49–85% for fungi and 18–69% for 16S bacteria) and were detected in a higher number of plant species (31 of 32 for fungi and 27 of 43 for 16S bacteria). The higher variability in the distribution of core bacterial ASVs can be visualized in the ridge plots representing the relative abundance of each ASV in all the samples of the dataset (Fig. 5a vs c).

Part 2: Specific patterns associated with each plant species

Plant-specific patterns in community composition In order to explore the influence of plant species on seed microbiota composition, we performed ordinations based on Bray–Curtis dissimilarities between samples. On the 16S-V4 dataset, the plant species was a driver of seed prokaryotic community composition (Fig. 6a). However, there was not a strong clustering by plant species and we observed a high variation in community composition of seeds even if they were harvested from the same plant species (e.g. bean or rapeseed). On a subset of four plant species (bean, radish, rapeseed, rice), we found that the studied plant species significantly explained 22% of the variance in seed community structure (Fig. S5A). The influence of plant species was stronger on the fungal community with a clearer clustering of samples (Fig. 6b). For instance, a separation between bean and rapeseed samples can be observed along the PCoA axis 1 and these two species are separated from radish samples along the PCoA axis 2. When analyzing only these three plant species (bean, radish, rapeseed) with the highest number of observations, we found that they significantly explained 30% of the variance in seed community structure (Fig. S5B). However, in many cases an important overlap was observed between samples from different plant species, suggesting that they share similar community composition.

We also assessed the influence of the plant family on seed community structure. We found a significant but lower signal than plant species, with 11% of the variance explained for the prokaryotic community and 17% for the fungal community (Fig. S6D, E).

Additionally, we explored the influence of the country of origin of the seeds on microbiota structure but no clear clustering by country could be observed for both 16S-V4 and ITS1 datasets (Fig. S14).

For the 16S-V4 and *gyrB* datasets, we also performed metagenome predictions based on KEGG orthologues (KOs) using PICRUST2 to determine if we could observe a structuring of seed microbiota based on their putative functional profiles. These metagenome predictions showed a limited effect of plant species (Fig. S15) on seed microbial community composition data with no clear clustering of samples by plant species and a large variability in functional profiles between samples. The statistical analyses on a subset of well sampled plant species and families indicated that plant species explained 13% of the variance (Fig. S5C) and plant families explained 12% of the variance of predicted metagenomes of the 16S rRNA gene-V4 dataset (Fig. S6C).

We next explored if specific patterns at the phylum or class levels existed by plant species (Fig. 6c,d). At the bacterial phylum level, we observed that Proteobacteria were dominant in the majority of plant species, with the exception of several Brassicaceae species that had a high abundance of Firmicutes (e.g. *A. thaliana*, *Erophila verna*; Fig. 6c). Bean also presented a significantly higher relative abundance of Firmicutes, compared to radish, rapeseed and rice (Fig. S5F). Actinobacteria relative abundance was significantly higher in Poaceae plants (Figs 6c, S6F). Most fungal communities were dominated by taxa from the Dothideomycetes class, followed by Tremellomycetes that were present in almost all plant species with sometimes a high relative abundance (e.g. carrot, *Cardamine hirsuta*, grass of Parnassus; Fig. 6d). The differences observed in fungal microbiota structure between bean, radish and rapeseed (Fig. 6b) were driven mainly by large differences in the contribution of multiple classes: Dothideomycetes (very high in radish, low in bean), Leotiomycetes and Tremellomycetes (high in rapeseed), Mortierellomycetes, Eurotiomycetes and Sordariomycetes (high in bean; Fig. S5F).

Identification of a seed core and flexible microbiota specific to each plant species Next, we identified the seed core and flexible microbiota specific to some plant species for which a sufficient number of independent studies ($n \geq 3$) and of samples ($n \geq 50$) were available for a robust analysis. Only four plant species responded to these criteria for the 16S rRNA gene-V4 dataset (bean, radish, rapeseed and rice) and three plant species for the *gyrB* and ITS1 datasets (bean, radish, rapeseed). To identify the core taxa specific to these plant species, we used the arbitrary criteria that an ASV should be detected in a minimum of two independent studies and be present in at least half of the samples (minimum prevalence 50%). Consistent with the core taxa analysis across all samples (Part 1; Fig. 4), we found that the core taxa identified were both extremely prevalent and abundant on seeds as this can be seen on abundance–occupancy curves (Fig. 7a–c). The core bacterial microbiota identified with the 16S rRNA gene-V4 dataset was composed of a small number of ASVs ranging from six ASVs for bean to 15 for radish (Fig. 7d). Still, these few ASVs collectively represented a large fraction of the entire bacterial community with a cumulative relative abundance of 28% (bean) to 85% (radish; Fig. 7d). The four plant species shared a similar core microbiota affiliated to the *Pantoea* and *Pseudomonas* genera but some core ASVs were plant-specific, like *Bacillus* for bean, *Sphingomonas* for rapeseed and rice, or *Rhizobium* and *Xanthomonas* for rice and radish (Fig. 7g). The core bacterial microbiota identified with the *gyrB* dataset presented similar patterns with the 16S rRNA gene dataset (Fig. 7b,e,h). The higher taxonomic resolution of the *gyrB* marker enabled us to identify the core ASVs at the species level (Fig. 7h). The core bacterial microbiota was dominated by two *P. agglomerans* ASVs and *E. persicina* (not identified in the 16S rRNA gene dataset) for the three plant species. Rapeseed and radish seeds also possessed two core ASVs affiliated to *P. viridiflava*. Two *P. fluorescens* core ASVs and a *Serratia marcescens* ASV were specific to rapeseed plants, whereas radish plants exhibited two specific *P. agglomerans*

