Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Comparison of targeted and shotgun animal gut metagenomics

Relatore:
Prof. Daniel Remondini

Presentata da:
Francesco Durazzi

Correlatore:
Dott. Claudia Sala
Dott. Alessandra De Cesare

Anno Accademico 2017/2018

Abstract (italiano)

La comprensione della composizione e delle funzioni dell'ecosistema intestinale è di particolare utilità per la valutazione e il miglioramento dello stato di produttività e salute degli animali da allevamento, come ad esempio il pollame[6]. Grazie al sequencing del *metagenoma*, che rappresenta il materiale genetico recuperato direttamente da campioni di ecosistemi come alcune sezioni del sistema digerente, gli scienziati tentano di ricostruire l'abbondanza dei microorganismi che vivono nell'intestino degli animali, in modo da ottenere informazioni sull'interazione con l'ospite. Lo scopo di questo lavoro è di confrontare, mediante l'analisi statistica dei dati, l'affidabilità di due diverse tecniche di sequencing, chiamate *metatassonomica* e *metagenomica*, le quali costituiscono entrambe un valido strumento per la ricostruzione delle popolazioni batteriche nel gut microbioma[14]. Sebbene la metagenomica, basata su uno shotgun sequencing dell'intero metagenoma, sia spesso ritenuta la migliore opzione per ottenere i profili di abbondanza batterica[15][16], alcuni studi recenti hanno raggiunto ottimi risultati servendosi di sequencing di frammenti di rRNA amplificato (metatassonomica); quest'ultima tecnica si basa sull'individuazione e il riconoscimento di particolari regioni del gene 16S del rRNA. Nel nostro studio, abbiamo a disposizione un dataset ben strutturato composto da 78 campioni metagenomici, provenienti dal cieco e dall'ingluvie di 40 polli, i quali sono stati studiati a differenti giorni di vita(1,14,35) e sono stati sottoposti (o no) a un probiotico aggiunto all'acqua potabile. Lo studio dei profili di abbondanza ottenuti separatamente mediante metagenomica e metatassonomica, mette in luce significative differenze fra le due tecniche sia in termini di capacità di riconoscere i generi più rari, sia di individare connessioni con dei marker biologici. Si è evidenziato in particolare che lo shotgun sequencing riconosce all'incirca cinque volte più generi rispetto a quelli osservati in comune con entrambe le metodologie, anche se alcuni set shotgun presentano un basso numero di sequenze metagenomiche. Inoltre, usando i silhouette score per valutare la segmentazione dello spazio dei profili di abbondanza in uno spazio PCoA a 2 dimensioni in confronto ai metadati biologici, notiamo che i batteri poco abbondanti, osservati solo nei set sequenziati con lo shotgun, contengono informazioni biologiche non trascurabili, nascoste al sequencing del gene 16S.

Abstract

The understanding of the composition and functions of the intestinal environment is particularly useful to evaluate and improve productivity and health of farmed animals, such as chickens[6]. By sequencing the *metagenome*, that represents the genetic material recovered directly from enviromental samples such as gut sections, scientists attempt to retrieve the abundances of microorganisms that inhabit the gut of animals, in order to access information about the interaction with the host. Our purpose is to compare, with a statistical approach, the reliability of two sequencing techniques, called *metataxonomics* and *metagenomics*, that can both provide a solid approach to investigate the populations of bacteria in gut microbiome[14]. Although metagenomics, based on shotgun sequencing of the full metagenome, is usually known as the best suited option to recover abundance profiles of bacteria[15][16], recent studies have highlighted remarkable results using amplicon sequencing, that targets and recognizes particular regions of 16S rRNA gene. In our study, we take advantage of a well-structured dataset of 78 samples collected from caeca and crops of 40 chickens, at different days of life(1,14,35) and fed (or not) with a probiotic supplemented to drinking water. The study of abundance profiles retrieved by metagenomics and metataxonomics separately, highlights several differences between the two techniques, in terms of detection of rare genera and connection to biological markers. Shotgun sequencing detects around five times more genera than those commonly detected by both techniques, even if several shotgun sets have low coverage. Furthermore, using silhouette scores to evaluate the space segmentation of abundance profiles in a 2-dimensional PCoA space according to biological metadata, we observe that low-abundance bacteria detected only by shotgun contain important biologic information, hidden to 16S sequencing.

Contents

Introduction

Recent studies have suggested that the gut microbiome performs numerous important biochemical functions for the host, and that disorders of the microbiome are associated with many and diverse disease processes[5]. Hence, the understanding of the composition and functions of the intestinal environment is particularly useful to evaluate and improve productivity and health of farmed animals, such as chickens.
Systems biology approaches based on next generation "omics" technologies are now able to describe the gut microbiome at a detailed genetic and functional (transcriptomic, proteomic and metabolomic) level, providing new insights into the importance of the gut microbiome in health, and they are able to map microbiome variability between species, individuals and populations. This has established the importance of the gut microbiome in the disease pathogenesis for numerous systemic disease states, as well as health status and productivity of poultry[6]. Thus, understanding microbiome activity is essential to the development of future personalized healthcare strategies, as well as potentially providing new targets for drug development.

In particular two approaches to sequencing can be offered to solve this problem, named *metataxonomics* and *metagenomics*. While the first relies on the targeting and recognition of a specific gene (16S rRNA) whose sequences are amplified (*amplicon sequencing*), the second attempts to a random sequencing of the full metagenome (*shotgun sequencing* or *Whole Genome Sequencing*). To choose between two methods, taking into account that 16S sequencing is noticeably cheaper, one has to assess the goodness of the evaluation of the effective abundances of bacteria and the resolution of rare species detection. Numerous studies have enlighted the advantages of shotgun sequencing[15][16] in having lower bias in abundance estimation, while recent works on large environmental metagenomes showed better resolution for 16S sequencing, leaving an open mark on a general answer to the initial question. The main complication is that scientists do not often know a-priori the real composition of the microbiome, so it is difficult to score which techniques is effectively more suitable, unless you build an artificial dataset[15].

In this work, we looked at the problematic from a different perspective, thanks to a wide and well structured database, richly furnished of biological metadata. We had metagenomes of around 40 chickens available, at different days of life and fed with different concentration of a probiotic supplemented to water, from which we collected metagenomes both from caeca and crop (78 datasets for shotgun and 78 for 16S). Then, we applied statistical and big data tools in order to assess correlations between the abundance profiles generated from both techniques and to determine how well these features enable to recognise the a-priori known biomarkers and classes (such as organ of collection, day of life of chickens and probiotic concentration).

In Chapter 1, we provide information about the gut microbiota, especially in chickens, explaining the importance of probiotics in the alteration of the intestinal environment. Basics of DNA sequencing are shown too, along with an introduction to metataxonomics and metagenomics and the state of the art in 16S/shotgun comparison.
In Chapter 2, we give some details about sample preparation, sequencing and building of abundance profiles with MG-RAST[19]. We also provide a full description of the dataset, before listing the main statistical and bioinformatical tools adopted for data analysis.
In Chapter 3, we show data analysis and results, paired with commentary and conclusions. For the sake of legibility, we moved to Supplementary Sections some less relevant results that we obtained from the analysis, that are not included in the main conclusions.

Chapter 1

Gut microbiota and sequencing of metagenomes

In this chapter we are going to provide informations about gut microbiome populations in human and chickens, to understand the importance of microbiome studies. Then we will report the basics of DNA sequencing, focusing on next-generation sequencing of metagenomes through *shotgun metagenomic sequencing* and 16S rRNA targeted sequencing.

1.1 Gut microbiota

The gut microbiome is the term given to describe the vast collection of symbiotic microorganisms in the gastrointestinal system and their collective interacting genomes[1]. In fact, mammals possess an "extended genome" of millions of microbial genes located in the intestine: the microbiome. This multigenomic symbiosis is expressed at the proteomic and metabolic levels in the host and it has therefore been proposed that humans represent a vastly complex biological "superorganism" in which part of the responsibility for host metabolic regulation is devolved to the microbial symbionts.
In most animals, the gut microbiome is dominated by four bacterial phyla that perform various tasks: Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria[2]. In reality there is tremendous variation in the composition of the gut microbiomes of mammalian species and even among individuals of the same species. Both genetic and environmental factors, such as sex, geography, diet and disease state, contribute to differences in microbial community composition among individuals[3]. The gut environment is subject to a constant influx of microbial colonizers, and yet, mammalian species harbour distinct microbiomes and can be readily differentiated based on their resident microbes. Among very recently diverged species with similar diets, and even in cases where co-occurring species participate in the microbial transfer, it is possible to partition host species based on their microbiomes[3].

