Table of contents

Table of figures

Table of tables

Glossary

Alpha diversity:	The mean species diversity within one habitat
Amplicon Single Variant (ASV):	A sequence obtained through PCR and sequencing process; only a single nucleotide polymorphism is needed to detect two different ASV
Beta diversity:	The mean species diversity among different habitats
Gamma-diversity	The total species diversity in a landscape The gamma-diversity is determined by the alpha and the beta diversity
Operational Taxonomic Unit (OTU):	A taxonomic unit obtained by merging different sequences that have 97% of similarity
Mock community:	An artificially composed bacterial community with known bacterial strain that then undergoes the same amplification, sequencing and data processes than other samples
Rarefaction:	For each sample, reads are randomly chosen between the ones present in that sample and according to their relative abundance until the number of read per sample reach an arbitrary threshold, chosen according to the rarefaction curve

Abbreviations

ASV: Amplicon Single Variant

CFU: Colony Forming Unit

CIRM-CFBP: International Centre of Microbiological Resources – French Collection of Plant Bacteria

INRA: National Institute of the Agronomic Research

IRHS: Horticulture and Seeds Research Institute

OTU: Operational Taxonomic Unit

SFR QUASAV: Research Federative Structure in Plant Quality and Health

UMR: Joint Research Unit

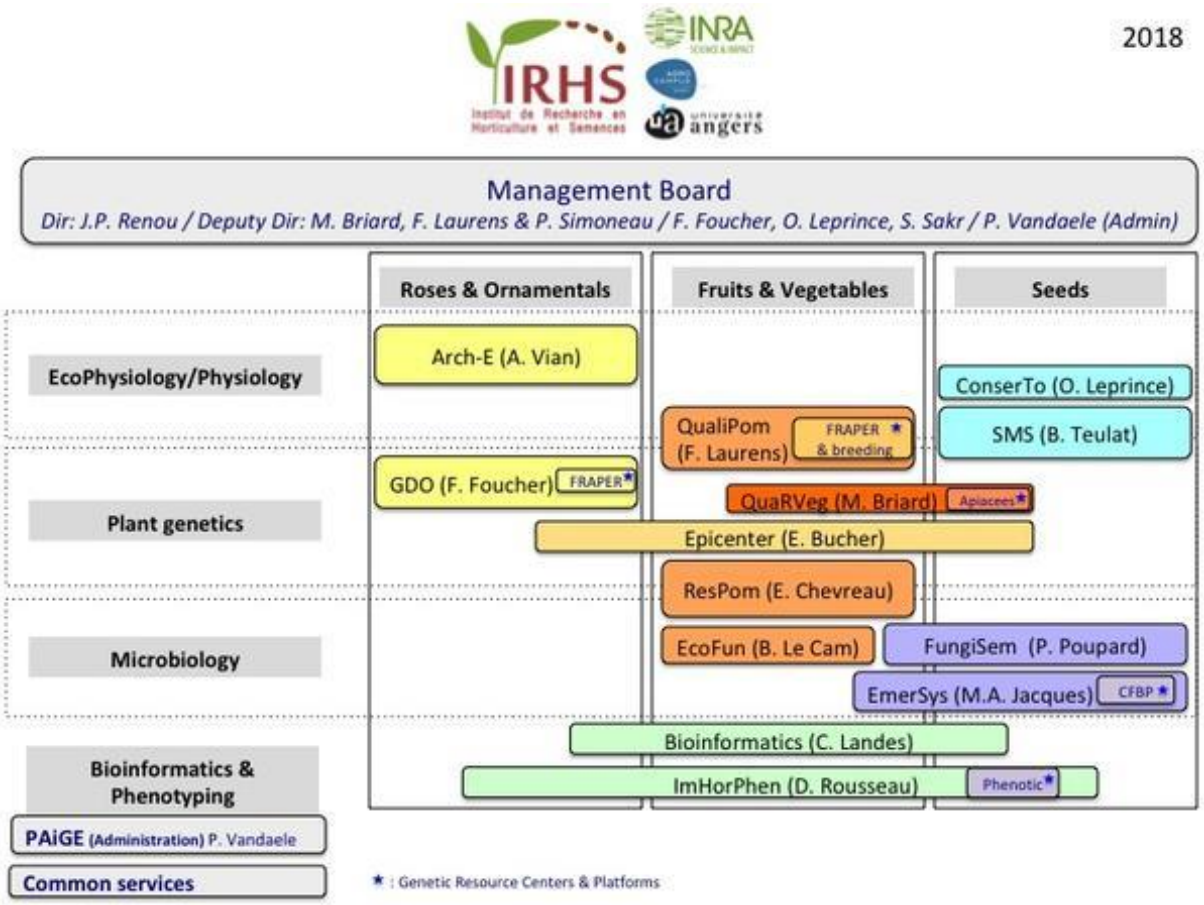


Figure 1: Organization chart of the IRHS (from https://www6.angers-nantes.inra.fr/irhs_eng/The-Research-Institute-on-Horticulture-and-Seeds)

Comparative analysis of bacterial community structure associated to different plant seeds

1. Introduction

1.1. Practicum site

My internship had last two and a half months in the Emersys team at the IRHS (Horticulture and Seeds Research Institute) in Beaucouzé (France). The IRHS is part of a larger structure, the SFR QUASAV (Research Federative Structure in Plant Quality and Health).

The SFR Quasav (Appendix I) combines scientific teams from three different institutions: i) the INRA (National Institute of the Agronomic Research), ii) the engineer school Agrocampus Ouest, and iii) the Angers University. In addition to these three institutions, the SFR Quasav also federate every other plant biology teams from the region of Pays de la Loire. Since 2008, this lab team clustering allows different institutions to aim one federative scientific project and to pool resources. This SFR gathers 380 people including 150 researchers and 60 PhD students. The federative scientific project is split into three research axis: the sustainable management of plant health, the seed biology, quality and health and the horticultural plant product quality. Almost, four technical facilities and two platforms are mutualized (Appendix I).

The IRHS (Figure 1) is one of the joint research unit (UMR) of the SFR Quasav. It is one of the biggest partner of the SFR Quasav with its 220 employees. The IRHS is specially focused on the horticultural plant biology and on the seed production. Moreover, this institute share its technological resources and expertise's between thirteen joint research teams from the INRA, the engineer school Agrocampus Ouest and the Angers University.

Among this united means, the Emersys team cope with the Emergence, systematics and ecology of the pathogenic bacteria, thus its name. This team is currently composed of 22 members, including researchers, lab techs, post-docs, PhD students and interns. In addition, the team includes one bio-informatics member specially dedicated to data processing. Indeed, the ecology of bacteria research field needs a lot of data processing. The team researches are focused on the plant associated bacteria. Three main axis are studied in the group. The first one is the identification of the processes leading plant disease emergence. Within the second one, they study the molecular mechanisms involved in the transmission of the bacteria from and to the seeds. And last, but not least, they transfer and share their results.

Moreover, the team hold a genetic resource center (Figure 1): the CIRM-CFBP, i.e. the French Collection of Phytopathogenic Bacteria. Most of the bacteria studied here belong to the *Xanthomonas* genus. Therefore, the Emersys team develop national and international collaborations related to *Xanthomonas* like the FNX (French Network on Xanthomonads) and the *Xanthomonas* Genomics Conference. The Emersys team is also part of other research programs such as the SEEDS project. This

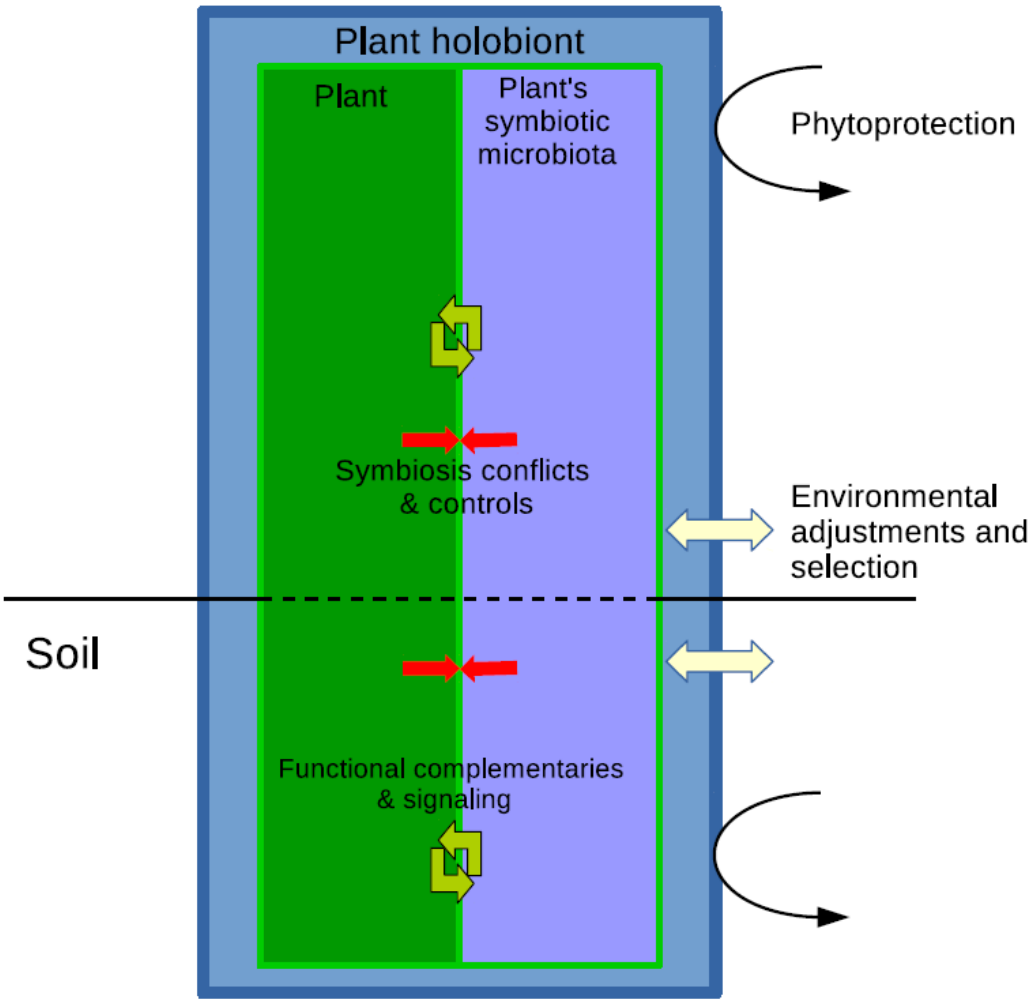


Figure 2: Scheme of the plant holobiont and related key interaction aspects both in term of evolution and functioning. The red arrows represents the symbiosis conflicts and controls. The green arrows represent the functional complementarities and signalling. The black curve arrows represent the phytoprotection. The white arrows represent the environmental adjustments and selection. All the arrows are present both above and below ground. both (from Vandenkoornhuyse *et al.*, 2015)

Table I: Plant associated microbial habitats (from Shade *et al.*, 2017)

Habitat name	**Plant organ concerned**
Antosphere	Microbial habitat associated to flowers
Carposphere	Microbial habitat associated to fruits
Caulosphere	Microbial habitat associated to stems
Endosphere	Microbial habitat located within plant tissues
Phyllosphere	Microbial habitat associated to leaf Including phylloplane (on leaf surface) and endosphere
Rhizosphere	Microbial habitat associated to roots Including rhizoplane (on root surface) and endosphere
Seed	Microbial habitat associated to seeds (not germinating)
Spermosphere	Microbial habitat associated to germinating seeds

project studies the evolution of the bacterial community of seeds in partnership work with the seed company Vilmorin and the Berkeley University.

1.2. Current knowledge about plant and seed microbiota

Plants are not only made of vegetal cells but can be considered as holobiont (Figure 2, Vandenkoornhuyse *et al.*, 2015). Indeed, they shelter and interact with many other organisms, including bacteria, viruses, fungi and archaea, both inside and outside their tissues. Together, all of these microbes associated to the plant, is what is considered as the plant microbiota (Bulgarelli *et al.*, 2013). Different microbiota can be deemed separately owing to plant organs in which there is different living standards. In that respect, we can examine from eight (Shade *et al.*, 2017) to 17 (Nelson, 2017) different plant-associated microbial habitats (Table I). A habitat can be defined as "a specific place occupied by a community of organisms for growth and reproduction" (Bulgarelli *et al.*, 2013).

The study of the plant microbiota is of interest because plant associated microbes can have many positive effects to the plants like resistance against biotic or abiotic stresses or nutrient acquisition and biomass accumulation (Sugiyama *et al.*, 2012). For example, the plant growth-promoting rhizobacteria (PGPRs) (Spaepen *et al.*, 2009) can help the plant to assimilate nutrients such as nitrogen, phosphorus or iron. Moreover, they can synthetize phytohormone such as auxin and interfere in its activity (Bulgarelli *et al.*, 2013). Some microorganisms, instead of directly promote plant growth, stimulate PGPR activity (Combes-Meynet *et al.*, 2010). Further, other members of the plant microbiota can provide biocontrol against biotic stresses such as pathogens (Mendes *et al.*, 2011; Bulgarelli *et al.*, 2013; Santhanam *et al.*, 2015; Busby *et al.*, 2016). In this line, microbial community can produce antimicrobial compounds against other micro-organisms (Emmert & Handelsman, 2006; Weller, 2007; Berg, 2009; Pérez-García *et al.*, 2011), or they can also activate what is name as the induced systemic resistance which increase the plant resistance against a broad spectrum of pathogens using the ethylene or jasmonate pathway (De Vleesschauwer & Höfte, 2009; Zamioudis & Pieterse, 2011). The seed microbiota influences the seed life as it affects the seed preservation (Chee-Sanford *et al.*, 2006), the release of seed dormancy (Goggin *et al.*, 2015) and the germination rate (Nelson, 2017). Thus, it is essential to better know the seed core microbiota as it influences the primordial very first step of the plant life cycle, not only at a lowest taxonomic rank but also at a community functional level (Shade & Handelsman, 2012). The detailed knowledge of the assembly and the composition of the seed microbiote provides promising agricultural approaches such as microorganism introduction. The microbiota manipulation can provide plant-growth promoting effects or biocontrol activity. In order to manipulate the seed core microbiote, a detailed knowledge is needed and thus accurate analysis methods.

