University of Milan-Bicocca

Department of Biotechnology and Biosciences
Ph.D. in Biology – Cycle XXVI

Biodiversity in the Era of Big Data.

On the problem of taxonomy assignment and the distribution of diversity in complex biological systems.

PhD thesis by:
Anna Sandionigi
Matr. N 745174

Project Supervisor: Maurizio Casiraghi PhD

Contents

List of Figures

List of Tables

1 General introduction

The study of complex biological matrices is a remarkable hot topic in biology. Soil, water, gut content are some of these matrices characterized by a prominent number of organisms living in tight connection. Hundreds or thousands of species and/or strains could be present in the same sample coming from different habitats (e.g. soil ecosystem) and showing inter-relationships, mainly energetic, to guarantee their ecosystem health functioning (Gaston and Spicer, 2009; Brussaard, 1997; Wagner et al., 2011; Tylianakis et al., 2008).

The discrimination and/or identification of the different biological entities, at least for the eukaryotic components, using traditional morphological approaches is relatively complicated, requiring a specialist in each taxonomic groups and generally an appropriate long time to achieve a correct identification and classification (Mace, 2004; Hajibabaei et al., 2011). On the other side, the identification of Archaeal and Bacterial domains has been shown to be impossible using these approaches (Hedrick, 2011).

To overcome these limitations, molecular approaches have demonstrated to be valid alternatives where PCRs, cloning, DNA sequencing and bioinformatics analysis of sequence differences were used as the standard protocol (Janda and Abbott, 2007; Huang et al., 2009). Nowadays, the genomic massive sequencing revolution, generated by the heterogeneous techniques collectively known as Next Generation Sequencing (NGS), has become the new gold standard (Deutschbauer et al., 2006; Bik et al., 2012; Shokralla et al., 2012). These new platforms can provide billions of sequence reads in a single experiment, which corresponds to an improvement of at least five orders of magnitude when compared to traditional Sanger sequencing using capillary electrophoresis (Taberlet et al., 2012).

NGS techniques can be used to investigate the biodiversity in complex matrices using two different approaches: (i) a metagenomic approach (Riesenfeld et al., 2004; Xie et al., 2010; Wooley and Ye, 2009) or (ii) a massive sequencing of few selected molecular markers.

The first approach implements a suite of genomic technologies and bioinformatics tools to directly access the genetic content of entire communities of organisms. It does not take into account taxonomical information of organisms, selecting only random pieces of the genomes present in the matrix independently from their origin. The second approach was also defined metabarcoding (Buee et al., 2009; Creer et al., 2010; Fonseca et al., 2010; Hajibabaei et al., 2011; Yu et al., 2012) and show a more limited and specific goal than metagenomic.The basis of this method is to use the target marker/s to achieve the taxonomic discrimination of the living organisms present in complex matrices, hopefully at the species rank, in a fast, reliable and reproducible way. Currently, in metabarcoding many molecular markers are used, exploiting NGS technologies, to investigate both prokaryotic (mainly) and eukaryotic biodiversity in a given environment. The 16S rRNA gene information has been and is being widely employed in many prokaryotic biodiversity studies of human gut (Sogin et al., 2006; Dethlefsen et al., 2008) microbial communities in the deep seas (Mason et al., 2012) and food microbiome (Humblot and Guyot, 2009). Similarly, *coxI* (the mitochondrial cytochrome c oxidase subunit-one) barcode region and 18S nuclear small subunit (18S rDNA or nSSU) has been used to get biodiversity profiling through NGS technologies mainly in metazoans biomonitoring programs (Fonseca et al., 2010; Hajibabaei et al., 2011).

With the exception of few DNA barcoding studies, where the entire *coxI* barcode (Yu et al., 2012) or a fragment of it (mini-barcode) (Hajibabaei et al., 2011) were sequenced on NGS platform, the vast majority of DNA barcoding studies, aiming at eukaryotic species identification, are still conducted with the traditional sequencing technology (i.e. Sanger method). The present thesis consists of four sections which overpass detailed aspects of the analysis of biodiversity and the issues associated and elaborate both promises and pitfalls of coupling DNA

barcode approach with high-throughput pyrosequencing in two different cases of biodiversity assessment. In the following a brief description of each section is provided:

Section 1: Introduction to the biodiversity analysis problem
In this section an analysis of the main methods used to investigate the biodiversity and their related problems has been addressed. Emphasis has been put on the integrated approach between the classic DNA barcoding approach and the advantages of high processivity guaranteed by next generation sequencing technologies. Furthermore, the state of art regarding bioinformatics methods for species assignment and biodiversity patterns elaboration including phylogenetic diversity analysis are described.

Section 2: Targeted Sequencing on Metazoan Communities
In this section, the precision and the accuracy of denoising procedure and the candidate parameters able to reduce sequence error rate are investigated. This work also proposed an innovative taxon assignment pipeline. In addition, a novel library preparation method allowing the sequencing of the entire *coxI* barcode region (approximately 700 bp) on 454 pyrosequencing platform (Roche Life Science) is proposed. To address the objectives, metazoan communities coming from complex environmental matrix (soil) were considered.

Section 3: Microbiota invasion mediated by *Varroa destructor* to *Apis mellifera*
The starting hypothesis of this section is that varroa mites play a fundamental role in the alteration of bacterial composition of honey bee larvae, acting not only as a vector, but also as a sort of an open "door" through which exogenous bacteria alter the mechanisms of primary succession in the "simple" honey bee larval microbiome. To explore these dynamics a classical microbial communities analysis approach and a new approach considering the phylogenetic entropy as a measure of biodiversity were tested. The varroa and honey bee bacterial communities were studied through barcoded amplicon pyrosequencing

method , taking advantage of the NGS methods (Blow, 2008; Metzker, 2010) and the opportunity to detect uncultured and uncultivable bacteria allowed by such techniques.

Section 4: general conclusions and perspectives
General conclusions and future promises highlighted by the above mentioned experiments are illustrated in this section.

Notes for the reader.
The case studies presented in this Thesis are investigated with a multidisciplinary approach that required the team-work contribution of people with different expertises. In particular I collaborated with Saverio Vicario and Bachir Balech (ITB Bari, Italy) for the bioinformatics part of analysis. For both case studies, I actively contributed to the definition of the scientific aims and scopes of the research and to the planning of the experimental design while I indirectly followed the laboratory analysis which were conducted by other colleagues. Then, I leaded the bioinformatic and ecological analysis of the data, focusing my attention to the problematic of this step. Finally I contributed to the discussion and interpretation of the biological problems.

2 Section 1

Introduction to the biodiversity analysis problem

2.0.1 Definition of Biodiversity

The Millennium Ecosystem Assessment (MA) defines biodiversity as: "...the diversity among living organisms in terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are part. It includes diversity within and between species and the diversity of ecosystems". Additionally, biodiversity encompasses all forms, levels and combinations of natural variation and thus serves as a broad unifying concept (Gaston and Spicer, 2009).

Concerning estimates of biodiversity, is still debated what exactly should be counted and how can be accomplished such a count. This uncertainty stems mainly from the fact that biodiversity is a multilevels concept whose basic building blocks are still the subject of debate (Faith, 2002). These blocks can be divided into three groups: (i) genetic diversity, (ii) organismal or phenotypic diversity, and (iii) ecological or relational diversity. Genetic diversity encompasses genetic structures (nucleotides, genes, chromosomes) and variation in the genetic makeup between individuals within a population and between populations.

Organismal diversity includes taxonomic hierarchy and its components, from individuals upwards to species, genera and beyond. Ecological diversity embraces the scales of ecological differences from populations, across niches and habitats, up to biomes. Although presented separately, these groups are closely related, and in some cases share ele-

ments in common (e.g. populations concept appear in all three) (Gaston and Spicer, 2009).

The most important target of the biodiversity monitoring is the species identification and count. Species richness is directly correlated to the role that one species or its population exhibit in a given ecosystem (Matthews et al., 2011). For example, the ecosystem productivity, can be influenced by the number of species (Loreau, 2010), the phylogenetic diversity of species (Cadotte et al., 2008) and the evolutionary history of species (Gravel et al., 2010). On the whole, the different ways in which species interact (e.g. through foraging, defense, territoriality and so on) could drive changes in community dynamics that, in turn, affect ecosystem functions (Loreau, 2010).

Moreover, the complexity of ecosystems comprises various components including both biotic (plants, vertebrates and invertebrates) and local abiotic factors (soil and water conditions, humidity, temperature, etc.). These components show inter-dependencies emphasizing many potential direct and indirect interactions that may occur among them and their environment. For this purpose, in order to understand the relationships between species in an ecosystem is essential to identify correctly as many species as possible.

Traditionally, species identification and classification have been the domain of specialists in taxonomy, providing a nomenclature and a several key prerequisites for numerous biological studies but this traditional method is a bottleneck. Indeed, it has been estimated that the total number of eukaryotic species on the earth to approximately 8.7 million.

Despite the big efforts on taxonomic classification of new species, around 86% of existing species on earth and 91% of species in the ocean still wait description (Mora et al., 2011). This means that given to the average productivity of a standard taxonomist and the time needed to properly describe a species, up to 100,000 taxonomists should be required simply to sustain the ability to recognize them the still unknown biodiversity.

2.0.2 Molecular biodiversity

The DNA barcoding idea

The use of molecular data to discriminate living beings is quite old, and has one of the founders in Carl Woese and his work in identifying prokaryotic species back in the '70s (Woese and Fox, 1977). In recent years, molecular approaches such as DNA barcoding have become very popular as a valid support for species identification. This approach was formally introduced in 2003 by Hebert et al. (2003) and is based on the idea that through the analysis of the molecular variability of standard DNA region(s) it is possible to discriminate all the biological entities. A single gene approach is consistent with the fact that DNA barcoding is a generalist identification meant to be used on a vast range of unknown samples, and species identification is not necessarily the unique result of such an identification system (see below).

An international network, the Consortium for the Barcode of Life (CBOL[1]), was soon established with the aim of coordinating the efforts of laboratories throughout the world, towards the definition of a database of documented and vouchered reference sequences. The actual evolution of CBOL is the International Barcode of Life[2]. The database will be the universal reference library for comparisons of any unidentified taxa (Ratnasingham and Hebert, 2007). Several local and international projects were launched (see a full list of these projects at CBOL). Nowadays, the DNA barcoding approach is applied with success in prokariots[3] (Zhou et al., 2008), in several groups of metazoans (Hebert et al., 2004; Smith et al., 2005) and fungi (Seifert et al., 2007), and in recent years it has been rapidly extended to plants (Hollingsworth et al., 2009) and algae (Saunders, 2005).

The success of DNA barcoding resides in the conjugation of three

[1] http://www.barcoding.si.edu

[2] iBOL, http://www.ibol.org

[3] Is correct to point that in the bacterial sphere the term DNA barcoding is not used. But the identification of bacterial entity is largely based on the same principle of the DNA barcoding in use for eukaryotes (the use of a single marker for identifying).

innovations: the molecularization of taxonomy, the computerization of data by using informatics supports and standardization by using a common DNA region for species. The main goal of Hebert et al. (2003) was to provide a fast and standardized method to perform reliable species identification, even in the absence of expert taxonomists. Some of the initial claims are nowadays interpreted in a different way. For instance, it seems clear that it is almost impossible to separate the molecular identification system from morphological characterization. For these reasons, in database construction the association between different data sources (morphological, chemical and molecular) should be carefully established and this is still a limit when considering the data from the prokaryotic world.

DNA barcoding became a successful tool for biodiversity investigation because of its simplicity. But this simplicity can also be a limit, because it is easy to confuse simplicity with inaccuracy. Indeed, there is sometimes the feeling that DNA barcoding is an inaccurate tool because at the present state of the art, in several cases, identification is impossible or largely approximate (i.e. identification of high rank taxa or association with phylogenetically distant taxa) (Taylor and Harris, 2012). The problem is real, but is mainly due to the fact that reference databases are still scarcely populated for many parts of the Tree of Life.

Species identification and the role of reference databases

Being a multidisciplinary approach, DNA barcoding has been recently displayed as an integrated taxonomy toolkit. DNA barcoding needs to firmly associate a molecular variability to some kind of morphological variability. In this way it is possibile to identify unknown organisms by comparing them with known oness'. To perform all these tasks the method need to integrate ecological, genetic, and morphological data that guarantee robustness and precision in species assignment analysis (Miller, 2007; Smith and Fisher, 2009; Ferri et al., 2009; Padial et al., 2010).

It is well known that no identification method (morphological, biochemical, genetic or what so ever based) can truly identify species, because species are entities in continuous evolution and it is theoretically impossible to define statically such dynamic matter (Casiraghi et al., 2010). Since the advent of molecular-based taxonomy, many studies contributed to define a plethora of new taxonomic entities. These entities identified by molecular approaches have been named in several ways: "Genospecies"; "Phylospecies", *sensu* (Eldredge and Cracraft, 1980); "Recognizable Taxonomic Units", RTUs, *sensu* (Oliver and Beattie, 1993); "Phylotypes" *sensu* (Moreira and López-García, 2002); "Molecular Operational Taxonomic Units", "MOTUs", *sensu* (Floyd et al., 2002). In molecular approaches, one of the most relevant entities is the Operational Taxonomic Unit (OTU) (Sokal and Rohlf, 1962) that was firstly defined in a non- molecular context. In its original use, the OTU is defined using as much characters as possible, even without knowing the "real" taxonomic value of each character. In such a context, DNA sequences are the typical data that can be used to define OTUs, because each sequence can be considered as a group of characters, not a priori weighted (Galimberti et al., 2012). Being therefore the sequence the subject at the basis of the identification, it is obvious the role of the genes reference database. Most of the incorrect assignments reside more in the completeness of data sets rather than in the data analysis system (Puillandre et al., 2009; Bruni et al., 2010; Burgess et al., 2011; Virgilio et al., 2012).

Nowadays, GenBank[4] is still the largest repository of sequences for all markers used DNA barcoding (Federhen, 2012). GenBank comparisons through the algorithm of the Basic Local Alignment Search Tool, or BLAST (Altschul et al., 1990), have been the primary fellow of geneticists and molecular biologists for many years. However, GenBank can only be a good tool for the first screening on newly generated sequences, but in the present form it is not the appropriate tool for species identification. Indeed, the reference sequences stored in this database show a high level of incorrect taxa assignment (Bidartondo et al., 2008).

There are many reasons to explain this situation, among which, a

[4]http://www.ncbi.nlm.nih.gov/genbank/

role of primary importance is played by the degree of inaccuracy of researchers, because several taxonomic details available in GenBank are incorrect or out of dated (McDonald et al., 2012). A second important reason is that in GenBank many sequences represent the only existing entry for certain species. In such a situation, it is easy to assume as real the monophyly of a species, but this can lead to an underestimation of phenomena like introgression and incomplete lineage sorting following recent speciation, the major causes of species level polyphyly (Rosenberg, 2003; Elias et al., 2007; Austerlitz et al., 2009).

On the other hand, with the present state of the art, the most relevant DNA barcoding database for metazoa is the Barcode of Life Data Systems (BOLD[5]). BOLD is still in constant evolution and updating, but it has already reached a good level of standardization and accuracy among metazoans (Ratnasingham and Hebert, 2007). In recent years, the bacterial sequences produced by large-scale environmental surveys have invaded NCBI increasing dramatically the rate of sequences with uncertain phylogenetic affiliation. This shortcoming has been addressed by several dedicated 16S databases, including the Ribosomal Database Project (Cole et al., 2009), Greengenes (DeSantis et al., 2006) and SILVA (Pruesse et al., 2007), that classify a higher proportion of environmental sequences. However, improvements are still needed because many sequences remain unclassified and numerous classification conflicts exist between the different 16S rDNA databases (DeSantis et al., 2006). It is indeed clear that the creation of a correct database for query comparison is a necessary step before performing any kind of analysis. An essential prerequisite for the proper construction of a reference DNA barcoding library is an adequate sampling coverage to fully evaluate both the intraspecific and interspecific variations (Moritz and Cicero, 2004; Meier et al., 2006; Casiraghi et al., 2010; Bergsten et al., 2012). The sampling has to be performed from distant sites to maximize the chance to observe intraspecific geographic variation among conspecifics. In such a condition, it is possible to test the hypothesis of species-level monophyly. Besides sampling coverage, taxon coverage has also to be taken into account in database construction. Unbalanced

[5]http:// www.boldsystem.org

representation of certain species within a given group is likely to greatly affect the analysis (Meyer and Paulay, 2005; Bergsten et al., 2012).

2.0.3 Contribution of the Next generation sequencing to the study of biodiversity

For the past 30 years, the Sanger method has been the dominant approach and gold standard for DNA sequencing. The commercial launch of the first massively parallel pyrosequencing platform (Roche 454) in 2005 ushered in the new era of genomic analysis now referred to as next-generation sequencing (NGS) (Voelkerding et al., 2009) or high-throughput DNA sequencing.

All commercially available NGS technologies differ from automated Sanger sequencing in that they do not require cloning of template DNA into bacterial vectors avoiding cloning difficulties and biases (Nowrousian, 2010). Moreover, these sequencing technologies provide an opportunity to generate very large amounts of sequence data in a very short time and at low cost. One of the most important applications of this technology is the ability to identify large numbers of species from complex communities (Leininger et al., 2006; Sogin et al., 2006; Mocali and Benedetti, 2010).

The increasing availability of different NGS platforms has allowed in recent years to overcome the limitations given by the classical DNA barcoding approach on identifying species within bulk environmental samles. The ability to automate a biodiversity survey of, for example, bulk macroinvertebrate samples can revolutionize large-scale biomonitoring programs that are costly, labour-intensive and time-consuming to implement across large geographic regions (Hajibabaei et al., 2011).

At the moment, sequencing platforms can produce up to 6 billions of sequence reads of 100 bp per run, with the possibility to implement paired-end experiments (http://www.molecularecologist.com/next-gen-fieldguide-2013/). Thus, it is not any more a problem to obtain several thousands of sequence reads per amplicon, and the length of the sequence reads is already fully compatible with the short fragment

lengths required for eDNA (environmental DNA) metabarcoding. There is no doubt that the technology will improve still further. As a consequence, NGS has the potential to provide an enormous amount of information per experiment from in-depth sequencing of uniquely tagged amplicons (Binladen et al., 2007; Valentini et al., 2009).

Metabarcoding - single locus sequencing

As described before the main goal of DNA metabarcoding is to identify taxa and it should be clearly differentiated from metagenomics that "describes the functional and sequence- based analysis of the collective microbial genomes contained in an environmental sample" (Riesenfeld et al., 2004). Today, metabarcoding is considered as the next generation single locus sequencing approach more prevalent among the large-scale molecular biodiversity studies. In fact, numerous studies use single locus 454 pyrosequencing approach to describe a given environment. The most informative loci published to date are: *coxI* in animals (Rougerie et al., 2009), 16S rDNA in prokariots (Singh et al., 2012), ITS in fungi (Seifert et al., 2007), *rbcL* and *matK* in plants (Chase et al., 2007; Hollingsworth et al., 2009), and 18S rDNA in protist, nematodes and algae (Powers et al., 2009; Hajibabaei et al., 2011).

Since long sequences means better taxonomic resolution, the 454 pyrosequencing platform has been up to last years the most preferable sequencing technology in molecular biodiversity studies (Hajibabaei et al., 2011) but in the very recent months other platforms are available (i.e. MySeq Illumina). 454 platform offers the longest sequence reads over the other available platforms (Margulies et al., 2005; Metzker, 2010; Nowrousian, 2010) as it is capable of providing 400-800 base long sequence reads. In the case of barcoding biodiversity, this sequence length can potentially cover the entire *coxI* sequence.

Recent study involving pyrosequencing of the standard metazoans barcode region *coxI* has accessed accurately the biodiversity content of freshwater benthic macroinvertebrate taxa in both natural sample and pooled one with known species (Hajibabaei et al., 2011). In addition,

pyrosequencing of the nuclear 18S small subunit (nSSU) highlighted the diversity of meiofaunal biosphere in marine and rainforest habitats (Creer et al., 2010) and marine metazoan biodiversity in Scottish temperate benthic ecosystem (Fonseca et al., 2010). Furthermore, a survey conducted on forest soil fungi based on ITS-1 has revealed unexpected biodiversity in Ascomycota and Basidiomycota fungi phyla, validating the effectiveness of high-throughput 454 sequencing technology in studying fungal communities in forest ecosystems (Buee et al., 2009). Also the studies of complex microbial communities have advanced considerably in recent years due to the use of high-throughput DNA sequencing technologies that yield detailed information on the composition of microbial communities (Sogin et al., 2006).

2.0.4 Bioinformatics approach to taxonomy assignment

Summarizing the previous section, the main steps in the analysis of biodiversity are: (i) the identification of the entities present in the environment under examination; (ii) the estimation of their abundances and (iii) consequently the distribution of variability. The puzzle is to have a system that can discriminate among closely related species, for which it can be expected that the distribution of variability in the chosen markers will be not clearly evident. Several publications in recent years have tried to perform a detailed comparison of the different method performances (Little and Stevenson, 2007; Ross et al., 2008; Ferri et al., 2009; van Velzen et al., 2012). The main result is that the discriminating performance of identification methods at the rank of genus or family, these performances are rather similar.

Differences among the approaches became clear when the identification of the species rank is reached. In DNA barcoding studies species are identified accordingly to the genetic information present in the standard DNA barcode region. Several methods, ranging from genetic distance, phylogeny and species delimitation are widely used in species assignment. Furthermore, based on identified species richness and abundance, diversity measures and/or phylogenetic diversity indices are calculated to accurately interpret and illustrate biodiversity of one

or more environment. On the whole, the methods used to identify taxonomic entities can be divided into three big categories: (i) similarity methods/pairwise distance; (ii) hierarchical clustering methods/tree-based; (iii) character-based and diagnostic methods. (Table 2.1 summarizes these methods, giving details and references).

Table 2.1: Details of the main DNA barcoding tools available at the present state of the art.

Class	Methods	Software	URL	Reference
Symilarity	Similarity	BLAST	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/	(Altschul et al., 1990)
	Pairwise Distance	Taxon DNA	http://taxondna.sf.net/	(Meier et al., 2006)
	Pairwise Distance	JMOTU	http://www.nematodes.org/bioinformatics/jMOTU/index.shtml	(Jones et al., 2011)
Hierarchical clustering	Neighbour Joining	i.e. R package APE	http://cran.r-project.org/web/packages/ape/index.html	(Paradis et al., 2004)
	Parsimony	i.e. Phylip	http://evolution.genetics.washington.edu/phylip.html	(Felsenstein, 1985)
	Bayesian inference	SAP	http://fisher.berkeley.edu/cteg/software/munch	(Munch et al., 2008)
	Bayesian inference	RDP	http://rdp.cme.msu.edu/	(Wang et al., 2007)
Character based	Diagnostic	CAOS	http://boli.uvm.edu/caosworkbench/caos_barcoder.php	(Sarkar et al., 2008)
	Diagnostic	BLOG	http://www.ibarcode.org	(Bertolazzi et al., 2009)
Other	Portal of Data analysis	Web browser	http://www.ibarcode.org	(Singer and Hajibabaei, 2009)
			http://greengenes.lbl.gov/cgi-bin/nph-index.cgi	(DeSantis et al., 2006)
			http://www.arb-silva.de/	(Pruesse et al., 2007)
	Complete package analysis	QIIME	http://qiime.org/	(Caporaso et al., 2010)
		SPIDER	http://spider.r-forge.r-project.org/SpiderWebSite/spider.html	(Brown et al., 2012)

Even if tools can perform more than one analysis, they are assigned to a category based on their most frequent use found in literature.