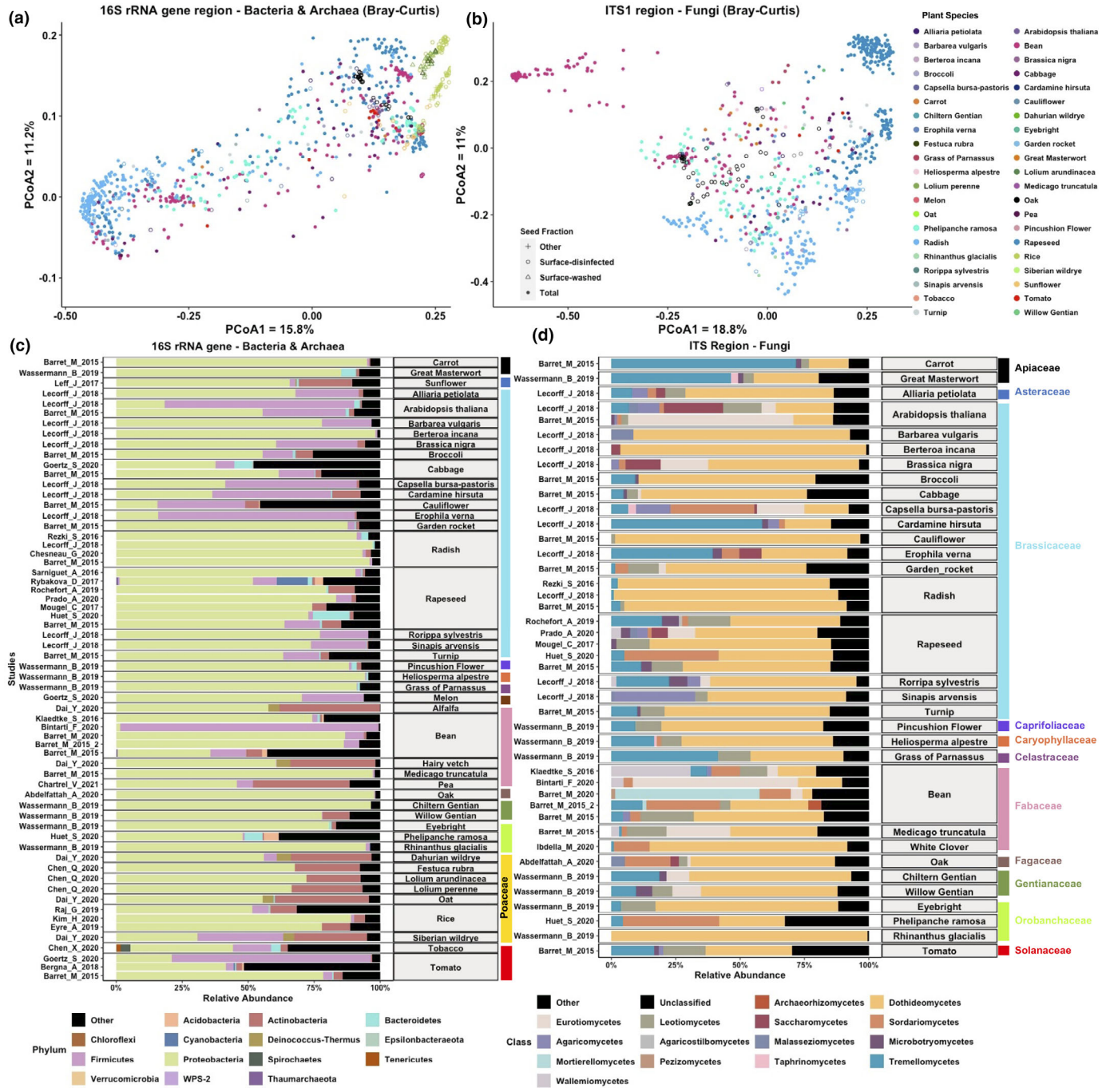


Fig. 6 Plant-specific patterns in seed microbiota composition. Ordinations to explore the influence of plant species on seed community structure on (a) the 16S ribosomal (r)RNA gene-V4 dataset and (b) the internal transcribed spacer (ITS)1 region dataset. Average relative abundance of (c) major bacterial phyla (16S-V4 dataset) and (d) fungal classes (ITS1 dataset) by study for the different plant species. Species are ordered by plant family from Apiaceae at the top to Solanaceae at the bottom.

ASVs and a *Pseudomonas syringae* ASV. The core fungal microbiota identified with the ITS1 dataset generally was more diverse than its bacterial counterpart, with nine ASVs (bean) to 30 ASVs (rapeseed; Fig. 7c,f). These core ASVs represented an important fraction of the fungal community with a cumulative relative abundance of 35% (bean) to 87% (rapeseed; Fig. 7f). The three plant species shared a similar core microbiota affiliated to *Cladosporium* but many core ASVs were plant-specific, such as *Mortierella* or *Fusarium* for bean, *Alternaria* or *Filobasidium* for rapeseed and radish (Fig. 7i).