The assembly of the microbiota shape its future composition. The main difference between root and leaf microbiota lie in the assembly of these microbial communities. Indeed, the source of inoculum are different in the phyllosphere and the rhizosphere. the phyllosphere seems to have several source of inoculum (Bulgarelli *et al.*, 2013). One hypothesis is that the phyllosphere microbes comes from the aerosols since air hold 10^1 to 10^5 cells per cubic meter (Fahlgren *et al.*, 2010; Lymperopoulou *et al.*, 2016).

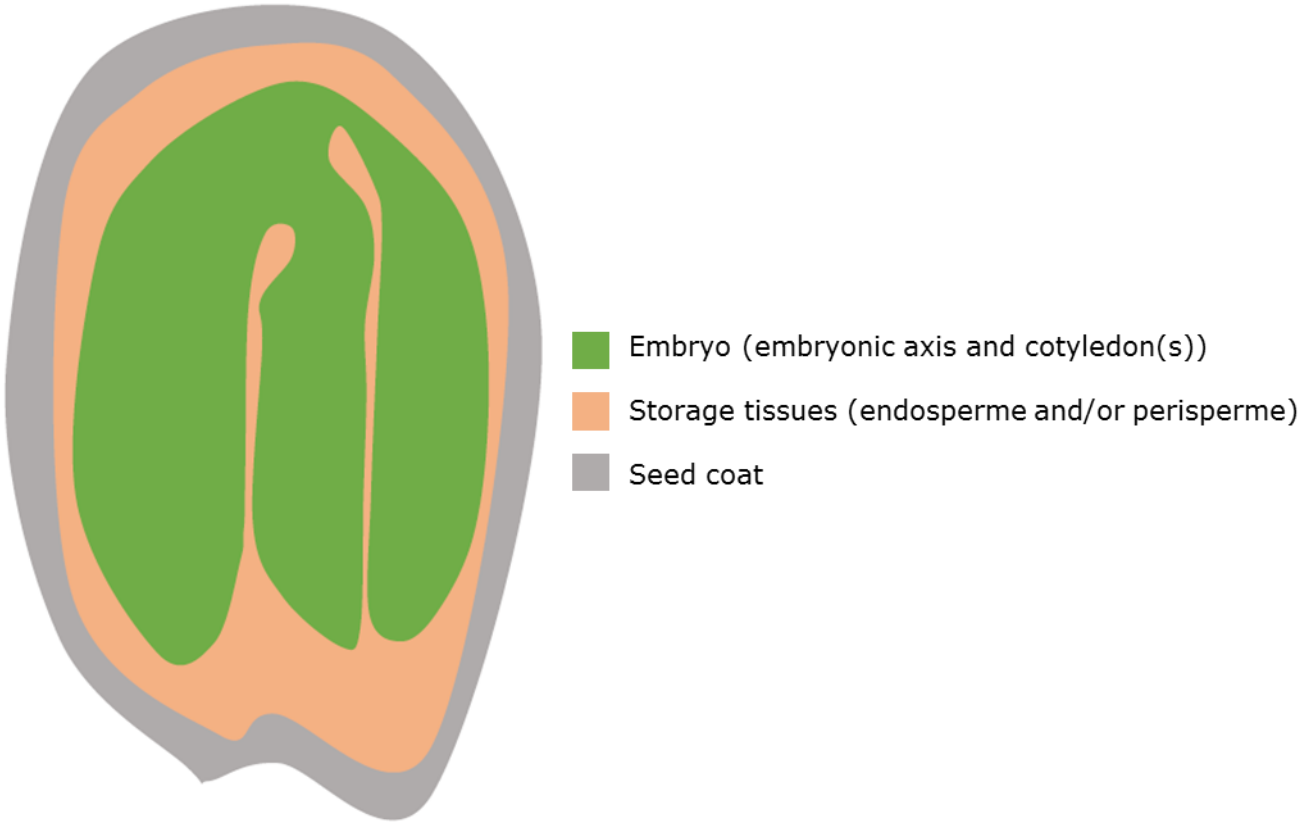


Figure 3: Schematic representation of the seed tissues. The endophytic microbiome colonise the embryo (green) and the storage tissues (orange) while the epiphytic microbiome colonise the seed coat (grey). Adapted from Shade *et al.*, 2017

However, only 2 out 28 taxa present in air were detected in the phyllosphere of nine Mediterranean perennial plant species (Vokou *et al.*, 2012). These results suggest other sources of inoculum such as flooding water (Knief *et al.*, 2012) or rainfall (Vacher *et al.*, 2016). The origin of the rhizosphere microbiota seems to be less variable. Indeed, Bulgarelli *et al.* (Bulgarelli *et al.*, 2013) suggest a two-step model for the rhizosphere microbiota assembly. The soil biome microbial community surrounding roots is mostly determined by the edaphic conditions. Firstly, the rhizodeposits and the cell wall features benefit to organotrophic bacteria, modifying therefore the soil biome community. Secondly, the bacteria of the rhizosphere are selected by factors depending on the host genotype.

At the moment, both root (rhizosphere) and leaf (phyllosphere) are the most studied plant microbial habitats. But, with a closer look to their preventive effect on plant disease, it is intuitively lighted that some pathogens colonise plants before rhizosphere and phyllosphere have been formed, on early developmental stages. Indeed, seeds are both the starting point and the ending point of the plant life cycle and the seed microbiota may be a major influencer of the whole plant microbiota (Shade *et al.*, 2017), making its exploration a promising field of study (Müller *et al.*, 2016; Nelson, 2017). Namely, seed-borne pathogens are vertically transmitted from the infected plant to its seed and then colonise the new born plant. Therefore, it seems quite obvious that the seed microbiota play a key role in the vertical transmission of seed-borne pathogens (Barret *et al.*, 2016). Seed-borne pathogens such as *Xanthomonas campestris* pv. *campestris* can also colonise and been vertically transmitted to non-host plant, providing a pathogenic agent reservoir (Darsonval *et al.*, 2008; Darrasse *et al.*, 2010).

The seeds can be anatomically divided in three compartments, the embryo, the endosperm and the seed coat (Figure 3, Shade *et al.*, 2017). Thus, two seed microbiotas can be distinguished: the endophytic microbiota (colonising the embryo and the storage tissue) and the epiphytic microbiota (colonising the seed coat surface) (Nelson, 2017). These two seed microbiotas can have two different origins: vertically (from the mother plant to its seed) or horizontally (from the environment) transmissions. The horizontally transmission can not only occurs within several environmental factors but also during the seed dispersal, even though this plant life stage is much more important in natural systems than in agricultural systems (Nelson, 2017). As highlighted above, the seed microbiota can also be vertically transmitted (Truyens *et al.*, 2015; Shade *et al.*, 2017). The vertical transmission can occur through the vascular system, the stigma or the fruit of the mother plant (Maude, 1996).

The composition of the microbiota can influence its effects on plant fitness. The phyllosphere microbiota is mostly composed of diverse, well-adapted to a tough environment bacteria compare with fungi and archaea (Lindow & Brandl, 2003; Delmotte *et al.*, 2009; Bulgarelli *et al.*, 2013). Indeed the phyllosphere is poor in nutrients and full of solar radiations (Hirano & Upper, 2000). Even though the rhizosphere is much more abundant in nutrients, the rhizosphere microbiota diversity is comparable to the phyllosphere one (Bulgarelli *et al.*, 2013). The richness of these habitats are also similar: the phyllosphere contained 10^6 to 10^7 bacterial cell per cubic centimetre (Fahlgren *et al.*, 2010) whereas the rhizosphere contain 10^6 to 10^9 CFU per gram (Spaepen *et al.*, 2009). As stated by several studies, the bacterial seed microbiota is mostly composed of Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes (Johnston-

Table II : Summary of samples variables by experience (mock communities represent all the mock communities used in these 12 studies)

Experiences (references)	Samples	Plant species	Seed harvest method	Type of inoculation	Harvest year	Production site	Process (native or disinfection)	Development stage	Production country	Plant taxonomic family	Plant taxonomic genus	Plant variety	Pollination type
Barret_2015 (Barret *et al.*, 2015)	84	12	1	1	5	4	1	3	7	4	8	20	2
Bee2seed (Torre-Cortés *et al.*, in prep)	54	2	2	2	2	2	2	7	2	2	2	2	2
Bnapus (unpublished)	83	1	1	1	1	1	1	1	1	1	1	11	1
F2S_Y1 (unpublished)	54	2	1	1	1	1	1	1	1	2	2	2	2
FNAMS_Y1 (Rezki *et al.*, 2016)	93	1	2	3	2	2	3	1	2	1	1	1	1
FNAMS_Y2 (Rezki *et al.*, 2016)	91	1	2	3	1	1	1	1	1	1	1	1	1
IDEAThodes_run1 (unpublished)	23	1	1	2	1	1	1	1	1	1	1	1	1
IDEAThodes_run2 (unpublished)	52	1	1	2	1	1	1	1	1	1	1	1	1
Klaedtke (Klaedtke *et al.*, 2016)	45	1	1	1	1	6	1	1	4	1	1	6	1
Navarro (unpublished)	12	1	1	1	1	1	1	1	1	1	1	2	1
Sweet (unpublished)	15	1	1	1	1	1	1	1	1	1	1	1	1
Vivanco (unpublished)	22	1	1	1	1	1	1	1	1	1	1	1	1
Mock communities	13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Total levels	641	16	2	3	10	15	2	9	8	5	9	40	2

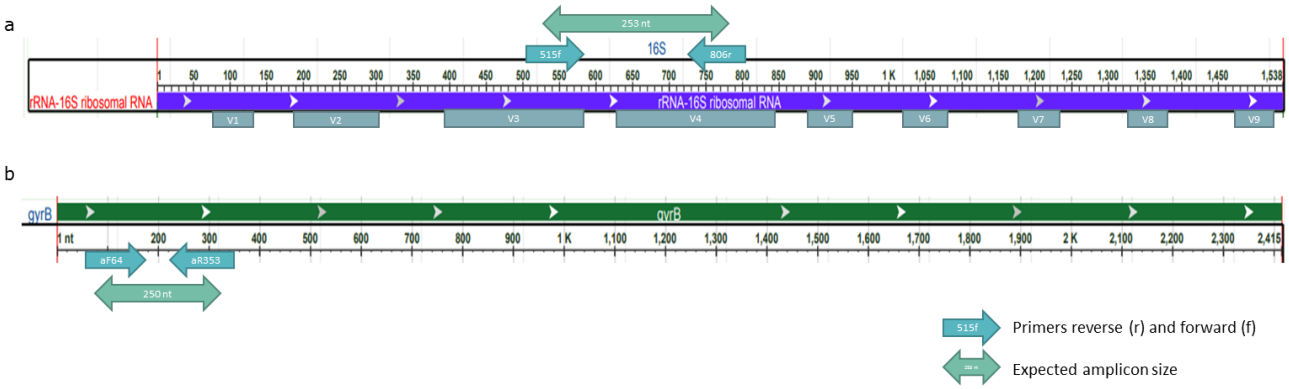


Figure 4: Schema of the 16S rRNA gene (a) and the *gyrB* gene (b). 515f & 806r 16S rDNA primers and the aF64 & aR353 gyrB primers were used for the amplifications of the 16S rRNA v4 region and a portion of *gyrB* which encodes the subunit B of the gyrase. The expected amplicon size is 253 nucleotides for the 16S rRNA gene and 250 nucleotide for *gyrB*.

Monje & Raizada, 2011; Lopez-Velasco *et al.*, 2013; Malfanova, 2013; Links *et al.*, 2014; Truyens *et al.*, 2015; Barret *et al.*, 2015; Adam *et al.*, 2016; Rezki *et al.*, 2016; Klaedtke *et al.*, 2016; Rybakova *et al.*, 2017). The seed microbiota also includes filamentous fungi, especially Dothideomycetes (Rodriguez *et al.*, 2009; Porras-Alfaro & Bayman, 2011; Barret *et al.*, 2015), and oomycetes (Thines, 2014). Epiphytic and endophytic bacterial seed microbiota can be distinguished (Nelson, 2017). Indeed endophytic communities harbour significant differences among plant genus whereas epiphytic communities are similar (Links *et al.*, 2014; Nelson, 2017). The four main bacterial phyla found on seeds are also the most dominant phyla in soil (Fierer *et al.*, 2012) and aquatic environment (Shafi *et al.*, 2017). These differences between epiphytic and endophytic as well as the four main phyla found in the environment and on seeds provides important clues on the assembly.

Although the microbiota shifts during the transition from seed to seedling (Barret *et al.*, 2015), this seed microbiota could influence the future plant microbiota, not only by its composition but also by its function. Here, we focused on the composition of the seed microbiota using an amplicon sequencing approach to detect and identify the membership of the microbial community.

1.3. Objective of this study and strategy

In this study, we'll focus on the bacterial seed microbiota. Datasets from different seed associated bacterial communities from different plants were studied and compared. We have defined two main aims for this work.

The first aim is to identify the main factors driving the composition of the seed microbiota. Since we have data from seeds from different plants and environmental conditions, in this first objective, we would like to analyse the possible relationships between the different treatments and seed microbiota. Thus, we will compare the observed richness, the alpha and the beta diversities.

The second aim is to identify some ubiquitous strain strains and to establish seed specific associated bacterial taxa. This would represent the bacterial taxa that are present in all the seeds from different plants and would be the seed core microbiota. This taxonomic composition analysis will be implement with two house-keeping genes: the v4 region of the 16S rRNA gene and the *gyrB* portion of the bacterial gyrase gene.