Similarity methods/pairwise distance

These methods are based on the analysis of similarity among query sequences compared to a reference data set. They discriminate entities exceeding a certain level of variability called "threshold value". Threshold approaches rely on the assumption that intraspecific sequences variation does not exceed a certain distance value; otherwise they are considered as different species (Casiraghi et al., 2010). As a general comment, these methods assume that conspecific samples will be more similar to

each other than to samples of any other species. However, this is an arbitrary assumption, which considers only the superficial population structure or phylogenetic relationships. Indeed, while gene variation represents a product of evolution, an arbitrary cut-off value does not reflect what is known about the evolutionary process responsible for this variation (Little, 2011). The best known program based on similarity methods/pairwise distance is BLAST (Altschul et al., 1990). BLAST uses a heuristic method locating short matches between two sequences. BLAST algorithm allows making fast searches at the expense of a higher accuracy. Indeed, the algorithm can produce ambiguous identification when a query sequence equally matches more than one sequence in the reference database (Little, 2011; van Velzen et al., 2012). A web and local version of BLAST are available, but the second is commonly used with large data sets (i.e. like those one deriving from environmental massive sequencing).

The similarity methods/pairwise distances have two clear advantages: they are fast and do not require huge calculation power. Consequently, they are the best choice as the first level analysis, in particular in the case of investigation on large data sets. However, they can have some shortcomings when a high level of accuracy is needed (i.e. discrimination of closely related species). The greater criticism is that similarity methods imply a loss of information due to the fact that the sequences are compared as whole units in the generated distance matrix, rather than by using each character separately (Ferguson, 2002; DeSalle et al., 2005; Desalle, 2006; Rach et al., 2008).

Hierarchical clustering methods/tree-based

These methods are the first answer to overcome the limits of similarity methods (i.e. the use of a threshold, the missing phylogenetic and population details). Tree-based approaches are hierarchical clustering methods that identify the groups through the analysis of a reconstructed relationships tree. On the other side, these methods are subjected to a certain level of criticisms among some DNA barcoding users.

Indeed, DNA barcoding is not strictly considered a phylogenetic method, but on the contrary, it is a simple identification method. Being precise, to identify is not to reconstruct the life history, even if the distinction is subtle (Casiraghi et al., 2010). Tree-based methods deeply rely on a multiple alignment, and the topology of the tree created, which identify the biological entities belonging hopefully to the rank of species, is used as a guide to identify the query sequence: the identification is reached when the query sequence falls within the unique specific cluster of the biological entities on the tree. The principal methods using tree-based approaches are Neighbor Joining (NJ) (Saitou and Nei, 1987; Howe et al., 2002; Munch et al., 2008) maximum parsimony (PAR) (Cavalli-Sforza and Edwards, 1967), and Bayesian Inference (BI) (Huelsenbeck et al., 2001). NJ followed by Kimura 2 parameters (K2P) correction is the most widely used method in DNA barcoding analysis in the published literature and is now included in the standard analysis package of the BOLD system (Ratnasingham and Hebert, 2007). The main reason for this success is the high speed of the analysis on large data sets, and the very limited investment in understanding the parameters of the analysis. NJ is a bottom-up clustering method based on the minimum evolution criterion (i.e. the shortest total length of the tree branches are considered the best). When queries are assigned to cluster with conspecific DNA barcodes, the identification is considered successful (Will and Rubinoff, 2004). NJ method has been tested in many works (Little and Stevenson, 2007; Ross et al., 2008; Austerlitz et al., 2009; Little, 2011; van Velzen et al., 2012), obtaining a similar, and not astonishing, performance. The almost universal NJ-K2P adoption in DNA barcoding works raised many criticisms (Will and Rubinoff, 2004; Meier et al., 2008): the principal is that distance matrix is an oversimplified representation of the reality that can lead to an incorrect visualization of the relationships between species (Ferguson, 2002; DeSalle et al., 2005; Desalle, 2006).

PAR methods do not employ distance matrices, and are considered character-state methods. Making assumption on each character, PAR methods use algorithms to calculate the most parsimonious tree. These methods are classically used to reconstruct phylogeny, but when used

for species identification they can be problematic. The reason relies on the fact that often, several trees are equally parsimonious or differences among different trees are really minimal. On the other hand, when we are analyzing many sequences, we are, in some way, forced to employ heuristic approaches, to keep the calculation time low. Heuristic approaches have typically a problem with reproducibility of the analyses (Will and Rubinoff, 2004; Meier et al., 2008). In spite of these problems, the performance of PAR methods is superior to NJ, in particular when unambiguous alignment (such as the case of *rbcL* in plants) is used (Little, 2011).

BI methods are other widely used tools for phylogenetic reconstructions, but in the field of taxonomy assignment is mostly common in bacteria analysis. RDP Classifier (Wang et al., 2007) is the most extensively used bioinformatics programs for 16S rRNA classification. Its success is probably due to its speed, to the fact that does not require sequence alignment, and works well with partial sequences. Moreover it is capable of classifying to the genus level near-full-length and 400-base segments with an overall accuracy above 88.7%. All these features made it a competitive tool for the analysis of big data.

This software is a naı́ve Bayesian classifier that provides taxonomic classification from domain to genus. The classifier is trained on the known type strain 16S sequences. The frequencies of all sixty-four thousand possible eight-base subsequences (words) are calculated for the training set sequences in each of the approximately 880 genera (with a range of probability values of assignment between 0% or 100%). Also Munch et al. (Munch et al., 2008) proposed a system for sequence assignment using BI for metazoan. The idea is quite innovative, because it introduces a statistical measure of confidence, but not widespread.

Character-based and diagnostic methods

These methods are a different answer to the problems raised by similarity methods. Character-based and diagnostic methods focus their discrimination power on the presence/absence of discrete characters or

combinations of these. They are based on phylogeny in which classification algorithms consider just diagnostically informative states along phylogenetic tree nodes. The four main tools developed are: CAOS (Sarkar et al., 2008) , BRONX (Little, 2011), DNA-BAR (DasGupta and Konwar, 2005) and BLOG (Bertolazzi et al., 2009).

2.0.5 Distribution analysis of of diversity

The taxonomic assignment is the first step of the work, but understanding differences in the composition of organismal communities is of major importance in molecular ecology. Biological and ecological systems are the result of a number of deterministic and stochastic processes as well as historical constraints. The most relevant historical coercion is that species are not independent entities, but their functional and ecological similarities are rather shaped by patterns of common ancestry. In the field of community ecology, only recently researchers have incorporated historical constraints represented as phylogenies in their analyses motivated by the fact that species interact within the community based on their traits, and traits have an evolutionary history (Webb et al., 2002; Faith, 2002). Phylogenetic comparative methods pioneered by the method of phylogenetic independent contrast (PIC) meant the first substantial effort to address the statistical non-independence among species due to common ancestry (Felsenstein, 1985; Ackerly, 2009). Phylogenetic distance (PD) measures take into consideration all components of biodiversity: species richness, species abundance and phylogenetic distances among species (Chao et al., 2010; Faith, 2002; Ricotta, 2007; Cardoso et al., 2009), so can provide far more power compared to more traditional ecology measure (i.e *Sorenson* and *Jaccard* indices of group overlap (Magurran, 2004)) because they exploit the degree of divergence between different sequences (Lozupone et al., 2007). Indeed, PD is more inclusive that a simple count of species richness because, considering the sum of the branch lengths from the members within a community, it quantifies the evolutionary history, and is also believed to correspond to the number of evolutionary derived traits within a biological community. Hence, PD will be higher when there

are more distantly related species in an assemblage.

Phylogenetic diversity measures

Diversity measures predicted from biodiversity patterns serve as generalization comparison at both spatial and geographical levels (Gaston and Spicer, 2009). The studies on the structure of communities focus on the total numbers of taxa or unique lineages found in individual samples (that is, α diversity), the relative abundances of individual taxa or lineages and the extent of phylogenetic or taxonomic overlap between communities or community categories (that is, β diversity). α diversity measures (for example, richness and coverage estimators, rarefaction curves) yield estimates of biodiversity and its limits in different environments (Hughes et al., 2001; Curtis et al., 2002; Sogin et al., 2006). Likewise, multivariate statistical techniques such as clustering and ordination, allowed ecologists to describe β diversity patterns, revealing how biotic and abiotic variables control the community composition. For example, analyses of β diversity patterns have revealed how microbial communities are structured across a wide range of natural habitats (Lozupone et al., 2007; Auguet et al., 2012; Barberán et al., 2012), the spatial and temporal variability of microbial communities on and in the human body (Fierer et al., 2008; Costello et al., 2009), and the factors structuring soil bacterial communities (Lauber et al., 2009).

Alpha, beta and gamma diversities can be obtained from their entropies calculation: H_α, H_β, H_γ, where: $H_\alpha + H_\beta = H_\gamma$. Entropies are reasonable indices of diversity giving the uncertainty in a sampling process outcome. The most common diversity measure is the Shannon-Wiener index or Shannon entropy, The Shannon entropy can be calculated according to the equation 2.1:

$$H = -\sum_{i=1}^{S} p_i \log_b p_i \tag{2.1}$$

where S is the number of species in a sample, p_i is the relative abundance of the *ith* species and b is the logarithm base. It is important

to note that the entropy gives the uncertainty in species identity of a sample, not the number of species in the community. For this reason, it is important to distinguish between entropy (H) and true diversity (D). Entropies, not diversities, and their mathematical behavior usually do not correspond to biologists theoretical or intuitive concept of diversity (Jost, 2006). An enormous difference in diversity can be sometimes expressed by two very close entropy values which mask the real statistical significance of diversity. However, entropy value can be transformed into true diversity value by applying an exponential transformation. The seminal paper of Hill (1973) already showed the relationship between diversity and entropies parameterized by a factor q 2.2:

$$H_q = \sum_{i=1}^{S} p_i^q \qquad D_q = (H_q)^{\frac{1}{1-q}} \tag{2.2}$$

where q is the order of diversity. The order of diversity indicates its sensitivity to common and rare species. The diversity of order zero ($q = 0$) is completely insensitive to species frequencies and is known as species richness. All values of q less than unity give diversities that disproportionately favor rare species, while all values of q greater than unity disproportionately favor the most common species. The critical point that weights all species by their frequency, without favoring either common or rare species, occurs when $q = 1$. At $q = 1$ 2.2 is the exponential of Shannon entropy (2.3):

$$D_1 = \exp\left(-\sum_{i=1}^{S} p_i ln p_i\right) = \exp(H) \tag{2.3}$$

Using 2.3, the relation between alpha, beta, and gamma diversity is obtained by 2.4

$$\exp(H_\alpha + H_\beta) = exp(H_\gamma) \tag{2.4}$$

Despite the widespread use of Shannon entropy in ecological studies, an additional similarity measure (called Kullback-Leibler divergence), has been proposed by Ludovisi and Taticchi (Ludovisi and

Taticchi, 2006) regarding beta diversity as a dissimilarity estimator between two or more communities. Kullback– Leibler divergence between two communities composed of the same species is given by 2.5, where $P = \pi_1, \pi_2, \ldots, \pi_S$ and $Q = \theta_1, \theta 2, \ldots, \theta_S$ are the probability distribution of communities P and Q

$$J(P:Q) = \sum_{i=1}^{S} (\pi_i - \theta_i) ln \frac{\pi_i}{\theta_i} \tag{2.5}$$

The same equation can be written in another form and gives 2.6

$$\begin{aligned} J(P:Q) &= \sum_{i=1}^{S} \pi_i ln \pi_i + \sum_{i=1}^{S} \theta_i ln \theta_i - \sum_{i=1}^{S} (\pi_i ln \theta_i + \theta_1 ln \pi_i) \\ &= -H_P - H_Q + H_P Q \end{aligned} \tag{2.6}$$

The same authors illustrated Kullback-Leiber divergence between two communities having different number of species. They took into account both the proportion of singletons and the probability of unobserved species within each community according to Chao and Shen (Chao and Shen, 2003) correction. Therefore, the probability of observed species is calculated according to the following equation 2.7,

$$p'_{iP} = \frac{N_{iP}}{N_p} - \left(1 - \frac{f_{1P}}{N_P}\right) for i = 1, 2, \ldots, k_P \tag{2.7}$$

where f_{1P} is the number of species present with a singleton, N_{iP} the number of individuals of the ith species, N_p the total number of counts in the sample of the community P and k_p is the number of species detected in the community P Consequently, the divergence between community pairs can be calculated as 2.8:

$$\widehat{J}(P:Q) = n \sum_{i=1}^{\widehat{S}} \left(p'_{iP} - p'_{iQ}\right) ln \frac{p'_{iP}}{p'_{iQ}} = -\widehat{H'_P} - \widehat{H'_Q} + \widehat{H'_{PQ}} \tag{2.8}$$

The terms $\widehat{H'_P}$, $\widehat{H'_Q}$, $\widehat{H'_{PQ}}$ represent modified estimators H_P, H_Q and H_{PQ} respectively.

Allen et al. (Allen et al., 2009) generalized the Shannon entropy to take into account phylogenetic differences. For a rooted tree, their phylogenetic entropy H_p is 2.9:

$$H_p = -\sum L_i a_i \log a_i \tag{2.9}$$

where the summation is over all branches, is the length of branch i, and denotes the abundance descending from branch i. The index proposed by Allen was Hill-behaved, but Chao et al. (2010) (Chao et al., 2010) proposed a new correction which framed the index within a generalization phylogenetic aware of the Hill numbers. 2.10:

$$PD_q(\bar{T}) = \left\{ \sum_{i \in B_{\bar{T}}} L_i \left(\frac{a_i}{\bar{T}} \right)^q \right\}^{\frac{1}{1-q}} \tag{2.10}$$

where $\bar{T}$ is mean evolutionary change and B are tree branches. Finally, once phylogenetic diversity indices are obtained, diversity in lineage per base change per site between and across samples can be detected.

3 Section 2

Targeted Sequencing on Metazoan Communities: A Seedbed For Further Laboratory and Computational Investigations

3.1 Introduction

In this section the success of sequencing the entire *coxI* barcode on 454 (Roche Life Science) platform was addressed. The main limitation affecting the reliability of biodiversity profiles in NGS reads assignment, is the intrinsic error level introduced by the platform used (Hoff, 2009) (Balzer et al., 2011). The pre-sequencing steps of library preparation (Bik et al., 2012; Schloss et al., 2011), resulting in a cascade-effect of misinterpretations during down-stream analyses is another source of errors. Thus, sequence reads denoising (e.g. AmpliconNoise, (Quince et al., 2011)) and chimera or sequence artifacts removal (e.g. Uchime, (Edgar et al., 2011)) are fundamental in environmental single locus sequencing frameworks. Moreover the produced sequences are not likely to be assembled and therefore any sequencing error will result in sequences seemingly coming from different organisms (Schloss et al., 2011). Following error detection and removal, taxonomic profiles can be produced through "Taxon Assignment" step in order to assess taxonomic diversity of organisms within the samples under study. Different approaches can be adopted for "Taxon Assignment" such as: (i) naíve Bayesian classification (e.g. RDP classifier, (Wang et al., 2007)), (ii) classification based on diagnostically informative sites along a phylogenetic tree (Sarkar et al., 2008), (iii) monophyly based on a phylogenetic infer-

ence including reference sequences (Munch et al., 2008), (iv) species delimitation using an intraspecific distance threshold (Blaxter et al., 2005; Jones et al., 2011), (v) combination of molecular and morphological characters (Rach et al., 2008), and (vi) genealogical methods based on the coalescent theory using demographic explicit genetic models with maximum likelihood/bayesian algorithms (Abdo and Golding, 2007; Nielsen et al., 2009).

In this study, the precision and accuracy of denoising procedure and the candidate parameters able to reduce sequence error rate were investigated, as well as an innovative taxon assignment pipeline. In addition, a novel library preparation method allowing the sequencing of the entire *coxI* barcoding region (approximately 700 bp) on 454 pyrosequencing platform (Roche Life Science) was proposed. To address these objectives, metazoan communities coming from complex environmental matrix of the chestnut soil sampled from different areas in Italy was considered.

The adopted bioinformatics methods were assessed by a comparative approach between two identical control samples, differing by biomass equilibration, and comprising taxonomically classified organisms at order or class taxonomic ranks and at species level for individuals of Carabidae (Coleoptera) family.

3.2 Material and Methods

3.2.1 Samples description

A total of eight samples were collected from three Italian chestnut soil forests situated in northern (province of Milan), central (province of Rome), and southern (province of Catania) parts of the peninsula (Table 3.1).

The samples consisted of one final soil sample per locality (coming from different pooled samples of the same site) and two pitfall traps ones taken from central Italy. Soil samples were taken from the first 30

Table 3.1: Samples provenience, the methods used to collect the organisms and categorize them.

Locality	North		Center				South	
Collection Method	BMF	BRF	BMF	BRF	PFT	PFT	BMF	BRF
Type	Hydrophilic	Aerophilic	Hydrophilic	Aerophilic	Soil Litter	Soil Litter	Hydrophilic	Aerophilic
Sample ID	NH	NA	CH	CA	MPE4	MPE5	SH	SA

BMF: Baerman funnel, **BRF**: Berlese funnel, **PFT**: Pitfall traps.

cm of soil surface by a soil sampler probe. From each sample, approximately 500 g of soil were pulled out and conserved in sterile plastic bags at room temperature to be used for hydrophilic and aerophylic biota extraction procedures. Aerophylic biota corresponds to the organisms living in aerobic spaces in the soil, while hydrophilic organisms occupy wet space between soil particles. Following the extraction of hydrophilic biota using Baerman funnel and aerophylic biota by Berlese funnel, each locality was represented by two samples one hydrophylic (NH = northern hydrophilic, CH = central hydrophilic, SH = southern hydrophilic) and another aerophylic (NA = northern aerophylic, CA = central aerophylic, SA = southern aerophylic). Furthermore, pitfall traps (PFT) were used to collect soil litter's macrofauna in sampling area of central Italy, where five traps were positioned at a distance of 10 m from each other and placed at soil surface to deliver the first control sample called MPE5. The organisms collected in this sample were classified, based on their morphology at species levels for the family Carabidae (Coleoptera) and biomass weighed. Regarding the other organisms present in MPE5 it was possible to classify them only at order or at class taxonomy ranks (Table 3.2). In addition, Carabidae and Isopoda organisms were sequenced for the *coxI* barcode region by Sanger sequencing to form our local control database (LocalDB). The second PFT sample, called MPE4, was derived from MPE5 and consisted on equal biomass content of all the classified organisms (see Table 3.2 for organismal content of MPE4 and MPE5 and their corresponding biomass).

Table 3.2: Taxonomically identified organisms in Pitfall traps samples and their corresponding biomass.

Taxonomic Group	Total Biomass (g)	Species Name	Biomass (g)
Coleoptera	22.89	*Carabus (Tomocarabus) convexus dilatatus*	12.94
		Carabus (Chaetocarabus) lefebvrei bayardi	4.85
		Calathus fracasii	0.58
		Abax parallelepipedus	0.29
		Laemostenus latialis	0.23
		Pterostichus micans	0.18
		Calathus montivagus	0.07
Diptera	4.12		
Orthoptera	2.44		
Blattodea	1.02		
Myriapoda	0.82		
Isopoda	0.7		
Arachnida	0.64		
Scorpiones	0.63		
Hymenoptera	0.21		
Lepidoptera	0.1		
Collembola	0.02		

*Class taxonomy rank; Organisms classified at species level belong only to the family Carabidae (Coleoptera).

3.2.2 DNA extraction and amplification of *coxI* barcode

Total genomic DNA was extracted from the eight collected samples by using different commercial kits according to manufacturers' instructions. The entire *coxI* DNA barcode was amplified from DNA extracts using the universal primer pair: forward-LCO1490 (5'- GGTCAACAAATCATAAA
GATATTGG-3'), and reverse-HCO2198 (5'-TAAACTTCAGGGTGACCAA
AAAATCA-3') (Folmer et al., 1994). PCR reactions were carried out in 50μl reaction volumes containing: 1.5mM $MgCl_2$, 250nM of each primer, 200μM of each dNTP, 1x of Phusion HF Buffer, 1U of Phusion DNA polymerase (M0530S, NEB) and 2μl of DNA extracts, using a thermocycling profile of one cycle of 60 s at 94°C, five cycles of 60 s at 94°C, 90 s at 45°C, and 90 s at 72°C, followed by 35 cycles of 60 s at 94°C, 90 s at 50°C, and 60 s at 72°C, with a final step of 5 min at 72°C. PCR products along with 100 bp DNA Ladder (Fermentas, Life Sciences) were visualized on a 1% agarose gel stained with 0.005% of ethidium

bromide. Although with some unspecific products and impurities, the expected amplicons of approximately 700bp were successfully amplified in NH, CH, SH and MPE5, while for NA, CA, SA and MPE4 the amplification did not succeed probably due to some inhibitors activity (e.g. soil's humic acid and insects' intestine content). For this reason, PCR reactions with different DNA dilutions (1/10, 1/20, 1/50) were tested in order to promote the amplification and in the same time to eliminate unspecific products amplification. Since unspecific products and impurities were evident even with dilution assays, dilutions that gave the highest amplicons yield were used in subsequent PCR reactions. Consequently, dilutions of 1/10 were used for CA and SA, 1/50 for NA and MPE4 while the raw DNA extract was used for the remaining samples. PCR products were gel purified using QIAquick Gel Extraction Kit (Qiagen) and their corresponding concentrations pre and post-purification were measured by densitometry.

3.2.3 Sequencing libraries preparation and 454 pyrosequencing

Given that standard protocols suitable to sequence long amplicons on next generation sequencing platforms are still missing, sequencing the entire *coxI* DNA barcode region (700bp) was considerably challenging. For this reason, a pre-requisite procedure enclosing amplicons ligation was necessary in order to build up shotgun libraries compatible with the following sequencing steps. This protocol was adapted on amplicons products since it was assembled and its conditions were optimized in our laboratories for cDNA molecules (Patent: RM2010A000293-PCT/IB2011/052369). Hence, 100ng of purified amplicons were ligated, using DNA Ligase (Roche), to obtain compatible fragments for ϕ29 polymerase amplification. Therefore, the concatenated amplicons were then amplified by ϕ29 polymerase using random primer pairs. The use of ϕ29 polymerase guaranteed strand displacement activity, proof- reading high-fidelity activity, high processitivity giving ideal product for nebulization and high yields of amplified products promoting the selection of rare molecules. The amplified concatenated amplicons were

checked on 0.8% agarose gel and quantified by densitometry. Rapid sequencing libraries were prepared using the amplified concatenated products and consisted of two main steps: (i) nebulization of the amplified ligated amplicons, (ii) fragment end repair followed by adaptor ligation to the nebulized fragments. 500 ng of the amplified ligated amplicons were submitted to nebulization using liquid nitrogen for 70 sec at 30 psi (2.1 bar). The fragment end repair process of nebulized products, conducted following the Rapid library preparation standard protocol, allowed the successful ligation of 454 Adaptors on both 3′ and 5′ ends. Libraries quality was assessed on an Agilent Bioanalyzer 2100 High Sensitivity DNA chip, where all profiles showed fragment populations between 300 and 1000bp, while libraries quantification was carried out by fluorospectrometer (Nanodrop 3300, Thermo Scientific). Following adaptors ligation, the DNA was enriched in emulsion PCR to be deposited on an 8-lane PicoTiterPlate (PTP) wells (Roche/454) and sequenced on GS FLX Titanium pyrosequencing platform.

3.2.4 Bioinformatics pipeline for sequence analysis

The bioinformatics pipeline for sequence reads analysis, illustrated in Figure 3.1, is divided into four main steps: (i) 454 standard filtering process, (ii) pattern search, (iii) denoising and (iv) taxon assignment.

454 filtering

The first analysis step, performed on the obtained sequence reads, was a standard filtering process suggested by the 454 platform manufacturer by means of stringent filtering algorithms to capture and discard poor quality reads. This filtering practice, comprising read rejecting and read trimming filters, was computed using GS Run Processor V2.4 (Roche 454 Life Sciences software package).