In reality, one of the earliest factors that can have a profound influence on the microbiota composition is the maternal environment. Several studies have shown that genetically identical mice from the same litters have a more similar microbiota than mice from different litters, even though they may be reared in adjacent cages[4].
This 'maternal effect' occurs when mouse pups are born vaginally and the birth mother's microbiota is their primary inoculum. Maternal effects can influence bacterial β-diversity (measured by UniFrac) regardless of host genotype, as well as affecting the relative abundances of phylotypes[4]. As a consequence, the maternal effect can be a major confounding factor when comparing the microbiota of mice with different genotypes or under different treatments.

1.1.1 Human gut microbiome

The human gastrointestinal tract harbors the most complex human microbial ecosystem (intestinal microbiota). The comprehensive genome of these microbial populations (intestinal microbiome) is estimated to have a far greater genetic potential than the human genome itself. Among the microbial communities that colonize human beings, the most rich and complex microbial consortium resides in the GIT, reaching a bacterial concentration of 100–200 billion cells/gram of feces (dry weight), so that the number of bacterial inhabitants within the gut lumen can reach 10^{14}[5].
Furthermore differences between individuals are known to be more marked among infants

than in adults, but later in life the gut microbiome converges to more similar phyla[1]. The colonization of the human gut begins at birth and is characterized by a succession of microbial consortia, the composition of which is influenced by changes in diet and by life events. The diversity and richness of the microbiota reach adult levels in early childhood, and the composition is thought to then remain relatively stable and resilient to stresses, such as antibiotic treatments[4].

Correlations between changes in composition and activity of the gut microbiota and common disorders, such as inflammatory bowel diseases, obesity, diabetes, and atopic diseases, have been proposed and proved, increasing the interest of the scientific community in this research field. In this perspective, a comprehensive and detailed view of the human gut microbiota, in terms of phylogenetic composition as well as genetic and metabolic potential, is essential to understand the dynamics and possible mechanisms of the cause/effect relationships between gut microbiota and pathology[5].
A growing number of studies highlight the fact that certain microbiota can be harmful to host health. Dysbioses of the microbiome are associated with an expanding list of chronic diseases that includes obesity, inflammatory bowel disease (IBD) and diabetes[4].
These types of correlative observations raise the question of whether the microbiota has a causative role in disease, or whether dysbiosis is a by-product of the disease. For several diseases, recent work shows the answer to be that the microbiota does contribute to disease. Transplantation experiments in which the microbiota of a diseased animal is grafted into a germ-free healthy recipient have demonstrated that several disease phenotypes could be transferred by the microbiota. These include excess adiposity, metabolic syndrome and colitis, all of which are traits of complex diseases that are also affected by host genetic and environmental factors[4].

1.1.2 Gut microbiome in chickens

The domestic chicken, *Gallus gallus domesticus*, with a global population exceeding 40 billion individuals per year has a unique status as "both the model and the system", which means that chickens are common model organisms for human biological research and also comprise an economically valuable global protein industry.

Recent advances in the technology available for culture-independent methods for identification and enumeration of environmental bacteria have invigorated interest in the study of the role of chicken intestinal microbiota in health and productivity. Chickens harbour unique and diverse bacterial communities that include human and animal pathogens. Increasing public concern about the use of antibiotics in the poultry industry has influenced the ways in which poultry producers are working towards improving birds' intestinal health. Effective means of antibiotic-independent pathogen control through competitive exclusion and promotion of good protective microbiota are being actively investigated[6]. With the realisation that just about any change in environment influences the highly responsive microbial communities and with the abandonment of the notion that we can isolate and investigate a single species of interest outside of the community, came a flood of studies that have attempted to profile the intestinal microbiota of chickens under numerous conditions[6].

The role of the GIT microbiota in both productivity and health is subject to intensive study. The microbiota within the GIT also has important roles in protection from pathogens, detoxification and modulation of immune system development. It harbours a very diverse microbiota that aids in the breakdown and digestion of food and comprised over 1000 species of bacteria, with a population density that can reach about 10^{11} cells/g digesta[7].

Two major groups of culture-independent methods, community fingerprinting and sequencing-based methods, are used for characterising microbial communities. In our study we are going to focus on the latter technique (Chapter 1.2), since with the rapid advances in the affordability and capacity of DNA sequencing technologies, the sequencing of 16S rRNA genes has rapidly replaced fingerprinting methods as the method of choice for community profiling[6].

Anyway, colonisation of the gastrointestinal tract is thought to start immediately after hatching, and therefore, the hatching environment has a major influence on a chicken's microbial profile. Differently from other animals common in production systems, poultry are somewhat unusual in that the young are generally separated from the parents, and hence, there is a markedly reduced parental influence on the development of microbiota post-oviposition; in fact once eggs have been washed or fumigated prior to hatching, there is no contact with adults during incubation[6]. Within commercial hatcheries, hygiene measures reduce the bacterial load in the hatching environment to limit the spread of bacterial pathogens. As a consequence, newly hatched chicks are exposed to a diverse range of bacteria from environmental sources such as human handlers, bedding material, feed and transport boxes, rather than from parental sources.
While chickens life goes on, their gut microbiome varies for several reasons. Amit-Romach et al. (2004)[8] found that temporal fluctuations of the groups investigated continued beyond day 4 to day 25. The results indicated that in young chickens the most abundant genus present in the small intestines and caeca was *Lactobacillus*, with a *Bifidobacteria* population becoming more dominant in the caeca at older age. *Clostridium* was detected in some segments of the small intestine in young chicks. In older chickens, *Salmonella*, *Campylobacter*, and *E. coli* species were found in the caeca.

The profile of the gut sections

In Figure 1.1 we show first of all an illustration of the anatomy of a chicken, since we have available data collected from two particular sections of the GIT.

The bacterial communities originating from different sections of the chicken GIT are so different that it has been suggested that they should be considered as separate ecosystems[6]. They are, however, highly connected, and they seed and influence microbiota both up and downstream in the GIT. Additionally, the profiles of different GIT sections differ significantly between studies due to differences in bird genetics, sex, diet, use of antimicrobials, housing and also technique-imposed differences such as primers used, method sensitivity, DNA extraction protocol etc. It is therefore difficult to define typical microbial profiles for any sections of the GIT. Even general measurements such as the ratio of *Firmicutes* to *Bacteriodes* can vary greatly.
In our work we collected data from two intestinal sections in particular, that are:

- **Crop:** that is usually[6] mostly populated by Lactobacillus (dominant) Clostridiaceae, Bifidobacterium, Enterobacteriaceae, Enterococcus.

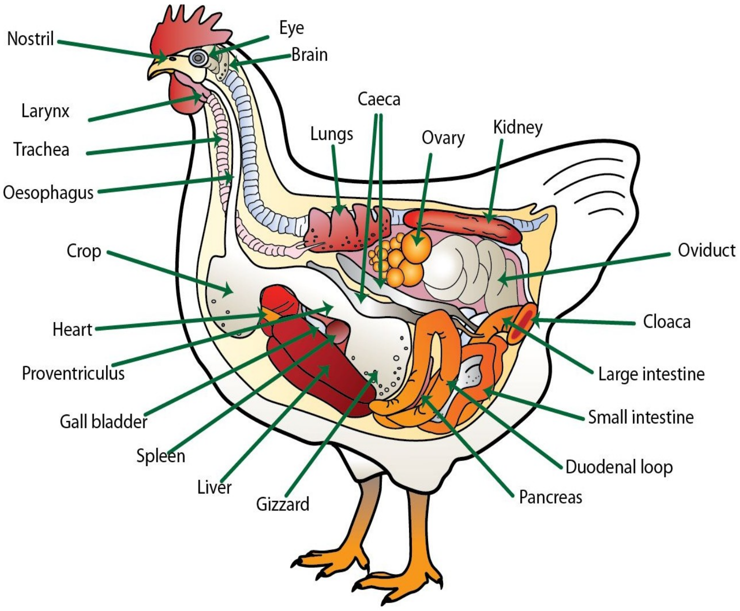


Figure 1.1: Illustration of chicken anatomy, from www.poultryhub.org .

- **Caecum:** rich in unknown and uncultured bacteria, and mostly composed of Lactobacillus, Bacterioides, Clostridium, Bifidobacterium.

So the crop is used for food storage and fermentation and is dominated by *Lactobacilli* and *Clostridiaceae.*
Caeca, that in mammals have a negligible role in digestion, in birds are an important site of fermentation instead[6], influencing animal health and performance. That is why the caecal microbiota profiles are widely investigated. Culture-free insights into caecal microbial profiles confirmed the most dominant genera identified through culturing methods but also pointed to much higher abundance of phylotypes with very low sequence similarity to known culturable isolates.
Firmicutes, Bacteroides, and Proteobacteria are the most common phyla in the chicken caeca, with Actinobacteria accounting for the remainder, as in Figure 1.2[9].