2. Material and Methods

2.1. Initial data

All the data analysed in this report have been collected from seven different studies done in the hosting group. To analyse these data together, the prerequisite was that they came from the same amplified gene portion. A total of 641 samples representing 16 different plant species and nine different organs or development stages have been gathered. These different plants were grown in 16 different sites of 8 different countries, during 10 different years (Table II).

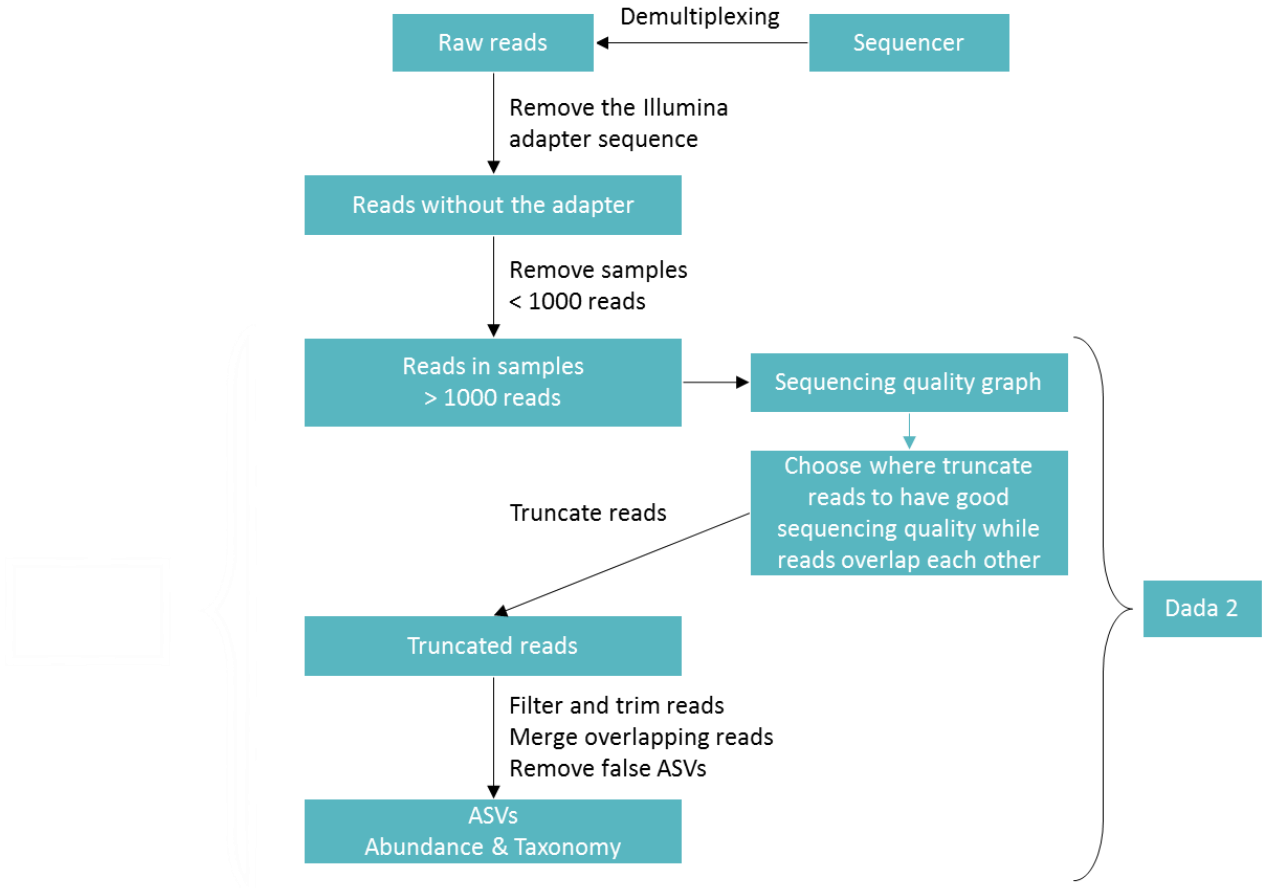


Figure 5: Dada2 pipeline i.e. data processing steps from the sequencer output to the ASVs rds files run in Rstudio afterwards

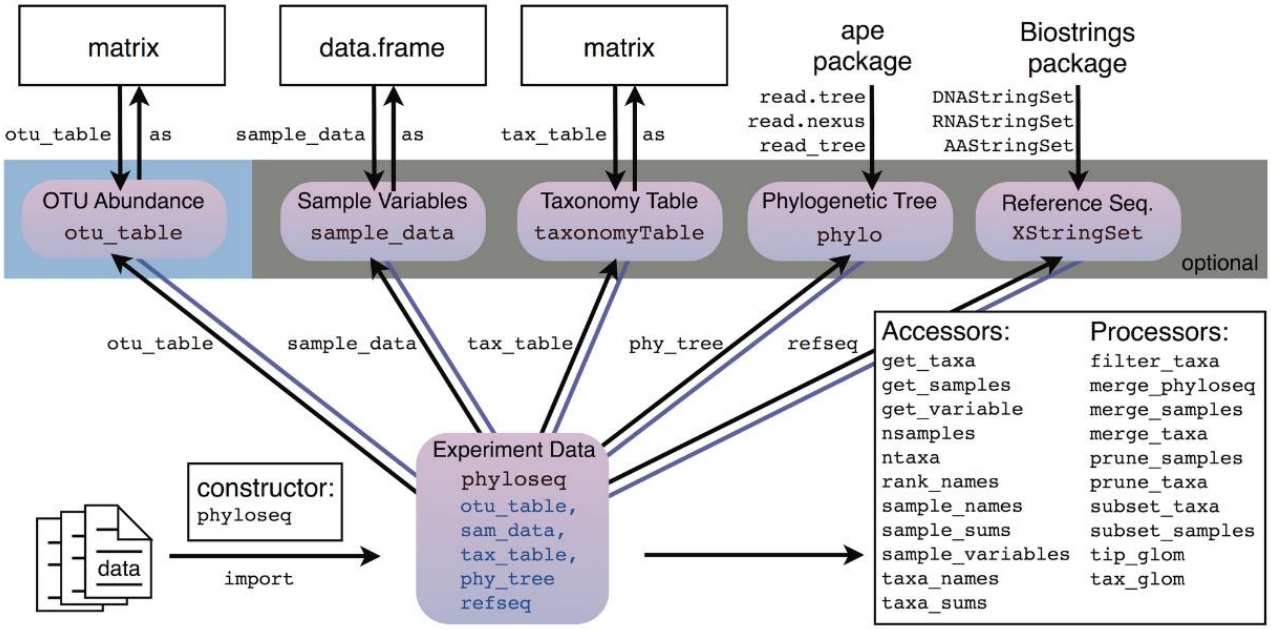


Figure 6: Schema of a phyloseq object and the different process it can undergoes, (from https://joey711.github.io/phyloseq/import-data.html)

For each sample, one or two genes were amplified, with the same primers in every studies, and sequenced: the v4 region of 16S (Figure 4) and the *gyrB*. The 16S gene is the ribosomal sub-unit gene. The amplification has been performed with the primers 515f/806r (Caporaso *et al.*, 2011). The v4 region is hypervariable despite it have well conserved flanked region. This allows good distinction between closely related organisms and also an easy PCR primer design. However, there is two main drawbacks to using this 16S gene: you can have coamplification of chloroplast and you cannot resolve bacteria taxa below the genus level. Consequently, Barret *et al.* (2015) designed primers for the gene *gyrB* that have been use in the other studies analysed here. This gene encodes the beta subunit of the DNA gyrase and permits the affiliation of the sequences at the species level.

The finally sequences were obtained by a Miseq system workflow('16S Metagenomic Sequencing Library Preparation'; 'System Specification Sheet: MiSeq® System'; Caporaso *et al.*, 2012).

2.2. Clustering MiSeq reads into Amplicon Sequence Variants (ASVs)

The data analysis pipeline is divided in different steps (Figure 5). In the initial step (demultiplexing), raw reads are assigned to their original sample by the sequencer. In the second step, the Cutadapt software (version 1.16) (Martin, 2011) is used to remove the Illumina adapter sequence from the reads and to match each read to one gene, if there is several amplified genes in the run (e.g. 16S & *gyrB*). This produces a fastq file per gene for each sample that was used as an input file for Dada2 (version 1.6).

Then Dada2 (Callahan *et al.*, 2016) defines how many reads are in each fastq file. Subsequently, each sample file with less than 1000 reads is manually erased. Afterwards, Dada2 produce a sequencing quality graph. From this graph, we notify the position where to cut the amplicons in Dada2. Dada2 recommends that only the nucleotides with a sequencing quality score higher than Q30 have to be conserved. It must be noted in this step that some nucleotides with less than a Q30 quality score were conserved to allow the assembly of the forward and reverse primers. Then, with the cut sequences, the reads are filter, trimmed and merged to produce the ASVs. The interest of Dada2 lies in its error correction model which seems to be the more accurate so far (Callahan *et al.*, 2016). During this step, the Dada2 algorithm removes ASVs that it considers as false ASVs, i.e. ASVs produced by sequencing errors. Finally, two rds files are produced: the first rds file gather all the ASVs of the run with their corresponding abundance; the second rds file gather all the ASVs of the run with their corresponding taxonomy. The taxonomy was assigned according to the 16S RDP database. In this study, the obtained sequences were analysed as ASVs (Callahan *et al.*, 2017).

2.3. Data subsetting

Rds files were run on R Studio (version 3.4.4) with the Phyloseq package (version 1.22.3; McMurdie & Holmes, 2013). A Phyloseq object (Figure 6) is an association between an ASVs abundance table (called the otu_table), a taxonomy table (called the tax_table) and a design made with samples variables (called

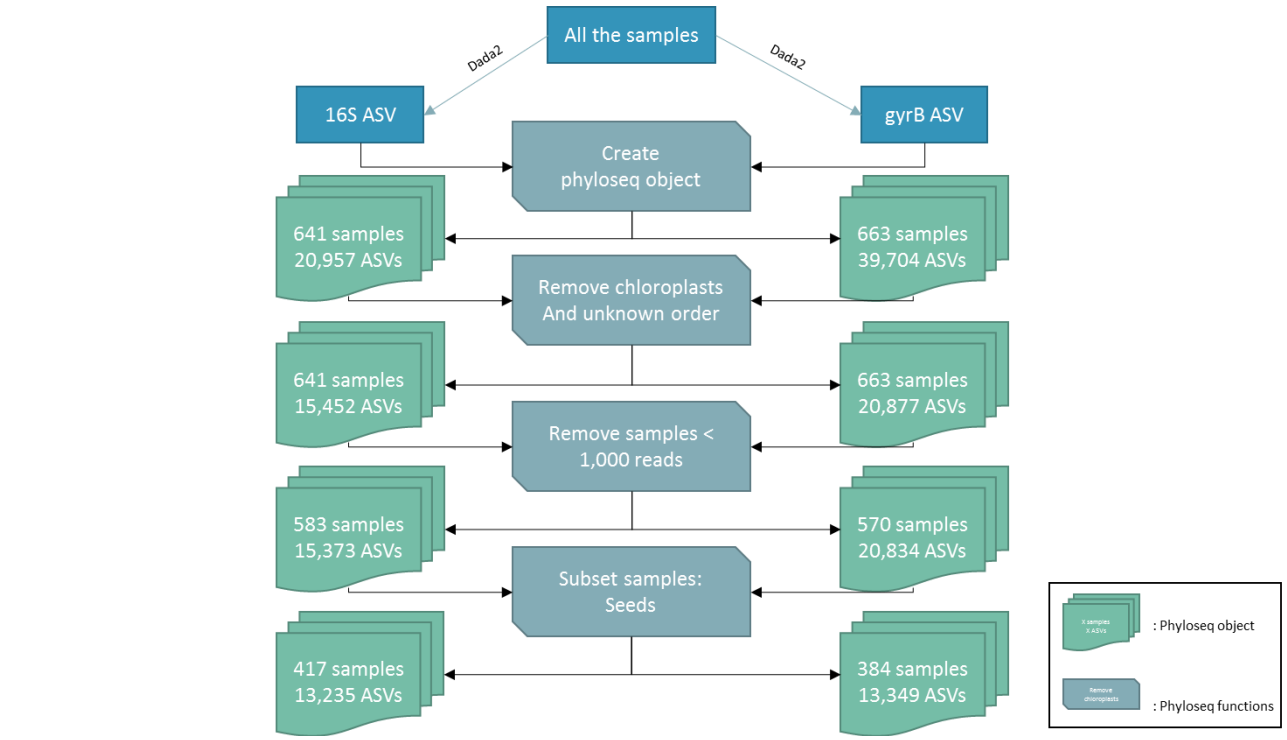


Figure 7: Pipeline from the Dada2 output to the Phyloseq objects with only the seed samples. The Phyloseq objects contain only 16S amplified sequences (left) or only gyrB amplified sequences (right).