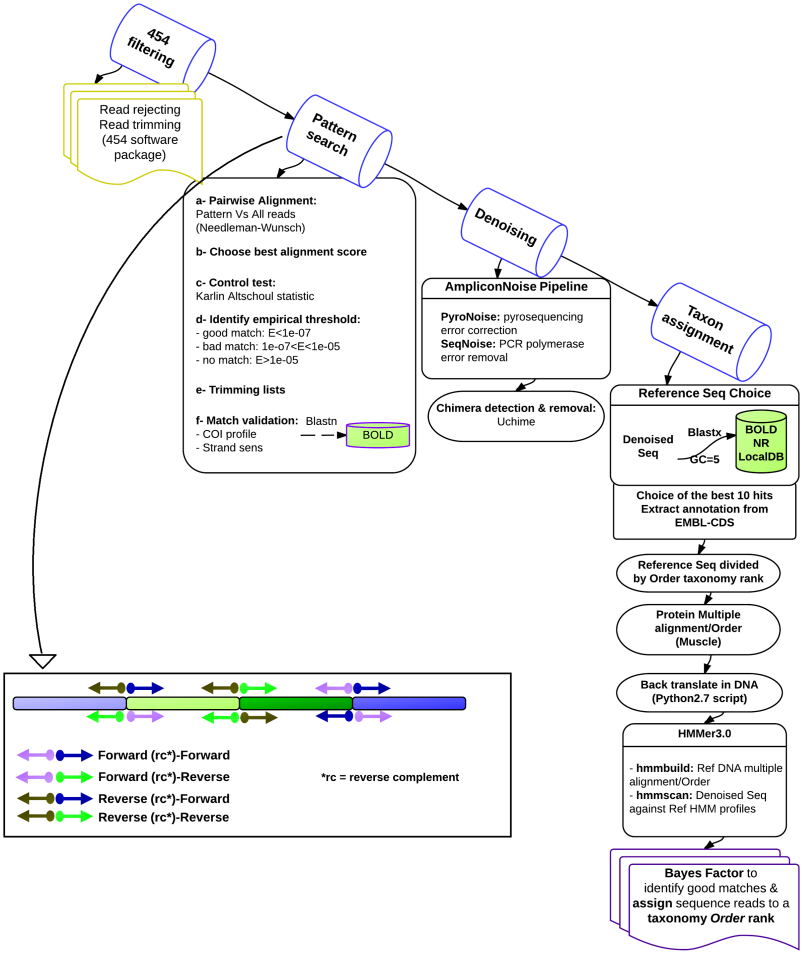


Figure 3.1: Bioinformatics sequence analysis pipeline illustration.
The pattern search consisted on finding the four possible combinations of PCR primers within sequencing reads.

Pattern search

Due to amplicons ligation and subsequent nebulization steps carried out for sequencing library preparation, different combinations of PCR primers were expected to be found within sequence reads. For that, it was necessary to conduct a pattern search analysis of four possible primer combinations: (a) primer reverse (reverse complement) + primer forward, (b) primer forward (reverse complement) + primer reverse, (c) primer reverse (reverse complement) + primer reverse, (d) primer forward (reverse complement) + primer forward (Figure 3.1). These analyses were computed by means of Python2.7 script to execute: (i) pairwise global alignment of the four patterns against all sequences using *Needleman– Wunsch* algorithm (from emboss package), (ii) comparison between the four alignment scores that considers the highest one as best match, (iii) computation of a modified *Karlin-Altschul* statistic ($E = mne^{-\lambda S}$; where mn is the size of the search space, λ=0.27, S is the alignment score) that classifies the best scores in three categories: a) good match for $E < 1e-07$, b) bad match for $1e-07 < E < 1e-05$, c) no match for $E > 1e-05$. The two thresholds were chosen looking at the calculated statistic frequency distribution on which the region between the two modes was considered as an ambiguous match. In the case of no match category, sequences were considered ready for downstream analysis without the need of pattern removal, while bad match was discarded from further analyses. Sequences belonging to good match category were spliced up- and down-stream of pattern position. A further match validation using blastn, with $E-value < 10e-03$ on good match and no match categories, was conducted against the public BOLD (Ratnasingham and Hebert, 2007)) *coxI* database and checked for correct sequences content (match with coxI profile and strand sense). Once this last validation terminates, the script outputs four trimming lists indicating pattern position when present. Sequence reads trimming has generated for each sample two separate data sets at 5′ and 3′ *coxI* barcode.

Denoising

On the above mentioned data sets, denoising protocol using "AmpliconNoise" pipeline (Quince et al., 2011) able to correct both pyrosequencing errors invoking PyroNoise algorithm and PCR single base substitution ones raised to SeqNoise, was conducted separately in order to avoid clustering biases of the 3′ and 5′ regions. Both Pyro- and Seq-Noise use a clustering distance parameter value (σ) used by expectation maximization algorithm to eliminate unreal sequences and construct the real ones. In addition, it was investigated both the precision and accuracy of Pyro- and Seq-Noise over four (60, 50, 40, 30) and three σ values (25, 10, 5) respectively and consequently σ = 60 was set up for PyroNoise while σ = 25 for SeqNoise (see results). The optimal σ values were chosen according to the results obtained on the control sample (MPE5) where organismal content was taxonomically classified prior to DNA extraction. Finally, denoising pipeline has resulted in sequence clusters represented by unique sequences, which were submitted to chimera detection and removal routine of Uchime software using the recommended parameters of denoised data available in its manual (Edgar et al., 2011).

Taxon assignment: Hidden Markov Model (HMM)

To address the comparative analysis of species composition of chestnut soil across the three geographical areas under study, all data sets of the same category (aerophylic, hydrophylic, PFT) were merged and separated data sets that map both ends of *coxI* DNA barcode region were excluded. In addition, each sequence ID was tagged according to its geographical provenance to keep track sequence origin information. In summary, the following six data sets were generated: 1) aerophylic-5′, 2) aerophylic-3′, 3) hydrophylic-5′, 4) hydrophylic-3′, 5) PFT-5′ and 6) PFT-3′. These data sets were used in taxon assignment workflow which encompassed three main steps: a) choice of reference sequences, b) reference HMM profiles building and c) taxon assignment.

In order to choise of reference sequences to achieve higher assign-

ment accuracy and sensibility, several sources of *coxI* reference sequences were taken into account in species assignment workflow execution. First of all, the LocalDB (see samples description), corresponding to 154 isopods and carabids individuals, was generated by sequencing the *coxI* DNA barcode of all morphologically identified specimens recovered from the collected samples. Additionally, public *coxI* sequences were retrieved from BOLD and the non-redundant database (NR-NCBI). Given that querying large databases is computationally expensive, the above mentioned databases were sampled by Blastx (Altschul et al., 1990) for the closest 10 matches using our sequence reads as queries. Blastx was run using the mitochondrial invertebrate genetic code (NCBI genetic code number: 5) for the reason that the majority of the expected organisms belong to metazoans. Blastx outputs were parsed using Biopython1.57 and the first 10 best matches ID were retrieved and tagged at taxonomical Order rank delivered by NCBI taxonomy. All matched references where inserted in a structured local database using Structured Query Language (SQL) that readily allowed further filtering and querying procedures. The later filtering was executed using MySQL and consisted of sampling all reference sequences having the same taxonomical order rank of the best match and discard all the others. Finally, to get nucleotide and protein sequences of the filtered references, EMBL-CDS was interrogated by means of web service fetch tool (EMBL fetch tools: wsdbfetch.py) using the chosen reference sequence IDs as queries. In the case of BOLD and LocalDB entries, nucleotide sequences were translated into their corresponding protein ones by means of a Python2.7 script. Note that, all reference sequences of each sample category were joined to custom a final reference data set tagged by its corresponding taxonomical order rank.

Reference HMM profiles building Protein reference sequences belonging to each taxonomical Order rank were multiple aligned using Muscle3.8.31 (Edgar, 2004) and then back-translated into their equivalent nucleotide ones using a Python2.7 script. Hidden Markov Model (HMM) nucleotide alignments profiles were built using hmmbuild (HMMer3.0) with its default parameters (Finn et al., 2011).

Taxon assignment Sequence clusters of the above described six data sets were assigned to one of the nucleotide HMM profiles using the algorithm hmmscan (HMMer3.0) with its default E-value threshold set to 10.0 over which the target profile is considered as a false positive assignment. The six hmmscan outputs were then parsed using a Python2.7 script which classifies the assigned sequence clusters in four categories:

- Unclassified assignment: sequence-profile match output an E-value higher than hmmscan threshold.
- Bad assignment: one sequence-profile hit found. In this case it is assumed that the sequence does not match a *coxI* profile and consequently it is a false positive match.
- Good assignment: the best match bit score passes the threshold of Bayes Factor (BF) set to 3.0. BF 3.1 was computed by subtracting the best bit score from the natural logarithm of the sum of all remaining exponential bit scores.

$$ln(BF) = S_{i=1} - ln\left(\sum_{i=2}^{n} exp\,(S_i)\right) \tag{3.1}$$

  where S is the score of the *ith* element and n is the number of scores
- Ambiguous assignment: sequence-profile hit does not pass BF test.

Except for the good assigned sequence clusters, all the other categories were joined and labeled as unclassified sequences. Good assigned sequence clusters were aligned against their corresponding HMM profile using hmmalign (HMMer3.0). Nucleotide multiple alignments were checked manually for the presence of poorly aligned sites and long insertions/deletions. Subsequently, by means of a Python2.7 script, sequences having long inserts longer than two nucleotides and not homopolymeric were eliminated completely while sites in sequence clusters not present in reference and that introduce one gap in the alignment were just removed to avoid frame shift error.

Denoising precision screening

The final yield of denoising pipeline over the 12 σ parameter values combination was considered for further precision screening in order to determine the best σ value suitable for taxon assignment at order and species taxonomic categories. This was possible by observing species loss rate, and the linearity between biomass weight and sequence counts per detected order and species. The *a priori* identified carabid species were used and the organisms classified at order and/or class ranks present in MPE5, their corresponding biomass and sequence reads number that matched them with the taxon assignment procedure described above.

3.3 Result

3.3.1 Sampling unit and sequencing yield

The *coxI* barcode from eight environmental soil samples 3.1, collected from three Italian chestnut forests (north, center, south), was sequenced on 454 GS FLX Titanium pyrosequencing platform (Roche Life Science). This platform was chosen because it provides the longest read length among the other available NGS platforms to the time when the experimental design has been decided.

The sampling unit comprised two identical control samples in terms of taxonomic complexity but different by organismal biomass equilibration (MPE4: equilibrated, MPE5: original collection of pitfall traps). The organisms in the control samples were taxonomically classified at the possibly identifiable rank. In addition, the *coxI* DNA barcode was sequenced by Sanger for the organisms of Isopoda and Coleoptera (Carabidae) to create a local database (LocalDB) used in bioinformatic pipeline optimization. The sequencing run produced a total of 1,303,314 reads, of which 613650 passed filter reads were considered eligible for downstream analysis. Moreover, signal intensity quality score was assigned to each obtained base in order to facilitate the selection of high

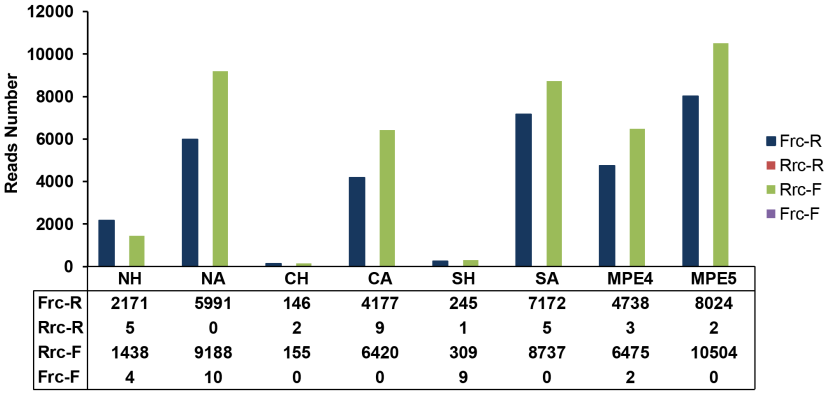
quality bases during sequence analysis. Quality scores, ranging from zero to 40, were mainly distributed between 11 and 40 where 27.10% of bases obtained the maximum score, 2.13% a score of 11 and a negligible number of bases had a quality score less than 11. Passed filter reads represented read lengths extending from 40 to 637 with an average of 346.15 bases.

3.3.2 Bioinformatics analysis pipeline results

In order to execute denoising process and successively downstream analyses, four possible patterns, expected to be found within sequence reads, were searched and consequently reads were trimmed at both start and end patterns position. These patterns were formed due to amplicon ligation following the amplification of the *coxI* DNA barcode region (for details see materials and methods). The four patterns were: a) RrcF, b) FrcR, c) RrcR and d) FrcF (Figure 3.1), where "F" and "R" are forward and reverse primers respectively used in PCR amplification, while "Frc" and "Rrc" are their corresponding reverse complements. Pattern search filtering produced a noticeable reduction to almost the half of sequencing reads number for all samples, 152,531 at 5′ and 174,600 at 3′ *coxI* DNA barcode out of 613,650 original reads number. Furthermore, these results showed a significant dominance of FrcR and RrcR patterns in all samples over the remaining two patterns which were absent or present at very low occurrence (Figure 3.2).

3.3.3 Denoising and chimera removal output

Denoising protocol using "AmpliconNoise" pipeline (Quince et al., 2011) was executed for all samples in order to correct both pyrosequencing (PyroNoise) and PCR polymerase (SeqNoise) errors. This procedure was conducted on homologous sequences, therefore for 3′ and 5′ regions separately. Both Pyro- and Seq-Noise use a main clustering distance parameter value (σ) resulting in the construction of final sequence clusters represented by unique sequences. Previously to AmpliconNoise execution on all samples, the precision and accuracy of Pyro- and



	NH	NA	CH	CA	SH	SA	MPE4	MPE5
Frc-R	2171	5991	146	4177	245	7172	4738	8024
Rrc-R	5	0	2	9	1	5	3	2
Rrc-F	1438	9188	155	6420	309	8737	6475	10504
Frc-F	4	10	0	0	9	0	2	0

Figure 3.2: Pattern search results distribution over the eight samples following reads trimming at pattern position.
Frc-R: Forward reverse complement-Reverse, Rrc-R: Reverse reverse complement-Reverse, Rrc-F: Reverse reverse complement-Forward, Frc-F: Forward reverse complement- Forward.

Seq-Noise over all combinations of four (60, 50, 40, 30) and three (25, 10, 5) σ values respectively were investigated. The optimal σ values were chosen according to the results obtained on the control sample (MPE5), where organismal content was taxonomically classified (Table 3.2) prior to DNA extraction, and consequently σ = 60 was set up for PyroNoise while σ = 25 for SeqNoise. Finally, denoised sequences were submitted to chimera detection and removal method of Uchime software (Edgar et al., 2011).

3.3.4 Sigma (σ) screening and identification of its best value

In order to determine the best value of σ for Pyro- and Seq-Noise, the taxon assignment step on the 12 resulting data sets was performed (Figure 3.1). The metadata provided by MPE5, namely individuals biomass of morphologically identified species and that of organisms belonging

to the same taxonomic order or class were used (see samples description). Figure 3.3 illustrates a qualitative overview of the relationship between σ parameter and specific species detection, reads abundance assigned to them and their biomass. Evidently, the parameter σ of SeqNoise has further contribution in species recognition as it masked *Carabus (Chaetocarabus) lefebvrei bayardi* at $\sigma = 5$ (s5). In addition, the heterogeneous distribution of read abundance in relation with species biomass failed to explain the linearity of *coxI* copies increase in superior biomass individuals. This fact is evident for both 5′ and 3′ data sets where *Carabus (Tomocarabus) convexus dilatatus* having the highest biomass falls under the remaining species for many σ values. A similar behavior can be noticed for *Carabus (Chaetocarabus) lefebvrei bayardi* embracing in almost all cases considerably less assigned reads than *Calathus fracasii*.

At the contrary, Figure 3.4(A) displays clearly a homogeneous distribution of reads abundance and orders biomass for all tested σ values in both 5′ and 3′ *coxI* data sets. It is important to note that, at σ = 5 ("s5") of SeqNoise the orders Myriapoda, Isopoda and Lepidoptera were not detected in the 5′ data set while for a value of σ = 25 ("s25") all the orders were found and a higher quality biomass-reads count correlation was observed. Regarding 3′ *coxI* data set, a relative reduction of the biomass-reads count correlation for SeqNoise parameter σ = 5 (Figure 3.4(B)). Moreover, many species were uncommonly found, namely, individuals belonging to Blattodea, Hymenoptera, Isopoda and Myriapoda. Based on these results shown above, the values of 60 and 25 ('*s*60_*s*25′) were chosen for σ parameter of Pyro- and Seq-Noise respectively.

Towards a quantitative validation, reads count was plotted against organisms' biomass. As expected for the higher taxonomy ranks (order, class), a positive increase between the two variables was evident and the number of assigned reads for the detected taxonomical groups are adjacent comparing 3′ and 5′ *coxI* barcode except for Diptera (Figure 3.5). On the contrary, at species level, reads count and species biomass are quite distant (Figure 3.6) pointing 3′ and 5′ *coxI* DNA barcode region giving a potential inconsistency and low sensitivity of reads assignment

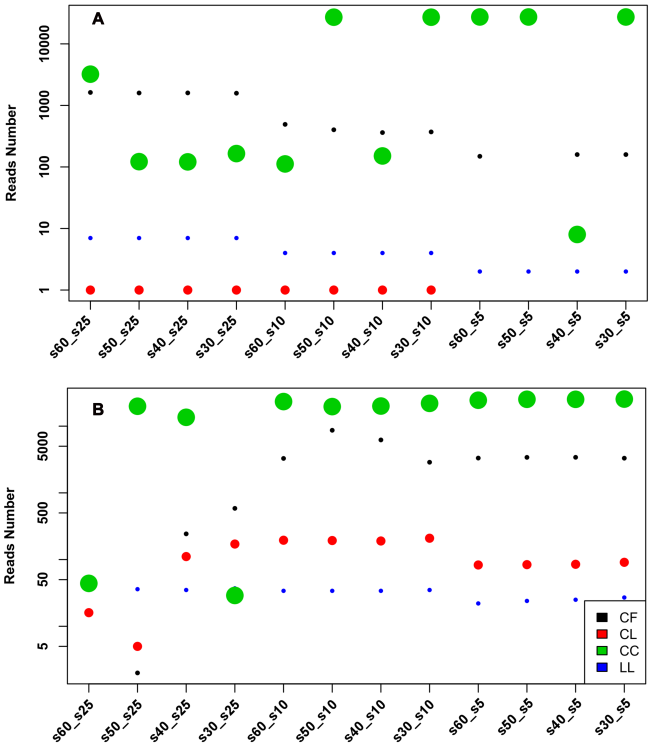


Figure 3.3: Influence of σ on assigned reads number at species level in MPE5 control sample.
Qualitative plot of the 12 σ values combinations of Pyro and Seq-Noise against the assigned reads abundance to known organisms at species level and their corresponding biomass. A) 5′ *coxI* barcode and B) 3′ *coxI* barcode. The bullets dimension are proportional to the biomass of each taxonomical group name (see Table 3.2). s=σ: clustering distance parameter value used by expectation maximization algorithm of Pyro- and Seq-Noise. CF *Calathus fracasii*; CL: *Carabus (Chaetocarabus) lefebvrei bayardi*; CC *Carabus (Tomocarabus) convexus dilatatus*; LL: *Laemostenus latialis*.

at low taxonomical category. Consequently, following the above mentioned assessment conducted on the control sample (MPE5), Ampli-

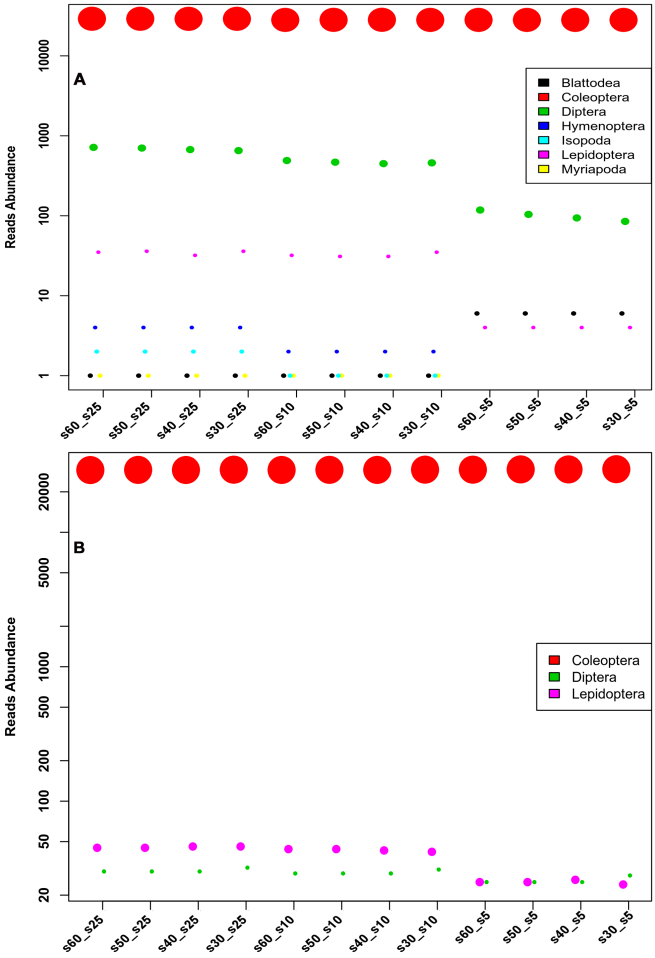


Figure 3.4: Influence of σ on assigned reads number at order level in MPE5 control sample.

Qualitative plot of the 12 σ values combinations of Pyro and Seq-Noise against the assigned reads abundance to known organisms at high taxonomical category (Order, Class) and their corresponding biomass. A) 5′ *coxI* barcode and B) 3′ *coxI* barcode. The bullets dimension are proportional to the biomass of each taxonomical group name (see Table 3.2). s=σ: clustering distance parameter value used by expectation maximization algorithm of Pyro- and Seq-Noise.

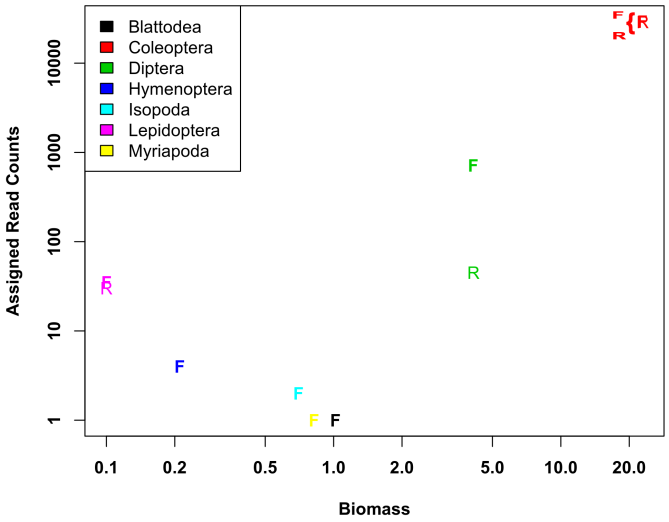


Figure 3.5: Quantitative plot of organisms' biomass and their assigned reads number at Order and Class levels.
Taxon assignment was conducted on AmpliconNoise output using s=σ parameters of 60 and 25 for Pyro and Seq-Noise respectively. The 'F' represent the reads mapping at 5' coxI barcode while the 'R' denote those mapping at 3'

conNoise workflow was executed for all eight samples at '$s60\_s25$' and the assigned sequence read clusters was annotated by Hidden Markov Model (HMM) classifier at order taxonomy rank.