This analysis also enabled us to characterize the diversity and relative abundance of the flexible seed microbiota that varied widely between plant species. Bean seeds presented the most flexible microbiota representing > 65% of the relative abundance and a high ASV diversity (12 135 ASVs for 16S-V4 and 1026 ASVs for ITS1). Radish seeds had the least flexible microbiota,

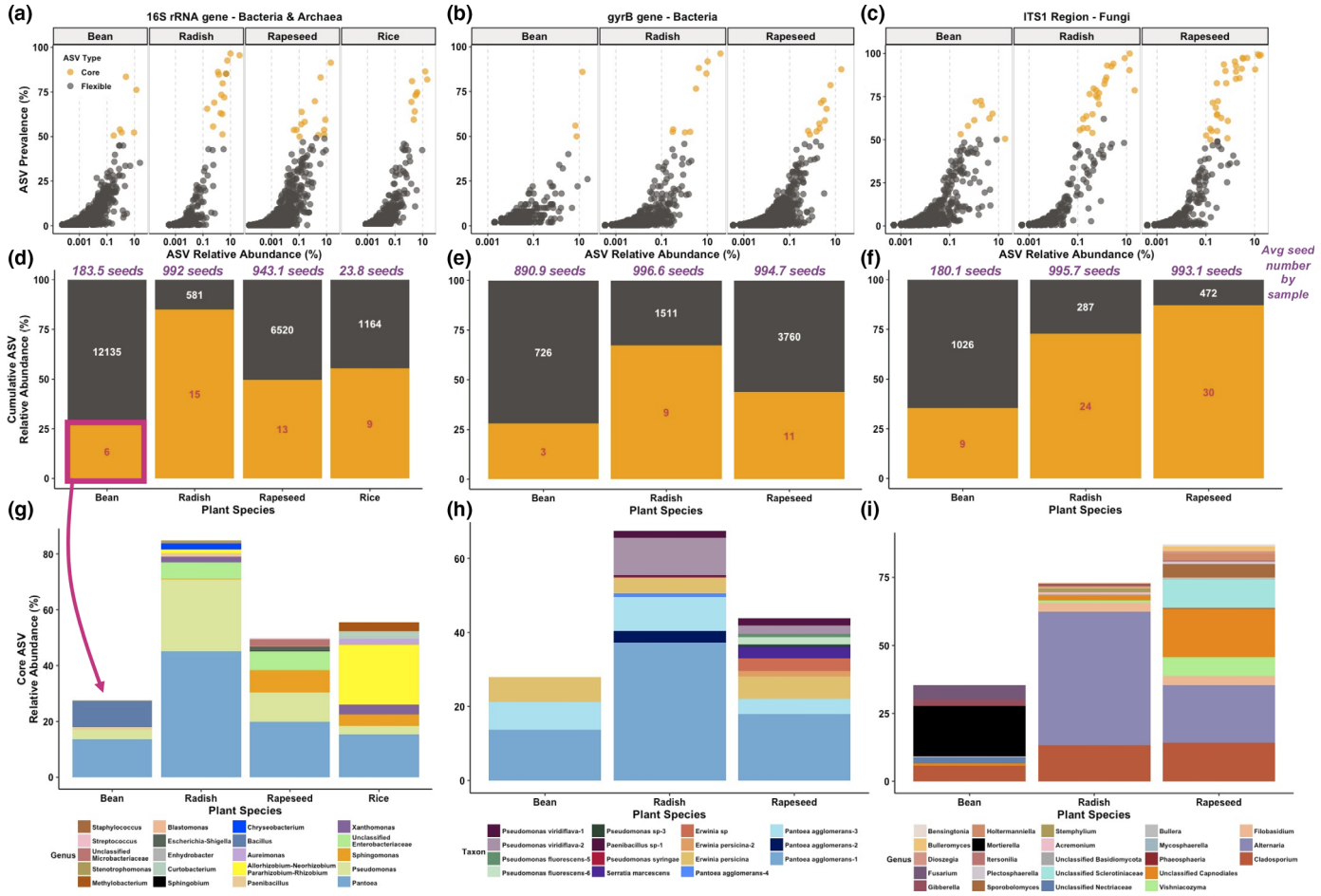


Fig. 7 Core and flexible seed microbiota by plant species. The top panels (a–c) represent the abundance-occupancy curves of all amplicon sequence variants (ASVs) by plant species for the three different marker genes (left, 16S ribosomal (r)RNA-V4 gene; middle, *gyrB*; right, internal transcribed spacer (ITS)1 region). This analysis was performed on four species for the 16S-V4 dataset and three species for the other two datasets based on the criterion of three independent studies available. Core ASVs (i.e. detection in a minimum of two studies and ≥ 50% prevalence) are highlighted in yellow. The middle panels (d–f) display the cumulative relative abundance of core and flexible ASVs by plant species. The number inside each bar indicates the number of ASVs for each group (in red for core ASVs, in white for flexible ASVs). The purple numbers on top of the graphs indicate the average number of seeds present in the seed samples of the different plant species to characterize the microbiota composition. The bottom panels (g–i) represent the taxonomic composition of the core microbiota of each plant species, at the genus level for the 16S-V4 and ITS1 datasets and at the species level for the *gyrB* dataset.

probably because most of the studies available originated from our research group and a single variety (Flamboyant5). The full list of core and flexible taxa by plant species is available in Dataset S2.

These results confirm that seed microbiota share many similarities across plant species, especially on the bacterial side with no clear clustering by plant species and several core taxa shared. For the fungal community, we observed more dissimilarities between plant species, especially with some compositional differences between bean, radish and rapeseed seeds.