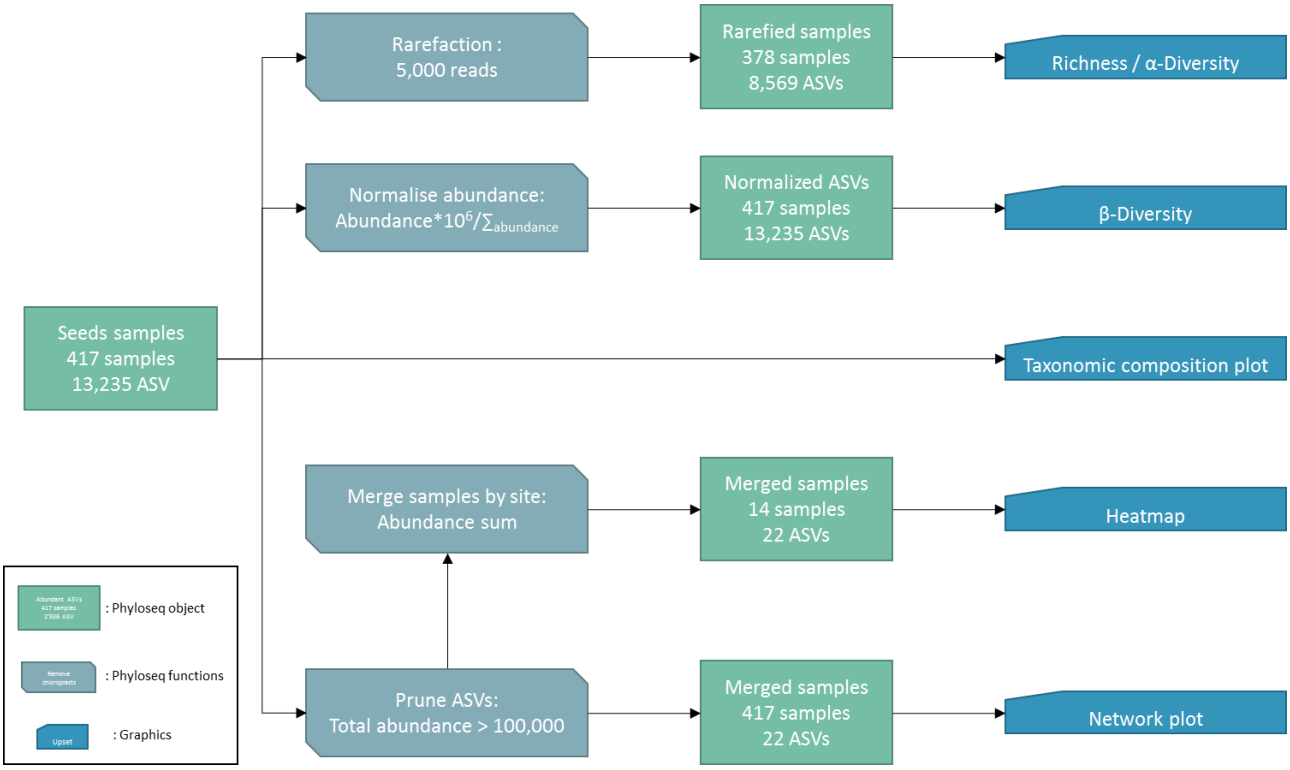


Figure 8: Pipeline to analyse the seed samples of the Phyloseq object containing the 16S amplified sequences

the sample_data). A phylogenic tree and the ASV sequences can also be added to that object. The taxonomy table we used has ranged up to the Genus rank. The design we used compiles 13 variables for each sample: the study were the sample come from, the plant species, the plant variety, the plant taxonomic family, the plant taxonomic genus, the seed harvest method, the inoculation or not of the plant, etc. (as resume in Table I).

Subsequently, in the first Phyloseq object created (Figure 7), not only seed samples were gathered but also flower and leaf samples. First, ASVs identified as chloroplasts were removed. Indeed, as outlined above, the 16S gene amplification on vegetal biological material, nearly in all cases, lead to chloroplast coamplification. After the chloroplasts ASVs were removed, some samples had only a few ASVs remaining. Therefore, we removed the samples with less than 1000 remaining reads. This 1000 reads arbitrary threshold have been chosen to be consistent with the Dada2 previous threshold. As a result, samples that had more than 90% of chloroplasts within their ASVs have been removed. After the filtering process, the Phyloseq object was divided between the different developmental stage (seed, germinating seed, seedling, leaf, nectar, pollen) thanks to the design (or sample_data on Figure 6). In this project, we have focused on the mock communities and the seeds samples. As explained in the glossary, the mock community evaluates the analysis potential to recognise bacterial strains.

2.4. Phylogenetic analysis of mock communities

To phylogenetically analyse the mock communities, we first extracted their ASVs sequences in fasta file (mock fasta file). In addition, a fasta file with the 16S gene v4 region sequences of each strain presented in the mock community samples, was also retrieved from the NCBI (reference fasta file). This two fasta files were combined and aligned using the Clustal software (version 2.1) (Chenna *et al.*, 2003). The alignment was visualize and editied using Jalview (version 2.10.4) (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009). all the sequences present in the alignment were trimmed to the same length (253 ntd).

Still using Jalview, the alignment was used to build a phylogenetic tree by the Neighbour Joining calculation. This tree that was exported as newick file. The newick tree file was visualized with the Figtree software (version 1.4.3) (Rambaut, 2007). This tree was used to manually list the ASVs and reference sequences that match at 100% sequence identity. Finally, we deduced the number of references that have been detected by ASVs. These data are not shown because of their size.

2.5. Seed microbial community analysis

Different ecological indexes have been used in this project to study the structure and composition of the seed microbiota (Hill, 1973): i) the observed richness that correspond to the number of detected ASVs and ii) the Shannon and iii) inverse Simpson's index reflexing the alpha diversity (Figure 8). The alpha diversity represents the species diversity in one habitat or in one condition. For both these indices, the higher they are, the higher the diversity between species is. At the same time, these indices are affected by differences between sample sizes. Therefore, we had to homogenize the sample sizes by rarefying at 5,000 reads. The rarefaction curve was obtained with the rarecurve function of the Phyloseq

package. Differences in richness and alpha-diversity were evaluated as whole by a Kruskal-Wallis test with post hoc Dunn test between each variable.

Beta diversity represent the species diversity among different habitats or conditions. It was investigated by Bray-Curtis dissimilarity matrix and Jaccard dissimilarity (Whittaker, 1972). These two indices were used because they do not highlight the same part of the diversity. Indeed, Jaccard index is based on a presence/absence matrix while Bray-Curtis index is based on abundance. Both these indices were calculated on normalised ASVs abundance i.e. ASVs count were divided by the number of reads per sample and multiply par 10^6. To evaluate the impact of the variables on the dissimilarity, a principal coordinate analysis was performed with the capscale function of the vegan package (version 2.5-1) on the following model: "distance ~ Site + Plant + Genotype + Experience + Harvest+ Inoculation + Plant Family + Plant Genus + Pollination +Process + Year". To assess the significance of constraints, a permutation test was performed on the model with anova.cca function of the vegan package. To assess the importance of each variable on the dissimilarity, permutated multivariate analysis of variance (PERMANOVA; Anderson, 2001) was implemented with the adonis function of the vegan package on R studio. Both dissimilarity indices were then ordinate using a Principal Coordinate Analysis (PCoA) with the plot_ordination function of the Phyloseq package, highlighting one variable effect.

To analyse the taxonomic composition, the rarefaction of the data is not needed. The taxonomic composition of a community refers to the abundance and prevalence of each species or Phylum compare to the others. The abundance of an ASVs in the number of reads detected for this ASV. The relative abundance is the percentage of the reads detected for one ASVs compared with all the reads. The prevalence of an ASVs is the number of samples is detected in. The relative prevalence is the percentage of the whole samples is detected in. In this study, the taxonomic composition is conveyed by the plot composition and the heatmap functions of the Phyloseq package. To produce a viewable heatmap, only the most abundant ASVs (aASVs) were kept by removing the ASVs with a total abundance lower than $1x10^6$ reads. This arbitrary threshold was defined to keep some twenty ASVs to produce a clearer graphic. Then, the samples were merged by site, summing by site the abundance of each ASVs.

3. Results

3.1. Mock communities

Mock communities are routinely used in sequencing projects to analysed taxa detection by the different sequencing protocols. In our project, we analysed the percentage of taxa detected by Dada2 in the mock communities composed of 69 bacterial strains. According to our analysis, 81 16S rRNA ASVs were detected in the mock communities represented 88.46% of taxa present in the mock community. The fact that we identified more ASV than the number of strains associated to the mock community can be explained by the fact that some bacterial strains have more than one copy of the 16S gene in their genome (e.g. *Bacillus* sp., *Rhodococcus* sp. & *Erwinia* sp.) and that these copies may present some polymorphism.

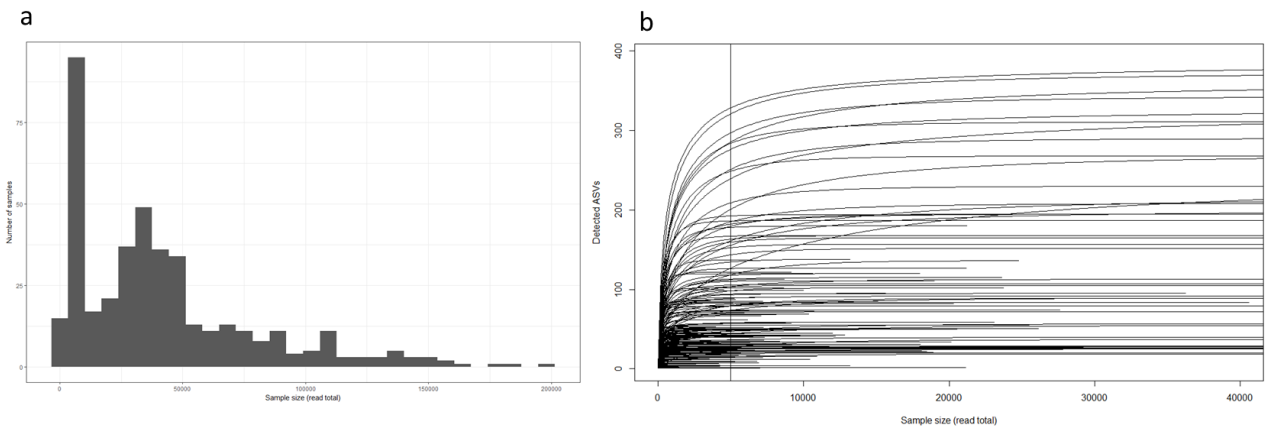


Figure 9: Sequencing depth (a) and Rarefaction curve (b) of the seed samples. The sequencing depth histogram (a) represent the number of samples as a function of the number of reads. The rarefaction curve (b) represents the number of ASVs as a function of the number of reads. The vertical black line indicates the 5,000 reads rarefaction threshold.

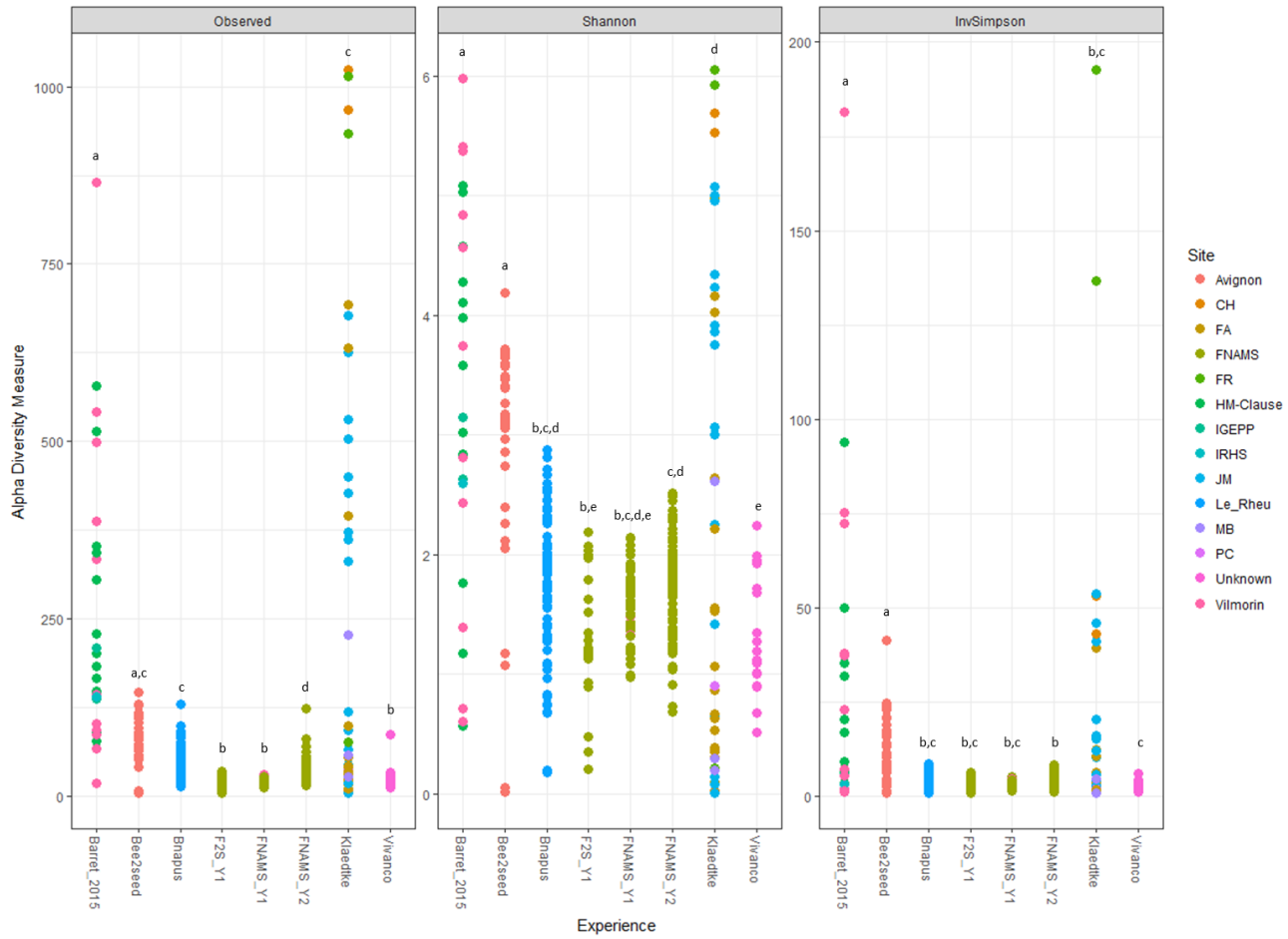


Figure 10: Alpha Diversity measures for each experience with the observed diversity (number of ASVs per experience), the Shannon index and the inverse Simpson index. Measures for bacterial community with 16S sequence. Alpha diversity was assessed with the number of ASVs rarefied at 5,000 reads per sample. Each dot correspond to a seed sample collected in different production site (one color for each site). Letters from "a" to "e" denote significant differences between experience averages (p-value ≤ 0.05/2, Dunn test).