At deeper levels, the most abundant groups in the chicken caeca were found to be Clostridiaceae, Bacteroidaceae, Lachnospiraceae, Lactobacillus, Proteobacteria, butyrate producing cluster and unknown Firmicutes, with an abundance of Clostridium, Ruminococcus, Eubacterium, Faecalibacterium and Lactobacillus species among a number of unknown and uncultured phylotypes[6].

In studies of human and other mammals, faecal samples are mostly used as representatives of intestinal microbiota. The ease in acquiring the sample is the main reason for this; the subject remains in good health after sampling, and samples can be taken daily for any period of time. In chickens, however, most of the studies focus on caecum as the chicken caeca are considered to be of highest importance in chicken health and major pathogen reservoirs.
Sekelja et al. (2012)[10] investigated chicken faecal samples to compare them with other GIT sections. They used a statistical approach by employing a calibration-free multivariate technique on faecal samples to look for the influence of other GIT sections over 16 days. They found that faecal microbiota is directly influenced by periodic emptying of different GIT sections and thus varies greatly between the time points. They proposed

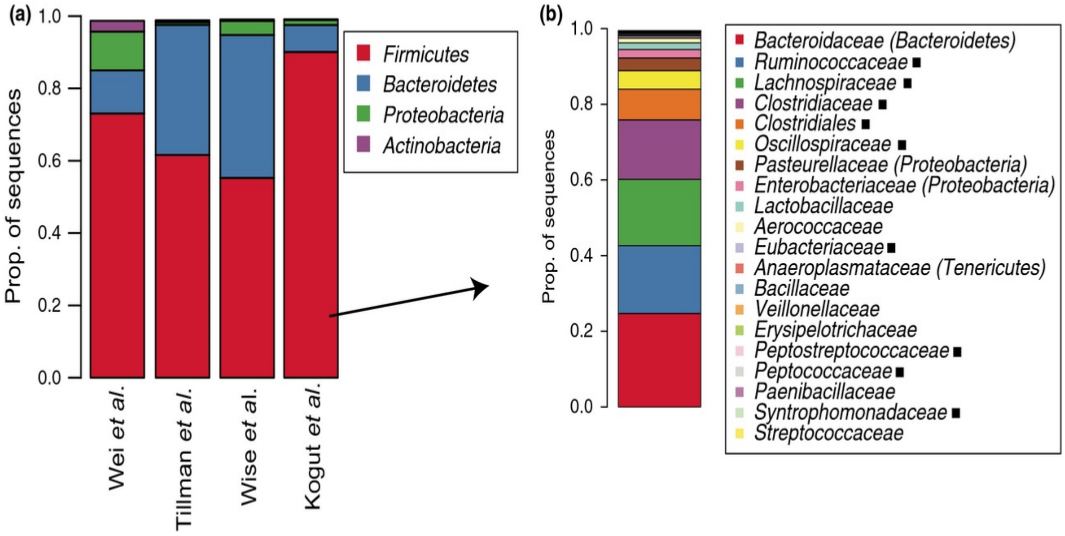


Figure 1.2: Relative proportions of bacterial phyla (a) and families (b) found in chicken caeca. Data from Wei et al. (2013) represent publically available sequences retrieved as described. Data from Tillman et al. (2011) and Wise & Siragusa (2007) are re-analyzed from data included in (Oakley et al., 2013) representing 8 and 10 birds, respectively. Kogut et al. data are unpublished, collected, and analyzed as previously described (Oakley et al., 2012b, 2013) representing 20 birds and c. 20 000 sequencing reads. Data for each of these three flocks are from 3 weeks posthatch. Sequences from Wei et al. were additionally screened by removing sequences with ambiguous base calls, and all sequences were classified against a reference database of type strains from SILVA v115 (Pruesse et al., 2007). Many of the sequences reviewed in (Wei et al., 2013) do not contain metadata regarding bird age, which can have strong effects on community composition and structure. For (b) families belong to the phylum Firmicutes unless otherwise noted; families followed by black squares belong to the Clostridiales.

that temporal shifts in faecal microbiota are a consequence of this periodic emptying of different GIT sections. Therefore, fecal samples may not be properly representative of the gastrointestinal tract due to differential mixing effects and to the less frequent voiding of the caeca compared to the rest of the gastrointestinal tract[9].

Probiotics

Probiotics are defined as viable microorganisms used as a food supplement with proven beneficial effects on health, able to promote or support a good balance of GIT microbial populations. Major molecular mechanisms of therapy with probiotics include the following: restoration of a beneficial consortium in the GIT including increased beneficial/-pathogen ratio, outcompeting pathogens for binding sites on intestinal epithelial cells, modulation of immune activity, stimulation of epithelial health and inhibition of tumour necrosis factor in intestinal epithelial cells[6].
Lactobacillus strains are among the most important and widely used probiotics. A number of strains have made their way into food as a supplement for humans and agricultural animals alike. It has been shown that previously used strain phenotypic identification does not correspond with 16S ribosomal RNA gene sequencing analysis showing that Lactobacillus species are not easily distinguishable.
Chicken indigenous Lactobacillus strains possess high antibiotic resistance, and genetic exchange may occur between native GIT strains. This also needs to be taken into con-

sideration when choosing future chicken probiotic strains.

The timing of probiotic administration may influence the onset of the beneficial effect. Nakphaichit et al. (2011)[11] administered Lactobacillus reuteri only during the first week post-hatch to find that the probiotic had no measurable effect at 3 weeks of age; however, at 6 weeks of age, delayed effects were shown through the increase in diversity and abundance of Lactobacillus and suppression of pathogen conferring groups of bacteria. They also noticed a positive influence on performance only during probiotic administration.

In general *Lactobacillus* strains have been described as beneficial additives because of their effects in promoting poultry production performance. However, kind of probiotic strain, dosage (i.e., colony forming unit (cfu)/bird/day), which should be modulated according to the flock health status and/or the farm hygienic conditions, as well as treatment duration, are among the critical factors influencing a probiotic efficacy. In particular, it has been shown[12] that the supplementation with *Lactobacillus acidophilus* D2/CSL (CECT 4529) at the recommended dietary dosage feed in broiler chickens significantly improved body weight at 28 days (commercial weight of 1.5 kg) and feed conversion rate from 0 to 41 days; and an overall positive effect of the supplementation with *Lactobacillus acidophilus* was observed in relation to the metabolic functions in the treated group, with particular reference to the higher abundance of β-glucosidase, improving animal performances and health.

1.2 Sequencing

In this section we are going to explain what is generally referred as *sequencing* and its most recent applications, focusing in particular on gut metagenomics and the techniques used in our study for the investigation of this particular type of metagenome (16s rRNA and shotgun sequencing).

In general term *sequencing* means the reconstruction of a biopolymer of nucleic acids. In particular we are referring to DNA (DeoxyriboNucleic Acid) and RNA (RiboNucleic Acid), that consist on long chains of units called *nucleotides*. As in Figure 1.3, each nucleotide is composed by a nitrogenous base, a sugar and a phosphate group bond together.

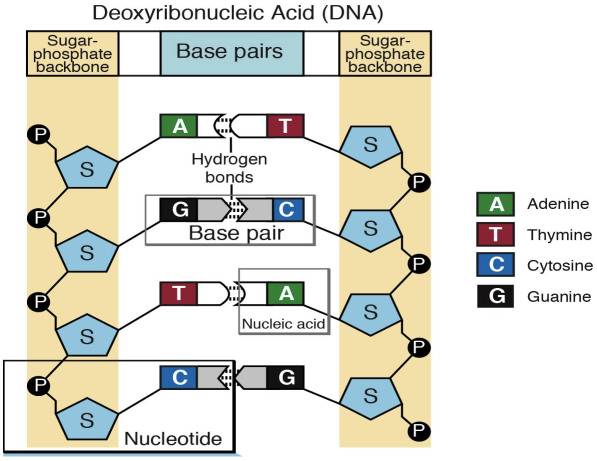


Figure 1.3: DNA portion, with two strand of 4 nucleotides held together by hydrogen bonds. Image courtesy of the National Human Genome Research Institution.

While in DNA the sugar is deoxyribose, in RNA it is ribose, and both are pentose (five-carbon sugar). Adjacent nucleotides are joined by a phosphodiester linkage, which consists of a phosphate group that links the sugars of two nucleotides. This bonding results in a backbone with a repeating pattern of sugar-phosphate units.
Differently from DNA in Figure 1.3, RNA molecules usually exist as single polynucleotide chains. Only certain bases in the double helix are compatible with each other. Adenine (A) always pairs with thymine (T), and guanine (G) always pairs with cytosine (C). Thus, the two strands of the double helix are complementary. Note that in RNA, adenine (A) pairs with uracil (U). A human genome contains approximately 3.2×10^9 of those base pairs, distributed among 22 paired chromosomes[13], but these numbers vary a lot among animals and other organisms.

For the aim of our study, once the structure of DNA reads has been recovered, one can try to assess which organism the DNA strain belongs to, and this is the main goal of *metagenomics* and *metataxonomics* (Tab 1.1).