3.3.5 Error rate evaluation

To compare base call errors of sequence reads with and without denoising a naïve clustering of un-denoised sequences was performed and compared with denoised ones obtained from AmpliconNoise pipeline. The naïve clustering consisted on collapsing all identical sequences of the same length and composition to unique representative ones. The

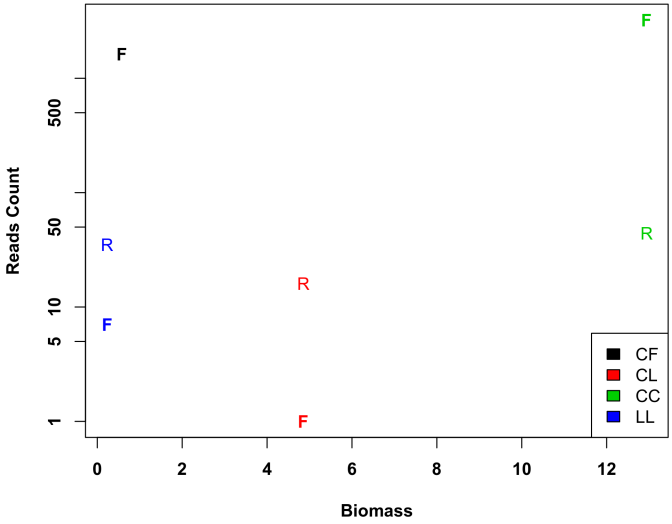


Figure 3.6: Quantative plot of organisms biomass and their corresponding assigned reads number at species level.
Taxon assignment was conducted on AmpliconNoise output using s=σ parameters of 60 and 25 for Pyro and Seq-Noise respectively. The 'F' represent the reads mapping at 5' *coxI* barcode while the 'R' denoten those mapping at 3'. CF *Calathus fracasii*; CL: *Carabus (Chaetocarabus) lefebvrei bayardi*; CC *Carabus (Tomocarabus) convexus dilatatus*; LL *Laemostenus latialis*.

error rate of both pyrosequencing and PCR polymerase was evaluated using Shannon entropy (H) (Shannon and Weaver, 1963) information calculated on naïve, Pyro- and Seq-Noise sequence clusters. A comparison of entropy information between sequence clusters produced by Pyro- and Seq-Noise showed that error rate was heterogeneously originated from both pyrosequencing and PCR polymerase. A predominance of PCR errors over pyrosequencing ones was evident in NH-5', CA-5', SH-5', SA-5' and MPE4-5' samples while the opposite was found in the remaining samples (Figure 3.7). This can provide a serious idea about error rate within sequence reads that can negatively interfere with a precise species assignment. Sequence number, entropy values

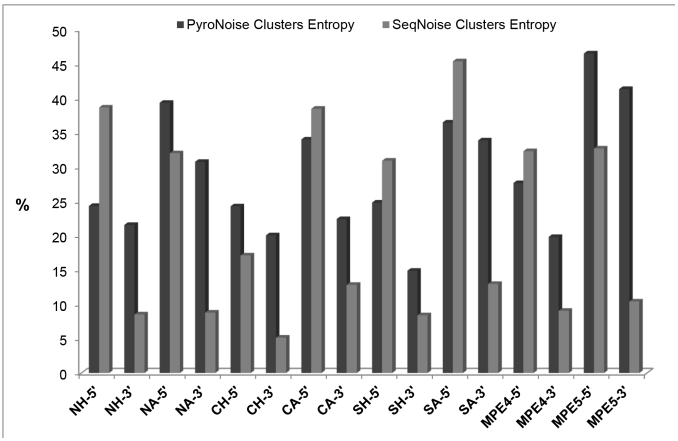


Figure 3.7: Percentage of lost entropy within read clusters following Pyro- and Seq-Noise computation.
The percentages were calculated taking naïve clusters as reference.

and their corresponding percentages over all denoising pipeline from pattern-based trimming to the final denoising step are in Table 3.4.

Comparing Figures 3.7 and 3.8, it is possible to observe that 3′ samples have bigger variation left after denoising than 5′ ones. This difference seems to be caused by a greater amount of variation removed by denoising in 5′ samples. The hypothesis is that the origin of this difference in the intensity of denoising procedure could be connected to the effect of GC content on the ability of sequencing reaction to work correctly. To test this hypothesis, the contribution of expected biological variation and the GC content on denoised read variation was tried to parse, taking into consideration the PFT sequences assigned to Coleoptera (the taxonomic group with more reads and a best set of reference sequences). In this perspective, ANOVA was computed using (i) the values of the mean exponential entropy value (Jost, 2006) calculated for each site of reference Coleoptera multiple alignment, (ii) the mean percentage of GC content obtained from the GC percentages at a sliding window of 100 nucleotides and (iii) the mean exponential entropy value

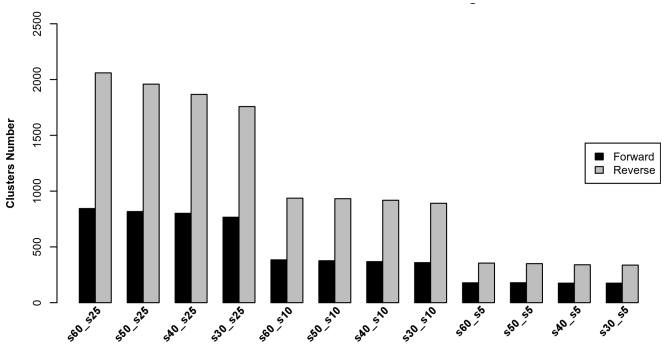


Figure 3.8: Clusters number at 5′ and 3′ *coxI* barcode for all σ combination of MPE5 sample.
Sequence clusters were obtained subsequently to denoising process where each cluster is represented by a unique sequence representing one (singleton) or more sequences.

calculated for each site of the multiple alignment of sequence clusters assigned to Coleoptera and belonging to MPE5 sample. The latter was calculated separately for the 5′ *coxI* DNA barcode, taking just the first 400bp of the multiple alignment, and the 3′ one starting from the 200th site until the end.

As expected, ANOVA results (Table 3.3) showed that the diversity index (expressed in exponential entropy value) of the assigned sequence clusters is linearly explained by that of reference sequences. In addition, the variability of entropy value in sequence clusters was statistically significant given GC content at both 5′ and 3′ *coxI* DNA barcode; but, the sum of squares values indicated that this significant variability due to GC content is larger at 3′ coxI barcode (38 corresponding to 17% of total read variance) than that at 5′ (4.5 equivalent to 2.4%). These results indicate some shortcomings of AmpliconNoise given by its incapacity to remove the excess of errors due to GC. In fact, if this was false, a minor difference in variation between the two groups would be expected at the end of denoising process. Chimeric sequences, detected by Uchime algorithm, were distributed homogeneously at both 5′ and 3′ ends of *coxI*

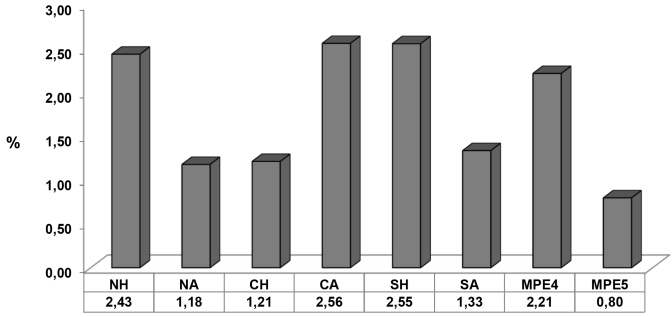


Figure 3.9: Chimeric sequences percentages per sample.

Table 3.3: ANOVA results.

	5′ *coxI* barcode analyzed sites: 1:400			3′ *coxI* barcode analyzed sites: 200-end		
	GCmean	**exp(RefEnt)**	**Residuals**	**GCmean**	**exp(RefEnt)**	**Residuals**
Degree of freedom	1	1	562	1	1	562
Sum of squares	4.518	58.766	123.808	37.999	50.025	135.407
Mean square	20.507	266.754	0.220	37.999	50.025	0.241
F-value	20.507	266.754	-	157.71	207.63	-
Pr(>F)	**7.257e-06** ***	< 2.2e-16 ***	-	**< 2.2e-16** ***	< 2.2e-16 ***	-

ANOVA of conservation pattern expressed in mean exponential entropy of assigned sequence clusters to Coleoptera order against average GC content and conservation pattern of Coleoptera reference sequences.
Significance codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ exp=exponential; RefEnt=reference entropy.

DNA barcode region. Chimerism results show that the most abundant chimeras were singletons obtained after denoising process. This fact can justify the identity of some singletons as amplification artifacts and not as true organisms’ sequence. However, the percentage of chimeric sequences was relatively low and ranged from 0.8 to 2.5% (Figure 3.9) which can be attributed to the optimized conditions used in sequencing libraries preparation.

Table 3.4: Sequence reads number along denoising process.

Sample	Reads	RFrag	Lenght Filtered RFrag	Naive Clusters	PyroNoise Cluster	Ampliconoise Clusters	Chimera Free Clusters
NH_F	54591	7763	7276	5476	2541	883	740
NH_R		11479	10642	9410	4676	3042	2895
NA_F	84402	29919	27761	20673	5895	1049	787
NA_R		29601	27767	19788	6518	3172	3009
CH_F	92598	1301	1253	1226	553	275	257
CH_R		4109	6436	3888	2741	2011	1949
CA_F	78938	20497	18909	15655	5090	887	695
CA_R		24634	23042	19197	7544	3192	3027
SH_F	56019	5461	5212	5226	2537	731	605
SH_R		8677	7906	7789	4218	2676	2565
SA_F	86936	29118	27187	21078	6163	919	660
SA_R		31919	29659	23056	7813	3771	3573
MPE4_F	67672	23449	21837	19856	7656	1675	1271
MPE4_R		22942	20396	19339	8126	4357	4097
MPE5_F	92494	35023	32479	24736	5116	844	650
MPE5_R		38534	34036	26100	5009	2060	1945

F = forward (5′ coxI barcode).
R = forward (3′ coxI barcode).
Read = Reads obtained from 454 (Roche life science) Sequencing machine.
Rfrag = Read Fragments obtained from trimming based on the found pattern.
Length Filtered Rfrag = Amplicon Noise filtering during Flowgram clustering and filtered at Flowgrams length longer than 100. **Naive Clusters** = unique sequences in lenght and composition (zero indels).
PyroNoise clusters = Clusters obtained after executing PyroNoise (the first step of AmpliconNoise pipeline). **AmpliconNoise Cluster** = Clusters obtained after executing AmpliconNoise pipeline. By substracting the number of AmpliconNoise Clusters from PyroNoise ones, we obtain the number of SeqNoise clusters.
Chimera Free Clusters = Clusters remained after Chimera removal with Uchime.

3.3.6 Taxon assignment

Hydrophilic samples

Following read clusters assignment to taxonomical orders by HMM classifier (HMMer3.0), all sequences that passed the threshold of Bayes Factor (BF) were considered potentially assigned to a taxonomical order name while the remaining ones were categorized as unclassified. This assignment revealed that CH (central hydrophilic) embraced the highest taxonomical diversity pattern as its representative sequences belong to 55 taxonomical orders. The less taxonomical diversity was encountered in NH (northern hydrophilic) to which 37 taxonomical orders were attributed while SH (southern hydrophilic) presented an intermediate diversity pattern of 44 orders when compared to the other

samples. Comparing abundance distributions across reads with taxonomical assignment, it is evident that samples having more singletons (CH followed by SH and NH) represented a higher taxonomical diversity. These singletons might probably represent rare species not yet discovered or catalogued in the taxonomy. To represent clearly taxonomic profiles for each analyzed data set, taxonomical assignments at both 3′ and 5′ ends were summarized together in a unique assignment pattern. 74.5% of hydrophilic sequence clusters were assigned to a total of 66 orders while the remaining 25.5% were unclassified. The most abundant taxonomical orders rank, having more than 50 assigned read clusters, are listed as follows in decreasing order of abundance: Haplotaxida, Rhabditida, Amphipoda, Hemiptera, Ascaridida, Tylenchida, Pyrenomonadales, Pantopoda, Philodinida, Rhodobacterales, Isoptera, Coleoptera, Siphonophora, Rhodospirillales, Mesostigmata, Rickettsiales, Orthoptera, Acoela, Ploimida. An important taxonomic assignment aspect can be observed in hydrophilic samples where 25 orders were shared between the three geographical areas while 21 between two, and 20 were unique for one of the three areas.

Aerophilic samples

As in the case of hydrophilic samples, read clusters that passed the threshold of Bayes factor were assigned to the corresponding order on which they matched. Sequences belonging to SA (southern aerophilic) showed the occurrence of 33 orders while 29 and 24 orders were checked in NA (northern aerophilic) and CA (central aerophilic) respectively. Samples with abundant singletons, NA at 5′ and SA at 3′, depicted a higher taxonomical diversity than CA whereas the latter represented the highest number of clusters assigned to the same order. Taxonomic identity of 57.7% sequence clusters in aerophilic samples were classified in 48 orders while 42.3% were Unclassified. Orders sorted in decreasing abundance and having more than 50 assigned sequence clusters were: Coleoptera, Oribatida, Hemiptera, Mesostigmata, Trichoptera, Diptera, Amphipoda, Araneae, Orthoptera, Decapoda, Collembola, Neuroptera, Astigmata. In addition, 13 orders were shared between the three geo-

graphical sites while 14 between two and 21 unique for one of the three sites. Note that, the shared orders have different abundance across sites and sometimes their occurrence within a site is represented by just few individuals or singleton sequence as in the case of Haplotaxida and Strepsiptera.

Pitfall traps samples

The assigned sequence clusters to a known taxonomical order confirmed a higher taxonomic complexity in MPE4 harboring 34 taxonomical orders than that of MPE5 embracing only 22. In addition, cluster counts across the same orders were in most cases more abundant in MPE4 when compared to those of MPE5. As expected, organisms belonging to the order Coleoptera were found at high but different abundance in both MPE4 and MPE5 (Table 3.5). Carabids species morphologically identified were not all found in PFT samples, where three species were missed: *Abax parallelepipedus*,*Pterostichus micans* and *Calathus montivagus*. This fact can be related to a certain amplification bias of universal primers within a bulk DNA sample is still under examination and needs further consideration. Eighteen shared orders were found in both samples with Coleoptera the most abundant one. As expected, the majority of taxonomic orders found in MPE5 were present in MPE4 being the same sample as the previous but equilibrated for organisms biomass. However, MPE5 ascertained three unique orders represented by singletons (Lithobiomorpha, Eunicida, Hemiptera) while MPE4 13 with low to intermediate abundance (Table 3.5). A higher amount of unclassified sequences were established in MPE4 (32.92%) compared to MPE5 (23.31%). In summary, the most represented eukaryotic taxonomy orders sorted by their abundance in PFT samples, with more than 50 assigned sequence clusters, are distributed as follows: in MPE4: Coleoptera, Hymenoptera, Diptera, Archaeognatha, Trichoptera, Isopoda, Heteronemertea, while in MPE5: Coleoptera, Diptera, Trichoptera.

3.4 Discussion

Taxonomic profiling based on high-throughput sequencing of species discriminant locus has been and still being used in numerous microbial biodiversity studies where it showed its feasibility in monitoring microbial species of clinical or ecological relevance (Fonseca et al., 2010; Buee et al., 2009; Brulc et al., 2009; Bittner et al., 2010; Singh et al., 2009). However, apart from few studies regarding meiofaunal biodiversity assessment through ultrasequencing of 18S rDNA (Creer et al., 2010), there is a single study that used the *coxI* pyrosequencing technology in biomonitoring few macroinvertebrates taxa of river benthos (Hajibabaei et al., 2011). In our study, a large-scale biodiversity assessment of metazoan taxa in chestnut soil ecosystem is addressed for the first time. The effectiveness of coupling coxI DNA barcoding with 454 (Roche Life Science) pyrosequencing technology is demonstrated. In addition, the importance of biomass equilibration to unveil the true diversity picture of chestnut soil environment is evaluated. Moreover, the relevance of parameter setting in sequence reads denoising and the need of an internal control sample to model errors distribution is confirmed. Several promises and considerations have emerged throughout the present study especially regarding sequencing libraries preparation and sequence reads filtering, sampling and organisms extraction procedures related to taxonomic profiling and taxon assignment method.

Sequence reads denoising and GC content

Sequence reads denoising showed its effectiveness in correcting base errors coming from pyrosequencing and/or PCR polymerase during DNA enrichment. This fact was evident where naïve sequence clusters entropy was higher for all samples than that of AmpliconNoise clusters obtained from the combined effect of Pyro- and Seq-Noise. In addition, *de novo* chimera removal algorithm of Uchime showed its high performance, in terms of computation cost and precision, in detecting and removing sequencing and amplification artifacts. Sigma (σ) screening, the magnitude clustering parameter of Pyro- and Seq-Noise, has revealed

several important facts concerning sequencing and PCR polymerase errors. In relation to the results obtained by Hoff (Hoff, 2009), our investigations on the effect of GC content on sequence diversity was evident from ANOVA outcomes (Table 3.3) where GC content has contributed to the increase of unreal sequences especially at 3' end *coxI* DNA barcode. These results highlighted the reliable sensitivity of AmpliconNoise pipeline in 454 sequence reads correction. The accuracy and precision of AmpliconNoise over the 12 σ values combinations showed a scarce performance following taxon assignment at species level while an acceptable protocol pointing taxonomy order rank. Moreover, qualitative plots as well as quantitative ones approved a straight correlation between reads number with their corresponding organisms' biomass just at taxonomy Order rank but never at species one (Figures 3.3 3.6). Nevertheless, the impossible detection of some organisms belonging to a pre-identified higher taxonomical group (order and/or class, Table 3.2) at the best σ parameter values could suggest the use of an internal error distribution by sample. This distribution can be generated by the control DNA usually sequenced together with the sample itself and setup it for PyroNoise; while PCR polymerase error distribution can arise from the construction of an artificial library of a taxonomic complexity similar to that of the empirical sample with known sequence coming from sequenced clone products of the genomic region of interest.

Sequencing libraries preparation

In this study, the final yield of sequence reads suitable for taxonomic profiling analyses was affected by several pre-processing procedures starting from pattern-based trimming to chimera removal. Pattern-based trimming which was a necessary operation for shotgun sequencing libraries seriously reduced the final number and the overall sequence reads length. A considerable importance of the sequencing libraries preparation used is that it promoted the harvest of whole coxI barcode region which is currently rare to get by the NGS platforms. A plausible alternative library preparation method recommends the

preparation of amplicon libraries sequencing (Buee et al., 2009; Yu et al., 2012; Zhou et al., 2011; Lee et al., 2011) to reduce read loss and keep read length intact. For this purpose, sequencing platforms might reach the capacity to sequence long amplicons to prevent reads loss and increase sequencing coverage fundamental for more reliable results in downstream analyses.

Taxonomic profiling

Taxonomic profiles corresponded mainly to metazoan organisms and encompassed 66, 48 and 34 taxonomical orders rank for hydrophilic, aerophilic and pitfall traps respectively. These results approved the capacity of Folmer primer set in amplifying the *coxI* DNA barcode of a large spectrum of metazoan species (Rougerie et al., 2009; Folmer et al., 1994; Zhou et al., 2011; Raupach et al., 2010). An additional feasibility feature of these primers is the amplification of bacterial (Richestsiales, Rhodospirillales and Legionellales), fungal (Saprolegniales) and algal (Cryptomonadales) individuals. These analyses highlighted also a consistent group of unclassified sequences that can be justified by the low number of described Eukaryotic species estimated to just 14% on the Earth (Mora et al., 2011). An important aspect that characterized all analyzed data sets is the higher variability of *coxI* DNA barcode at its 3′ end. This fact was explained by GC content and suggests a GC content normalization of denoising protocols based on a reference training data set having similar characteristics as the sample under study. The unusual presence of marine organisms (e.g. Pantopoda, Decapoda, Siphonophora), even at low abundance could be introduced either by a shortcoming in denoising normalization that introduced a systematic error along the taxon assignment step or by a low discriminative capacity of *coxI* DNA barcode for these organisms. Moreover it is also proper to think of a possible airborne contamination. We still know too little about the dynamics of small organism and DNA dispersions on the environment in order to exclude this hypothesis.

Conservation patterns or intersections of taxonomic orders across sites and between hydrophilic and aerophilic sample categories put in

evidence the interaction of these organisms with the microhabitat of chestnut soil. The differences in individual abundances within the conserved taxonomic orders across hydrophilic and aerophilic samples express so far the role these organisms exhibit in different habitats as well as the influence of the global climate on them. For instance, species belonging to Coleoptera order were the most abundant in aerophilic samples because they are inhabitant of soil litter while Haplotaxida were those most abundant in Hydrophilic samples due to their affinity towards water films in the second 10cm of soil. Note that, Coleoptera can be directly affected by soil and climate conditions as they are soil dwelling crawlers and consequently indicating the forest ecology conditions. This fact justify the differences in their abundance among the three sampled geographical sites and consequently their use as environmental indicators for forest ecology (Jeffery et al., 2010). The presence of Haplitaxida species is an indicator of soil health given by their main activity in soil aeration and organic matter release. These conservation aspects reflect strongly the importance of organismal extraction methods in separating organisms belonging to different micro-habitats (soil strata) and eventually are part of a specific micro-ecosystem.

Another example of forest conditions interaction with the soil organisms content across sites is the presence of Rhabditida and Tylenchida species, known to be phytoparasitic, plausibly correlated to the presence of different vegetation in the sampled chestnut soil forests highly influenced by the regional climate. In aerophilic samples, the prevalence and the difference in abundance of Coleoptera, Collembola, Trichoptera and Neuroptera, members of the macrofauna, is reasonably due to their direct relation with environmental conditions in terms of weather and soil conditions. For instance, Collembola are feeder of fungus and organic matter which can vary between two different chestnut forests; while, members of Neuroptera are predators of some Hemiptera species that live and eat on plant foliage. The comparison between controls samples, the pitfall traps collected MPE5 and the biomass- equilibrated (MPE4) derived from it, highlighted relevant details for sampling and PCR amplification procedure optimization. In details, equilibrating biomass and diluting DNA extract at 1/50 in

MPE4 revealed higher number of taxonomic orders even at low or intermediate abundance potentially due to the reduction of inhibition activity within diluted DNA sample; while the amplification success of DNA extract without dilution in MPE5, where the highest biomass of Carabids dominated, is actually correlated to the higher mitochondrial presence and eventually *coxI* in greater biomass content. The detection of 13 additional orders in biomass-equilibrated sample when compared to its naturally occurring one provides insights about the importance of biomass balance used in environmental samples suspected to sequencing. Therefore, biomass dominance within samples seemed to alter the true taxonomic profile by masking significant organisms in a forest and soil environment such as Oribatid mites and organisms belonging to Orthoptera, Collembola and Archaeognatha. The actual significance between the number of detected orders and their corresponding sequence reads abundance may be due to a random loss of rare species across experimental manipulation, mainly the universal primer amplification bias that potentially masked existing individuals. This aspect will be explained and demonstrated in an on-going publication regarding the same control samples. To conclude, discovering and monitoring soil metazoan biodiversity has been a real scientific challenge mainly due to the pronounced number of un-described species within soil environment. Coupling DNA barcoding approach with high-throughput 454 pyrosequencing showed to be a robust and fast method in accessing information about chestnut soil metazoan biodiversity and can be suggested for biodiversity studies to understand species population dynamics and their corresponding interactions within their ecosystem. The practical value of this coupled approach allowed the maximization of species recovery from samples and the construction of taxonomic profiles both shared between samples as well as unique for each site. The unique taxonomic orders found in all sample categories (hydrophilic, aerophilic and pitfall traps) might be due to a true diversification among geographical areas or to the shortcomings of denoising procedure not considering GC content and therefore influence directly taxon assignment methods. To address this challenge, is recommend the consideration of introducing an additional parameter regarding GC content in denoising pipeline, assessing a phylogenetic taxon assign-

ment method and compare it with the present one to ensure its consistency and reliability.