Discussion

Seed microbiota are diverse and extremely variable in the number of microbial taxa capable of colonizing seed samples, from one to thousands of Amplicon Sequence Variants (ASVs), with a median of a hundred ASVs (prokaryotes + fungi). This variability in the assembly of seed microbiota has been reported by different studies and can be observed even at the single-seed level (e.g. 2–28 bacterial ASVs per oak seed or 1–45 bacterial ASVs per bean seed) (Abdelfattah *et al.*, 2021; Chesneau *et al.*, 2021). However, the variability observed in this meta-analysis also could be partly methodological because of the variable number of seeds used in different samples to characterize seed microbiota (from one to 1000 seeds).

The seed taxa originate primarily from four bacterial phyla (Proteobacteria, Actinobacteria, Firmicutes and Bacteroidetes) and three fungal classes (Dothideomycetes, Sordariomycetes and Tremellomycetes), which is similar to other plant compartments (Trivedi *et al.*, 2020). However, our findings show that seeds generally present similar proportions of bacterial and fungal diversity, which contrasts with other plant compartments that generally are highly dominated by bacterial diversity (Hacquard *et al.*, 2015; Simonin *et al.*, 2020; Trivedi *et al.*, 2020). This might be a consequence of the specific characteristics of the seed habitat (high desiccation, low resource availability, high pressure)

compared to other plant compartments that may favour the survival of a great diversity of fungal taxa with adapted structures (e.g. spores). This meta-analysis highlights that seed microbiota present stable (i.e. core) and variable (i.e. flexible) microbial fractions across samples, as already hypothesized for the plant microbiome (Vandenkoornhuyse *et al.*, 2015) and observed in other microbial ecosystems (Hernandez-Agreda *et al.*, 2016; Björk *et al.*, 2018).

We observed a significant influence of the plant species and plant botanical family on the diversity and structure of seed microbiota, on a subset of well-characterized plants. These effects were stronger on the fungal community, with clear variations in the relative abundance of fungal orders between Brassicaceae (dominated by Dothideomycetes) and Fabaceae. For the bacterial community, Poaceae seeds were significantly enriched in Actinobacteria, whereas bean seeds had a higher relative abundance of Firmicutes compared to other species. Future studies should aim to understand the mechanisms responsible for these different assemblies of seed microbiota between different plant species, whether they depend on the physicochemical composition of the seeds, the contribution of vertical transmission of taxa from the mother plants, or the influence of different types of pollination and reproductive structures.

An interesting result of this meta-analysis is that approximately 30 bacterial and fungal taxa are shared between highly contrasting plant species and detected in samples from all over the world. In particular, core ASVs affiliated to the *Pantoea*, *Pseudomonas*, *Sphingomonas*, *Cladosporium* and *Alternaria* genera appear as extremely abundant and prevalent seed-borne microorganisms. Several of these seed-borne taxa have already been identified to promote plant fitness, such as seed endophytes *Cladosporium* favouring the establishment of root symbiosis (Ridout *et al.*, 2019) or *Sphingomonas* conferring resistance to a bacterial disease (Matsumoto *et al.*, 2021). It is worth noting that this study confirms that arbuscular mycorrhizal fungi are not frequently transmitted to seedlings through seeds, with very few Glomeromycetes ASVs detected and at a very low prevalence (four ASVs, 0.3% of seed samples) (M. Chen *et al.*, 2018). By contrast, two *Rhizobium* ASVs were identified as core ASVs across all samples with high prevalence and generally low abundances, indicating that transmission through seeds is probably a prevalent pathway for these potential nitrogen-fixing bacteria (Mora *et al.*, 2014). The core taxa identified represent a short-list of microbial taxa of interest to investigate their potential role in plant fitness, such as seed maturation, disease tolerance, germination or emergence, and to identify their environmental origin (horizontal or vertical transmission). It should be kept in mind that the core taxa listed here were identified in samples composed of multiple seeds ($\leq$1000 seeds). This methodological constraint may overestimate the prevalence of taxa and it is likely that the core taxa are not present in every single seed analyzed or that various cooccurrence patterns exist between these core taxa.

We encourage future studies to attempt to isolate and characterize core taxa using genomics, synthetic community reconstruction and thorough plant phenotyping to identify their adaptation to the seed habitat and their potential selection by plant hosts. Still, the majority of seed microbiota belongs to the 'flexible' fraction that is reflective of the abiotic and biotic fluctuations of the seed environment. These taxa also are adapted to the seed habitat but generally harbour lower abundances and prevalence. This fraction should not be neglected as it represents an important diversity reservoir that can be transmitted to seedlings (Rochefort *et al.*, 2021; Walsh *et al.*, 2021) and provides microbial taxa that are adapted to specific local constraints (Hernandez-Agreda *et al.*, 2016).