Metagenomics and metataxonomics have emerged as the most powerful sequence-driven approaches to study the composition and the genetic potential of gut microbiota, and efforts in this direction have been smoothed by the implementation of next generation sequencing platforms.

Technique	Advantages and challenges	Main applications
Metataxonomics using amplicon sequencing of the 16S gene	+ Fast and cost-effective identification of a wide variety of bacteria and eukaryotes − Does not capture gene content other than thetargeted genes − Amplification bias − Viruses cannot be captured	Profiling what is present Microbial ecology rRNA-based phylogeny
Metagenomics using random shotgun sequencing of DNA or RNA	+ No amplification bias + Detects bacteria, archaea, viruses and eukaryotes + Enables de novo assembly of genomes − Requires high read count − Many reads may be from host − Requires reference genomes for classification	Profiling of what is present across all domains Functional genome analyses Phylogeny Detection of pathogens

Table 1.1: Metataxonomics and metagenomics strategies[14].

Since correlations have been suggested between the host's health and the composition and activity of the bacterial communities, especially the GIT microbiota, human microbial ecology is receiving increasing interest, thanks to technological advances in culture-independent methods. In fact, traditional microbiology requires laborious and time-consuming cultivation of microorganisms, and allows the recovery of less than 20–30% of the total bacterial richness because of the insufficient anaerobic cultivation technologies, as well as the poor knowledge about the specific carbon source requirements[5].
For this reason the vast majority of the biodiversity of the microbiota remains uncultured, and the assessment of the microbial composition and abundance of such a dense and complex microbial community needs to be performed through molecular techniques, i.e. culture-independent approaches based on the extraction of the bacterial DNA directly from the samples (i.e. feces or intestinal biopsies) and analysis of the 16S ribosomal RNA (rRNA) gene.

Metataxonomics

Metataxonomics relies on the detection and recognition of 16S rRNA gene, that consists of about 1,500 nucleotides and contains regions conserved among all the bacteria, interspersed with 9 regions (V1 to V9 in Figure 1.4) that are highly variable among bacterial "phylotypes", defined as a group of 16S rRNA sequences having 97–99% of sequence identity[5].

Conserved regions can be used as targets for PCR primers with almost universal bacterial specificity. The variable regions have different discriminatory power depending on the groups of microbes and amongst the short target regions (<300 bp), the hypervariable region 4 (V4) was generally the most informative. Phylotype identification is obtained by comparative sequence analysis of the amplicons using available databases, such as the Ribosomal Database Project (http://rdp.cme.msu. edu), though different platforms and pipelines have spread through the years, such as MG-RAST[19].
Metataxonomics is an invaluable tool for microbial ecology. rRNA gene sequences are the

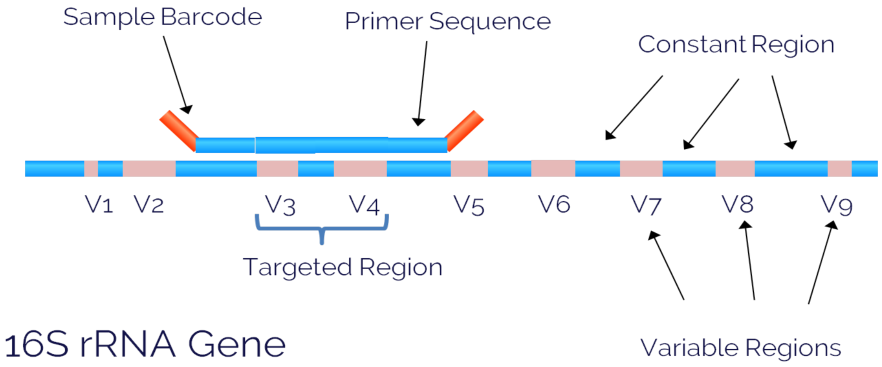


Figure 1.4: Scheme of 16s rRNA gene and its regions. From www.lcsciences.com .

most widely used marker sequences; these include the 16S rRNA gene for bacteria, the 18S rRNA gene for eukaryotes, and the internal transcribed spacer (ITS) regions of the fungal ribosome for fungi[14]. These markers work well for phylogenetic profiling because they are ubiquitously present in the population, they have hypervariable regions that differentiate species and they are flanked by conserved regions that can be targeted by 'universal' primers.

The workflow for 16S analysis typically includes quality filtering, error correction (sometimes called de-noising), removal of chimeric sequences, clustering of reads into 'Operational Taxonomic Units' (OTUs) based on sequence similarity and classification of the OTUs.

Marker gene sequencing does have some drawbacks, which explains (in part) the rising popularity of metagenomics. First, marker gene-based methodologies do not capture viruses, which have no conserved genes analogous to 16S gene[14]. The use of the 16S rRNA gene itself is imperfect as well: for the recently described Candidate Phyla Radiation, which comprises up to 15% of the bacterial domain, it was estimated that $> 50\%$ of the organisms evaded detection with classical 16S amplicon sequencing. Furthermore the short reads produced by next-generation sequencers further limit analysis at the species level.

Metagenomics

Metagenomics refers to the random 'shotgun' sequencing of microbial DNA, without selecting any particular gene. Many strategies can be used for analysis of metagenomics shotgun data. A common first step is to run a variety of computational tools for quality control, which identify and remove low-quality sequences and contaminants[14].

After quality control, the reads can either be assembled into longer contiguous sequences called contigs or passed directly to taxonomic classifiers. Taxonomic classification of every read is a form of binning because it groups reads into bins corresponding to their taxon ID.

When the analysis only returns the estimated abundances of the different taxa (instead of a classification of each read), we call it taxonomic profiling. The choice of assembly-based analyses versus direct taxonomic classification of reads depends on the research question. Compared with marker gene-based community profiling, metagenomic shotgun sequencing alleviates biases from primer choice and enables the detection of organisms across all domains of life, assuming that DNA can be extracted from the target environment[14].

1.2.1 State of the art in shotgun/16S comparison

The choice of shotgun or 16S approaches for microbiome analyses is usually dictated by the nature of the studies being conducted. For instance, 16S is well suited for analysis of large number of samples, i.e., multiple patients, longitudinal studies,etc. but offers limited taxonomical and functional resolution.

Jovel et al. (2016)[15] built an artificial bacterial population using DNA from 11 species and constructed 16S and shotgun libraries in parallel using the NEXTflex 16S V4 Amplicon-Seq (BioO Scientific)and the Nextera XT (Illumina) kits, respectively, before assigning taxonomy with UCLUST for 16S and MetaPhlAn for shotgun.
For 16S rRNA sequences, a consistent over-representation of sequences in the *Clostridium* and *Lactobacillus* genera was found. These two genera contain sequences that are perfectly complementary to the primers used for amplification, while at least one mismatch is found in the rest of genera included in their experimental (mock) bacterial population. This demonstrates how subtle differences in primer binding sites within the 16S rRNA gene sequences lead to biased estimates of relative abundance.
As for shotgun, all species included in the mock populations were correctly classified and a good approximation to their expected relative abundance was provided too.

Campanaro et al.(2018)[16] have recently performed an in-depth comparative evaluation of three widely used sequencing methods to investigate the taxonomic composition specifically focused on the anaerobic digestion microbiome. The microbial communities under investigation were grown in three laboratory scale Continuous Stirred Tank Reactors (CSTR) operated at thermophilic conditions ($54 \pm 1°C$) and fed with cattle manure. Both DNA and RNA were extracted using the same kit and protocols used for Illumina sequencing were very similar for all the samples[16].
It was demonstrated that the classical 16S rRNA amplicon sequencing is biased by two main effects, which are the limited number of hypervariable regions investigated (V4 in the present study) and, at to a lesser extent, the failure of universal primers to match all the 16S rRNA targets. These two biases influenced different taxonomic groups and, more specifically, amplification drawbacks were more problematic for *Euryarchaeota* and *Spirochaetes.*
Interestingly, analysis of shotgun DNA reads performed using a group of clade-specific marker genes other that 16S rRNA confirms that the use of this marker gene can lead to the underestimation in abundance of *Euryarchaeota* in the AD system. This finding also indicates that the use of multiple marker genes, or analysis at transcriptional level, could improve the evaluation of abundance for crucial taxonomic groups. Moreover, it is concluded that the absolute abundance level of different taxa is markedly influenced by the selected hypervariable region.

Not all studies are so unilateral on the better performances of shotgun sequencing. Tessler et al.(2017)[17], in fact, carried out a large-scale study on biodiversity in water samples across four of Brazil's major river floodplain systems. Their sequencing procedure used 454 GS Junior for 16S rRNA and the Illumina HiSeq 2500 for shotgun reads.
They found that less than 50% of phyla identified via amplicon sequencing were recovered from shotgun sequencing, clearly challenging the dogma that mid-depth shotgun recovers

more diversity than amplicon-based approaches. At family level, taxonomical classification revealed even less overlap between the two approaches. The amplicon approach resulted in the classification of 56 families, while the shotgun approach recognized 41 families, but only 18 families showed overlap between the two strategies. Furthermore, the amplicon data were overall more robust across both biodiversity and community ecology analyses at different taxonomic scales.