Table 3.5: Taxonomic Orders rank in Pitfall traps

Shared Taxonomic Orders between MPE4 and MPE5		
Order Name	MPE4	MPE5
Coleoptera	1794	1493
Hymenoptera	370	3
Diptera	209	163
Trichoptera	146	79
Isopoda	141	2
Heteronemertea	51	1
Rickettsiales	38	14
Amphipoda	37	3
Mantodea	33	21
Neuroptera	22	6
Pyrenomonadales	18	9
Plecoptera	10	12
Lepidoptera	8	41
Primates	7	13
Blattodea	6	1
Glomerida	5	2
Decapoda	3	1
Sphacelariales	3	1
Unique Taxonomic Orders for MPE4		
Legionellales	215	-
Archaeognatha	147	-
Alteromonadales	54	-
Orthoptera	35	-
Collembola	27	-
Oribatida	15	-
Tremellales	10	-
Dictyotales	4	-
Araneae	3	-
Ephemeroptera	3	-
Oceanospirillales	3	-
Capnodiales	3	-
Perciformes	2	-
Ascaridida	2	-
Haplotaxida	1	-
Mesostigmata	1	-
Onygenales	1	-
Unique Taxonomic Orders for MPE5		
Eunicida	-	1
Hemiptera	-	1
Lithobiomorpha	-	1

Taxonomic Orders rank in Pitfall traps (PFT) samples and their corresponding sequence clusters abundance per sample type.

4 Section 3

Observing microbiota invasion mediated by *Varroa destructor* to *Apis mellifera*

.

4.1 Introduction

The interpretation of host-parasite interactions is one of the most intriguing themes in biological studies. In a symbiotic relationship the partners reciprocally influence their physiology and, in general, their evolution. Nowadays, the characterization of the microbiome (intended as the sum of microscopic living beings found in a symbiotic relationship in different host body compounds, ranging from the gut to the skin), is considered pivotal to undestand physiological changes in a symbiotic relationship (Mazmanian et al., 2005). In recent years the parasitism, a of symbiosis, has become one of the most studied, with the publication of several papers describing the microbiome composition of different hosts (Sanchez et al., 2012; Meriweather et al., 2013; Dimitriu et al., 2013). In spite of such interest, the understanding of the mechanisms influencing microbial diversity and its distribution between the host and its symbiont is essential to describe the dynamics occurring in a symbiosis.

To explore these dynamics a classical approach for the analysis of microbial communities was tested against a new approach considering the phylogenetic entropy as a measure of diversity (Chao et al., 2010). The

phylogenetic entropy is a generalization of Shannon entropy that takes into account the fact that the different categories observed are not all equally different from each other but have a similar structure that could be modeled using a phylogenetic tree. The use of Shannon entropy is quite new in community ecology studies. Indeed, researchers have just started incorporating historical constraints represented as phylogenies into their analyses. This innovation is motivated by the aim to bridge the gap between evolutionary and ecological analyses (Lozupone et al., 2011) Aware of the great potentials of this new vision, was decided to test it on the biological model constituted by the honey bee (*Apis mellifera*) and its parasitic mite varroa (*Varroa destructor*). The rationale is that the analysis of the microbiome of both varroa parasites and honey bee larvae could open new perspectives concerning the role of varroa on the health of honey bee colonies and the phylogenetic entropy approach would became a new standard in the analyses of bacterial communities.

Varroa destructor (Arachnida: Varroidae) has been described as a major cause affecting honey bee colonies. Many studies documented the role played by *V. destructor* in increasing the incidence of deformed wing virus (Möckel et al., 2011)and as a vector of bacterial pathogens, such as those responsible for European foulbrood (e.g. *Mellisococcus plutonis*) (Forsgren, 2010; Evans and Schwarz, 2011) *V. destructor* parasites mainly honey bee larvae in their brood cells, where female mites feeds on honey bee hemolymph, and lays eggs. Hatched varroa males and females mate, and when the honey bee emerge from the brood cell, the fecundated females of the parasite start the phoretic phase on adult bees, until they reach a new brood cell in the same nurse or disperse in other hives using worker adults (Pernal et al., 2005). Mites have a large dispersal capability and in absence of reiterate chemical and/or antibiotic treatments, infested honey bee colonies typically collapse in few years.

For these reasons, the occurrence of varroa has serious consequences in the ecological, social and economic contexts (Rinderer et al., 2010; Rosenkranz et al., 2010; Annoscia et al., 2012; Guzman-Novoa et al., 2012). The characterization of microbial communities involved in A. mellifera biology has proven to be a good indicator of its state of health

(Martinson et al., 2012). Nevertheless, the ecological dynamics of the honey bee varroa parasitic symbiosis are still largely unknown.

The studies conducted so far on adult honey bees showed a characteristic microbiome whose structure has been confirmed in various papers (Jeyaprakash et al., 2003; Dillon and Dillon, 2004; Mohr and Tebbe, 2006; Martinson et al., 2011; Sabree et al., 2012). However,the microbiome of the larval stages as well as that of the parasitic mite are largely unknown (see for examples, (Martinson et al., 2012; Cornman et al., 2010)). As a consequence, apart from few dedicated works on the trasmission of specific pathogens (Mouches et al., 1984; Forsgren, 2010), it is still unclear if and how bacterial communities of honey bee and varroa affect each other.

The subject of of work is highly up do date, indeed the general interest in microbiomes is supported by several recent studies. For instance, there are evidences that humans and mice subjected to different kind of stress (such as diseases, parasites, ecological factors) are characterized by intense modifications in their own microbiomes in terms of initial colonization, final composition and overall stabilization (Candela et al., 2012; Lozupone et al., 2012). Given these premises, it's reasonable to expect alterations of honey bee microbiome due to the symbiosis with varroa.

Previously published data indicate a peculiar pattern of microbiome dynamics over the life cycle of the insect. The pupa is almost sterile, as a consequence of the physiologic characteristics of the gut tract and the diet of mature larvae during the six days before capping (i.e. the closure of the brood cell) (Martinson et al., 2012). The larva retains its faeces from the early days of development, due to the temporary absence of a connection between the large midgut and the hindgut. The mature larva defecates just before spinning a cocoon, when the capping has already happened. Since the cocooned pupa obviously does not eat, we can assume that there is no further colonization by bacteria present in the brood cell. Through these mechanisms the early microbiome characterizing honey bee larvae is maintained constant in composition and ubiquitous in space (Jeyaprakash et al., 2003; Mohr and Tebbe, 2006). On the whole, it is reasonable to assume that the bacterial load within

the brood cells partially reflects the total bacterial count of the hives and that the microbial communities characterizing the hives are partially present in the cells even after the capping (Martinson et al., 2012). But what happens when varroa alter this equilibrium? The disturbance of varroa in the developmental phase, and the consequent formation of the nutrition hole caused by the parasite, could lead to the intrusion of external bacterial into the larva, with a substantial modification of the bacteria community.

The hypothesis is that varroa mites play a fundamental role in the alteration of bacterial composition of honey bee larvae, acting not only as a vector, but also as a sort of an open "door" through which exogenous bacteria alter the mechanisms of primary succession in the "simple" honey bee larval microbiome. To validate this hypothesis, varroa and honey bee bacterial communities were studied through barcoded amplicon pyrosequencing methods, taking advantage of the NGS methods (Blow, 2008; Metzker, 2010) and the opportunity to detect uncultured and uncultivable bacteria allowed by such techniques. The results showed a significant alteration of the microbiome of parasitized honey bees and an advantage of this new method in terms of capability to recognize the OTUs that can discriminate the categories examined.

4.2 Materials and Methods

A schematic vision of the experimental pipeline is shown in Figure 4.1

4.2.1 Sampling

Honey bees larvae and varroa mites were sampled directly from capped brood cells in 8 apiaries in Northern Italy. A total of 43 samples were used for the molecular analysis. 21 individuals of honeybee larvae from 7 different apiaries and, for each one, the varroa mites found in the same brood cell; as a negative control, a pool of 5 healthy larvae from a non infected site was analyzed. Opercula of cells were opened with

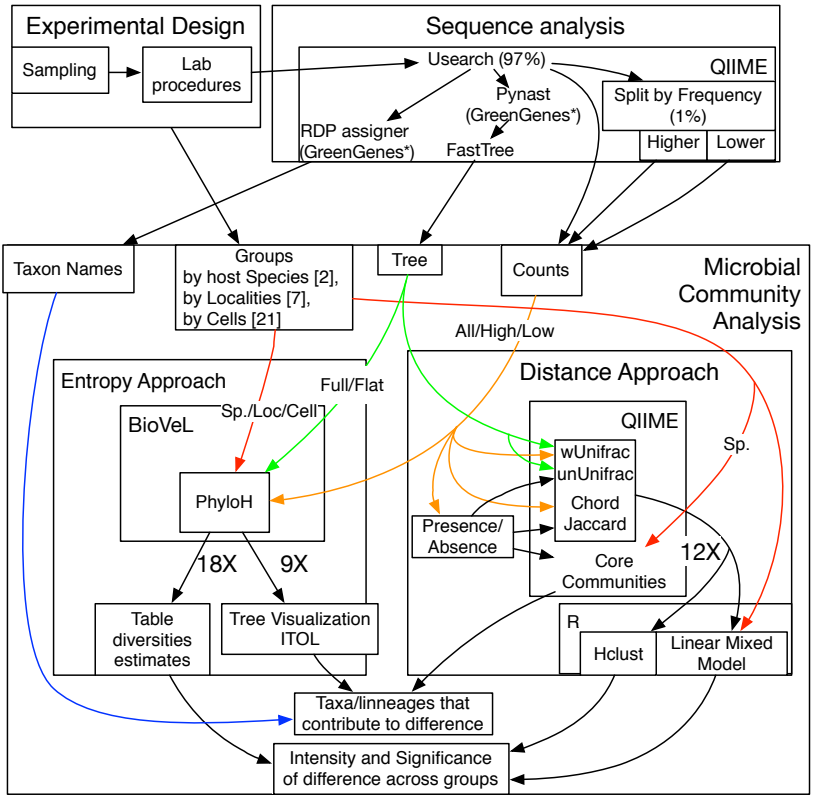


Figure 4.1: A schematic vision of the experimental pipeline
In the figure is summarized all steps analysis used to explore diversity in the bacterial communities. Different colours were used to describe the variability of the approach involved in the analysis.

sterile instruments; honeybee larvae and varroa were immediately removed and put in 2 ml tubes filled with absolute ethanol. The samples were stored at $-20°C$ until the DNA extraction.In the study area, *V. destructor* is abundant and widespread; for this reason, a only one non-infested apiary was found. This apiary was determined to be healthy

after a careful inspection of all the hives.

4.2.2 DNA extraction

All the extraction steps were performed in a sterile biological hood. After the removal of the head, only the first segments of *A. mellifera* specimens were used for the extraction, while for *V. destructor* DNA was extracted from the whole organism. The dissections were made with a sterile scalpel in a Petri dish. Every sample was then rehydrated for 4h in sterile water at room temperature, and mechanically grinded with the scalpel. Total DNA was then extracted using a commercial kit (DNeasy blood and tissue kit; Qiagen), and eluted in 50 μ l sterile water. A pretreatment of Qiagen columns was performed to wash away any trace of contaminating bacterial DNA (Evans et al., 2003; Mohammadi et al., 2005). Following the DNA extraction from individual mites, DNA extracts from mites coming from the same brood cell were pooled together. Similarly, extracts from five larvae in the non-infested apiary were pooled.

4.2.3 16S rRNA amplification and pyrosequencing

A 16S rRNA gene fragment corresponding to the V3 hypervariable region was PCR-amplified with Roche 454 FLX (Titanium reagents) using forward primer 341F (5'-CCTACGGGAGGCAGCAG-3') and reverse primer 518R (5'-ATTACCGCGGCTGCTGG-3') (Watanabe et al., 2001). A second PCR step using the products of the firsts as template was then performed. The first reaction was performed in a 20 μl volume with the following reagents: Taq-buffer with MgCl2 1X dNTPs 2 mM, forward and reverse primers 1 $pmol/\mu l$ each, Taq polimerase 0,5 $U/\mu l$, DNA 1 μl, milliQ water. The thermal cycle was: 94° for 90 s, 29 cycles at 94° for 20 s, 94° 30 s, 94° for 20 s and then 94° for 10 s and 60° for 5 min. The second PCR reaction step was performed with the use of 52 bp primers, comprising pyrosequencing primers A or B, MID identifiers and 518R or 341F primers. The reagent concentrations were

the same except for the primers (0.5 $pmol/\mu l$). The thermal cycle was 94° for 90 s, 40 cycles at 94° for 20 s, 58° for 30 s, 72° for 20 s and then 72° for 10 s and 60° for 5 min. For every sample a unique combination of MIDs on the primer forward and reverse was used. PCR products were quantified using Bioanalyzer 2100 (Agilent) in order to put in the pool to be sequenced the same number of DNA amplicons coming from every sample. 454 pyrosequencing was performed by BMR Genomics service at the Interdepartmental Biotechnology Centre of the University of Padua (CRIBI).

4.2.4 Sequences analysis

Acacia software version 1.52 (Bragg et al., 2012) was used for pyrosequencing noise removal considering Balzer Error model and a Maximum k-mer distance between reads of 13 (default parameters for error corrections). All reads were trimmed, filtered and assigned to the corresponding sample according to their tag. Sequences shorter than 100 bp with a quality value less than 30 or containing unresolved nucleotides were removed from the dataset. The detection of chimera reads was performed using a pipeline based on USEARCH (Edgar, 2010) and UCHIME (Edgar et al., 2011) included in Quantitative Insights Into Microbial Ecology (QIIME) pipeline software (version 1.6.0) (Caporaso et al., 2010). UCLUST wrapper was used to cluster the sequences in Operational Taxonomic Units (OTUs), based on 97% sequence similarity. The cluster centroid for each OTU was chosen as the OTU representative sequence.

To estimate diversity and reduce noise in patterns of beta diversity, singleton OTUs were removed before community analysis (Zhou et al., 2011). Using a Python script the Greengenes 16S rRNA database prefiltered at 97% identity(McDonald et al., 2012) was merged with a bacterial OTUs dataset constituted by pathogens or symbiont (i.e. pathogens, mutualists and commensals) previously described in studies conducted on *Apis mellifera* (Mohr and Tebbe, 2006; Martinson et al., 2011; Mattila et al., 2012; Martinson et al., 2012; Moran et al., 2012; Sabree et al., 2012). All sequences used to create the bacterial OTUs dataset were

downloaded after querying GeneBank database for published Accession Number or Taxonomy and clustered at 97 % sequence similarity.

The taxonomic attribution of representative sequence was carried out using RDP Bayesian Classifier (Wang et al., 2007) retrained on the new merged dataset, using a 0.8 confidence level as suggested for default parameters in QIIME. OTUs were assigned by the RDP classifier, considering the fifth and sixth taxonomic levels where possible, which, in most cases, corresponded to family and genus ranks. If RDP assigned the OTU with a probability between 0.8 and 0.9, the representative sequence was sought in NCBI nucleotide database. When a perfect match with BLAST was found the NCBI taxonomy was used.

4.2.5 Microbial Community Analyses

The community abundance profile produced by UCLUST and labeled by RDP Bayesian Classifier was split in two groups depending if their global frequency was lower or higher than 1%. This produced three data sets: All (where all OTUs found are considered, AS), Low (where only the OTUs with frequency lower than 1% are considered; Low Frequency Cluster, LFC) and High frequency clusters (where only the OTUs with frequency Higher than 1% are considered; High Frequency Cluster, HFC). This allowed to explore the effect of dominant and rare organism in the differentiation among the two hosts (honey bee and varroa). Changes in community diversity in relationship with environmental parameters, and organism groups that contributed the most to the differentiation were examined using two different approaches one based on distance method and the other based on a measure of entropy. Both approaches take into account the phylogenetic structure of the data, albeit in very different manner. Linear Mixed model is a well known statistical framework in microbial community analysis that allows to compare the effect of different explanatory variable together, but require rarefaction of the data because sensitive to unbalanced sampling design.

To cope with these limits for the first time in Microbial Community

Analyses was applied partitioning phylogenetic diversity (Chao et al., 2010) approach. This method is being framed within Information Theory and can deal directly with discrete value, without producing distance matrix. It can also incorporate information deriving from unbalanced sampling, thus it does not need a step of data rarefaction.

4.2.6 Distance and Mixed Models approach

All the analyses were performed on the rarefied OTU tables to permit comparisons of diversity patterns within and between communities. The number of OTUs (based on 97% sequences similarity) was determined for each sample. Community analyses were performed with QIIME (Caporaso et al., 2010) and R environment for statistical computing (R Development Core Team, 2011) . To explore different perspectives on the main factors that have an impact on microbial community composition involved it was decided to use both a qualitative (*Jaccard* and *Unweighted unifrac* (Lozupone et al., 2011)) and a quantitative (squared-chord (Cavalli-Sforza and Edwards, 1967; Orlóci L, 1967) and *weighted Unifrac* (Lozupone et al., 2011)) analyisis approach. *Jaccard* and *squared-chord* distances were chosen as a complementarity metrics to Unifrac metrics to test how the signal changes with or without phylogenetic information. The squared-chord distance metric was chosed because was identified in previous works as a metric that fit well at an exploratory analysis of communities where sampling was conducted blindly (see for example(Legendre and Gallagher, 2001)).

To interpret the distance matrix we used UPGMA hierarchical clustering method and we tested the robustness of results with a Jackknifing analysis (1000 permutations). Furthermore, we tested the possibility of a microbial flow between honeybee and its parasite varroa using a *Linear Mixed Models (LMMs)*. The model used was the following:

$$Unifrac_{H,V} \sim A_{(ce_H == ce_V)} + B_{(hi_H == hi_V)} + \epsilon_{ce_H} + \epsilon_{ce_V} + \epsilon_{hi_H} + \epsilon_{hi_V}$$

with *(un)weighted Unifrac* distances rarefied as the response variable, as fixed effect $A_{(ce_H == ce_V)}$ and $B_{(hi_H == hi_V)}$, respectively the pair of honeybees and parasite of the same cell and the pair of all honeybees and

all the parasites belonging to the same hives. Single honeybees (ϵ_{ce_H}) and parasites (ϵ_{ce_V}) were treated as random effects as the single hives (ϵ_{hi_H} and ϵ_{hi_V}). Models were fitted with the lmer procedure in the lme4 package in R 2.8.1 (Bates and Maechler, 2009).

The core OTUs found in the honey bee and in the mite was identified using *compute_core_microbiome.py* script from QIIME. OTUs were grouped according to their presence in a specific percentage of the total samples. The groping-steps were defined as elevens threshold between 0.5 and 1, corresponding to the 50% and 100% of the samples, respectively. This allowed to define the core community of each host species. Specifically, to recognize the OTUs that are present in the majority of the samples of a given host. Some figures, showing the taxonomic assignment and the abundance distribution, were made with the aid of ggplot2 package in R (Wickham, 2009).

4.2.7 Partitioning Phylogenetic Diversity

Following the framework proposed by (Jost, 2007) it is possible to parse the total entropy of a data set, named γ, in intra-groups, called α, and inter-groups, called β components. Jost clearly distinguish between entropy measures, that have as unit bits (or nats, or bans, depending from the base of logarithm), and diversity measure that have as unit the equivalent number of equally abundant categories.

Partitioning operation are performed using entropies while the final result is transformed in diversity by elevating to the base of the used logarithm. Assuming that cluster label of observation is collected in vector X and that group label is collected in vector Y, this framework allows to define H_γ as entropy of X, H_α as conditional entropy of X conditional to Y, and H_β entropy as $H_\gamma - H_\alpha$ or also as the mutual information between X and Y.

It is important to notice that the β diversity (D), the exponential of H_β , has as unit the equivalent number of equally abundant and different sample (the categories of Y), while D_γ and D_α are measured in equivalent number of equally abundant cluster (the categories of X).

Within microbial community analysis, the interest lays generally in estimating the D_β or some transformation of it as the species turnover or the species overlap. To assess if this measure is significantly different from 1 (the group are as different as if they were only one group) the realized statistics were compared with a null distribution obtained by a permutation of X values on to Y ones. This procedure keeps number of observation per group constant, allowing to account potential different sampling effort per group. As described here this procedure do not take into account the phylogenetic structure that links the categories of the vector X.

This is possible using the phylogenetic entropy. The phylogenetic entropy is a generalization of Shannon entropy that takes into account the fact that the different categories observed are not all equally different from each other but have a similar structure that could be modeled using a phylogenetic tree. Following Chao et al. (2010), and assuming that variable of the categories, here the cluster defined by UCLUST, were organized with a phylogenetic structure t. Entropy measure could be defined as follows:

$$H_p(X) = -\sum_{i \in B_t} \frac{L_i}{T} p_i \log p_i$$

where B_t is the set of branches of the tree t and p_i is the frequency of the descendant of branch i. Once this point set it is easy to generalize the partitioning of diversity to include phylogenetic information. Phylogenetic entropy γ ($H_{p\gamma}$) is equal to $H_p(X)$, while phylogenetic entropy α is equal to the weighted by observation mean of the phylogenetic entropy per group more formally expressed as follow:

$$H_p\alpha = \sum_{y \in Y} p_y \sum_{x \in X} H_p(X|Y = y)$$

while phylogenetic entropy beta ($H_{p\beta}$) remain defined as the $H_{p\gamma} - H_{p\alpha}$. It is important to notice that although phylogenetic entropy is a generalization of Shannon entropy and such do not guarantee to keep

all its properties is possible to show (Appendix I) that $H_{p\beta}$ as defined here could be obtained both using this formula or using a phylogenetic generalization of the Kullback-Leiber distance. This matches the different way to estimate the classical mutual information. Given that the $H_{p\beta}$ s a difference of two summations in which each term is relative to a branch in the tree is possible to reorder the term and obtain the contribution of each branch to the final $H_{p\beta}$.

The approach was applied using the tree obtained from FastTree and the another tree with the same topology but internal branch with length zero and terminal branch with length 1. This last setting allow to perform the analysis as it was Shannon entropy but it using same software implementation. These two alternative set up allow to evaluate the importance of the phylogeny in defining the results. The method is implemented in a python script (PhyloH[1]) and wrapped in a Web Service[2] that is used within a workflow[3] that parse UCLUST output in the correct input for the script. The full service is available on a portal[4] as web application (Figure 4.2).

4.3 Results

4.3.1 Sequences Analysis

After sorting sequence reads for quality scores, sequencing errors and chimeras, our dataset consisted of 34,816 sequences. An average of 809 DNA sequences (range: 250-1650 length: > 100 bp) were then available for further analyses. UCLUST returned 295 OTUs (All Sequences, AS), where 21 exceeded the threshold of 1% of minimum total observations count (Hight Frequency Cluster, HFC)(tot sequences: 24,005) and 274 (Low Frequency Cluster, LFC) were defined as a rare OTUs (tot sequences: 10,811). The complete list of OTUs found is in the Table 4.2.

[1]https://github.com/svicario/phyloH
[2]https://www.biodiversitycatalogue.org/rest_s_do5(m)ethods/143
[3]http://www.myexperiment.org/workflows/3570.html
[4]https://portal1.at.biovel.eu/workflows/81

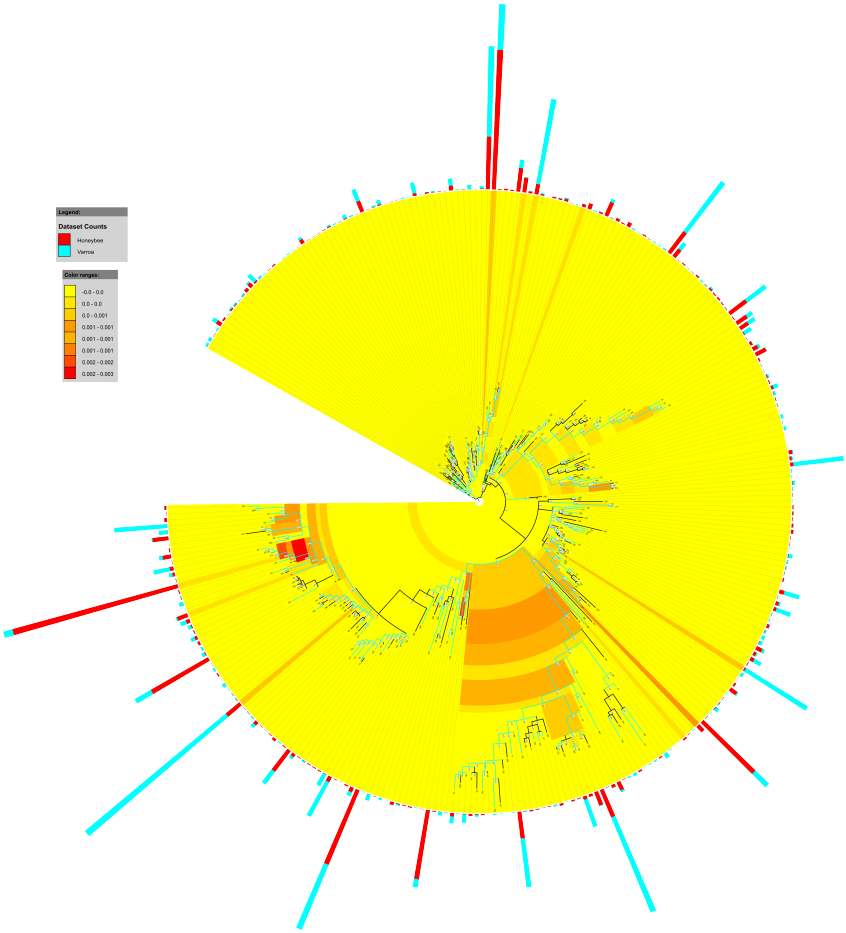


Figure 4.2: An example of PhyloH output.
The output of python script is an interactive .html. This file is impossible to visualize in a standard page. It is possible to traverse and navigate the tree. At the base the colors ranging from yellow to red, indicating the contribution of a particular lineage in discriminating variable considered (where yellow zero contribution and red maximum contribution). The bars on the terminal nodes of the tree correspond to the relative abundance of corresponding OTU.The colors are representative of the groups examined.