This meta-analysis not only provides new insights on seed microbiota diversity and composition, but also highlights key knowledge gaps. This study gathered data on only 50 plant species (including 30 crops), with often only one study by plant species and few seed samples analyzed. This means that the seed microbiota of a majority of crops and natural plant communities have not been characterized. More investigations also are required to assess the influence of geographical sites and plant genotypes on seed microbiota. Additionally, this meta-analysis included mainly seed microbiota characterization from multiple seeds (> 3 seeds in sample) with > 1000 reads. More work is needed at the single-seed level (Abdelfattah *et al.*, 2021; Bintarti *et al.*, 2021; Chesneau *et al.*, 2021) to assess seed-to-seed variation and the proportion of seeds that do not have detectable microbiota (i.e. sterile seeds) (Newcombe *et al.*, 2018; Ridout *et al.*, 2019). Moreover, seeds can vary greatly in their chemical composition, size, anatomic features and physiology, yet we still do not know how these seed attributes influence microbiome composition. Unfortunately, the diverse data gathered in this meta-analysis cannot robustly address this question and future studies specifically designed to characterize the links between seeds traits and microbiota are needed. Another major knowledge gap in the seed microbiota literature regards the characterization of microorganisms other than bacteria and fungi, especially Oomycetes and other protists, Archaea and viruses. The most common primers used for amplicon sequencing of seed microbiota completely miss these taxonomic groups (Oomycetes, protists, viruses) or partially characterize them (Archaea), despite knowledge on their presence on seeds from cultivation-based approaches and phytopathological studies (Sastry, 2013; Thines, 2014; Taffner *et al.*, 2020). To allow the Seed Microbiota Database to grow in the future, we encourage new studies to perform multi-marker analyses on a minimum of five replicate seed samples and to deposit their data on public databases with detailed metadata to allow reusability. These future studies will help in gaining a more comprehensive view of seed microbiota and will accelerate discoveries using seeds as a vector of agricultural innovations, especially for plant microbiome engineering.

Acknowledgements

We are deeply grateful to the authors whose data contributed to this database and meta-analysis. Without their efforts to make their datasets available with good metadata, this meta-analysis would not have been possible. We thank all our collaborators who contributed unpublished datasets to this study, in particular Josiane Le Corff, Muriel Marchi and Marie-Anne Le Moigne for the Endowild project (Le Corff_J_2018 study, IRHS lab,

University of Angers, INRAE), Simon Goertz (Goertz_S_2020 study, NPZ Innovation GmbH), Christophe Mougel for the Brassica Div Patho project (Mougel_C_2017 study, IGEPP lab, INRAE) and Brice Marolleaux for the Seed Microbiome project (Barret_M_2020 study, IRHS lab, INRAE). This research was conducted as part of the SUCSEED project (ANR-20-PCPA-0009) funded by the French ANR as part of the PPR-CPA program and in the framework of the OSMOSE project funded by regional program 'Objectif Végétal, Research, Education and Innovation in Pays de la Loire', supported by the French Region Pays de la Loire, Angers Loire Métropole and the European Regional Development Fund. The Seed Microbiome project (Barret_M_2020 study) was funded by USDA 2019-67019-29305, the Brassica Div Patho project was funded by the P10037 INRAE – Metaomics and Ecosystems Metaprogram, Plant Health division (Mougel_C_2017 study). The RhizoSeed project (Sarniguet_A_2016 study) was funded by the INRAE – Metaomics and Ecosystems Metaprogram, Plant Health division.

Author contributions

MS, M Barret and AS designed the study; CM, GC and AR performed the experiments on several studies included in the meta-analysis; MS, M Briand and M Barret analyzed the datasets and created the database; and MS prepared the figures and wrote the paper, and all co-authors reviewed and edited the manuscript.

ORCID

Matthieu Barret https://orcid.org/0000-0002-7633-8476
Martial Briand https://orcid.org/0000-0002-5822-2824
Alain Sarniguet https://orcid.org/0000-0001-6232-0200
Marie Simonin https://orcid.org/0000-0003-1493-881X

Data availability

The Seed Microbiota Database is available on the Data INRAE portal (10.15454/2ANNJM). The database includes all the metadata tables, ASV tables, taxonomy tables, FASTA files and phylogenetic trees associated to the five datasets corresponding to the different molecular markers. Two scripts are available to query the database: (1) the 'getASVInfosInDb.pl' to search the database for a specific ASV using its sequence; (2) the 'getASVFoundOnSpecies.pl' to get the list of all ASVs associated with a plant species. More details are available on the README file in the repository. The code and files to reproduce the bioinformatic analyses and figures are available at https://github.com/marie-simonin/Seed-Microbiota-Database. The list of accession numbers (e.g. NCBI-SRA, ENA) to download the raw data of the studies included in the database is available in the Dataset S1 – Study description.

References

Abdelfattah A, Wisniewski M, Schena L, Tack AJ. 2021. Experimental evidence of microbial inheritance in plants and transmission routes from seed to phyllosphere and root. *Environmental Microbiology* **23**: 2199–2214.

Bakker MG, McCormick SP. 2019. Microbial correlates of *Fusarium* load and deoxynivalenol content in individual wheat kernels. *Phytopathology* **109**: 993–1002.

Barret M, Briand M, Bonneau S, Préveaux A, Valière S, Bouchez O, Hunault G, Simoneau P, Jacques M-A. 2015. Emergence shapes the structure of the seed microbiota. *Applied and Environmental Microbiology* **81**: 1257–1266.

Ben-Jabeur M, Kthiri Z, Harbaoui K, Belguesmi K, Serret MD, Araus JL, Hamada W. 2019. Seed coating with thyme essential oil or *Paraburkholderia phytofirmans* PsJN strain: conferring septoria leaf blotch resistance and promotion of yield and grain isotopic composition in wheat. *Agronomy* **9**: 586.

Berg G, Raaijmakers JM. 2018. Saving seed microbiomes. *The ISME Journal* **12**: 1167–1170.