Two studies (Jovel et al.[15] and Tessler et al.[17]) used two different sequencing kits for amplicon and shotgun separately, though Jovel built in parallel both 16S and shotgun libraries, trying to avoid experimental bias caused by different tools.
The other study (Campanaro et al.[16]), that sequenced with the same kit both 16S rRNA and shotgun DNA, seems to be more reliable, though they only compared the abundance profiles produced by both methods without external reference; Jovel et al. built an artificial sample instead, in order to avoid this uncertainty, thus their results supporting shotgun sequencing appear very reliable, though it would be really interesting to introduce new and more quantitative criteria to assess how good the correlation between profiles found with different methods is, what the limitations of choosing one technique over the other are and how much biological information is lost when choosing a method over the other.

Chapter 2

Materials and methods

In this chapter we will provide details about the microbiome data we analysed, explaining which datasets we used and how we partitioned them into groups. Furthermore we are going to describe briefly some techniques of data analysis that helped us to get to the results in Chapter 3.

2.1 Data resume

All metagenomes are collected by the group of Prof. G. Manfreda, in particular by dott. A. De Cesare, Department of Agricultural and Food Sciences, Alma Mater Studiorum - University of Bologna (Ozzano).

2.1.1 Extraction of metagenomes and sequencing

We studied environmental metagenomes both from caeca and crop of poultry, collected from around forty chickens. The animals were divided into three groups, according to null (C), low (L) and high (H) dosage of a probiotic that contained $5 \times 10^{10} CFU/g$ of LA (*Lactobacillus acidophilus* D2/CSL), added to water with a concentration of $0.02g/d/bird$ for L and $0.02g/d/bird$ for H. For caeca[1] we detained 40 samples, in particular 4 from day 1 (dosed C), 16 from day 14 (dosed 5C, 6L, 5H) and 20 from day 35 (dosed 7C, 7L, 6H). For crop we had 38 samples, in particular 5 from day 1 (dosed C), 15 from day 14 (dosed 5C, 5L, 5H) and 18 from day 35 (dosed 6C, 6L, 6H).

The DNA was extracted from each caecum and crop content using a bead-beating procedure. Briefly, 0.25 g of caecal content were suspended in 1 ml lysis buffer (500 mMNaCl, 50mMTris-Cl, pH 8.0, 50mMEDTA, 4% SDS) with MagNA Lyser Green Beads (Roche, Milan, Italy) and homogenized on the MagNA Lyser (Roche) for 25 sec at 6500 rpm. The samples were then heated at $70°C$ for 15 min, followed by centrifugation to separate the DNA from bacterial cellular debris. This process was repeated with a second $300\mu l$ aliquot of lysis buffer. The samples were then subjected to $10Mv/v$ ammonium acetate (Sigma, Milan, Italy) precipitation, followed by isopropanol (Sigma) precipitation, 70% ethanol (Carlo Erba, Milan, Italy) washing and suspension in 100 ul 1X Tris-EDTA (Sigma). All samples were treated with DNase-free RNase (Roche) and incubated overnight at $4°C$, before being processed through the QIAmp® DNA Stool Mini Kit (Qiagen, Milan, Italy) according to manufacturer's directions with some modifications. DNA quantity and quality were measured on a BioSpectrometer® (Eppendorf, Milan, Italy).

For shotgun sequencing, DNA from each of the 78 samples (40 caeca and 38 crop) was fragmented and tagged with sequencing adapters using the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA). Whole genome sequencing was performed using the HiScanSQ sequencer (Illumina) at 100 bp in paired-end mode. Following sequencing, all reads were assessed for quality parameters and the paired end merged. The MG-RAST pipeline was used to identify the relative abundances of bacterial taxa performing a BLAST similarity search for the longest cluster representative against RDP database.

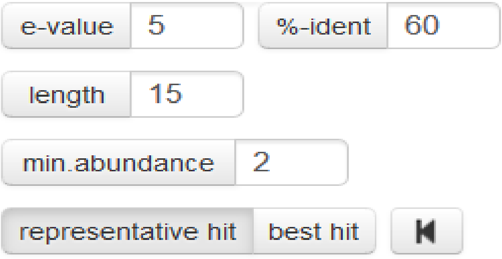
For amplicon sequencing, the libraries were prepared following the Illumina 16S Library preparation protocol, amplifying the variable V3 and V4 regions of the 16S rRNA

[1] Samples from caeca and crop were collected from the same chickens for the most part.

in order to obtain a single amplicon of approximately 460 bp. Sequencing was performed in paired-end in the Illumina MiSeq with the MiSeq Reagent kit v2 500 cycles, characterised by a maximum output of 8.5 Gb. When paired end sequencing is selected in the MiSeq, the ends of each read are overlapped to generate high-quality, full-length reads of 98 bp. The maximum output of the v2 kit is 15 million of reads per run, meaning approximately 187 500 reads per sample. All metagenomic sequences were deposited in MG-RAST (http://metagenomics.anl.gov/)

2.1.2 Processing in MG-RAST

Details of MG-RAST platform are provided in Section 2.2.1 or directly in the handbook[18]. The collected samples available had both shotgun and 16S abundance profiles, in fact we kept only those datasets who have been sequenced with both methods (78 datasets). Shotgun samples have the prefix *XT*, while amplicon samples are prefixed by *B*, and sets with the same number ID are collected exactly from the same sample (same chicken, same day, same organ). As we said before, data was collected both from caeca (40 samples) and crop (38 samples) of chicken, for period of time corresponding to 1 day, 14 days and 35 days from probiotic supplementation and chicken birth. Name of samples and number of sequences are displayed in Supplementary Tab S1 and S2.
Quality parameters set by default on MG-RAST are those on Figure 2.1, for samples sequenced in both ways.



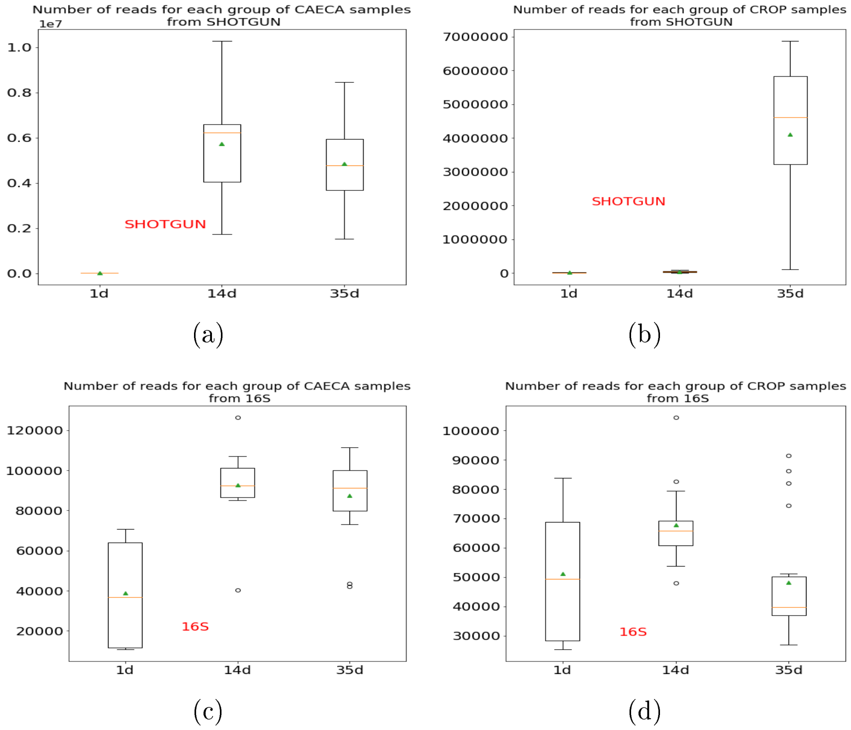
e-value 5
%-ident 60
length 15
min.abundance 2

Figure 2.1: Default MG-RAST parameters for quality of alignment of reads, both for 16S rRNA and shotgun.

In general, shotgun sequencing would have an amount of DNA reads higher than 16S before QC preprocessing (around few millions vs few hundreds of thousands), but for three out of six groups of samples, the mean number of shotgun reads is around the order of only 10^4, probably because of the difficult extraction of samples in young chickens. Since the percentage of reads passing the QC is roughly the same for both methods, the processing of data in order to obtain species abundance relies on a lot of more sequences for shotgun sets. So even if around 50% of these shotgun sequences escaped from QC are classified as unknown proteins, so not used for the taxonomy assignment, we can rely on around 10 times more shotgun sequences than 16S ones for the species annotation task for three out of six groups (caeca14day, caeca35day and crop35day), while they're 10 times lower for the remaining three groups (caeca1day, crop1day and crop14day).
We decided to remove from 16S those genera who appeared in only one sequence, because a OTU with only one occurrence is probably misannotated, while in shotgun samples we have by default that each individual has at least two reads. Then we have narrowed down the analysis only to those genera annotations marked as *Bacteria*, in order to study the relevant component of gut microbiota.