Another explanation is that some sequencing errors may remain within the ASVs (see the discussion below).

Within the analysis of the mock communities, we conclude that the experimental protocol, from the amplification to the Dada2 pipeline, is accurate enough to determine the bacteria present in the samples.

3.2. Data set display

In our study, there is 417 seed samples representing 13,235 ASVs (Figure 8). Seeds samples represent 63.15% of the ASVs of the initial data set. As a whole, seeds samples gathered $17.5x10^6$ reads. The number of read per seed sample ranges from 1,098 to 199,203 reads. The sample size median is at 33,609 with a standard deviation of 38,040. Highly variable ASVs abundance ranges from 1 to $5.4x10^6$ reads. The seed ASV abundance median is 14 with a standard deviation of 52,768.

These numbers reveal the important part of seed samples in the original dataset, allowing a meta-analysis of the seed microbiota with these data. In addition, the very high variability of sample size and ASVs abundance is clearly highlight here.

3.3. Factors influencing the richness and diversity of the seed microbiota

3.3.1. Seed production site influence seed microbiota richness and alpha diversity

Regarding to the sample size heterogeneity (Figure9a), prior to the richness and alpha-diversity study we had to homogenize the sample size. In that purpose, samples were rarefied at 5,000 reads (Figure 9b). Consequently, samples with less than 5,000 reads were removed. The 5,000 threshold was chosen according to the rarefaction curve (Figure 9b). The plateau of the curve means that even if there is more reads in the sample, there will not be more ASVs detected. The 5,000 threshold allows, here, to keep 90.65% of samples without reaching the plateau of the curve. Within the rarefaction, 35.26% of the ASVs have disappear: these species were only present on the removed samples or they have been lost in samples that did not reached their curve plateau (Figure 9b).

After homogenising the sample size, we analysed the richness and alpha diversity (Figure 10). The observed richness, i.e. the number of ASVs observed, ranged from 5 to 1024 and its median is at 37 with a standard error of 149.37. The Shannon index ranged from 0.01 to 6.05 and its median is at 1.77 with a standard error of 1.09. The inverse Simpson index ranged up from 1.00 to 192.73 and its median is at 3.42 with a standard error of 18.03. This reveals the high variability of the richness and the alpha diversity across samples.

To understand this heterogeneity, Kruskal-Wallis test were implemented on 15 variables (Appendix II). Regarding to the observed richness, statistical differences are found in 13 out of 15 variables; only plant families and pollination types do not show statistically significant differences between their groups. In the case of alpha diversity, regarding both Shannon and inverse Simpson indices, 10 out of 15 variables show statistically significant differences between their groups. These 10 variables are statistically different

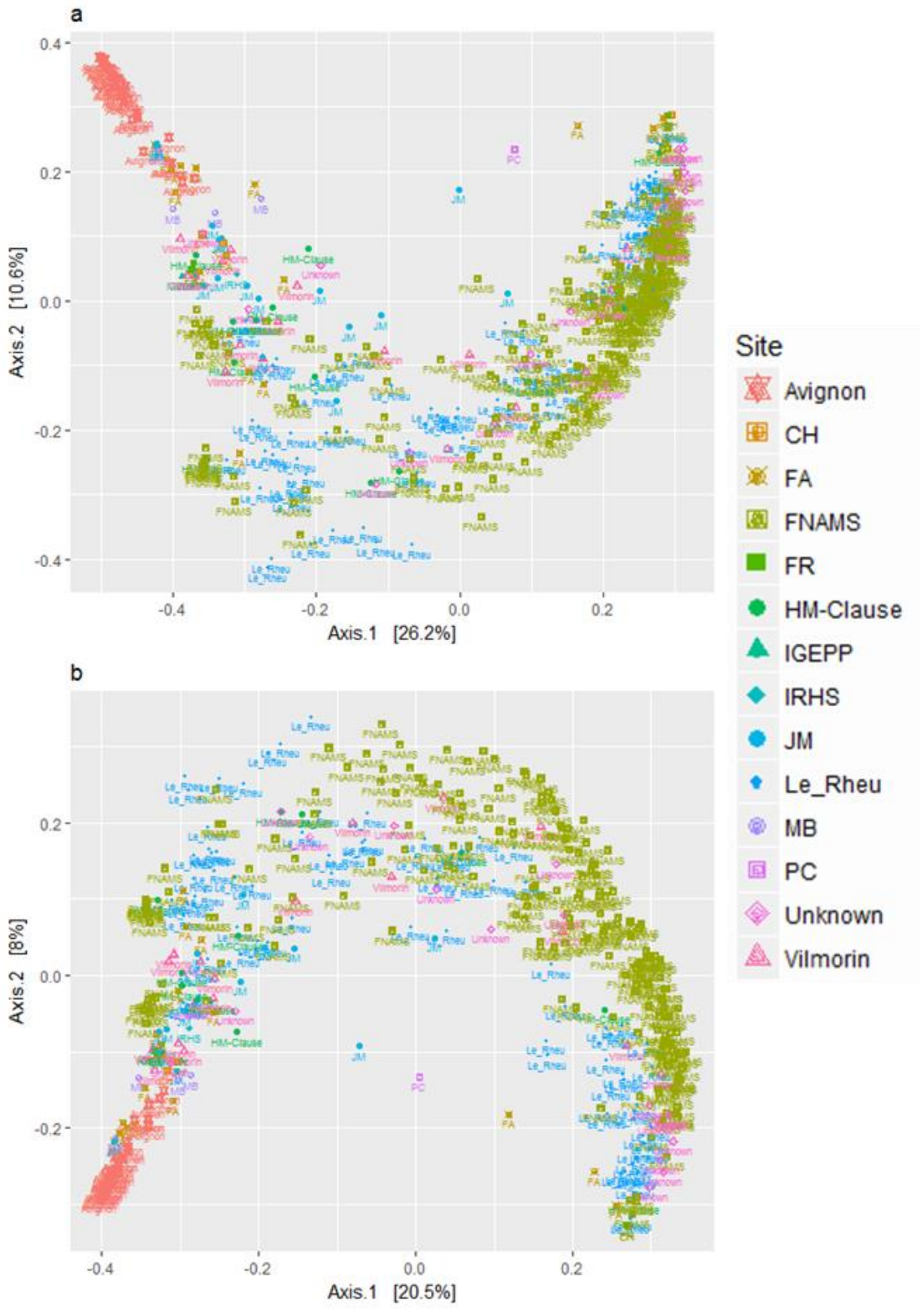


Figure 11: Bray-Curtis (a) and Jaccard (b) distances depending on the production site (colours and shapes). Production site names are written as we know them.

for the three indices. In the descending order of the p-value, we have found: the experience carried out, the harvest year, the plant variety, the number of plant species per experience, the number of plant varieties per experience, the number of production site per experience, the plant species, the plant genus, the production site and the production country.

Therefore, it seems that the experience carried is the most significant factor to explain the heterogeneity within the samples alpha diversity here (Figure 10, Appendix II). Moreover, the experiences that had seed harvested from different production site seems to have more richness and diversity variances than experiences with only one production site (Figure10).

3.3.2. Seed production site is the main factor influencing seed microbiota beta diversity

The beta-diversity between samples was estimate both with Jaccard and Bray-Curtis distances on the normalised seed samples. The canonical analysis of principal coordinates (CAP) allows to investigate further the relative contribution of whole the variables on the microbiota dissimilarity. The model tested was: dissimilarity index ~ Site + plant Species + plant Variety + Experience + Harvest + Inoculation + plant Family + plant Genus + Pollination + Process + Year. This model explains 23.39% of the Bray-Curtis dissimilarity and 20.85% of the Jaccard dissimilarity. The permutation test implemented on the CAP shows it is statistically significant (PERMANOVA, p-value ≤ 0.001).

To evaluate more precisely the influence of each variable on the dissimilarity, a PERMANOVA was directly implemented on the model (Appendix III). Within all the normalised seed samples, the production site explains 30.69% of the Bray-Curtis distance between samples and 22.50% of the Jaccard distance between samples. Among every variable, the production site have the biggest impact on the beta-diversity. As highlighted for the alpha-diversity, the influence of the experience carried out on the beta-diversity is also, but less, significant: 4.37% of the Bray-Curtis distance between samples and 4.01% of the Jaccard distance between samples are due to the experience carried out. The third and fourth variables are the plant species and the plant variety. Respectively, they explain 7.63% and 6.12% of the Bray-Curtis distance between samples and 6.38% and 6.18% of the Jaccard distance between samples for the variety and the species respectively.

As the production site is the most important factor, ordination of Bray-Curtis and Jaccard dissimilarities were performed with a principal coordinate analysis (PCoA) to visualise the production site clustering (Figure 11). This reveals a spatial separation between the microbial communities associated at each production site. Bray-Curtis ordinations have a higher explanatory value than Jaccard ones, respectively 26.2% and 20.5% for the x-axes and 10.6% and 8% for the y-axes. While the representation of the site production clusters are discrete for both dissimilarity indices. This suggests that the differences between sites in community structure can be attributed to taxa exclusively observed in each site.

Figure 12: Phyla relative abundance (a) and ASVs prevalence of each phyla as a function of its raw abundance (b) of the 16S ASVs. The x-axis represent the samples on the (a) graph. The colours used refers to the phyla (a) or to the genus (b).

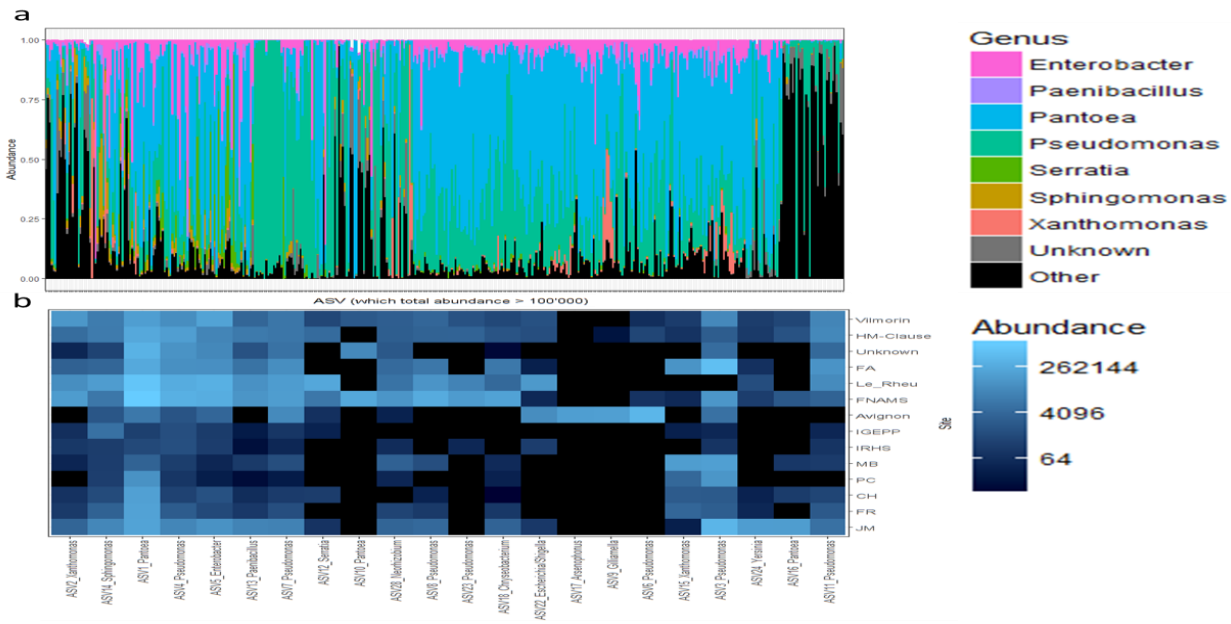


Figure 13: Variations in relative abundance of the whole bacterial community (a) with the 16S rRNA gene. Each colour represents one of the seven most prevalent bacterial genera. Unknown taxa represents ASVs that could not be assigned a taxonomy at the genus level. The x-axis represents the whole samples. Heatmap of abundance per production site with only aASVs (b). Abundant ASVs (aASVs) are ASVs with a total abundance > 100,000 reads. The abundance per each site and ASVs is visualise by a blue colour. The higher the abundance is, the lighter the blue is.

3.4. Taxonomic composition of the seed microbiota

3.4.1. Seed microbiota is composed by bacteria belonging to *Proteobacteria*, *Bacteroidetes*, *Actinobacteria* and *Firmicutes* phyla

To explore the taxonomic composition of the seed microbiota, we first investigated the distribution of phyla across samples. The four main phyla (out of 22 in total) of the seed microbiota are *Proteobacteria*, *Bacteroidetes*, *Actinobacteria* and *Firmicutes* (Figure 12), representing 76.80% of the ASVs (10,165 out of 13,235 ASVs) and 99.27% of the total abundance (17,425,725 out of 17,553,685 reads). Thus, the 18 other phyla represent 23.20% of all ASVs and only 0.73% of the total abundance. Furthermore, *Proteobacteria* are by far the most important phyla of seed microbiota. Indeed, they represent alone 35.34% of all ASVs and 87.71% of the total abundance. Thereby, Actinobacteria, Bacteroidetes and Firmicutes respectively represent 15.83%, 14.14% and 11.49% of all ASVs and 3.70%, 2.92% and 4.93% of the total abundance.

Regarding to their prevalence, i.e. the number of sample they are detected in, these four most abundant phyla are also the more prevalent (Figure 12b). Ranked in the descending order of relative prevalence, we found *Proteobacteria* (detected in 100% of the samples), *Firmicutes* (94.24%), *Bacteroidetes* (88.97%) and *Actinobacteria* (75.78%). The relative prevalence of the following phylum (*Acidobacteria*) declines to 23.74%. This difference between the relative prevalence of the four main phyla and the others really highlight their preponderance.