Table 4.1: Counts across 3 data sets

	Cluster	Reads
All (AS)	295	34090
High (HFC)	21	23922
Low (HFC)	274	10168

The first row show the count from the original output of UCLUST in term of number of cluster and number of read. Following row show the data set composed of only high frequency cluster and low frequency cluster using 1% as threshold.

4.3.2 Microbial community analysis

Distance and Mixed Models approach

There are three environmental variables considered: (i) "cell": single bee and the corresponding parasite (found in the same brood cell). (ii) "localities": all bees and relative parasites present in a hive. (iii) "status": the differences between the pools of healthy honey bees, parasitized honey bees and the mites. The UPGMA analysis considering Jaccard distance shows a strong a unique cluster for all samples belonging to the parasitized honey bees and the mites, and a separate cluster for the pool of healthy honey bees. However, with chord metric , which considers the abundances information of the OTUs, the analysis shows two different distribution between the parasitized honey bees and the mites (figure tree). Using the same metrics it did not find any distinctive traits in the microbial composition between cells or hives, in relation to the sampling localities (Figures 4.3.

These findings agree with the results of the same test performed using unweighted and weighted Unifrac distances. Taking into account only the 21 OTUs that exceeded the threshold of 1% of the total distribution, they were found both in the parasitized honey bees and in the varroa mites with different values of relative abundance (See Table 4.1 and Figure 4.4).

The coefficients of the LMMs model showed that the distance be-

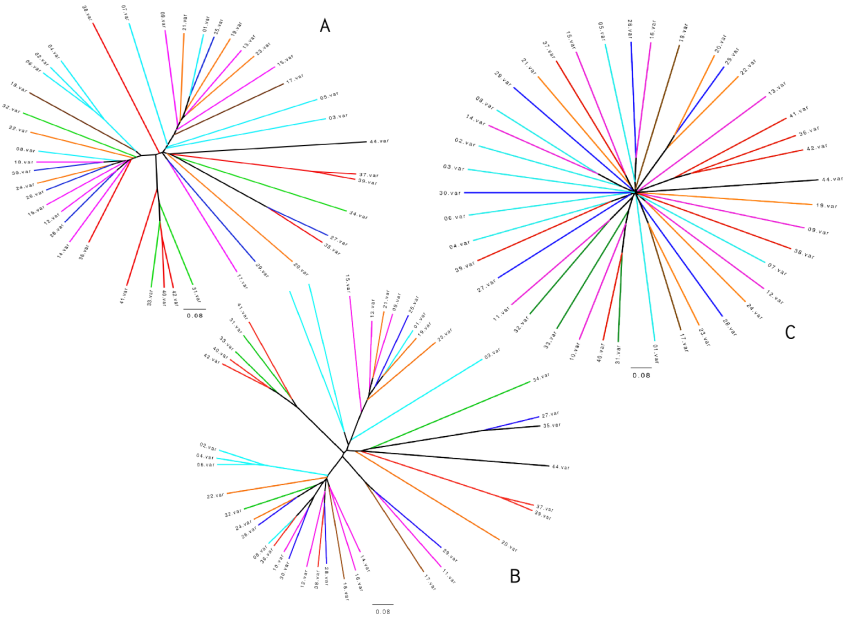


Figure 4.3: UPGMA trees of chord metric for localities.
Any colors indicates the menberships of a sample to one of the 8 locations considetate.
A) UPGMA tree of the AS B) UPGMA tree of the HFC c) UPGMA tree of the LFC

tween the honey bee and its varroa does not differ significantly from the distance between a honey bee and a varroa taken at random from another cell . The coefficients of the same model also indicate that the distance between the honey bees of a locality and the mites of the same locality does not differ significantly from the distance between the honey bees from another locality and the mites of a locality taken at random from the others (all coefficients are not significative ($p >$ 0.05). Based on these results the microbiomes characterizing the healthy honey bees, the parasitized honey bees and the mites were described. For a complete description of the distribution of the OTUs found associated with the three categories considered, refer to Figure 4.5.

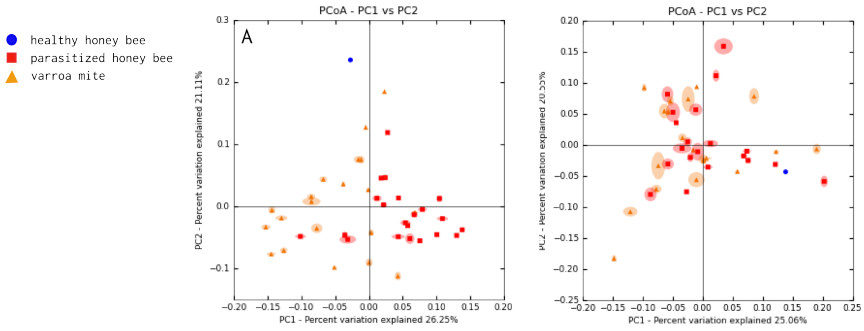


Figure 4.4: Jackknifing of PCoA analysis of host and parasite samples with weighted UniFrac.
Shown is a plot of the first two principal coordinate axes (factors) for PCoA with the resulting tree of FastTree tool. Point locations are the average location in the 100 jackknife replicates.Color ellipses represent the IQR for the 100 jackknife replicates. A) PCoA considering the 21 most represented OTU. B) PCoa considering 274 rare OTUs.

The core microbiome

Regarding the pool of healthy honey bees, only three OTUs were found and one of them (Proteo-7 a member of the genus Serratia) represents the 99% of presence in abundance. The remaining 1% is composed of OTUs Proteo-2 and Firmi-7, respectively identified as Achromobacter sp. and *Lactobacillus sp.* Analysing the curves (Figure 4.6 returned by QIIME script it was decided to take a 0.8 threshold to define the core microbiome of the host and the parassite, being the point from which both curves tends to shift down faster. This means that a single OTU must be present in the 80% of the samples of a species to be considered "core" for that species.

The honey bee larvae found in association with varroa show a set of bacteria more complex than the one found in the pool of healthy honey bees. All samples of parasitized honey bee are dominated by the Firmi-1 and Proteo-1 OTUs, respectively *Streptococcus sp* and *Hydrogenophilus sp*. An OTU (Firmi-3) belonging to Clostridiaceae family previously

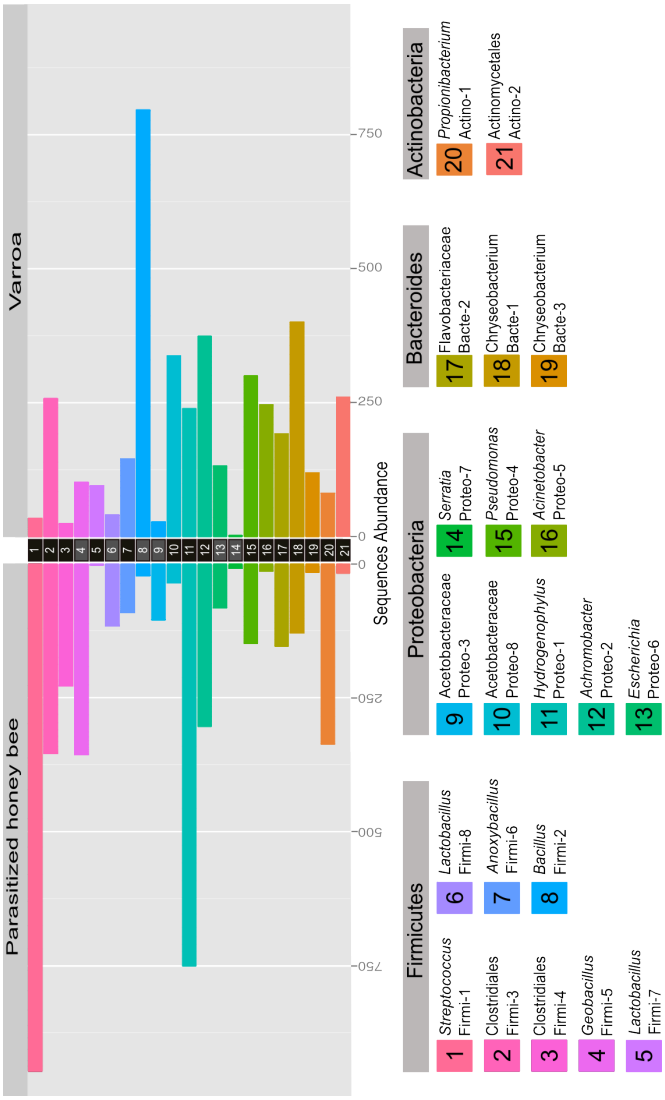


Figure 4.5: Distribution of the 21 most represented OTUs founds in the parisitez honey bee and in varroa mite.

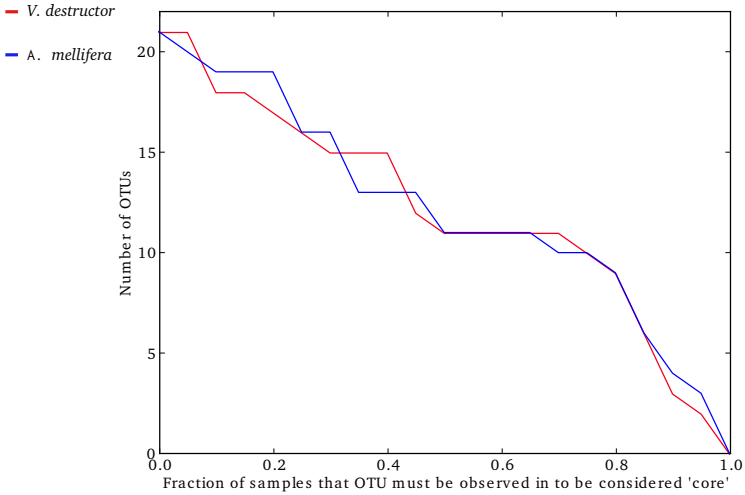


Figure 4.6: Core microbiomes inferred in parasitized *A.mellifera* and *V. destructor* Core microbiomes were inferred using QIIME script at a range of cutoffs (x-axis).

found in other study on honey bee microbiome was detected (Moran et al., 2008). This OTU with Firmi-5 and Actino-1 are evenly distributed between all samples. We also found four poorly represented OTUs, Proteo-6 of the *Escherichia* genus, Proteo-4 of the *Pseudomonas* genus, Proteo-2 of the *Acromobacter* genus and Firmi-2 of the *Bacillus* genus.

Considering the bacterial communities associated with *V. destructor*, it emerges that nine OTUs are present in at least 80% of the samples. As in the case of *A. mellifera* the phylotype most represented (Firmi-2) belong to the phylum of Firmicutes and in particular to the genus *Bacillus*. Most in general, the distribution of singles OTUs among all varroa samples is less homogeneous than in the honey bees (Figure 4.7).

A series of OTUs typical of different habitat such as soil or associated to flowers (Bacte-1, Bacte-2-Firmi-6-Proteo-1) were identified. Proteo-2

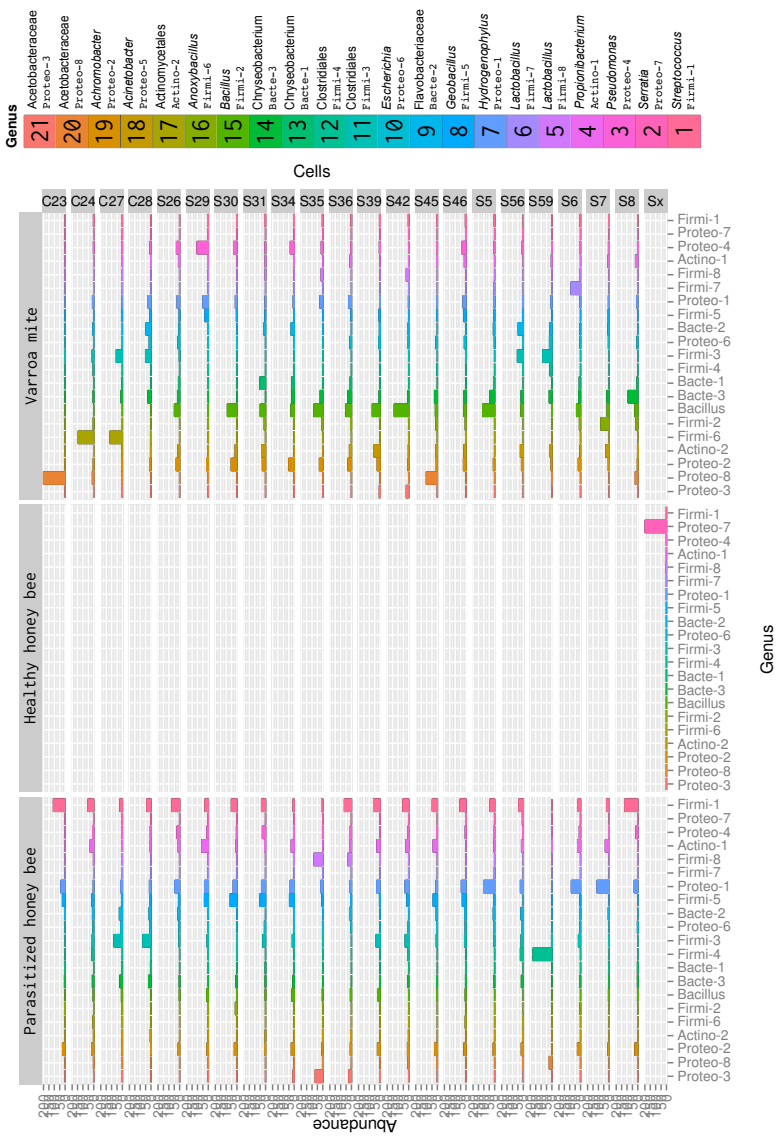


Figure 4.7: Cells taxon distribution.
Distribution of the 21 most represented OTUs founds in the three microbiome considered take in to account the membership of the host or of the parasite to a given cell. The last cell (Sx) indicates the pool of healthy honey bees.

(genus *Achromobacter*) was the second OTU most represented but is also the one highly shared with honey bee. As well as in the honeybee, an OTU belonging to the genera *Pseudomonas* (Proteo-4), *Escherichia* (Proteo-6) and *Streptococcus* (Firmi-1) was found. The complete descriptions of distribution of the OTUs belonging to the High frequency cluster are provided in (Figure 4.5).

Partitioning phylogenetic diversity

All 18 tested partitioning produce a significant mutual information between the sequence and each of the 3 types of environmental variables (host species, localities, cell in beehive) unresponsive of the type of data (all or low or high frequency cluster) and of the use of the phylogeny.

It is possible to observe several qualitative difference across the 18 analyses transforming the mutual information in percentage of overlap to allow a comparison across variables. As shown in Figure 4.8 the localities and the cell bee hive variables behave similarly unresponsive to the phylogenetic information: low frequency OTUs have similar low overlap across states of the variables, while high frequency OTUs are more similar across sample of the belonging to different groups.

It should also be noted that the fact that the percentage of overlap is always higher for the phylogenetic estimation is caused by the similarly acknowledged by the phylogeny of the different OTUs observed, while the classic Shannon-based approach assume each OTU totally different from each other. The variable host species produce a very different pattern from previous variables. Taking in account phylogeny high frequency OTUs differentiate across groups (overlap 87%), with the other two data sets showing about 91% overlapping. On the contrary, if we are not taking in account similarity across OTUs the pattern is similar to the other 2 variables with most difference observable in low frequency OTUs and the least difference observable in high frequency ones, although this pattern is expressed in a smaller range of frequency with all values around 60% of overlap. So the difference found in the high frequency OTUs seem to entail a stronger phylogenetic signal than the low frequency ones.

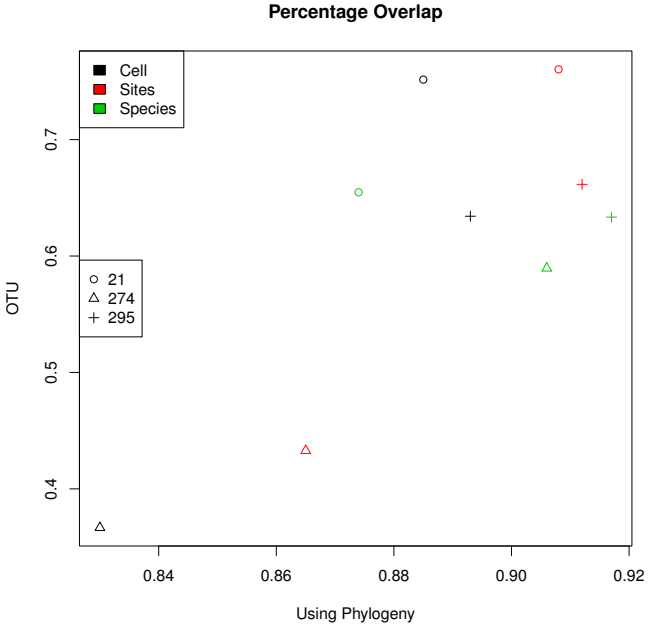


Figure 4.8: Overlap distribution of OTUs *versus* phylogeny information. see text for the details.

The contribution of the branch to the mutual information for the HFC with the host specie variable show that the 21 strains are well distributed on the tree, thus increasing their impact on the phylogenetic index, but it is possible to notice that Bacte-2, Bacte-3, and Bacte-1 are more typical of the varroa host and are all grouped in the same lineages (L220), similarly Proteo-5, Proteo-6, and Proteo-4 OTUs all belong to lineage L107 and are preferentially present in varroa. On the contrary Firmi-1 and Firmi-7, typical of honey bee are mixed with varroa's OTUs Firmi-6 and Firmi-5 (both descending from lineage L292). This lack of phylogenetic signal seems more caused by recent specialization given that Firmi-6 have a lower frequency sister taxa Firmi-43

present mainly in Varroa and Firmi-6 is similarly paired with low frequency OTU Firmi-14.

OTUn	Phylum	Class	Order	Family	Genus	Counts
Acido-1	Acidobacteria	Solibacteres	Solibacterales	Solibacteraceae	*Candidatus*	11
Acido-2	Acidobacteria	[Chloracidobacteria]	RB41	Ellin6075		6
Actino-1	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	*Propionibacterium*	1158
Actino-2	Actinobacteria	Actinobacteria	Actinomycetales			939
Actino-3	Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	*Pimelobacter*	261
Actino-4	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae		227
Actino-5	Actinobacteria	Actinobacteria	Actinomycetales			155
Actino-6	Actinobacteria	Actinobacteria	Actinomycetales			116
Actino-7	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	*Corynebacterium*	4
Actino-8	Actinobacteria	Actinobacteria	Actinomycetales			17
Actino-9	Actinobacteria	Actinobacteria	Actinomycetales			117
Actino-10	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	*Corynebacterium*	61
Actino-11	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae		65
Actino-12	Actinobacteria	Actinobacteria	Actinomycetales			62
Actino-13	Actinobacteria	Actinobacteria	Actinomycetales			46
Actino-14	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Rothia*	24
Actino-15	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Rothia*	38
Actino-16	Actinobacteria	Actinobacteria	Actinomycetales	Dietziaceae	*Dietzia*	40
Actino-17	Actinobacteria	Actinobacteria	Actinomycetales			25
Actino-18	Actinobacteria	Actinobacteria	Actinomycetales	Brevibacteriaceae	*Brevibacterium*	26
Actino-19	Actinobacteria	Actinobacteria	Actinomycetales			22
Actino-20	Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae		18
Actino-21	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	*Microbacterium*	12
Actino-22	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	*Corynebacterium*	16
Actino-23	Actinobacteria	Acidimicrobiia	Acidimicrobiales	C111		15
Actino-24	Actinobacteria	Actinobacteria	Actinomycetales			7
Actino-25	Actinobacteria	Actinobacteria	Actinomycetales			10
Actino-26	Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	***Marmoricola***	1
Actino-27	Actinobacteria	Actinobacteria	Actinomycetales			14
Actino-28	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae		11
Actino-29	Actinobacteria	Actinobacteria	Actinomycetales			1
Actino-30	Actinobacteria	Actinobacteria	Actinomycetales			7
Actino-31	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*Actinomyces*	6
Actino-32	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	*Corynebacterium*	6
Actino-33	Actinobacteria	Acidimicrobiia	Acidimicrobiales	C111		6
Actino-34	Actinobacteria	Actinobacteria	Actinomycetales			53
Actino-35	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*N09*	6
Actino-36	Actinobacteria	Actinobacteria	Actinomycetales			8
Actino-37	Actinobacteria	Rubrobacteria	Rubrobacterales	Rubrobacteraceae	*Rubrobacter*	5
Actino-38	Actinobacteria	Actinobacteria				9
Actino-39	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	*Propionibacterium*	11
Actino-40	Actinobacteria	Actinobacteria	Actinomycetales			42
Actino-41	Actinobacteria	Actinobacteria	Actinomycetales			32
Bacte-1	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]	*Chryseobacterium*	1679
Bacte-2	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae		962
Bacte-3	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]	*Chryseobacterium*	410
Bacte-4	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]	*Cloacibacterium*	191

Bacte-5	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	*Prevotella*	140
Bacte-6	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	133
Bacte-7	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	1
Bacte-8	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	113
Bacte-9	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Hymenobacter*	2
Bacte-10	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Pedobacter*	6
Bacte-11	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]		43
Bacte-12	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae		32
Bacte-13	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Sphingobacterium*	32
Bacte-14	Bacteroidetes	[Saprospirae]	[Saprospirales]	Chitinophagaceae		34
Bacte-15	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]	*Chryseobacterium*	20
Bacte-16	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	2
Bacte-17	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	23
Bacte-18	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Spirosoma*	1
Bacte-19	Bacteroidetes					21
Bacte-20	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Sphingobacterium*	21
Bacte-21	Bacteroidetes	[Saprospirae]	[Saprospirales]	Chitinophagaceae	*Chitinophaga*	18
Bacte-22	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Pedobacter*	16
Bacte-23	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	13
Bacte-24	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae		13
Bacte-25	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]	*Chryseobacterium*	12
Bacte-26	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae		8
Bacte-27	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	*Parabacteroides*	7
Bacte-28	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]		8
Bacte-29	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae		6
Bacte-30	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae		7
Bacte-31	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Pedobacter*	1
Bacte-32	Bacteroidetes	Flavobacteriia	Flavobacteriales	[Weeksellaceae]	*Cloacibacterium*	33
Bacte-33	Bacteroidetes	Sphingobacteriia	Sphingobacteriales			5
Chloro-1	Chlorobi	BSV26	PK329			6
Cyano-1	Cyanobacteria	Chloroplast	Streptophyta			23
Firmi-1	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Streptococcus*	2285
Firmi-2	Firmicutes	Bacilli	Bacillales	***Bacillaceae***	***Bacillus***	2545
Firmi-3	Firmicutes	Clostridia	Clostridiales			1912
Firmi-4	Firmicutes	Clostridia	Clostridiales			992
Firmi-5	Firmicutes	Bacilli	Bacillales	***Bacillaceae***	***Geobacillus***	1068
Firmi-6	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Anoxybacillus*	570
Firmi-7	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	673
Firmi-8	Firmicutes	Bacilli	Lactobacillales	***Lactobacillaceae***	***Lactobacillus***	536
Firmi-9	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Streptococcus*	257
Firmi-10	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae		216
Firmi-11	Firmicutes	Bacilli				126
Firmi-12	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae		156
Firmi-13	Firmicutes	Bacilli	Gemellales	Gemellaceae		156
Firmi-14	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	*SMB53*	117
Firmi-15	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	114
Firmi-16	Firmicutes	Clostridia	Thermoanaerobacterales	Caldicellulosiruptoraceae	*Caldicellulosiruptor*	98
Firmi-17	Firmicutes	Bacilli	Gemellales			90
Firmi-18	Firmicutes	Bacilli	Bacillales	Bacillaceae		92