2.1.3 Data exploration

We downloaded all the datasets available on MG-RAST relative to the study of de Cesare et al. about biodiversity of gut microbiota in chickens who have been given water supplemented with a probiotic (*Lactobacillus acidophilus* D2/CSL).
In Figure 2.2, we see that the number of reads is quite variable and biased according to day of sampling. In particular three out of six shotgun samples (day1caeca, day1crop and day14crop) are very poor of sequences respect to the rest of metagenomics data[2] (and respect to 16S too).



(a) (b) (c) (d)

Figure 2.2: Mean and standard deviation of the number of sequences per chicken used for taxonomic annotations for shotgun samples from caeca (a) and crop (b). Mean and standard deviation of the number of sequences per chicken used for taxonomic annotations for 16S samples from caeca (c) and crop (d). We are now taking into account all sequences selected with default MG-RAST parameters, at day 1, 14 and 35.

In Tab 2.1 we have quality indicators of the datasets, inclusive of mean sequences abundance among the samples relative to the same group, already displayed in Figure 2.2.

Inside each group that we show in Tab 2.1 and 2.2, there is a different amount of metagenomic samples (around 5 on average), so instead of the total sum of sequences in each group, we prefer to display their mean number for each sample, that has the meaning of annotated sequences for each chicken.

[2] The number of DNA sequences in shotgun metagenomics is quite subject to the experimental asset, and in particular the extraction of samples from a newborn chick is a hard task.

Shotgun	**Seq.**	**st.dev. of seq.**	**e-value**	**Align. length**	**Percent id.**
caeca1day	22 544	4 184	-6.87	31	80.5
caeca14days	5 741 705	2 285 882	-8.85	38	74.6
caeca35days	4 861 034	1 766 362	-8.35	36	75.5
crop1day	15 797	4 304	-7.07	31	81.5
crop14days	44 212	22 521	-7.33	32	80.7
crop35days	4 108 357	2 256 211	-8.09	35	78.5

Table 2.1: Mean and standard deviation of the number of sequences per chicken used for taxonomic annotations for shotgun samples, with default MG-RAST parameters, at day 1, 14 and 35. We also show mean value of e-value, alignment length and percent identity, computed by MG-RAST.

16S	**Seq.**	**st. dev. of seq.**	**e-value**	**Align. length**	**Percent id.**
caeca1day	116 380	82 915	-67.00	134	98.4
caeca14days	277 715	51 130	-40.41	86	98.9
caeca35days	262 079	55 384	-40.77	87	98.8
crop1day	153 334	68 025	-56.77	115	98.7
crop14days	203 211	39 252	-57.7	117	98.9
crop35days	144 515	59 928	-81.23	161	97.8

Table 2.2: Mean and standard deviation of the number of sequences per chicken used for taxonomic annotations for 16S samples, with default MG-RAST parameters, at day 1, 14 and 35. We also show mean value of e-value, alignment length and percent identity, computed by MG-RAST.

We can see that 16S samples (Tab 2.2) overcome the others in terms of e-value, alignment length and percent identity, whose meaning is described in next section (2.2.1). Since we have a lot of shotgun sequences, at first thought we may select among them only those that have a minimum percent identity and e-value (i.d. percent identity 93% and e-value -20 at least), in order to obtain a similar reliability for both sequencing methods. For example, one expects percent identity to be much higher in amplicon samples because they all belong to the same section of 16S gene (though hypervariable), so they distinguish only for small segments.

But, keeping in mind that amplicon and shotgun sequencing are structurally different, it's not granted that having the same quality features will lead to the same significance, so we opted for a tuning in order to find the parameters that yield the best reliability for shotgun sequences. In fact, since we can not rely on external confirmations about the correctness of the taxonomic assignation, we first believed that the optimal configuration of the sequences was the one that got the highest correlation, in terms of bacteria population, between 16S and shotgun abundance profiles for the same sample, but this tuning of e-value thresholds only led to a consistent loss of rare genera in 16S sets, as we see in Supplementary Section 3.4, so we kept MG-RAST default thresholds of -5 for both shotgun and 16S sequences as in Figure 2.1.

2.2 Methods

2.2.1 MG-RAST

MG-RAST[19] is a portal built to provide users of an exhaustive analysis of environmental DNA ("metagenomic sequences"), in terms of alignment of sequences and taxonomic (and functional) assignments.
Users can easily upload various types of data, according to their purpose[18]:

1. enviromental clone libraries (functional metagenomics), if they use Sanger sequencing instead of next-generation sequencing.

2. Amplicon metagenomics (16S rRNA).

3. Shotgun metagenomics.

4. Metatranscriptomics, that uses cDNA transcribed from mRNA.

The system provides answers to several crucial questions, and in particular for our purpose it helps to identify the composition of a microbial community either by using amplicon data for single genes or by deriving community composition from shotgun metagenomic data using sequence similarities.
In order to do so, the MG-RAST pipeline (in Figure 2.3) performs quality control, protein prediction, clustering and similarity based annotation on nucleic acid sequence datasets using several bioinformatics tools.

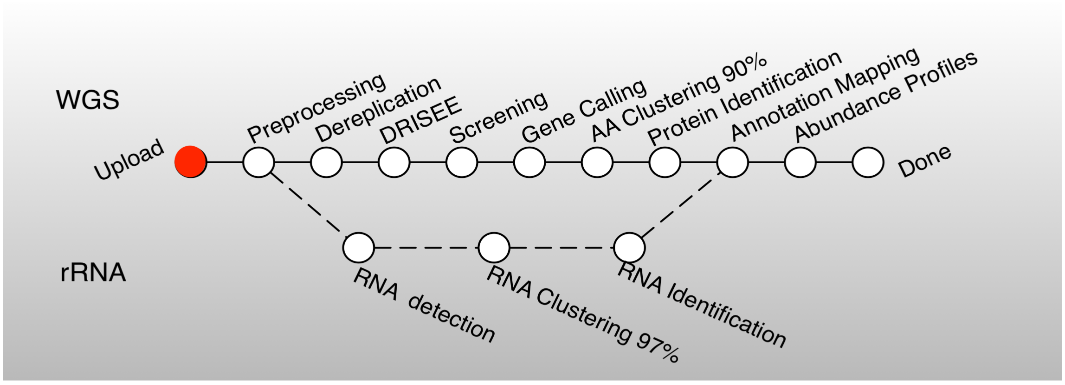


Figure 2.3: Details of the analysis pipeline for MG-RAST version 3[18].

The processes can be summarized in five steps, that we explain briefly[18]

1. **Data hygiene:** quality control and removal of artifacts. It is composed by *preprocessing* that trims low-quality regions from FASTQ data and discard sequences whose length is more than two standard deviation away from the mean read length, *dereplication* that removes Artificial Duplicate Reads by identifying all 20 character prefix identical sequences and *screening* that removes reads that are near-exact matches to the genomes of a handful of model organisms, including fly, mouse, cow, and human.

2. **Feature identification:** for shotgun samples a machine learning approach with FragGeneScan performs predictions of coding regions in DNA sequences and identifies proteins. For amplicon samples, ribosomial RNA is identified using a search

against a reduced database, built from a 90% identity clustering of SILVA, Greengenes and RDP databases.

3. **Feature annotation:** for shotgun samples, MG-RAST builds clusters of proteins at the 90% identity level preserving relative abundances, then a representative of each cluster is subjected to similarity analysis with sBLAT, an implementation of BLAT algorithm[20], in order to reconstruct the putative species composition of the sample by looking at the phylogenetic origin of the database sequences hit by the similarity searches. For amplicon samples, the rRNA-similar reads are clustered at 97% identity using cd-hit, and the longest sequence is picked as the cluster representative, then a BLAT similarity search is operated against the databases.

4. **Profile generation:** in the final stage, MG-RAST generates abundance profiles, that represent a pivoted and aggregated version of the similarity files.

To comprehend and manage appropriately the abundance profiles produced so far, one has to understand the meaning of quality cut-off that can be set and how the abundances are counted.

The statistics of sequence comparison

The threshold for annotation transfer can be set using the following parameters: e-value, percent identity, and minimal alignment length. While the two latter indicators are quite obvious to understand, e-value computation requires some explanation.
The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance[20].
The lower the E-value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query sequence. These high E values make sense because shorter sequences have a higher probability of occurring in the database purely by chance.