As the v4 region of the 16S rRNA gene allows a taxonomic determination until the genus rank, we can visualise the most abundant genus using their relative abundance (Figure 13a). Unsurprisingly, the seven most abundant genus (out of 847 genus) belong to the *Proteobacteria* phylum, except *Paenibacillus* sp. which belong to the *Firmicutes*. The two most abundant genus of the seed microbiota are *Pantoea* sp. and *Pseudomonas* sp. that represent respectively 33.76 % and 29.06% of the total abundance. However these two genus does not represent the majority of the ASVs: they cumulate respectively 0.14% (19 ASVs) and 1.30% (172 ASVs) of all ASVs. It means that only 2 genera with 191 ASVs in total (1.44% of all the ASVs) rack up 62.82% of the total abundance.

3.4.2. *Pantoea* and *Pseudomonas* genera are the main members of the seed core microbiota

As ASVs allows differentiation of a single nucleotide polymorphism between amplicons, we can define the community members more thinly (Figure 13b). An arbitrary threshold of 100’000 total abundance was defined to select the most abundant ASVs (aASVs). These aASVs represents 77.74% of the total abundance. There are 22 aASVs, 15 of which are in the seven main genus.

To further analyse the community memberships and structure, the shared taxa among all variables were determined. None of the ASVs have been detected in every sample. However, as the production site is the main driver of community composition, the shared ASVs between sites have been sought. As result, five ASV have been found in, at least, one sample of each production site (Appendix IV). Moreover, these five ASVs are detected in more than 75% of the whole seed samples and their total abundance is higher

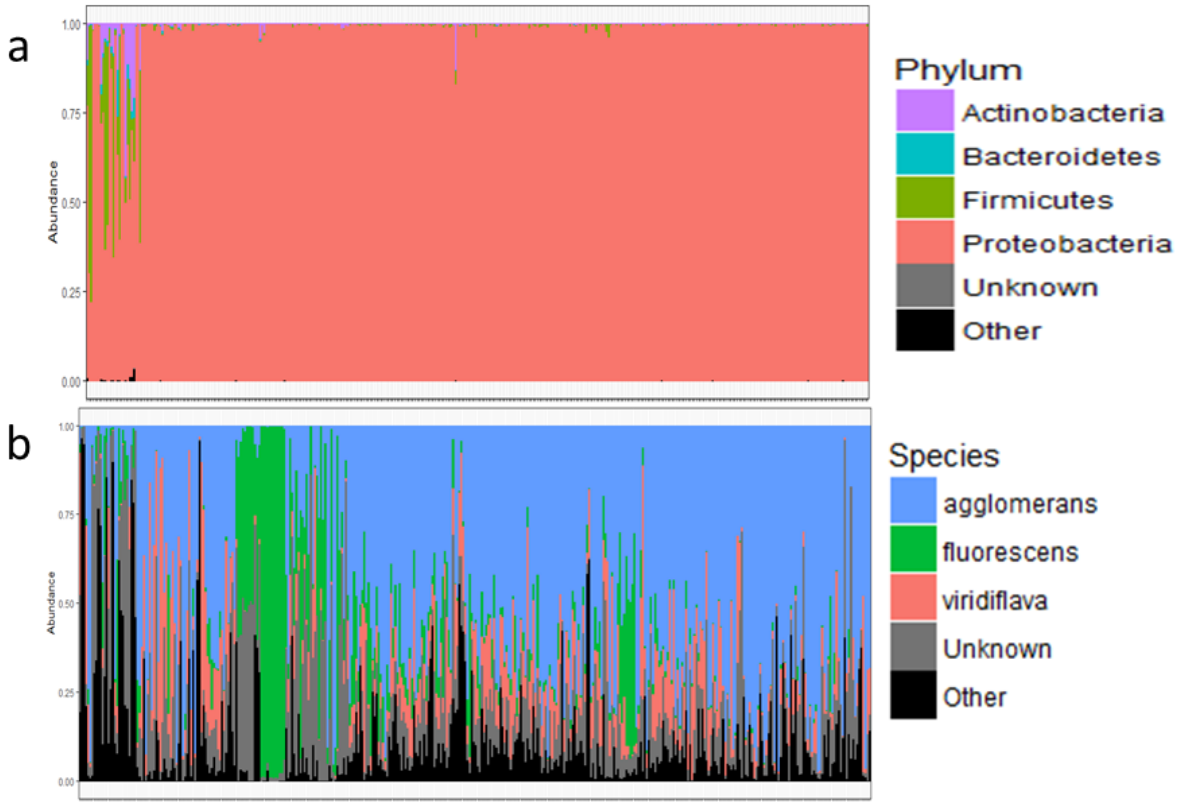


Figure 14: Variations in relative abundance of the whole bacterial community at the phylum level (a) and at the species level (b) of the *gyrB* ASVs. Each colour represents one of the seven most prevalent bacterial phyla (a) or species (b). Unknown taxa represents ASVs that could not be assigned a taxonomy at these levels. On both graphics, the x-axis represent the whole samples.

than 1'000'000 reads. These five ASVs are from the genera *Pantoea* sp., *Enterobacter* sp., *Pseudomonas* sp. (2 ASVs) and *Sphingomonas* sp.. All these genera belong to the *Proteobacteria* phylum.

3.4.3. *Pantoea agglomerans* and *Pseudomonas viridflava* are the main representative species of their genera in the seed core microbiota

Finally, the same taxonomic analysis than for 16S ASVs have been implemented on *gyrB* ASVs. Here, only 15 phyla were detected with *gyrB* (22 were detected with 16S). However, the same main four phyla have been found i.e. *Proteobacteria*, *Actinobacteria*, *Bacteroidetes* and *Firmicutes*. They cumulate 98.31% of the ASVs and 99.99% of the total abundance. In fact, *Proteobacteria* alone gather 69.28% of the ASVs and 99.20% of the total abundance. *Proteobacteria* are by far the main phyla of the seed microbiota with both 16S and *gyrB* analysis.

Regarding to their relative prevalence, *Proteobacteria* (100%), *Firmicutes* (92.45%), *Actinobacteria* (89.84%) and *Bacteroidetes* (69.53%) are the more prevalent taxa. *Verrucomicrobia* is the following phyla by descending order with 10.94% of relative prevalence. Once more, there is a huge difference between the relative prevalence of the four main phyla and the other ones.

At the genus rank, we only observed 355 genera within the *gyrB* ASVs while there were 857 detected genera within the 16S ASVs. The seven most abundant genera are not the same than the seven most abundant genera detected with the 16S rRNA. Only two genera out of the seven most abundant were found both within 16S and *gyrB* ASVs: *Pantoea* and *Pseudomonas*. Within the *gyrB* ASVs, *Pantoea* represents 1.71% of the ASVs and 54.45% of the total abundance while *Pseudomonas* gather 12.11% of the ASVs and 36.21% of the total abundance. Together, these two genera gather 13.82% of the ASVs and rack up 90.66% of the total abundance. The other abundant genera in *gyrB* were also found in the 16S analysis but with a lower abundance and vice-versa.

As the *gyrB* permits the affiliation of the sequences at the species level, 848 different species were assigned to ASVs. Three main species were detected (Figure 14): *Pantoea agglomerans*, *Pseudomonas viridiflava* and *Pseudomonas fluorescens*. These three main species gather 3.56% of all the *gyrB* ASVs and represent 76.42% of the total abundance. *Pantoea agglomerans* alone represents 54.33% of the total abundance with only 0.82% of all the ASVs and is present in 99.22% of the samples. *Pseudomonas viridiflava* and *Pseudomonas fluorescens* respectively represent 11.52% and 10.59% of the total abundance with 0.95% and 1.79% of all the ASVs and are present in 93.23% and 89.59% of the samples.

As ASVs allows a finer taxonomic determination than the species rank, the most abundant ASVs (aASVs) were determined. The same abundance threshold was implement for both 16S and *gyrB.* Thus, 22 ASVs out of 13,349 have more than 1,000,000 of total abundance and are considered as aASVs. Within these 22 ASVs, two are detected in at least one sample of each production site. These two ASVs belong to *Pantoea agglomerans* and *Pseudomonas viridiflava* species.

4. Discussion

In this study we compare seven seed associated microbial community studies. These studies used the v4 region of the 16S rRNA gene to detect ASVs. We compare the richness, the diversity and the

taxonomic composition within the seven studies. The taxonomic composition analysis was also implemented with another gene, the *gyrB* gyrase subunit gene. By using Dada2, that provides the ASVs table, we based our work on ASVs rather than on OTUs. The ASVs are sequences distinguished at the level of a single-nucleotide difference (Callahan *et al.*, 2017). This accuracy provides a biological reality contrary to OTUs analysis. Indeed, OTUs are clusters of reads distinguished by an arbitrary threshold: most commonly 97% of similarity. These threshold does not reflect a biological reality and can therefore lead to bias. Moreover, OTU cluster threshold can change among studies whereas ASVs are reusable across studies thus making this meta-analysis possible. ASVs are as accurate that some have only been detected by a single read. In fact, the minimum abundance is one and the median is 14 with a standard deviation of 52,768. Some of these ASVs may be the result of potential sequencing mistakes. These false ASVs overestimate the richness and the diversity. However, most of the sequencing errors have been removed by Dada2. Dada2 is so far the more accurate error correction model (Callahan *et al.*, 2016). Therefore, most part of this very low abundant ASVs could really correspond to bacteria strains present in a low amount. Indeed, the species abundance distribution of microbial communities follow a hollow curve with a long "tail of low-abundance species" (Nemergut *et al.*, 2013). In other words, in a microbial community, just a few species are very abundant and most of the species are present at a very low abundance.

Regarding diversity, we can see that the richness and the diversity are very variable among studies, especially the alpha-diversity. The observed richness, estimating the microbial population size, is influence by nearly all the variable tested. In literature, the determination of the bacterial population on and in seeds have been estimated several times but all these estimations are also highly variable. Indeed, these estimations range from 10^1 to 10^8 CFU/g seed for the endophytic population and from 10^4 to 10^8 CFU/g seed for the epiphytic population (Nelson, 2017). These numbers only support the hypothesis that the epiphytic population size is higher than the endophytic one.

We noticed that several factors influence the observed richness, the alpha and the beta diversity. Among all the variables tested, the production site, the plant species and the harvest year seems to have an important impact on the seed microbiota structure.

The production site seems to be the most important factor shaping the seed microbiota, explaining from 22.50% to 30.69% of the bacterial diversity (Jaccard and Bray-Curtis indices respectively). This influence of the production site on seed microbiota of different plant species have already been reported but with only 12.2% of the bacterial diversity (Bray-Curtis index) explained by the farm site (Klaedtke *et al.*, 2016). On maize, the influence of different production site on seed microbiota as also been reported (Johnston-Monje & Raizada, 2011).

Within our analysis, the plant genus, the plant species and the plant variety influence the richness and the alpha-diversity. However, only the plant species and the plant variety influence the beta-diversity, respectively from 6.38% to 7.63% and from 6.12% to 6.19%. Therefore, considering the influence of the host plant on the microbiota, the species of the host plant seems to be the more influent factor shaping the microbial community. This host-driven selection was reported, especially in roots but with less impact

than the environment-driven selection (Lundberg *et al.*, 2012; Bulgarelli *et al.*, 2013; Peiffer *et al.*, 2013; Edwards *et al.*, 2015; Dombrowski *et al.*, 2017).

The harvest year is also a influent factor of the seed microbiota. Regarding the observed richness and the alpha diversity, the harvest year is the second most significant factor explained the richness and diversity differences among samples. However, the harvest year only explain 1.1% of the beta diversity dissimilarities.

Regarding the principal coordinate analysis on beta-diversity, the most clustered samples were the ones from the Avignon production site (Figure 11). This site provides the seeds from only one plant genotype and harvested in one single year. This leads to a confounding effect since we cannot determine if it is the site, the plant genotype or the harvest year that influence the most the diversity within these samples. Confounding effect can introduce bias in the analysis of influencing factors. However, one other study with a similar confounding effect does not present samples as clustered. The particularity of the Avignon samples is that the seed mother plants have been pollinated only by bees belonging to one species and one hive.

This underlines the limits of this kind of meta-analysis. In fact, all the studies gathered here were not designed with the aim of being analysed together. In fact, the variables of each experiment were merged together without a predefine design. In this way, samples used in this study came from different production sites and countries, different plant genotypes, harvest years and methods, inoculation and process types. All of these differences allow us to formulate the following hypothesis: if there is some ASVs detected in all these samples within all of these variables, they will be ubiquitous on seed and thus they would be part of the seed core microbiota.

In this meta-analysis, four main phyla were detected among all the samples: *Proteobacteria*, *Bacteroidetes*, *Actinobacteria* and *Firmicutes*. Other seed microbiota studies have detected these phyla (Johnston-Monje & Raizada, 2011; Barret *et al.*, 2015; Rezki *et al.*, 2018). The same phyla were also detected within other habitats such as phyllosphere and rhizosphere (Knief *et al.*, 2012). Hence, it seems that these four phyla, and especially *Proteobacteria*, are not only present in the whole plant holobiont but they are shared among individuals. Moreover, these four phyla were described as the dominant phyla in soil (Fierer *et al.*, 2012) and in aquatic environment (Shafi *et al.*, 2017) supporting the hypothesis that these phyla are more likely to colonize the seeds first (Nelson, 2017).

Within Proteobacteria, *Pantoea* and *Pseudomonas* genera are the most prevalent bacteria. More precisely, *Pantoea agglomerans* and *Pseudomonas viridiflava* seems to be the two-main bacteria species and appear to be common to every seed. In other words, they are two fully-fledge members of the seed core microbiota. These two species have already been described on seeds of different plant species (Links *et al.*, 2014; Truyens *et al.*, 2015; Barret *et al.*, 2015; Rezki *et al.*, 2016, 2018).

The seed core microbiota is composed of bacteria species that are present on every seed. To be exhaustive, at least one seed of every plant species should be analysed to define the seed core microbiota. As every plant species are not known so far, seeds from the main plant family should be analysed to define the seed core microbiota. In this study, most of the seeds come from the *Brassicacaea* and the *Fabacaea*.