Bergna A, Cernava T, Rändler M, Grosch R, Zachow C, Berg G. 2018. Tomato seeds preferably transmit plant beneficial endophytes. *Phytobiomes Journal* **2**: 183–193.

Bertolini E, Teresani GR, Loiseau M, Tanaka FAO, Barbé S, Martínez C, Gentit P, López MM, Cambra M. 2015. Transmission of '*Candidatus* Liberibacter solanacearum' in carrot seeds. *Plant Pathology* **64**: 276–285.

Bintarti AF, Kearns PJ, Sulesky A, Shade A. 2020. Abiotic treatment to common bean plants results in an altered seed microbiome. *bioRxiv*. doi: 10.1101/2020.06.05.134445.

Bintarti AF, Sulesky-Grieb A, Stopnisek N, Shade A. 2021. Endophytic microbiome variation among single plant seeds. *Phytobiomes Journal* **6**: 45–55.

Björk JR, O'Hara RB, Ribes M, Coma R, Montoya JM. 2018. The dynamic core microbiome: structure, dynamics and stability. *bioRxiv*. doi: 10.1101/137885.

Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory CJ. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**: 90.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F *et al.* 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**: 852–857.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**: 581–583.

Caruso V, Song X, Asquith M, Karstens L. 2019. Performance of microbiome sequence inference methods in environments with varying biomass. *mSystems* **4**: e00163-18.

Chartrel V, Dugat-Bony E, Sarthou A-S, Huchette S, Bonnarme P, Irlinger F. 2021. The microbial community associated with pea seeds (*Pisum sativum*) of different geographical origins. *Plant and Soil* **462**: 405–427.

Chen H, Wu H, Yan B, Zhao H, Liu F, Zhang H, Sheng Q, Miao F, Liang Z. 2018. Core microbiome of medicinal plant salvia miltiorrhiza seed: a rich reservoir of beneficial microbes for secondary metabolism? *International Journal of Molecular Sciences* **19**: 672.

Chen M, Arato M, Borghi L, Nouri E, Reinhardt D. 2018. Beneficial services of arbuscular mycorrhizal fungi – from ecology to application. *Frontiers in Plant Science* **9**: 1270.

Chen Q, Meyer WA, Zhang Q, White JF. 2020. 16S rRNA metagenomic analysis of the bacterial community associated with turf grass seeds from low moisture and high moisture climates. *PeerJ* **8**: e8417.

Chen X, Krug L, Yang H, Li H, Yang M, Berg G, Cernava T. 2020. *Nicotiana tabacum* seed endophytic communities share a common core structure and genotype-specific signatures in diverging cultivars. *Computational and Structural Biotechnology Journal* **18**: 287–295.

Chesneau G, Laroche B, Preveaux A, Marais C, Briand M, Marolleau B, Simonin M, Barret M. 2021. Single seed microbiota: assembly and transmission from parent plant to seedling. *bioRxiv*. doi: 10.1101/2021.05.31.446402.

Chesneau G, Torres-Cortes G, Briand M, Darrasse A, Preveaux A, Marais C, Jacques M-A, Shade A, Barret M. 2020. Temporal dynamics of bacterial communities during seed development and maturation. *FEMS Microbiology Ecology* **96**: fiaa190.

Dai Y, Li X-Y, Wang Y, Li C-X, He Y, Lin H-H, Wang T, Ma X-R. 2020. The differences and overlaps in the seed-resident microbiome of four Leguminous and three Gramineous forages. *Microbial Biotechnology* **13**: 1461–1476.

Darrasse A, Darsonval A, Boureau T, Brisset M-N, Durand K, Jacques M-A. 2010. Transmission of plant-pathogenic bacteria by nonhost seeds without induction of an associated defense reaction at emergence. *Applied and Environmental Microbiology* **76**: 6787–6796.

Darsonval A, Darrasse A, Durand K, Bureau C, Cesbron S, Jacques M-A. 2009. Adhesion and fitness in the bean phyllosphere and transmission to seed of *Xanthomonas fuscans* subsp. *fuscans*. *Molecular Plant–Microbe Interactions* **22**: 747–757.

Doherty JR, Crouch JA, Roberts JA. 2020. Elucidating the influence of resident seed and soil microbiota on the developing creeping bentgrass microbiome. *Agrosystems, Geosciences & Environment* **3**: e20038.

Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MG. 2020. PICRUST2 for prediction of metagenome functions. *Nature Biotechnology* **38**: 685–688.

Escobar Rodríguez C, Mitter B, Antonielli L, Trognitz F, Compant S, Sessitsch A. 2018. Roots and panicles of the C4 model grasses *Setaria viridis* (L). and *S. pumila* host distinct bacterial assemblages with core taxa conserved across host genotypes and sampling sites. *Frontiers in Microbiology* **9**: 2708.

Eyre AW, Wang M, Oh Y, Dean RA. 2019. Identification and characterization of the core rice seed microbiome. *Phytobiomes Journal* **3**: 148–157.

Fels-Klerx HJVD, Klemsdal S, Hietaniemi V, Lindblad M, Ioannou-Kakouri E, Asselt EDV. 2012. Mycotoxin contamination of cereal grain commodities in relation to climate in North West Europe. *Food Additives & Contaminants: Part A* **29**: 1581–1592.

Fitzpatrick CR, Salas-González I, Conway JM, Finkel OM, Gilbert S, Russ D, Teixeira PJPL, Dangl JL. 2020. The plant microbiome: from ecology to reductionism and beyond. *Annual Review of Microbiology* **74**: 81–100.

Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, McHardy A, Dangl J, Knight R, Ley R *et al.* 2015. Microbiota and host nutrition across plant and animal kingdoms. *Cell Host & Microbe* **17**: 603–616.

Hernandez-Agreda A, Leggat W, Bongaerts P, Ainsworth TD. 2016. The microbial signature provides insight into the mechanistic basis of coral success across reef habitats. *mBio* **7**: e00560-16.

Hesse C, Schulz F, Bull CT, Shaffer BT, Yan Q, Shapiro N, Hassan KA, Varghese N, Elbourne LDH, Paulsen IT *et al.* 2018. Genome-based evolutionary history of *Pseudomonas* spp. *Environmental Microbiology* **20**: 2142–2159.

Huet S, Pouvreau J-B, Delage E, Delgrange S, Marais C, Bahut M, Delavault P, Simier P, Poulin L. 2020. Populations of the parasitic plant *Phelipanche ramosa* influence their seed microbiota. *Frontiers in Plant Science* **11**: 1075.

Idbella M, Bonanomi G, De Filippis F, Amor G, Chouyia FE, Fechtali T, Mazzoleni S. 2021. Contrasting effects of *Rhizophagus irregularis* versus bacterial and fungal seed endophytes on *Trifolium repens* plant-soil feedback. *Mycorrhiza* **31**: 103–115.

Kim H, Lee KK, Jeon J, Harris WA, Lee Y-H. 2020. Domestication of *Oryza* species eco-evolutionarily shapes bacterial and fungal communities in rice seed. *Microbiome* **8**: 1–17.

Klaedtke S, Jacques M-A, Raggi L, Préveaux A, Bonneau S, Negri V, Chable V, Barret M. 2016. Terroir is a key driver of seed-associated microbial assemblages. *Environmental Microbiology* **18**: 1792–1804.

Lahti L, Shetty S, Blake T. 2017. *Tools for microbiome analysis in R.* MICROBIOME package v.0.99 88: 2012–2017.

Leck MA, Parker VT, Simpson RL, Simpson RS. 2008. *Seedling ecology and evolution*. Cambridge, UK: Cambridge University Press.

Liu Y, Xu P, Yang F, Li M, Yan H, Li N, Zhang X, Wang W. 2019. Composition and diversity of endophytic bacterial community in seeds of super hybrid rice 'Shenliangyou 5814' (*Oryza sativa* L.) and its parental lines. *Plant Growth Regulation* **87**: 257–266.

Matsumoto H, Fan X, Wang Y, Kusstatscher P, Duan J, Wu S, Chen S, Qiao K, Wang Y, Ma B *et al.* 2021. Bacterial seed endophyte shapes disease resistance in rice. *Nature Plants* **7**: 60–72.

McMurdie PJ, Holmes S. 2013. PHYLOSEQ: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**: e61217.

Mora Y, Díaz R, Vargas-Lagunas C, Peralta H, Guerrero G, Aguilar A, Encarnación S, Girard L, Mora J. 2014. Nitrogen-fixing rhizobial strains isolated from common bean seeds: phylogeny, physiology, and genome analysis. *Applied and Environmental Microbiology* **80**: 5644–5654.

Nelson EB. 2018. The seed microbiome: origins, interactions, and impacts. *Plant and Soil* **422**: 7–34.

Newcombe G, Harding A, Ridout M, Busby PE. 2018. A hypothetical bottleneck in the plant microbiome. *Frontiers in Microbiology* **9**: 1645.

Nilsson RH, Larsson K-H, Taylor AF S, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L *et al.* 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research* **47**: D259–D264.

O'Callaghan M. 2016. Microbial inoculation of seed for improved crop performance: issues and opportunities. *Applied Microbiology and Biotechnology* **100**: 5729–5746.

Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M. 2007. *VEGAN: community ecology package*, v. 2.5-7 [WWW document] URL https://cran.ism.ac.jp/web/packages/vegan/vegan.pdf [accessed 17 November 2020].

Paredes SH, Lebeis SL. 2016. Giving back to the community: microbial mechanisms of plant–soil interactions. *Functional Ecology* **30**: 1043–1052.

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

Prado A, Marolleau B, Vaissière BE, Barret M, Torres-Cortes G. 2020. Insect pollination: an ecological process involved in the assembly of the seed microbiota. *Scientific Reports* **10**: 3575.

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

Raj G, Shadab M, Deka S, Das M, Baruah J, Bharali R, Talukdar NC. 2019. Seed interior microbiome of rice genotypes indigenous to three agroecosystems of Indo-Burma biodiversity hotspot. *BMC Genomics* **20**: 1–16.

Reeves TG, Thomas G, Ramsay G. 2016. *Save and grow in practice: maize, rice, wheat—a guide to sustainable cereal production*. Rome, Italy: UN Food and Agriculture Organization.

Rezki S, Campion C, Iacomi-Vasilescu B, Preveaux A, Toualbia Y, Bonneau S, Briand M, Laurent E, Hunault G, Simoneau P *et al.* 2016. Differences in stability of seed-associated microbial assemblages in response to invasion by phytopathogenic microorganisms. *PeerJ* **4**: e1923.

Rezki S, Campion C, Simoneau P, Jacques M-A, Shade A, Barret M. 2018. Assembly of seed-associated microbial communities within and across successive plant generations. *Plant and Soil* **422**: 67–79.