In the limit of sufficiently large sequence lengths m and n, the statistics of high-scoring segment pairs (HSP) scores are characterized by two parameters, K and lambda. Most simply, the expected number of HSPs with score at least S is given by equation (2.1):

$$E = Kmne^{-\lambda S} \tag{2.1}$$

We call this the E-value for the score S.
This formula makes eminently intuitive sense. Doubling the length of either sequence should double the number of HSPs attaining a given score. Also, for an HSP to attain the score 2x it must attain the score x twice in a row, so one expects E to decrease exponentially with score. The parameters K and lambda can be thought of simply as natural scales for the search space size and the scoring system respectively.

One can think that a query is a priori more likely to be related to a long than to a short sequence, because long sequences are often composed of multiple distinct domains.

If we assume the a priori chance of relatedness is proportional to sequence length, then the pairwise E-value involving a database sequence of length n should be multiplied by N/n, where N is the total length of the database in residues. Examining equation (2.1), this can be accomplished simply by treating the database as a single long sequence of length N. The BLAST programs take this approach to calculating database E-value. Notice that for DNA sequence comparisons, the length of database records is largely arbitrary, and therefore this is the only really tenable method for estimating statistical significance.

Best hit and representative hit profiles

To understand the meaning of the abundance counts used for our measurements and graphs, we have first to remember that in some cases sequences are identical between different database records, e.g. version of E. coli might share identical proteins and it becomes impossible to determine the "correct" organism name. In those cases, the translation of those similarities (that are against an anonymous database, with merely MD5 hashes[3] used as identifiers) can be done in several different ways[18].

- **Best hit:** using one organism. The best hit classification reports the functional and taxonomic annotation of the best hit in the M5nr nonredundant protein database for each feature. In those cases where the similarity search yields multiple same-scoring hits for a feature, we do not choose any single "correct" label. For this reason they have decided to double count all annotations with identical match properties and leave determination of truth to our users. While this approach aims to inform about the functional and taxonomic potential of a microbial community by preserving all information, subsequent analysis can be biased because of a single feature having multiple annotations, leading to inflated hit counts. For users looking for a specific species or function in their results, the best hit classification is likely what is wanted.
- **Representative hit:** MG-RAST pick a random member of the group of identical sequences. The representative hit classification selects a single, unambiguous annotation for each feature. The annotation is based on the first hit in the homology search and the first annotation for that hit in our database. This approach makes counts additive across functional and taxonomic levels and thus allows, for example, the comparison of functional and taxonomic profiles of different metagenomes.

For our purpose of a comparative analysis, representative hit seems to fit better, because we prefer to assess consistent values for bacteria abundances in different metagenomes more than find some particular organisms.

So the MG-RAST v3 annotation pipeline does not usually provide a single annotation for each submitted fragment of DNA. Steps in the pipeline map one read to multiple annotations and one annotation to multiple reads. These steps are a consequence of genome structure, pipeline engineering, and the character of the sequence databases that MG-RAST uses for annotation.
The first step that is not one-to-one is gene prediction, because long reads can contain pieces of two or more microbial genes, and yet are annotated separately.
Then, as we already reported, an intermediate clustering step identifies sequences at 90%

[3] Cryptographic items corresponding to keys in a nonredundant protein database (M5nr) used for annotation of metagenomic sequences[18].

amino acid identity and performs one search for each cluster. Sequences that do not fall into clusters are searched separately. The “abundance” column in the MG-RAST tables, that we use directly in our study, presents the estimate of the number of sequences that contain a given annotation, found by multiplying each selected database match (hit) by the number of representatives in each cluster.

Each read is assigned taxonomically to a strain, but at strain and species level the results are quite sensitive to user choices in the pipeline; anyway we know for each read the entire phylogenetic profile, so from less fine level to finest: domain → phylum → class → order → family → genus → species → strain.

2.2.2 Data analysis

Here we list some of the statistical methods that helped us to explore the data and extract information from the abundance profiles downloaded from MG-RAST platform.

Preston plot

Preston, since 1948[22], argued that *Relative Species Abundance* (RSA) distributions were often bell-shaped curves, such that species having intermediate abundances were more frequent than very rare species. Preston actually noted that the distributions were lognormal and introduced a simple way to display this lognormal distribution of relative species abundance. He built doubling categories of abundance (1, 2, 4, 8, etc.), and counted the species having abundances falling in each category[23].
This classification of species into doubling abundance classes effectively log transforms the relative abundance data to the log base 2. He chose log base 2 for the simple practical expedient of spreading the distribution of species abundances over more categories to make its shape more apparent.

In the logseries, previously used to fit environmental populations, the expected number of species is always largest in the rarest abundance category, consisting of singleton species. However, in a small sample, one should observe only a truncated distribution of relative abundances, comprising only the most common species. This is because common species are generally collected sooner than rare species[23].

β-diversity

Beta (β) diversity considers the difference in bacterial community composition for different environments. There are two main approaches for quantifying β-diversity: those that take into account the evolutionary differences between communities, formally known as phylogenetic β-diversity, and those that do not, formally known as taxon-based or non-phylogenetic methods[15].

One of the most popular non-phylogenetic approaches to quantify β-diversity is the Bray-Curtis dissimilarity, that we used in our analysis. It is robust to the presence of zeroes in a count table, as often is the case for microbiome data (i.e., some bacterial taxa will be present in some but not all samples).
Bray-Curtis dissimilarity takes its minimum value (0) when two samples have no species

in common, irrespective of the precise abundances[24], as in:

$$BC(x,y) = \frac{\sum_i^n \mid x_i - y_i \mid}{\sum_i^n \mid x_i \mid + \sum_i^n \mid y_i \mid} \tag{2.2}$$

where x and y are two n-dimensional arrays, that in our case are two abundance profiles. Bray-Curtis dissimilarity between samples was computed by *metrics.pairwise_ distances*, from the Python package *sklearn*.

Principal Coordinate Analysis

Once distances/dissimilarities between samples (i.e., differences in bacteria abundance) have been computed, they can be positioned (ordinated) in a low-dimensional space (two or three orthogonal axes) to better appreciate how closely related they are to each other. The main assumption in all ordination methods is that there are a limited number of factors that greatly influence distribution and relative abundance of species. The two most commonly used ordination techniques in bacterial ecology are non-metric multidimensional scaling (NMDS) and principal coordinate analyses (PCoA), also known as metric multidimensional scaling. In particular, in PCoA the ordination attempts to faithfully match their original inter-sample distances, providing results that are more readily interpretable[15].

We implemented a simple PCoA algorithm, following the literature[29] for classical metric multidimensional scaling:

1. Set up the matrix of squared proximities $P^{(2)} = [p^2]$.
2. Apply the double centering: $B = -\frac{1}{2}JP^{(2)}J$ using the matrix $J = I - n^{-1}11'$, where n is the number of objects.
3. Extract the m largest positive eigenvalues $\lambda_1 \dots \lambda_m$ of B and the corresponding m eigenvectors $e_1 \dots e_m$.
4. A m-dimensional spatial configuration of the n objects is derived from the coordinate matrix $X = E_m \Lambda_m^{1/2}$, where E_m is the matrix of m eigenvectors and Λ_m is the diagonal matrix of m eigenvalues of B, respectively.

Silhouette score

In order to have an hint on how well the samples are divided into separate groups (based on organ of collection, day of life and probiotic dosage), one could use *Silhouette scores* (SS), tipically used to assess how good an algorithm has clustered the observations. In our case, cluster labels will be the effective name of the groups, known a-priori. We prefer to utilize an indicator like this more than an effective classifier as Discriminant Analysis, in order to be more general since SS describes only the compactness of a cluster and the distance from other ones, so it is not a trained classifier and thus it is more general.