To be more exhaustive, seeds from *Poacaea* should have been analysed. Some studies describe cereals microbiota (Yang *et al.*, 2017). *Proteobacteria*, *Firmicutes* and *Actinobacteria* were also part of the barley seed core microbiota with *Proteobacteria* as the main taxa.

5. Conclusions and perspectives

By comparing and summarizing different microbiota studies, we managed to state that production site is one of the main factors driving the seed microbiota and that there is a group of 10 bacteria taxa that are always present in all the seed samples analysed. It would be interesting to analyse further the functions of these 10 taxa in the community assembly on sterile seeds. For this, representative isolates of these taxa need to be isolated from seeds and tested in sterile seeds. By doing a similar approach in roots, Niu *et al.* (Niu *et al.*, 2017), observed that only the removal of one of the dominant taxa lead to the complete loss of the community. A similar approach be done in seeds by using part of the results coming out from this master thesis. This would be a nice system to study how bacterial interactions affect the assembly of the seed microbiota.

Moreover, production site is an important factor driving seed microbiota structure, but up to know we do not know the resilience of the seed microbiota in natural conditions. Thus, an interesting research line would be to analyse how seed microbiota would be replaced or not by the soil microbiota present in the production site.

As the taxonomic composition of the seed core microbiota could be defined, its function should be investigated also. Indeed, we stated here that *Pantoea* and *Pseudomonas* are present on every seed forming the dominant members of the seed core microbiota. This is certainly for a reason as that they may have a primordial function that confers selective advantages in the seed habitat. To analyse their functions, the total DNA should be extracted and sequenced. This would be a metagenomic approach as the total DNA from a habitat is a metagenome. In addition, an approach based on RNA could be implement also to study expression patterns. This would be a metatranscriptomic approach. Indeed, DNA amplicon sequencing allows us to reveal the bacteria present in an environment but not its functions and if these bacteria are active or not. And even if its active, which genes exactly are expressed. An RNA amplicon sequencing approach allows us to reveal which genes are expressed and by deduction which bacteria express these genes (Klappenbach et al., 2000). The limiting part of the metagenomic and metatranscriptomic analysis reside in the DNA and RNA extraction. For metagenomic analysis, you need a high concentration of microbial DNA that is not always easy to achieve. In the case of metatranscriptomic analysis, you have very often contamination with plant RNA and rRNA.

As stated before, there is different ways to study the microbiota. Even if we consider only a DNA sequencing approach, different techniques are available (Quail et al, 2012). However, to allow more meta-analysis as we have done here, only studies using the same methods and more precisely the same amplification primers can be gathered. However, regarding our results with the 16S gene and the *gyrB* gene, we noticed that the number of phyla and genus were lower for the *gyrB* even if it is supposed to allow a finer taxonomic determination. This fact is explained by the taxonomic database of *gyrB* which is

less filled out than the 16S one. Therefore, to systematise the kind of meta-analysis we have made here, some primers could be systematically used to allow the comparison between studies. In addition, other gene could be amplified also to allow a taxonomic determination until the lowest possible rank. But these other gene should have taxonomic database as filled out as possible. This kind of protocol could provide a large amount of data analysable together. Certainly, a large amount of data is needed to explore the abounding diversity of microbiota.

As seed are the starting and the ending point of a plant life cycle, an interesting approach could be to follow the ASVs and their dynamics through an entire life cycle (Nelson, 2017). To this purpose, a seed lot can be sown in a culture chamber to avoid the site effect. As the amplicon sequencing approach is destructive, the seed lot should be huge enough. Then, samples will be collected all along the entire plant life. The different organs can be investigated. This approach will allow to detect if there is ASV present all along the plant life cycle, if there is some that colonise the other organs from the seed and to compare the different habitats within one plant.

6. Bibliography

16S Metagenomic Sequencing Library Preparation. *Illumina*. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

Adam E, Bernhart M, Müller H, Winkler J, Berg G. **2016**. The Cucurbita pepo seed microbiome: genotype-specific composition and implications for breeding. *Plant and Soil* **422**: 35–49.

Anderson MJ. **2001**. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**: 32–46.

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.

Barret M, Guimbaud J-F, Darrasse A, Jacques M-A. **2016**. Plant microbiota affects seed transmission of phytopathogenic microorganisms. *Molecular Plant Pathology* **17**: 791–795.

Berg G. **2009**. Plant–microbe interactions promoting plant growth and health: perspectives for controlled use of microorganisms in agriculture. *Applied Microbiology and Biotechnology* **84**: 11–18.

Bulgarelli D, Schlaeppi K, Spaepen S, Themaat EVL van, Schulze-Lefert P. **2013**. Structure and Functions of the Bacterial Microbiota of Plants. *Annual Review of Plant Biology* **64**: 807–838.

Busby PE, Ridout M, Newcombe G. **2016**. Fungal endophytes: modifiers of plant disease. *Plant Molecular Biology* **90**: 645–655.

Callahan BJ, McMurdie PJ, Holmes SP. **2017**. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**: 2639–2643.

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.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, *et al.* 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**: 1621–1624.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. **2011**. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108**: 4516–4522.

Chee-Sanford JC, Williams MM, Davis AS, Sims GK. **2006**. Do microorganisms influence seed-bank dynamics? *Weed Science* **54**: 575–587.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. **2003**. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* **31**: 3497–3500.

Clamp M, Cuff J, Searle SM, Barton GJ. **2004**. The Jalview Java alignment editor. *Bioinformatics* **20**: 426–427.

Combes-Meynet E, Pothier JF, Moënne-Loccoz Y, Prigent-Combaret C. **2010**. The Pseudomonas Secondary Metabolite 2,4-Diacetylphloroglucinol Is a Signal Inducing Rhizoplane Expression of Azospirillum Genes Involved in Plant-Growth Promotion. *Molecular Plant-Microbe Interactions* **24**: 271–284.

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, Meyer D, Demarty M, Durand K, Bureau C, Manceau C, Jacques M-A. **2008**. The Type III Secretion System of Xanthomonas fuscans subsp. fuscans Is Involved in the Phyllosphere Colonization Process and in Transmission to Seeds of Susceptible Beans. *Applied and Environmental Microbiology* **74**: 2669–2678.

De Vleesschauwer D, Höfte M. **2009**. Chapter 6 Rhizobacteria-Induced Systemic Resistance. In: Advances in Botanical Research. Advances in Botanical Research. Academic Press, 223–281.

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C, Vorholt JA. **2009**. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 16428–16433.

Dombrowski N, Schlaeppi K, Agler MT, Hacquard S, Kemen E, Garrido-Oter R, Wunder J, Coupland G, Schulze-Lefert P. **2017**. Root microbiota dynamics of perennial Arabis alpina are dependent on soil residence time but independent of flowering time. *The ISME Journal* **11**: 43–55.

Edwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S, Eisen JA, Sundaresan V. **2015**. Structure, variation, and assembly of the root-associated microbiomes of rice. *Proceedings of the National Academy of Sciences* **112**: E911–E920.

Emmert EAB, Handelsman J. **2006**. Biocontrol of plant disease: a (Gram-) positive perspective. *FEMS Microbiology Letters* **171**: 1–9.

Fahlgren C, Hagström Å, Nilsson D, Zweifel UL. **2010**. Annual Variations in the Diversity, Viability, and Origin of Airborne Bacteria. *Applied and Environmental Microbiology* **76**: 3015–3025.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. **2012**. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences* **109**: 21390–21395.

Goggin DE, Emery RJN, Kurepin LV, Powles SB. **2015**. A potential role for endogenous microflora in dormancy release, cytokinin metabolism and the response to fluridone in Lolium rigidum seeds. *Annals of Botany* **115**: 293–301.

Hill MO. **1973**. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**: 427–432.

Hirano SS, Upper CD. **2000**. Bacteria in the Leaf Ecosystem with Emphasis onPseudomonas syringae—a Pathogen, Ice Nucleus, and Epiphyte. *Microbiology and Molecular Biology Reviews* **64**: 624–653.

Jimenez JJ. **2018**. Unidentified microscopic seed, surrounded by bacteria. *Shutterstock*. https://www.shutterstock.com/video/clip-12440729-stock-footage-unidentified-microscopic-seed-surrounded-by-bacteria.html

Johnston-Monje D, Raizada MN. **2011**. Conservation and Diversity of Seed Associated Endophytes in Zea across Boundaries of Evolution, Ethnography and Ecology. *PLOS ONE* **6**: e20396.

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.

Knief C, Delmotte N, Chaffron S, Stark M, Innerebner G, Wassmann R, Mering C von, Vorholt JA. **2012**. Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *The ISME Journal* **6**: 1378–1390.

Lindow SE, Brandl MT. **2003**. Microbiology of the Phyllosphere. *Applied and Environmental Microbiology* **69**: 1875–1883.

Links MG, Demeke T, Gräfenhan T, Hill JE, Hemmingsen SM, Dumonceaux TJ. **2014**. Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on Triticum and Brassica seeds. *The New Phytologist* **202**: 542–553.

Lopez-Velasco G, Carder PA, Welbaum GE, Ponder MA. **2013**. Diversity of the spinach (Spinacia oleracea) spermosphere and phyllosphere bacterial communities. *FEMS Microbiology Letters* **346**: 146–154.

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S, Tremblay J, Engelbrektson A, Kunin V, Rio TG del, *et al.* 2012. Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**: 86–90.

Lymperopoulou DS, Adams RI, Lindow SE. **2016**. Contribution of Vegetation to the Microbial Composition of Nearby Outdoor Air. *Applied and Environmental Microbiology* **82**: 3822–3833.

Malfanova NV. **2013**. Endophytic bacteria with plant growth promoting and biocontrol abilities. PhD thesis, Leiden University

Martin M. **2011**. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Maude RB. **1996**. *Seedborne diseases and their control: principles and practice.* Wallingford: CAB INTERNATIONAL.

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

Mendes R, Kruijt M, Bruijn I de, Dekkers E, Voort M van der, Schneider JHM, Piceno YM, DeSantis TZ, Andersen GL, Bakker PAHM, *et al.* 2011. Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. *Science* **332**: 1097–1100.

Müller DB, Vogel C, Bai Y, Vorholt JA. **2016**. The Plant Microbiota: Systems-Level Insights and Perspectives. *Annual Review of Genetics* **50**: 211–234.

Nelson EB. **2017**. The seed microbiome: Origins, interactions, and impacts. *Plant and Soil* **422**: 7–34.

Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, *et al.* 2013. Patterns and Processes of Microbial Community Assembly. *Microbiology and Molecular Biology Reviews* **77**: 342–356.

Niu B, Paulson JN, Zheng X, Kolter R. **2017**. Simplified and representative bacterial community of maize roots | PNAS. *PNAS* **114**: E2450–E2459.

Peiffer JA, Spor A, Koren O, Jin Z, Green Tringe S, Dangl JL, Buckler ES, Ley RE. **2013**. Diversity and heritability of the maize rhizosphere microbiome under field conditions | PNAS. *PNAS* **16**: 6548–6553.

Pérez-García A, Romero D, de Vicente A. **2011**. Plant protection and growth stimulation by microorganisms: biotechnological applications of Bacilli in agriculture. *Current Opinion in Biotechnology* **22**: 187–193.

Porras-Alfaro A, Bayman P. **2011**. Hidden Fungi, Emergent Properties: Endophytes and Microbiomes. *Annual Review of Phytopathology* **49**: 291–315.

Rambaut A. **2007**. FigTree, a graphical viewer of phylogenetic trees. http://tree.bio.ed.ac.uk/software/figtree/

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**.

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.

Rodriguez RJ, Jr JFW, Arnold AE, Redman RS. **2009**. Fungal endophytes: diversity and functional roles. *New Phytologist* **182**: 314–330.

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**.

Santhanam R, Luu VT, Weinhold A, Goldberg J, Oh Y, Baldwin IT. **2015**. Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. *Proceedings of the National Academy of Sciences* **112**: E5013–E5020.

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.

Shafi S, Kamili AN, Shah MA, Parray JA, Bandh SA. **2017**. Aquatic bacterial diversity: Magnitude, dynamics, and controlling factors. *Microbial Pathogenesis* **104**: 39–47.

Spaepen S, Vanderleyden J, Okon Y. **2009**. Chapter 7 Plant Growth-Promoting Actions of Rhizobacteria. In: Advances in Botanical Research. Advances in Botanical Research. Academic Press, 283–320.

Sugiyama A, Bakker MG, Badri DV, Manter DK, Vivanco JM. **2012**. Relationships between Arabidopsis genotype-specific biomass accumulation and associated soil microbial communities. *Botany* **91**: 123–126.

System Specification Sheet: MiSeq® System. *Illumina*. https://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf

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

Truyens S, Weyens N, Cuypers A, Vangronsveld J. **2015**. Bacterial seed endophytes: genera, vertical transmission and interaction with plants. *Environmental Microbiology Reports* **7**: 40–50.

Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE. **2016**. The Phyllosphere: Microbial Jungle at the Plant–Climate Interface. *Annual Review of Ecology, Evolution, and Systematics* **47**: 1–24.

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.

Vokou D, Vareli K, Zarali E, Karamanoli K, Constantinidou H-IA, Monokrousos N, Halley JM, Sainis I. **2012**. Exploring Biodiversity in the Bacterial Community of the Mediterranean Phyllosphere and its Relationship with Airborne Bacteria. *Microbial Ecology* **64**: 714–724.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. **2009**. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.

Weller DM. **2007**. Pseudomonas Biocontrol Agents of Soilborne Pathogens: Looking Back Over 30 Years. *Phytopathology* **97**: 250–256.

Whittaker RH. **1972**. Evolution and Measurement of Species Diversity. *Taxon* **21**: 213–251.

Yang L, Danzberger J, Schöler A, Schröder P, Schloter M, Radl V. **2017**. Dominant Groups of Potentially Active Bacteria Shared by Barley Seeds become Less Abundant in Root Associated Microbiome | Plant Science. *Frontiers in Plant Science*: 1005.