Firmi-19	Firmicutes	Clostridia	Clostridiales	Clostridiaceae		83
Firmi-20	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Streptococcus*	77
Firmi-21	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae		76
Firmi-22	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	*Enterococcus*	14
Firmi-23	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	56
Firmi-24	Firmicutes	Bacilli	Bacillales	Paenibacillaceae		56
Firmi-25	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	35
Firmi-26	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	*Blautia*	48
Firmi-27	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	*Thermoanaerobacterium*	43
Firmi-28	Firmicutes	Bacilli				32
Firmi-29	Firmicutes	Bacilli	Bacillales	Bacillaceae		42
Firmi-30	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae		35
Firmi-31	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Geobacillus*	25
Firmi-32	Firmicutes	Clostridia	Thermoanaerobacterales	Thermoanaerobacterales	*Thermovenabulum*	19
Firmi-33	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	*Finegoldia*	27
Firmi-34	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	29
Firmi-35	Firmicutes	Bacilli	Turicibacterales	Turicibacteraceae	*Turicibacter*	26
Firmi-36	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	25
Firmi-37	Firmicutes	Bacilli	Bacillales			9
Firmi-38	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	*Fructobacillus*	20
Firmi-39	Firmicutes	Bacilli				19
Firmi-40	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	*Peptostreptococcus*	17
Firmi-41	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	*Anaerococcus*	15
Firmi-42	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Lactococcus*	15
Firmi-43	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	8
Firmi-44	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Anoxybacillus*	172
Firmi-45	Firmicutes	Bacilli	Bacillales			15
Firmi-46	Firmicutes	Bacilli	Bacillales	Listeriaceae	*Brochothrix*	13
Firmi-47	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	*Thermoanaerobacterium*	3
Firmi-48	Firmicutes	Bacilli	Bacillales	Bacillaceae		13
Firmi-49	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Geobacillus*	18
Firmi-50	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Streptococcus*	9
Firmi-51	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	*Granulicatella*	11
Firmi-52	Firmicutes	Clostridia	Clostridiales			10
Firmi-53	Firmicutes	Bacilli	Bacillales	Listeriaceae		13
Firmi-54	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	*Granulicatella*	9
Firmi-55	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	*Brevibacillus*	9
Firmi-56	Firmicutes	Bacilli				58
Firmi-57	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	13
Firmi-58	Firmicutes	Bacilli	Bacillales			7
Firmi-59	Firmicutes	Bacilli	Bacillales			3
Firmi-60	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae		6
Firmi-61	Firmicutes	Clostridia				6
Firmi-62	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	6
Firmi-63	Firmicutes	Bacilli	Bacillales			6
Firmi-64	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae		18
Firmi-65	Firmicutes	Bacilli	Bacillales			10
Firmi-66	Firmicutes	Bacilli	Bacillales			15
Firmi-67	Firmicutes	Bacilli	Bacillales	[Exiguobacteraceae]	*Exiguobacterium*	5

Firmi-68	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	*Anaerococcus*	6
Firmi-69	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	4
Firmi-70	Firmicutes	Clostridia	Thermoanaerobacterales	Thermoanaerobacterales	*Thermovenabulum*	16
Firmi-71	Firmicutes	Bacilli	Bacillales	Paenibacillaceae		4
Frimi-72	Firmicutes					5
Proteo-1	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	*Hydrogenophilus*	2349
Proteo-2	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Achromobacter*	1813
Proteo-3	Proteobacteria	Alphaproteobacteria	***Rhodospirillales***	***Acetobacteraceae***		1225
Proteo-4	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	*Pseudomonas*	1128
Proteo-5	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	680
Proteo-6	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	***Escherichia***	573
Proteo-7	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	***Serratia***	508
Proteo-8	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae		407
Proteo-9	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		303
Proteo-10	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	*Pseudomonas*	221
Proteo-11	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Gluconacetobacter*	236
Proteo-12	Proteobacteria	Gammaproteobacteria	Pasteurellales			239
Proteo-13	Proteobacteria	Alphaproteobacteria	Rhizobiales			228
Proteo-14	Proteobacteria	Gammaproteobacteria	Pasteurellales			223
Proteo-15	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae		195
Proteo-16	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae		81
Proteo-17	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Haemophilus*	153
Proteo-18	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		171
Proteo-19	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae		151
Proteo-20	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Stenotrophomonas*	130
Proteo-21	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		71
Proteo-22	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	*Burkholderia*	134
Proteo-23	Proteobacteria	Gammaproteobacteria	Pasteurellales			116
Proteo-24	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Haemophilus*	102
Proteo-25	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Paracoccus*	39
Proteo-26	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		88
Proteo-27	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	*Shewanella*	95
Proteo-28	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		2
Proteo-29	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		16
Proteo-30	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	*Agrobacterium*	72
Proteo-31	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	42
Proteo-32	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	58
Proteo-33	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae		40
Proteo-34	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Delftia*	57
Proteo-35	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		5
Proteo-36	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae		41
Proteo-37	Proteobacteria	Alphaproteobacteria				58
Proteo-38	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae		59
Proteo-39	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		3
Proteo-40	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	***Hyphomicrobium***	47
Proteo-41	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Luteibacter*	38
Proteo-42	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Janthinobacterium*	26
Proteo-43	Proteobacteria	Gammaproteobacteria				43
Proteo-44	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae		42

Proteo-45	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	*Shewanella*	33
Proteo-46	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Haemophilus*	36
Proteo-47	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	*Wolbachia*	36
Proteo-48	Proteobacteria	Alphaproteobacteria	Rhizobiales			19
Proteo-49	Proteobacteria	Alphaproteobacteria	Sphingomonadales			27
Proteo-50	Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae		35
Proteo-51	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Janthinobacterium*	1
Proteo-52	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		32
Proteo-53	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae		33
Proteo-54	Proteobacteria	Alphaproteobacteria	Rhizobiales			30
Proteo-55	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Enhydrobacter*	26
Proteo-56	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Polynucleobacter*	29
Proteo-57	Proteobacteria	Gammaproteobacteria				11
Proteo-58	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae		30
Proteo-59	Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae		31
Proteo-60	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Sphingomonas*	46
Proteo-61	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae		26
Proteo-62	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae		11
Proteo-63	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		3
Proteo-64	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	*Mesorhizobium*	23
Proteo-65	Proteobacteria	Deltaproteobacteria				24
Proteo-66	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae		24
Proteo-67	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae		22
Proteo-68	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	*Neisseria*	21
Proteo-69	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Haemophilus*	21
Proteo-70	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae		21
Proteo-71	Proteobacteria	Alphaproteobacteria	Rhizobiales	Aurantimonadaceae		11
Proteo-72	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae		21
Proteo-73	Proteobacteria	Alphaproteobacteria	Rhizobiales			2
Proteo-74	Proteobacteria	Alphaproteobacteria	Rhizobiales			8
Proteo-75	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Dokdonella*	18
Proteo-76	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Methylibium*	18
Proteo-77	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Methylibium*	1
Proteo-78	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae		16
Proteo-79	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	*Cellvibrio*	16
Proteo-80	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	*Arcobacter*	16
Proteo-81	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Trabulsiella*	16
Proteo-82	Proteobacteria	Betaproteobacteria				16
Proteo-83	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	13
Proteo-84	Proteobacteria	Gammaproteobacteria				14
Proteo-85	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae		14
Proteo-86	Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	*Methylotenera*	14
Proteo-87	Proteobacteria	Alphaproteobacteria	Rhizobiales			2
Proteo-88	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae		16
Proteo-89	Proteobacteria	Gammaproteobacteria				1
Proteo-90	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae		12
Proteo-91	Proteobacteria	Gammaproteobacteria				12
Proteo-92	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	***Mangroveibacter***	11
Proteo-93	Proteobacteria	Alphaproteobacteria	Rhizobiales			13

Proteo-94	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Janthinobacterium*	12
Proteo-95	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		1
Proteo-96	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae		10
Proteo-97	Proteobacteria	Alphaproteobacteria				10
Proteo-98	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Simplicispira*	12
Proteo-99	Proteobacteria	Betaproteobacteria	Procabacteriales	Procabacteriaceae		9
Proteo-100	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		10
Proteo-101	Proteobacteria	Betaproteobacteria	Burkholderiales			8
Proteo-102	Proteobacteria	Gammaproteobacteria				8
Proteo-103	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	*Marinomonas*	8
Proteo-104	Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae		8
Proteo-105	Proteobacteria	Alphaproteobacteria				8
Proteo-106	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Rhodobacter*	7
Proteo-107	Proteobacteria	Betaproteobacteria				6
Proteo-108	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Achromobacter*	6
Proteo-109	Proteobacteria					5
Proteo-110	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		5
Proteo-111	Proteobacteria	Alphaproteobacteria				6
Proteo-112	Proteobacteria	Betaproteobacteria				5
Proteo-113	Proteobacteria	Gammaproteobacteria	Legionellales	Coxiellaceae	*Aquicella*	5
Proteo-114	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Achromobacter*	45
Proteo-115	Proteobacteria					32
Proteo-116	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	*C39*	6
Proteo-117	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		5
Proteo-118	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae		5
Proteo-119	Proteobacteria	Betaproteobacteria	Procabacteriales	Procabacteriaceae		5
Proteo-120	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		5
Proteo-121	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Pelomonas*	5
Proteo-122	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		4
Proteo-123	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	4
Proteo-124	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		4
Proteo-125	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Rhodanobacter*	6
Proteo-126	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	*Neisseria*	4
Proteo-127	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae		4
Proteo-128	Proteobacteria					8
Proteo-129	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Haemophilus*	5
Proteo-130	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae		1
Proteo-131	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae		9
Proteo-132	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Ralstonia*	4
Proteo-133	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae		4
Proteo-134	Proteobacteria	Gammaproteobacteria	Alteromonadales	[Chromatiaceae]	*Rheinheimera*	4
Proteo-135	Proteobacteria	Deltaproteobacteria	Spirobacillales			4
Proteo-136	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfohalobiaceae		4
TM7-1	TM7	TM7				23
TM7-2	TM7	SC3				15
TM7-3	TM7	SC3				8
TM7-4	TM7	TM7				7
TM7-5	TM7	TM7				5
Unw-1						67

Unw-2	13
Unw-3	13
Unw-4	4

Table 4.2: Taxonomy assignment of OTUs found in all samples

The table show the taxonomy assignment for OTUs found in all samples considered. First column are the acronym attributed to each OTUs. In the last column are indicated the number of sequences attributed to each OTUs. OTUs that RDP assigned with a probability between 0.8 and 0.9 were sought in NCBI nucleotide database. When a perfect match with BLAST was found the NCBI taxonomy was used. The BLAST match are indicated in bold

4.4 Discussions

Both approaches show that the differences observed in the bacterial communities can not be explained grouping the samples according to the brood cells or to the different sampling localities. According to this result, other studies (Sabree et al., 2012) have shown that geographical distance is not crucial in shaping the bacterial community in beehives. It can be assumed that distance influences more likely the distribution of rare OTUs. However, in this case it was not possible to identify unambiguously the information resulting from the two sub-categories of sampling.

If the species status is considered variable, the results change slightly. Both distance matrix and entropy methods can discriminate the three different categories, but in the case of QIIME's script all the information carried by LFC disappears. This because the script considers only the abundance distribution as a source of information. In addition, the OTUs found in the categories parasitized honey bees and varroa parasites have a high rate of overlap, which makes difficult the interpretation of single OTU role in the discrimination of the categories. Considering the results obtained, the comments on the biology of the OTUs are giving are broadly as possible, taking account of both approaches.

A. mellifera bacterial communities

When considering the microbiome of healthy larvae the results are in agreement with previous studies on honey bee larval microbiome (Martinson et al., 2011).

The variability of bacteria communities is extremely poor and it is completely different to that of a parasitized larvae. The most representative phylotype found (Proteo-7), belongs to the genus *Serratia*. This genus was isolated from the intestinal contents of healthy foraging worker honeybees (Jeyaprakash et al., 2003)

and is found as a symbiont in other insect taxa (Dillon and Dillon, 2004). The scarcity of bacteria in the larvae could be attributed both to

the particular morphology and physiology of honey bee larval gut and to the type of nutrition (as noted by Martinson data Martinson et al. (2012)). In addition, as suggested by the same author, the variation in the bacterial community of honey bee strains may explain why the presence of different bacteria in honey bee larvae is variable.

The different microbial communities characterizing the parasitized honey bees and the mite are identical from a qualitative point of view, but they differ when considering the abundances. This result, together with the fact that in the healthy bee only a few bacterial OTU were found, endorse the hypothesis that the attack on larva by the varroa mitereates a gate for bacteria colonization. The homogenization at the qualitative level in the HFC suggests an effect of dispersion by the most representative species already present at the level of varroa and at the hive. The differences in abundance, could be explained by a different ability and attitude of the bacterial species to colonize different environments (Hanson et al., 2012). The two most represented phyla in parasitized honey bees are Proteobacteria and Firmicutes. Particularly, among Firmicutes *Streptococcus sp* (Firmi-1) and two different OTUs of Clostridiales (Firmi-3,-4) both classified as bacteria than could also cause systemic infections were found.

A recent research by (Lozupone et al., 2012)showed the aptitude of batcteria belonging to the class Clostridiales to act as pioneering species, especially in cases of habitat in disequilibrium after traumatic events (e.g. diseases). This may in part explain the significant difference in abundance of these species compared to the others. In particular, OTU Firmi-4 was also found to be one of most represented in a recent research on the lepidoptera larvae (Tang et al., 2012) where the subjects of the study were intoxicated. The authors also noticed that when a disturb affects the microbiota of the lepidoptera larvae the rate of diversity increases as showed in our case.

A OTUs of *Lactobacillus* (Firmi-7), a genus of potential probiotic bacteria, was found as in many studies on microbiome of honey bees (Mohr and Tebbe, 2006; Evans and Schwarz, 2011; Moran et al., 2012). This Gram-positive organism has a primarily protective function (Mohr and Tebbe, 2006; Moran et al., 2012). It is interesting to note how its pres-

ence leads to a nearly total absence of the pathogenic genera (Figure 4.7).

Hydrogenophylus (Proteo-1), *Geobacillus* (Firmi-5), *Anoxybacillus* (Firmi-6) and *Propionibacterium* (Actino-1) are bacteria found in a wide range of environments. Firmi-5 in particular is a thermophilic bacterium that has been found associated with plants and in in hot environments such as thermal water (Dulger et al., 2004). Therefore, it is interesting to note that the highest presence of this bacterium was found in honey bees sampled in a period ranging from June to July. The phylotype Proteo-3 was previously found in studies on microbiome of insects (Prince et al., 2009; Crotti et al., 2010) as typical of this class. None of the bacterial species previously identified and held responsible for major diseases described affecting honey bees has been identified.

V. destructor bacterial communities

In the literature there is only a brief description of the bacteria that colonizing the varroa mite (Cornman et al., 2010). Unfortunately, it was not possible to use the 16S rDNA sequences for a direct comparison with the present data due to the fact that the sequences were not deposited individually in Gene bank, as soon as they are part of a larger metagenomic project. Two OTUs (Firmi-3, -7) of the phylum Firmicutes appear to have been found in previous studies on honey bee microbiome. The hypothesis is that this fact is due to the external presence of this two strains in the hive. As shown in Cornman (Cornman et al., 2010), the Actomycetales order is well represented in varroa, but in our case this OTU is not the most abundant. The role of Actinomycetales in this parasite should be examined in more in-depth studies. Instead, there is a significant presence of two OTUs belonging to the genus *Chryseobacterium* (Bacte-1,-3). Contrary to the majority of bacteria belonging to Flavobateria, typically found in soil and water environment, these two particular OTUs were found as pathogens of soft ticks (Buresová et al., 2006). Given the phylogenetic closeness between mites and ticks, it would be interesting to investigate the role of this bacterium in the mites.

Finally, a difference in percentage in the presence of *Pseudomonas* between honey bee and varroa was found. This is possible due to the fact that the bacteria belonging to genus *Pseudomonas* colonize the cuticle completely (Tang et al., 2012). Therefore it is possible to assume that during the experimental stage, the removal of the cuticle was less effective in the case of varroa, specially due to the small size of the organisms.

In this case study the use of an approach based on phylogenetic analysis allowed to highlight which OTU are more discriminated within the variable species. The definition of the core microbiome remains to this day one of the steps of analysis lacking from statistical point of view, based almost solely on subjective considerations. This new approach has proved to be a valuable tool in this type of analysis.

5 General discussion

The routinely use of new sequencing technologies reached in the recent time has made essential to develop new methods of analysis to fully exploit the data. This is related to nature itself of the new platforms and to the advancement and improvement of old approaches to the investigation of biodiversity.

The design and the development of pipeline (i.e. QIIME) that allows a more standardized approach to data analysis perfectly matches to the idea of standardization proposed by DNA barcoding. However, in the case of studies of metabarcoding the difficulties are amplified. Available data is exponentially grown and the need to greater integration between different scientific disciplines require a stronger experimental structuring, in addition to a greater effort of calculation.

The main aim of my thesis was to analyze the problems and the techniques already available and to seek new methods of analysis. In the first of the cases analyzed, the core of the problem was the construction of a pipeline that could in some way to cope with problems that may be encountered in large-scale studies of metazoans. Among the problems encountered during the process of analysis it is clear that the presence of a bias at the level of sampling could be reflected hamper all the following steps.

In constructing taxonomic profiles through species assignment, denoising pipeline proved to be a promising method in estimating the true sequence diversity produced by 454 pyrosequencing and discarding unreal sequences. This method significantly eliminated pyrosequencing and PCR polymerase errors. However, in each sequencing run it is strongly recommended the use of empirical internal control sample with known sequences to construct error distributions and consequently estimate error rate.

Contrary to the metazoan, an approach on a large scale for the study of bacteria is in use for the longest time (Caporaso et al., 2010). This fact has led to a pipeline that today are widely recognized (QIIME). This has allowed to obtain a fair degree of standardization as regards the steps involved in the process of the taxonomy assignment. What still remains a subject of debate is the analysis of the structure of these communities. This is mainly due to the lack of knowledge of ecological dynamics at the base of complex bacterial systems. I am convinced that the time for a significant contribution of microbial ecology to general ecology has arrived. However, it is becoming clear that microbial taxa display non-random environmental and geographical distribution (see recent reviews in Martiny et al. (2006); Ramette and Tiedje (2007); Fierer et al. (2008); Hanson et al. (2012)). Most of the patterns that support this evidence have been studied for decades in plant and animal communities but have largely been ignored by microbiologists. Over the past decade, microbial ecologists generated abundant molecular data from ribosomal surveys, and now is possible to combine bioinformatical and statistical tools with critical testing of ecological theory in order to integrate microorganisms into the broader field of ecology.

From these conditions came the need to explore as much as possible the means available from ecology and phylogenesis to investigate possible patterns of communities' distribution. In the case of symbiosis between *A. mellifera* and *V. destructor* (Section 3) shows how real is this necessity and how there is still a wide margin of exploration in this direction. The field of metabarcoding is still constantly changing. Being associated with a new technology is constantly evolving and it directly follows the trends. Since this project began to date, the chemical products of the machines currently available have changed, with significant effects in some fields. As regards the approach to the analysis of metazoans the reference technology remained the Roche 454 that with new developments in chemical sequencing succeeded in raising the length of the reads up to 800 base pairs. Unfortunately, the problems related to the bias of PCR and to the rate of presence of homopolymers and chimeras seems not to be significantly changed.

However, the proposal and the debate around new pipeline analysis

like the one proposed in my thesis are contributing effectively to the improvement in the process of data analysis and to increase in capacity on the detection errors of sequencing. As for the metabarcoding approach in the study of microbial communities the situation has changed. Over the past two years, the technology that is seen to establish for this type of studies is the Illumina HiSeq 2000 that produces more than 300 million reads. The length of the reads remains below that of Roche and the error rate higher, but in front of a production rate of the order of 100 times greater. This fact in a survey of the community of millions of individuals is proving to be the right compromise in order to obtain valid data even for ecological studies.

In general, therefore it can be concluded that environmental DNA metabarcoding has an enormous potential to boost data acquisition in biodiversity research. At the moment, we are living at the beginning of this approach that can to revolutionize the way to know and investigate the biodiversity around us.

Bibliography

Abdo, Z. and Golding, G. B. (2007). A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic biology*, 56(1):44–56.

Ackerly, D. (2009). Phylogenetic Methods in Ecology. In *Encyclopedia of life sciences*. John Wiley & Sons, Ltd.

Allen, B., Kon, M., and Bar-Yam, Y. (2009). A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *The American Naturalist*, 174(2):236–243.

Altschul, S., Gish, W., and Miller, W. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215:403–410.

Annoscia, D., Del Piccolo, F., and Nazzi, F. (2012). How does the mite Varroa destructor kill the honeybee Apis mellifera? Alteration of cuticular hydrcarbons and water loss in infested honeybees. *Journal of insect physiology*, 58(12):1548–55.

Auguet, J.-C., Triadó-Margarit, X., Nomokonova, N., Camarero, L., and Casamayor, E. O. (2012). Vertical segregation and phylogenetic characterization of ammonia-oxidizing Archaea in a deep oligotrophic lake. *The ISME journal*, 6(9):1786–97.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., and Laredo, C. (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC bioinformatics*, 10 Suppl 1:S10.

Balzer, S., Malde, K., and Jonassen, I. (2011). Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics (Oxford, England)*, 27(13):i304–9.

Barberán, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal*, 6(2):343–51.

Bates, D. and Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. *R package version 0.999375-32.*

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G. N., Ribera, I., Nilsson, A. N., Barraclough, T. G., and Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic biology*, 61(5):851–69.

Bertolazzi, P., Felici, G., and Weitschek, E. (2009). Learning to classify species with barcodes. *BMC bioinformatics*, 10 Suppl 1:S7.

Bidartondo, M. I., Bruns, T. D., Blackwell, M., Edwards, I., Taylor, A. F. S., Horton, T., Zhang, N., Kõljalg, U., May, G., Kuyper, T. W., and Others (2008). Preserving accuracy in GenBank. *Science*, 319(5870).

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in ecology & evolution*, 27(4):233–43.

Binladen, J., Gilbert, M., and Bollback, J. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*.

Bittner, L., Halary, S., Payri, C., Cruaud, C., de Reviers, B., Lopez, P., and Bapteste, E. (2010). Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biology direct*, 5(1):47.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943.

Blow, N. (2008). Metagenomics: exploring unseen communities. *Nature*, 453(7195):687–90.

Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P., and Tyson, G. W. (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nature methods*, 9(5):425–6.

Brown, S. D. J., Collins, R. a., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C. J., and Cruickshank, R. H. (2012). Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular ecology resources*, pages 1–4.

Brulc, J. M., Antonopoulos, D. A., Miller, M. E. B., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., Wacklin, P., Coutinho, P. M., Henrissat, B., Nelson, K. E., and White, B. A. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(6):1948–53.