Ridout ME, Schroeder KL, Hunter SS, Styer J, Newcombe G. 2019. Priority effects of wheat seed endophytes on a rhizosphere symbiosis. *Symbiosis* **78**: 19–31.

Risely A. 2020. Applying the core microbiome to understand host–microbe systems. *Journal of Animal Ecology* **89**: 1549–1558.

Rocha I, Ma Y, Souza-Alonso P, Vosátka M, Freitas H, Oliveira RS. 2019. Seed coating: a tool for delivering beneficial microbes to agricultural crops. *Frontiers in Plant Science* **10**: 1357.

Rochefort A, Briand M, Marais C, Wagner M-H, Laperche A, Vallée P, Barret M, Sarniguet A. 2019. Influence of environment and host plant genotype on the structure and diversity of the *Brassica napus* seed microbiota. *Phytobiomes Journal* **3**: 326–336.

Rochefort A, Simonin M, Marais C, Guillerm-Erckelboudt A-Y, Barret M, Sarniguet A. 2021. Transmission of seed and soil microbiota to seedling. *mSystems* **6**: e00446-21.

Rodríguez CE, Antonielli L, Mitter B, Trognitz F, Sessitsch A. 2019. Heritability and functional importance of the *Setaria viridis* bacterial seed microbiome. *Phytobiomes Journal* **4**: 40–52.

Rybakova D, Mancinelli R, Wikström M, Birch-Jensen A-S, Postma J, Ehlers R-U, Goertz S, Berg G. 2017. The structure of the *Brassica napus* seed microbiome is cultivar-dependent and affects the interactions of symbionts and pathogens. *Microbiome* **5**: 1–16.

Sastry KS. 2013. Ecology and epidemiology of seed-transmitted viruses. In: *Seed-borne plant virus diseases*. Delhi, India: Springer, 165–183.

Shade A, Handelsman J. 2012. Beyond the Venn diagram: the hunt for a core microbiome. *Environmental Microbiology* **14**: 4–12.

Shade A, Jacques M-A, Barret M. 2017. Ecological patterns of seed microbiome diversity, transmission, and assembly. *Current Opinion in Microbiology* **37**: 15–22.

Shahzad R, Khan AL, Bilal S, Asaf S, Lee I-J. 2018. What is there in seeds? Vertically transmitted endophytic resources for sustainable improvement in plant growth. *Frontiers in Plant Science* **9**: 24.

Simonin M, Dasilva C, Terzi V, Ngonkeu ELM, Diouf D, Kane A, Béna G, Moulin L. 2020. Influence of plant genotype and soil on the wheat rhizosphere microbiome: evidences for a core microbiome across eight African and European soils. *FEMS Microbiology Ecology* **96**: fiaa067.

Taffner J, Bergna A, Cernava T, Berg G. 2020. Tomato-associated archaea show a cultivar-specific rhizosphere effect but an unspecific transmission by seeds. *Phytobiomes Journal* **4**: 133–141.

Thines M. 2014. Phylogeny and evolution of plant pathogenic oomycetes—a global overview. *European Journal of Plant Pathology* **138**: 431–447.

Torres-Cortés G, Bonneau S, Bouchez O, Genthon C, Briand M, Jacques M-A, Barret M. 2018. Functional microbial features driving community assembly during seed germination and emergence. *Frontiers in Plant Science* **9**: 902.

Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. 2020. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18**: 607–621.

Vandenkoornhuyse P, Quaiser A, Duhamel M, Van AL, Dufresne A. 2015. The importance of the microbiome of the plant holobiont. *New Phytologist* **206**: 1196–1206.

Walsh CM, Becker-Uncapher I, Carlson M, Fierer N. 2021. Variable influences of soil and seed-associated bacterial communities on the assembly of seedling microbiomes. *The ISME Journal* **15**: 2748–2762.

Wassermann B, Cernava T, Müller H, Berg C, Berg G. 2019. Seeds of native alpine plants host unique microbial communities embedded in cross-kingdom networks. *Microbiome* **7**: 1–12.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Switzerland: Springer Nature.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Detailed information on the 63 studies included in the meta-analysis.

Dataset S2 List of flexible and core taxa identified in the different datasets.

Fig. S1 Table providing general information on the five marker gene datasets.

Fig. S2 Description of the 16S rRNA gene samples (both V4 and V5–V6 regions).

Fig. S3 Description of the gyrB gene samples.

Fig. S4 Description of the ITS samples.

Fig. S5 Beta- and alpha-diversity analyses on a subset of four plant species.

Fig. S6 Beta- and alpha-diversity analyses on a subset of four plant families.

Fig. S7 Pielou's evenness index of seed microbiota within and across plant species.

Fig. S8 Shannon diversity index of seed microbiota within and across plant species.

Fig. S9 Seed microbiota diversity in function of the seed fraction considered.

Fig. S10 Seed microbiota diversity in function of the seed preparation.

Fig. S11 Relationship between seed number in sample and the Amplicon Sequence Variant richness observed.

Fig. S12 Abundance–occupancy curves of all prokaryotic phyla for the 16S V4 dataset.

Fig. S13 Abundance–occupancy curves of all prokaryotic phyla for the gyrB gene dataset.

Fig. S14 Structuration of seed microbiota based on the country of origin.

Fig. S15 Metagenome prediction of seed microbiota based on KEGG orthologs.

Table S1 Detailed description of all the datasets and subsets used in the meta-analysis.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.