Zamioudis C, Pieterse CMJ. **2011**. Modulation of Host Immunity by Beneficial Microbes. *Molecular Plant-Microbe Interactions* **25**: 139–150.

List of appendixes

Appendix I: SFR Quasav Brochure

Appendix II: P-value obtained with a Kruskal-Wallis statistic test

Appendix III: Determination coefficient of each significant variable for two beta diversity indices (Jaccard and Bray-Curtis distances) with their corresponding p-value and significance codes

Appendix IV: Table of the ASVs ranked by their total abundance and that have a total abundance > 100,000

Appendix I: SFR Quasav Brochure

SFR 4207 Angers
Qualité et Santé du Végétal
Quasav

Federative Research Structure « Plant Quality and Health »

The Federative Research Structure Plant Health and Quality (SFR4207 QUASAV) was created in 2008 with the ambition of promoting research development in plant sciences in the Region Pays de la Loire by strengthening the cohesion between research units in this field and increasing their national and international visibility. The scientific objectives of the SFR focus on the understanding and management of key characteristics and processes that govern plant health and quality of products, in a sustainable plant production perspective.

Main partners

UMR IRHS
EA SIFCIR
EA SONAS
EA LBPV
UP GRAPPE
UP LEVA
UP EPHor
Unité BVO
UE Horticole

Associated partner
Team EGI of UMR IGEPP

Other partner
Team PRP of UR BIA
GEVES
Vegepolys

Three federative research topics:
- Sustainable management of plant health
- Seed biology, quality and health
- Horticultural plant product qualities

2 platforms

PHENOTIC
SEMENCES & PLANTES
Phenotyping service

Senso Veg
Sensory analysis and consumer perception

4 technical facilities

IMAC
Cellular Imaging

ANAN
Nucleic acid analysis

PIAM PHYTO
Phytochemicals and secondary metabolites analysis

COMIC
Microorganisms collection

Contact details:
Director : Marie-Agnès JACQUES, **Deputy director** : Thomas GUILLEMETTE
Contacts : marie-agnes.jacques@inra.fr, thomas.guillemette@univ-angers.fr
Information : https://www.sfrquasav-angers.org/

Appendix II: P-value obtained with a Kruskal-Wallis statistic test.

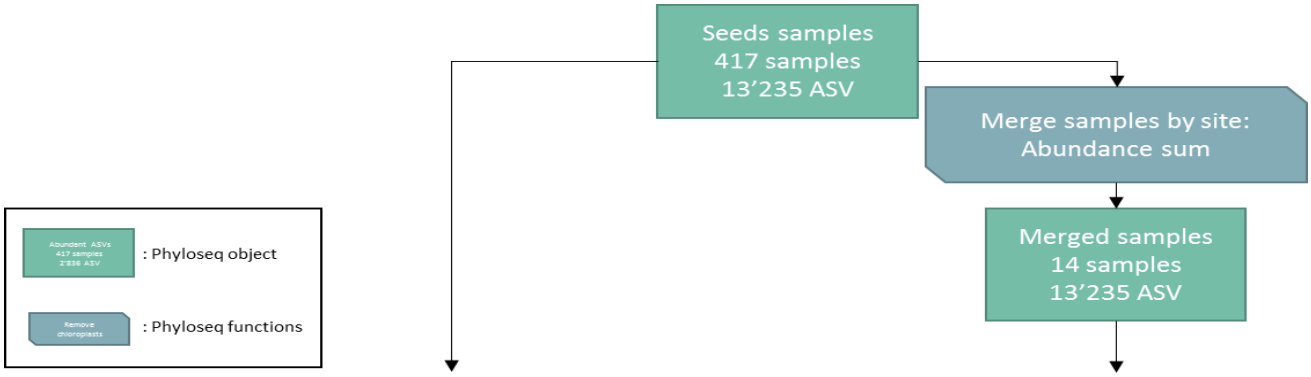
P-value < 5% (a) are statistically significant, other (b) aren't. In bold text, the variable has significant p-value for the three indices.

	Observed richness		Shannon's index		Inverse Simpson's index	
Experiences	**< 2.20E-16**	**a**	**1.66E-12**	**a**	**1.17E-09**	**a**
Number of production site per experience	**< 2.20E-16**	**a**	**1.53E-05**	**a**	**1.57E-04**	**a**
Number of plant species per experience	**< 2.20E-16**	**a**	**3.88E-07**	**a**	**4.36E-05**	**a**
Number of plant varieties per experience	**< 2.20E-16**	**a**	**5.51E-06**	**a**	**3.29E-04**	**a**
Plant species	**< 2.20E-16**	**a**	**1.20E-04**	**a**	**2.41E-03**	**a**
Seed harvest method	7.96E-05	a	2.76E-01	b	9.44E-01	b
Type of inoculation (including no inoculation)	7.83E-10	a	1.33E-01	b	1.80E-01	b
Harvest year	**< 2.20E-16**	**a**	**1.59E-10**	**a**	**9.23E-09**	**a**
Production site	**5.24E-04**	**a**	**5.05E-04**	**a**	**1.75E-03**	**a**
Process (native or disinfection)	1.87E-05	a	1.05E-01	b	1.22E-01	b
Production country	**5.24E-04**	**a**	**5.05E-04**	**a**	**1.75E-03**	**a**
Plant taxonomic family	1.14E-01	b	3.50E-01	b	4.56E-01	b
Plant taxonomic genus	**< 2.20E-16**	**a**	**4.48E-04**	**a**	**1.15E-02**	**a**
Plant variety	**< 2.20E-16**	**a**	**8.54E-10**	**a**	**5.29E-08**	**a**
Pollination type	3.25E-02	a	1.09E-01	b	1.18E-01	b

Appendix III: Determination coefficient of each significant variable for two beta diversity indices (Jaccard and Bray-Curtis distances) with their corresponding p-value and significance codes ('***': p-value≤0.001; '**': p-value≤0.01; '*': p-value≤0.05; '.': p-value≤0.1; '-': p-value≤1; ' ': not significant). These results were obtained after performing a PERMANOVA with or without the singleton samples. In bold text, the variable has significant p-value ≤ 0.0001 each time.

	Bray-Curtis		Jaccard	
Variable	R^2 (%)	P-value	R^2 (%)	P-value
Site	**30.687**	**≤ 0,0001 *** **	**22.503**	**≤ 0,0001 *** **
Plant	**7.627**	**≤ 0,0001 *** **	**6.378**	**≤ 0,0001 *** **
Genotype	**6.123**	**≤ 0,0001 *** **	**6.178**	**≤ 0,0001 *** **
Experience	**4.366**	**≤ 0,0001 *** **	**4.013**	**≤ 0,0001 *** **
Inoculation	0.240	0.0613 .	0.326	0.0190 *
Years	**1.086**	**0,0002 *** **	**1.111**	**≤ 0.0001 *** **

Appendix IV: Table of the ASVs ranked by their total abundance and that have a total abundance > 100,000. The mean, the standard deviations, the prevalence (raw and percentage) were calculated for both all seed samples and samples merged by site. . In bold text, the ASV is present in every site.



	Within all samples					Within samples merged by site			
ASV	Abundance sum	Mean	Standard deviation	Prevalence (raw)	Prevalence (%)	Mean	Standard deviation	Prevalence (raw)	Prevalence (%)
ASV1_Pantoea	**5,41E+06**	**1,30E+04**	**2,01E+04**	**393**	**94,24**	**3,87E+05**	**6,49E+05**	**14**	**100,00**
ASV3_Pseudomonas	1,97E+06	4,73E+03	2,17E+04	236	56,59	1,41E+05	3,00E+05	13	92,86
ASV5_Enterobacter	**1,12E+06**	**2,69E+03**	**9,49E+03**	**360**	**86,33**	**8,03E+04**	**1,37E+05**	**14**	**100,00**
ASV4_Pseudomonas	**1,07E+06**	**2,56E+03**	**5,83E+03**	**374**	**89,69**	**7,63E+04**	**1,30E+05**	**14**	**100,00**
ASV6_Pseudomonas	5,82E+05	1,39E+03	1,08E+04	25	6,00	4,15E+04	1,55E+05	4	28,57
ASV8_Pseudomonas	3,94E+05	9,44E+02	2,80E+03	290	69,54	2,81E+04	7,99E+04	9	64,29
ASV7_Pseudomonas	**3,81E+05**	**9,13E+02**	**2,86E+03**	**343**	**82,25**	**2,72E+04**	**4,25E+04**	**14**	**100,00**
ASV10_Pantoea	2,82E+05	6,76E+02	1,24E+03	187	44,84	2,01E+04	6,53E+04	3	21,43
ASV2_Xanthomonas	2,63E+05	6,30E+02	4,08E+03	269	64,51	1,88E+04	3,76E+04	12	85,71
ASV12_Serratia	2,60E+05	6,23E+02	2,79E+03	198	47,48	1,86E+04	6,46E+04	8	57,14
ASV14_Sphingomonas	**2,09E+05**	**5,01E+02**	**1,35E+03**	**320**	**76,74**	**1,49E+04**	**3,18E+04**	**14**	**100,00**
ASV15_Xanthomonas	2,03E+05	4,88E+02	6,50E+03	19	4,56	1,45E+04	3,64E+04	10	71,43
ASV18_Chryseobacterium	1,95E+05	4,69E+02	1,57E+03	215	51,56	1,40E+04	4,63E+04	11	78,57
ASV13_Paenibacillus	1,88E+05	4,51E+02	1,15E+03	325	77,94	1,34E+04	2,44E+04	13	92,86
ASV11_Pseudomonas	1,86E+05	4,46E+02	3,35E+03	152	36,45	1,33E+04	2,11E+04	12	85,71
ASV9_Gilliamella	1,44E+05	3,46E+02	1,64E+03	29	6,95	1,03E+04	3,86E+04	2	14,29
ASV23_Pseudomonas	1,44E+05	3,45E+02	1,47E+03	165	39,57	1,03E+04	3,09E+04	5	35,71
ASV22_Escherichia /Shigella	1,43E+05	3,43E+02	1,58E+03	113	27,10	1,02E+04	2,74E+04	8	57,14
ASV17_Arsenophonus	1,36E+05	3,26E+02	1,97E+03	27	6,47	9,72E+03	3,64E+04	1	7,14
ASV24_Yersinia	1,29E+05	3,09E+02	6,22E+03	37	8,87	9,21E+03	3,41E+04	8	57,14
ASV16_Pantoea	1,27E+05	3,05E+02	6,06E+03	20	4,80	9,09E+03	3,32E+04	6	42,86
ASV28_Neorhizobium	1,04E+05	2,49E+02	6,36E+02	311	74,58	7,43E+03	2,07E+04	12	85,71

RÉSUMÉ

Les graines portent différents assemblages microbiens dont les compositions et les fonctions restent largement méconnues. La méta-analyse présent dans ce rapport de master étudie et compare sept différentes études sur le microbiote des semences. Regroupant 417 échantillons, la richesse, la diversité et la composition taxonomique de la communauté bactérienne ont été analysées. L'étude de la composition taxonomique a été faite à l'aide de deux gènes, la région v4 du gène de la sous-unité 16S de l'ARN ribosomal et le gène *gyrB* de la sous-unité de la gyrase bactérienne. Il en ressort que le site de production est le facteur majeur influençant la structure de la communauté bactérienne associée aux semences. De plus, quatre phyla sont majoritaires et ubiquitaires : *Proteobacteria*, *Actinobacteria*, *Bacteroidetes* et *Firmicutes*. Plus précisément, deux espèces bactériennes ont été détectées dans tous les sites de productions : *Pantoea agglomerans* et *Pseudomonas viridiflava*. De plus, ces deux espèces bactériennes sont présentes à des abondances bien supérieures aux autres espèces, représentant jusqu'à plus de 90% de l'abondance totale. Nous émettons donc ici l'hypothèse que ces deux espèces sont les membres principaux du microbiote cœur des graines. Cette étude s'inscrit donc dans la suite de nombreuses autres menés sur la composition taxonomique du microbiote des semences et sur la détermination des facteurs influençant l'assemblage de ce microbiote. Afin de poursuivre le travail d'investigation du microbiote des semences, une étude suivant les taxons bactériens tout au long de la vie de la plante pourrait être menées. En complément de l'étude sur la structure du microbiote, des études de métagénomiques et de métatranscriptomiques devraient être menées afin de définir les fonctions potentielles du microbiote des graines et déterminer quels gènes sont exprimés.

Mots-clés : ASV ; 16S rRNA ; *gyrB* ; production site ; *Proteobacteria ; Pantoea ; Pseudomonas*
.

ABSTRACT

Seeds carry diverse microbial assemblages whose compositions and functions remain largely unknown. The meta-analysis present in this master thesis study and compare the microbial communities of seven different seed microbiota studies. Pooling 417 samples, richness, diversity and taxonomic composition of the bacterial community were analysed. The taxonomic composition analysis was implemented by two genes: the v4 region of the rRNA 16S sub-unit and the *gyrB* gene of the bacterial gyrase sub-unit. We stated here that the production site if the most influent factor shaping the structure of the seed bacterial community. Four main phyla are dominant and ubiquitous forming the seed core microbiota: *Proteobacteria*, *Actinobacteria*, *Bacteroidetes* and *Firmicutes*. Moreover, two bacterial species were detected in every production site: *Pantoea agglomerans* and *Pseudomonas viridiflava*. These two bacterial species were also present at really high abundance compare to the other species, representing more than 90% of the total abundance. We hypothesize that these two species are the main members of the seed core microbiota. Therefore, this study follows several other works carried out to determine the taxonomic composition of the seed microbiota. To pursue the investigative work on the seed microbiota, a study following specific taxa all along the plant life cycle could be carried out. In addition, metagenomic and metatranscriptomic approaches should be implemented to describe the potential functions of the seed microbiota and whose genes are really expressed.

Keywords: ASV; 16S rRNA; *gyrB* ; production site ; *Proteobacteria ; Pantoea*; *Pseudomonas*