Bruni, I., De Mattia, F., Galimberti, A., Galasso, G., Banfi, E., Casiraghi, M., and Labra, M. (2010). Identification of poisonous plants by DNA barcoding approach. *International journal of legal medicine*, 124(6):595–603.

Brussaard, L. (1997). Biodiversity and ecosystem functioning in soil. *Ambio*, pages 563–570.

Buee, M., Reich, M., Murat, C., Morin, E., Nilsson, R. H., Uroz, S., and Martin, F. (2009). 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, 184(2):449–456.

Buresová, V., Franta, Z., and Kopácek, P. (2006). A comparison of Chryseobacterium indologenes pathogenicity to the soft tick Ornithodoros moubata and hard tick Ixodes ricinus. *Journal of invertebrate pathology*, 93(2):96–104.

Burgess, K. S., Fazekas, A. J., Kesanakurti, P. R., Graham, S. W., Husband, B. C., Newmaster, S. G., Percy, D. M., Hajibabaei, M., and Barrett, S. C. H. (2011). Discriminating plant species in a local temperate flora using the rbcL+ matK DNA barcode. *Methods in Ecology and Evolution*, 2(4):333–340.

Cadotte, M. W., Cardinale, B. J., and Oakley, T. H. (2008). Evolutionary history and the effect of biodiversity on plant productivity. *Proceedings of the National Academy of Sciences*, 105(44):17012–17017.

Candela, M., Biagi, E., Maccaferri, S., Turroni, S., and Brigidi, P. (2012). Intestinal microbiota is a plastic factor responding to environmental changes. *Trends in microbiology*, 20(8):385–391.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., Mcdonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Publishing Group*, 7(5):335–336.

Cardoso, a., Serrano, a., and Vogler, a. P. (2009). Morphological and molecular variation in tiger beetles of the Cicindela hybrida complex: is an 'integrative taxonomy' possible? *Molecular ecology*, 18(4):648–64.

Casiraghi, M., Labra, M., Ferri, E., Galimberti, A., and De Mattia, F. (2010). DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in bioinformatics*, 11(4):440–53.

Cavalli-Sforza, L. L. and Edwards, A. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution*, 21:550–570.

Chao, A., Chiu, C.-H., and Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1558):3599–609.

Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, 10(4):429–443.

Chase, M. W., Cowan, R. S., Hollingsworth, P. M., Van Den Berg, C., Madriñán, S., Petersen, G., Seberg, O., Jø rgsensen, T., Cameron, K. M., Carine, M., and Others (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon*, pages 295–299.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Others (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research*, 37(suppl 1):D141—-D145.

Cornman, S. R., Schatz, M. C., Johnston, S. J., Chen, Y.-P., Pettis, J., Hunt, G., Bourgeois, L., Elsik, C., Anderson, D., Grozinger, C. M., and Evans, J. D. (2010). Genomic survey of the ectoparasitic mite Varroa destructor, a major pest of the honey bee Apis mellifera. *BMC genomics*, 11(1):602.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science (New York, N.Y.)*, 326(5960):1694–7.

Creer, S., Fonseca, V. G., Porazinska, D. L., GIBLIN-DAVIS, R. M., Sung, W., Power, D. M., Packer, M., Carvalho, G. R., Blaxter, M. L., Lambshead, P. J. D., and Others (2010). Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, 19(s1):4–20.

Crotti, E., Rizzi, A., Chouaia, B., Ricci, I., Favia, G., Alma, A., Sacchi, L., Bourtzis, K., Mandrioli, M., Cherif, A., Bandi, C., and Daffonchio, D. (2010). Acetic acid bacteria, newly emerging symbionts of insects. *Applied and environmental microbiology*, 76(21):6963–70.

Curtis, T. P., Sloan, W. T., and Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10494–9.

DasGupta, B. and Konwar, K. (2005). DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*, 21.

Desalle, R. (2006). Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conservation biology : the journal of the Society for Conservation Biology*, 20(5):1545–7.

DeSalle, R., Egan, M. G., and Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1905–1916.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, 72(7):5069–5072.

Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS biology*, 6(11):e280.

Deutschbauer, A. M., Chivian, D., and Arkin, A. P. (2006). Genomics for environmental microbiology. *Current opinion in biotechnology*, 17(3):229–235.

Dillon, V. M. and Dillon, R. (2004). The gut bacteria of insects: nonpathogenic interactions. *Annual Review of Entomology*, 49(98):71–92.

Dimitriu, P. a., Boyce, G., Samarakoon, A., Hartmann, M., Johnson, P., and Mohn, W. W. (2013). Temporal stability of the mouse gut microbiota in relation to innate and adaptive immunity. *Environmental Microbiology Reports*, 5(2):200–210.

Dulger, S., Demirbag, Z., and Belduz, A. O. (2004). Anoxybacillus ayderensis sp. nov. and Anoxybacillus kestanbolensis sp. nov. *Interna-*

tional journal of systematic and evolutionary microbiology, 54(Pt 5):1499–503.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–7.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, 26(19):2460–1.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–200.

Eldredge, N. and Cracraft, J. (1980). *Phylogenetic patterns and the evolutionary process. Method and theory in comparative biology*. New York.: Columbia Univ. Press.

Elias, M., Hill, R. I., Willmott, K. R., Dasmahapatra, K. K., Brower, A. V. Z., Mallet, J., and Jiggins, C. D. (2007). Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings. Biological sciences / The Royal Society*, 274(1627):2881–9.

Evans, G. E., Murdoch, D. R., Anderson, T. P., Potter, H. C., George, P. M., and Chambers, S. T. (2003). Contamination of Qiagen DNA Extraction Kits with Legionella DNA. *Journal of Clinical Microbiology*, 41(7):3452–3453.

Evans, J. D. and Schwarz, R. S. (2011). Bees brought to their knees: microbes affecting honey bee health. *Trends in microbiology*, 19(12):614–20.

Faith, D. P. (2002). Quantifying Biodiversity: a Phylogenetic Perspective. *Conservation Biology*, 16(1):248–252.

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic acids research*, 40(Database issue):D136–43.

Felsenstein, J. (1985). Phylogenies and thecomparative method.

Ferguson, J. (2002). On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*.

Ferri, E., Barbuto, M., Bain, O., Galimberti, A., Uni, S., Guerrero, R., Ferté, H., Bandi, C., Martin, C., and Casiraghi, M. (2009). Integrated taxonomy: traditional approach and DNA barcoding for the identification of filarioid worms and related parasites (Nematoda). *Frontiers in zoology*, 6:1.

Fierer, N., Liu, Z., Rodríguez-Hernández, M., Knight, R., Henn, M., and Hernandez, M. T. (2008). Short-term temporal variability in airborne bacterial and fungal populations. *Applied and environmental microbiology*, 74(1):200–7.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–37.

Floyd, R., Abebe, E., Papert, A., and Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology*, 11(4):839–850.

Folmer, O., Hoeh, W. R., Black, M. B., and Vrijenhoek, R. C. (1994). Conserved primers for PCR amplification of mitochondrial DNA from different invertebrate phyla. *Molecular Marine Biology and Biotechnology*, 3:294–299.

Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., Packer, M., Blaxter, M. L., Lambshead, P. J. D., Thomas, W. K., Others, and Creer, S. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature communications*, 1:98.

Forsgren, E. (2010). European foulbrood in honey bees. *Journal of invertebrate pathology*, 103 Suppl(2010):S5–9.

Galimberti, A., Spada, M., Russo, D., Mucedda, M., Agnelli, P., Crottini, A., Ferri, E., Martinoli, A., and Casiraghi, M. (2012). Integrated Operational Taxonomic Units (IOTUs) in Echolocating Bats: A Bridge between Molecular and Traditional Taxonomy. *PloS one*, 7(6):e40122.

Gaston, K. J. and Spicer, J. I. (2009). *Biodiversity: an introduction*. Wiley. com.

Gravel, D., Bell, T., Barbera, C., Bouvier, T., Pommier, T., Venail, P., and Mouquet, N. (2010). Experimental niche evolution alters the strength of the diversity-productivity relationship. *Nature*, 469(7328):89–92.

Guzman-Novoa, E., Emsen, B., Unger, P., Espinosa-Montaño, L. G., and Petukhova, T. (2012). Genotypic variability and relationships between mite infestation levels, mite damage, grooming intensity, and removal of Varroa destructor mites in selected strains of worker honey bees (Apis mellifera L.). *Journal of invertebrate pathology*, 110(3):314–20.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. a. C., and Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PloS one*, 6(4):e17497.

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., and Martiny, J. B. H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature reviews. Microbiology*, 10(7):497–506.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological sciences / The Royal Society*, 270(1512):313–21.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of Birds through DNA Barcodes. *PLoS biology*, 2(10):e312.

Hedrick, P. (2011). Molecular Approaches in Natural Resource Conservation and Management. *BioScience*, 61(4):330–331.

Hoff, K. J. (2009). The effect of sequencing errors on metagenomic gene prediction. *Bmc Genomics*, 10(1):520.

Hollingsworth, M. L., Andra Clark, A., Forrest, L. L., Richardson, J., Pennington, R. T., Long, D. G., Cowan, R., Chase, M. W., Gaudeul, M., and Hollingsworth, P. M. (2009). Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular ecology resources*, 9(2):439–57.

Howe, K., Bateman, A., and Durbin, R. (2002). QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*.

Huang, Y., Lai, X., He, X., Cao, L., Zeng, Z., Zhang, J., and Zhou, S. (2009). Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in Escherichia coli. *Marine Biotechnology*, 11(1):124–131.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science (New York, N.Y.)*, 294(5550):2310–4.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, 67(10):4399–4406.

Humblot, C. and Guyot, J.-P. (2009). Pyrosequencing of tagged 16S rRNA gene amplicons for rapid deciphering of the microbiomes of fermented foods such as pearl millet slurries. *Applied and environmental microbiology*, 75(13):4354–4361.

Janda, J. M. and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–2764.

Jeffery, G., Gardi, C., Jones, A., Montanarella, L., and Marmo, L. (2010). *European Atlas of soil biodiversity*. Publications Office of the European Union.

Jeyaprakash, A., Hoy, M. a., and Allsopp, M. H. (2003). Bacterial diversity in worker adults of Apis mellifera capensis and Apis mellifera scutellata (Insecta: Hymenoptera) assessed using 16S rRNA sequences. *Journal of Invertebrate Pathology*, 84(2):96–103.

Jones, M. O., Koutsovoulos, G. D., and Blaxter, M. L. (2011). iPhy: an integrated phylogenetic workbench for supermatrix analyses. *BMC bioinformatics*, 12(1):30.

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Jost, L. (2007). Partitionig diversity into independent alpha and beta components. *Ecology*, 88(10):2427–2439.

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and environmental microbiology*, 75(15):5111–5120.

Lee, O. O., Wang, Y., Yang, J., Lafi, F. F., Al-Suwailem, A., and Qian, P.-Y. (2011). Pyrosequencing reveals highly diverse and species-specific microbial communities in sponges from the Red Sea. *The ISME journal*, 5(4):650–64.

Legendre, P. and Gallagher, E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2):271–280.

Leininger, S., Urich, T., Schloter, M., and Schwark, L. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*.

Little, D. P. (2011). DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PloS one*, 6(8):e20552.

Little, D. P. and Stevenson, D. W. (2007). A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, 23:1–21.

Loreau, M. (2010). Linking biodiversity and ecosystems: towards a unifying ecological theory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):49–60.

Lozupone, C., Faust, K., Raes, J., Faith, J. J., Frank, D. N., Zaneveld, J., Gordon, J. I., and Knight, R. (2012). Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome research*.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *The ISME journal*, 5(2):169–72.

Lozupone, C. a., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5):1576–85.

Ludovisi, A. and Taticchi, M. I. (2006). Investigating beta diversity by Kullback–Leibler information measures. *Ecological Modelling*, 192(1-2):299–313.

Mace, G. M. (2004). The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444):711–719.

Magurran, A. E. (2004). Measuring biodiversity.

Margulies, M., Egholm, M., Altman, W., and Attiya, S. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*.

Martinson, V. G., Danforth, B. N., Minckley, R. L., Rueppell, O., Tingek, S., and Moran, N. a. (2011). A simple and distinctive microbiota associated with honey bees and bumble bees. *Molecular ecology*, 20(3):619–28.

Martinson, V. G., Moy, J., and Moran, N. a. (2012). Establishment of characteristic gut bacteria during development of the honeybee worker. *Applied and environmental microbiology*, 78(8):2830–40.

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Ovreå s, L., Reysenbach, A.-L., Smith, V. H., and Staley, J. T. (2006). Microbial biogeography: putting microorganisms on the map. *Nature reviews. Microbiology*, 4(2):102–12.

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S. G., Dubinsky, E. A., Fortney, J. L., Han, J., Holman, H.-Y. N., Hultman, J., Lamendella, R., and Others (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *The ISME journal*, 6(9):1715–1727.

Matthews, B., Narwani, A., Hausch, S., Nonaka, E., Peter, H., Yamamichi, M., Sullam, K. E., Bird, K. C., Thomas, M. K., Hanley, T. C., and Turner, C. B. (2011). Toward an integration of evolutionary biology and ecosystem science. *Ecology letters*, 14(7):690–701.

Mattila, H. R., Rios, D., Walker-Sperling, V. E., Roeselers, G., and Newton, I. L. G. (2012). Characterization of the active microbiotas associated with honey bees reveals healthier and broader communities when colonies are genetically diverse. *PloS one*, 7(3):e32962.

Mazmanian, S. K., Liu, C. H., Tzianabos, A. O., and Kasper, D. L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, 122(1):107–18.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–8.

Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. L. (2006). DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success. *Systematic Biology*, 55(5):715–728.

Meier, R., Zhang, G., and Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic biology*, 57(5):809–13.

Meriweather, M., Matthews, S., Rio, R., and Baucom, R. S. (2013). A 454 survey reveals the community composition and core microbiome of the common bed bug (Cimex lectularius) across an Urban Landscape. *PloS one*, 8(4):e61465.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46.

Meyer, C. P. and Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS biology*, 3(12):e422.

Miller, S. E. (2007). DNA barcoding and the renaissance of taxonomy. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):4775–6.

Mocali, S. and Benedetti, A. (2010). Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Research in Microbiology*, 161(6):497–505.

Möckel, N., Gisder, S., and Genersch, E. (2011). Horizontal transmission of deformed wing virus: pathological consequences in adult bees (Apis mellifera) depend on the transmission route. *The Journal of general virology*, 92(Pt 2):370–7.

Mohammadi, T., Reesink, H. W., Vandenbroucke-Grauls, C. M. J. E., and Savelkoul, P. H. M. (2005). Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *Journal of microbiological methods*, 61(2):285–8.

Mohr, K. I. and Tebbe, C. C. (2006). Diversity and phylotype consistency of bacteria in the guts of three bee species (Apoidea) at an oilseed rape field. *Environmental microbiology*, 8(2):258–72.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., and Worm, B. (2011). How many species are there on Earth and in the ocean? *PLoS biology*, 9(8):e1001127.

Moran, N. a., Hansen, A. K., Powell, J. E., and Sabree, Z. L. (2012). Distinctive gut microbiota of honey bees assessed using deep sampling from individual worker bees. *PloS one*, 7(4):e36393.

Moran, S., Turner, P. D., and O'Reilly, C. (2008). Non-invasive genetic identification of small mammal species using real-time polymerase chain reaction. *Molecular ecology resources*, 8(6):1267–9.

Moreira, D. and López-García, P. (2002). The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends in microbiology*, 10(1):31–38.

Moritz, C. and Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS biology*, 2(10):e354.

Mouches, C., Bové, J., and Albisetti, J. (1984). Pathogenicity of Spiroplasma apis and other spiroplasmas for honey-bees in Southwestern France. *Annales de l'Institut Pasteur / Microbiologie*, 135(1):151–155.

Munch, K., Boomsma, W., Willerslev, E., and Nielsen, R. (2008). Fast phylogenetic DNA barcoding. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1512):3997–4002.

Nielsen, L., Arctander, P., Jensen, T. H., Dietz, H.-H., Hammer, A. S., Banyard, A. C., Barrett, T., and Blixenkrone-Mø ller, M. (2009). Genetic diversity and phylogenetic analysis of the attachment glycoprotein of phocine distemper viruses of the 2002 and 1988 epizootics. *Virus research*, 144(1):323–328.

Nowrousian, M. (2010). Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryotic cell*.

Oliver, I. and Beattie, A. J. (1993). A possible method for the rapid assessment of biodiversity. *Conservation biology*, 7(3):562–568.

Orlóci L (1967). An agglomerative method for classification of plant communities. *The Journal of Ecology*, 55:193–205.

Padial, M., Miralles, A., De la Riva., I., and Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7:1–14.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–290.

Pernal, S. F., Baird, D. S., Birmingham, A. L., Higo, H. A., Slessor, K. N., and Winston, M. L. (2005). Semiochemicals influencing the host-finding behaviour of Varroa destructor. *Experimental & applied acarology*, 37(1-2):1–26.

Powers, T. O., Neher, D. a., Mullin, P., Esquivel, a., Giblin-Davis, R. M., Kanzaki, N., Stock, S. P., Mora, M. M., and Uribe-Lorio, L. (2009). Tropical nematode diversity: vertical stratification of nematode communities in a Costa Rican humid lowland rainforest. *Molecular ecology*, 18(5):985–96.

Prince, V., Simao-Beaunoir, A.-M., and Beaulieu, C. (2009). Amplified ribosomal DNA restriction analysis of free-living bacteria present in the headbox of a Canadian paper machine. *Canadian journal of microbiology*, 55(7):810–817.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Gl{\"o}ckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, 35(21):7188—-7196.

Puillandre, N., Strong, E. E., Bouchet, P., C., B. M., Couloux, A., and Samadi, S. (2009). Identifying gastropod spawn from DNA barcodes: possible but not yet practicable. *Molecular Ecology*, 9:1311–1321.

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, 12(1):38.

R Development Core Team (2011). R: a language and environment for statistical computing. R Foundation for Statistical Computing.

Rach, J., Desalle, R., Sarkar, I. N., Schierwater, B., and Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings. Biological sciences / The Royal Society*, 275(1632):237–47.

Ramette, A. and Tiedje, J. M. (2007). Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial ecology*, 53(2):197–207.

Ratnasingham, S. and Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7:355–364.

Raupach, M. J., Astrin, J. J., Hannig, K., Peters, M. K., Stoeckle, M. Y., Wägele, J.-W., and Others (2010). Molecular species identification of Central European ground beetles(Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Frontiers in Zoology*, 7(26):1–15.

Ricotta, C. (2007). A semantic taxonomy for diversity measures. *Acta biotheoretica*.

Riesenfeld, C. S., Goodman, R. M., and Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental microbiology*, 6(9):981–989.

Rinderer, T. E., Harris, J. W., Hunt, G. J., and de Guzman, L. I. (2010). Breeding for resistance to Varroa destructor in North America. *Apidologie*, 41(3):409–424.

Rosenberg, N. a. (2003). The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, 57(7):1465–1477.

Rosenkranz, P., Aumeier, P., and Ziegelmann, B. (2010). Biology and control of Varroa destructor. *Journal of invertebrate pathology*, 103 Suppl:S96–119.

Ross, H., Murugan, S., and Li, W. (2008). Testing the reliability of genetic methods of species identification via simulation. *Systematic biology*.

Rougerie, R., Decaëns, T., Deharveng, L., Porco, D., James, S. W., Chang, C.-h., Richard, B., Potapov, M., Suhardjono, Y., and Hebert, P. D. N. (2009). DNA barcodes for soil animal taxonomy. *Pesquisa Agropecuaria Brasileira*, 44(8):789–802.

Sabree, Z. L., Hansen, A. K., and Moran, N. a. (2012). Independent studies using deep sequencing resolve the same set of core bacterial species dominating gut communities of honey bees. *PloS one*, 7(7):e41250.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*.

Sanchez, L. M., Wong, W. R., Riener, R. M., Schulze, C. J., and Linington, R. G. (2012). Examining the fish microbiome: vertebrate-derived

bacteria as an environmental niche for the discovery of unique marine natural products. *PloS one*, 7(5):e35398.

Sarkar, I. N., Planet, P. J., and Desalle, R. (2008). caos software for use in character-based DNA barcoding. *Molecular ecology resources*, 8(6):1256–9.

Saunders, G. W. (2005). Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1462):1879–88.

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS one*, 6(12):e27310.

Seifert, K. a., Samson, R. a., Dewaard, J. R., Houbraken, J., Lévesque, C. A., Moncalvo, J.-M., Louis-Seize, G., and Hebert, P. D. N. (2007). Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proceedings of the National Academy of Sciences of the United States of America*, 104(10):3901–6.

Shannon, C. and Weaver, W. (1963). *The matematical theory of comunications*. University of Illinois Press.

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular ecology*, 21(8):1794–805.

Singer, G. a. C. and Hajibabaei, M. (2009). iBarcode.org: web-based molecular biodiversity analysis. *BMC bioinformatics*, 10 Suppl 6:S14.

Singh, J., Behal, A., Singla, N., Joshi, A., Birbian, N., Singh, S., Bali, V., and Batra, N. (2009). Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnology journal*, 4(4):480–94.

Singh, K. M., Shah, T., Deshpande, S., Jakhesara, S. J., Koringa, P. G., Rank, D. N., and Joshi, C. G. (2012). High through put 16S rRNA gene-based pyrosequencing analysis of the fecal microbiota of high FCR and low FCR broiler growers. *Molecular biology reports*.

Smith, M. A. and Fisher, B. L. (2009). Invasions, DNA barcodes, and rapid biodiversity assessment using ants of Mauritius. *Frontiers in zoology*, 6:31.

Smith, M. A., Fisher, B. L., and Hebert, P. D. N. (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1462):1825–34.

Sogin, M. M. L., Morrison, H. G. H., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103(32):12115–12120.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21:2045–2050.

Tang, X., Freitak, D., Vogel, H., and Ping, L. (2012). Complexity and variability of gut commensal microbiota in polyphagous lepidopteran larvae. *PloS one*, 7(7):1–9.

Taylor, H. R. and Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular ecology resources*.

Tylianakis, J. M., Rand, T. A., Kahmen, A., Klein, A.-M., Buchmann, N., Perner, J., and Tscharntke, T. (2008). Resource heterogeneity moderates the biodiversity-function relationship in real world ecosystems. *PLoS Biology*, 6(5):e122.

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E., and Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular ecology resources*, 9(1):51–60.

van Velzen, R., Weitschek, E., Felici, G., and Bakker, F. T. (2012). DNA barcoding of recently diverged species: relative performance of matching methods. *PloS one*, 7(1):e30490.

Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T., and De Meyer, M. (2012). Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. *PloS one*, 7(2):e31581.

Voelkerding, K., Dames, S., Durtschi, J., and Jacob, D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4):641—-658.

Wagner, V., Antunes, P. M., Ristow, M., Lechner, U., and Hensen, I. (2011). Prevailing negative soil biota effect and no evidence for local adaptation in a widespread Eurasian grass. *PloS one*, 6(3):e17580.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–7.

Watanabe, K., Kodama, Y., and Harayama, S. (2001). Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of Microbiological Methods*, 44(3):253–262.

Webb, C. O., Ackerly, D. D., McPeek, M. a., and Donoghue, M. J. (2002). Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*, 33(1):475–505.

Wickham, H. (2009). *ggplot2 Elegant Graphics for Data Analysis*.

Will, K. W. and Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, 20:47–55.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090.

Wooley, J. C. and Ye, Y. (2009). Metagenomics: Facts and Artifacts, and Computational Challenges. *Journal of Computer Science and Technology*, 25(1):71–81.

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., Huang, G., Li, Y., Yan, Q., Wu, S., and Others (2010). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *The ISME journal*, 5(3):414–426.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., and Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3(4):613–623.

Zhou, F., Olman, V., and Xu, Y. (2008). Barcodes for genomes and applications. *BMC bioinformatics*, 9(1):546.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., Van Nostrand, J. D., He, Z., and Yang, Y. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal*, 5(8):1303–13.