The Silhouette coefficient is calculated using the mean intra-cluster distance a and the mean nearest-cluster distance b for each sample. The Silhouette Score for a sample is:

$$\frac{(b-a)}{max(a,b)} \tag{2.3}$$

To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of, and a is the mean distance between that sample and the others in the same cluster. The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample is more near to samples from another cluster than to those of its own.
We computed Euclidean Silhouette scores on the previously filled PCoA space, with *metrics.silhouette_score*, from the Python package *sklearn.*

PQN (Probabilistic Quotient Normalization)

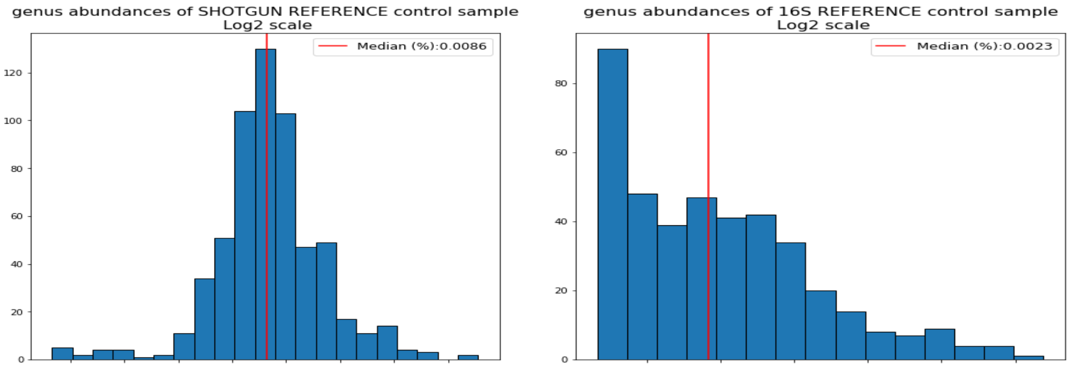
By now we normalized the reads in each sample by dividing the sequences assigned to each taxon by the total sum of reads in that sample and then multiplied by 100, having so obtained the percent abundance of each bacteria in the sample. This Total Sum Normalization (TSN) had a quite simple interpretation, that we used in Section 3.1 and 3.2 and it is generally better than not normalizing data[31], though is not granted that it is the most apt to preserve the real proportion between genera abundances. In particular the massive difference between some samples in terms of DNA sequences and the fact that we want to compare reads collected by different methods that leads to different proportions make us to consider to look for a different normalization that can overcome these complications.

In chromatography, different studies focused on the removal of the so called *size effect*[26], related to different samples volume and/or concentration, where signals do not carry any absolute information about the sample components. If the data comparison has to be performed based on sample fingerprints, then the size effect is undesired, and the shape effect is of main interest. With "shape", we refer to data information which is contained in the ratios between the variables. So far, different normalization methods have been applied to the removal of size effect.
Probabilist Quotient Normalization (PQN) seems to be the best option when we want to focus on the ration between variable[26][27], since it estimates the size effect by the median of the ratios of the elements of an observation and the corresponding elements of a preselected standard fingerprint.
In our case, each chicken is then referred as an array so computed:

$$x_i^{PQN} = [x_{i1}/s_i, ..., x_{in}/s_i] \text{ with } s_i = median(x_{i1}/x_1^{ref}, ..., x_{in}/x_n^{ref}) \tag{2.4}$$

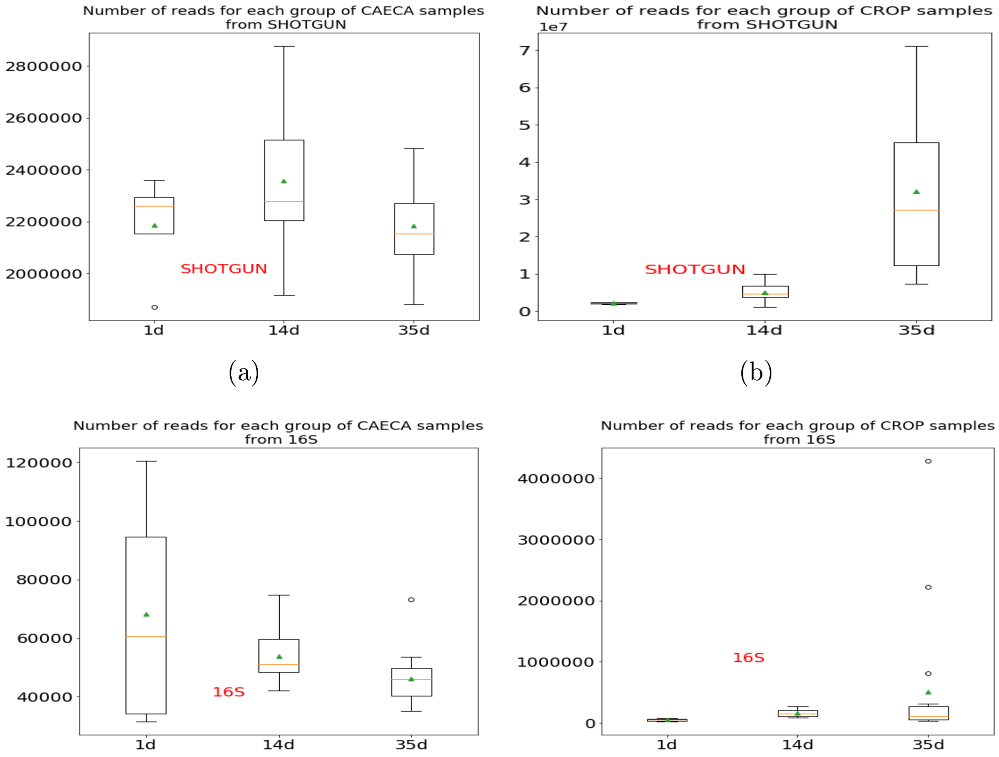
and $x^{ref} = [x_1^{ref}, ..., x_n^{ref}]$ are the value of a "reference", computed as the mean of all control samples collected from both organs but keeping separate the two sequencing methods. Also a "golden standard" made of two sequencing methods together could be made, even if it would appear quite strange, having mixed together information between sequencing methods that supplied a very different amount of sequences. We display abundances of these reference samples in Figure 2.4, but anyway its choice is not crucial [27] for the performances of the method.
The RSAs of the two populations are similar to the real ones in Figure S1 and S2. The main particularity of this artificial set is that since we have averaged the populations, the result is that all genera detected at least once are represented, so there is a consistent portion of very rare genera and their abundance is lower than one (minimum value when counting sequences instead).

Size differences between samples are now strongly decreased, as we see in Figure 2.5, so this normalization is situated mid-way between total sum and none normalization.



(a) (b)

Figure 2.4: Base 2 logarithm of genera abundances in reference control sample from shotgun (a) and amplicon (b) samples. The reference is computed by the mean of each genus among control samples taken from both organs, with e-value thresholds of $[-5, -5]$. Median is shown as percent genus abundance.



(a) (b)

(c) (d)

Figure 2.5: Mean and standard deviation of the number of sequences per chicken used for taxonomic annotations for shotgun samples from caeca (a) and crop (b), normalized by PQN. Mean and standard deviation of the number of sequences per chicken used for taxonomic annotations for 16S samples from caeca (c) and crop (d), normalized by PQN. We are now taking into account all sequences selected with default MGrast parameters, at day 1, 14 and 35.

Notes about the code

After the downloading of abundance profiles from MG-RAST, all data were processed and visualized thanks to around 5000 lines of original code written in Python appositely.

Mostly used packages were Pandas for DataFrame managing, Numpy for vectorized operations, Matplotlib and Seaborn for plotting and Sklearn for statistical data analysis. The most useful functions of the code were built for the purpose of:

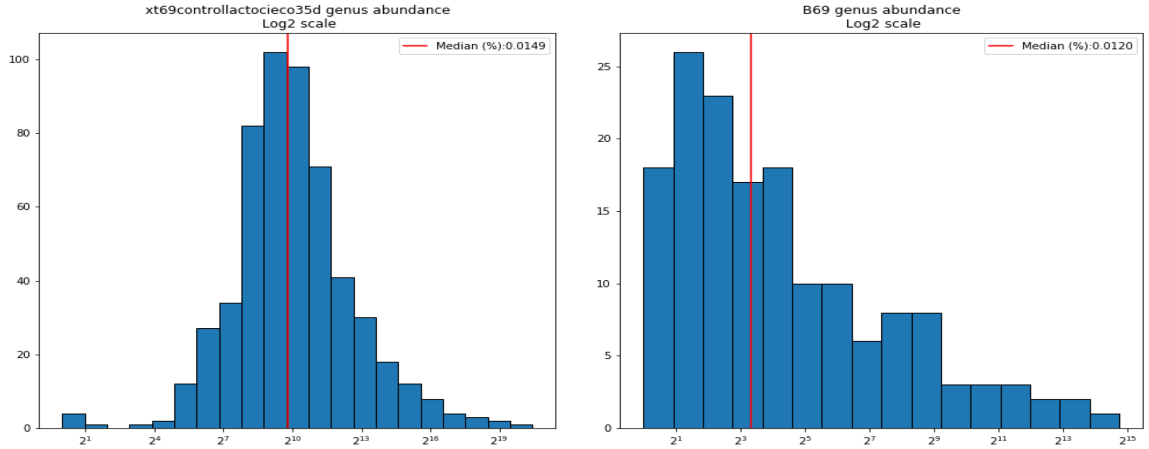
- plotting RSA of genera abundances into Preston plots;
- visualizing stacked bar graphs of genera abundances;
- calculating and visualizing the correlation between 16S and shotgun abundances for the same sample, both as scatter plot and full dataset heatmap;
- computing β-diversities between different sets in order to place the samples into a 2-dimensional PCoA space;
- calculating silhouette scores in order to assess the correspondence between space segmentation and biological markers.

Chapter 3

Results

3.1 Bacteria populations in shotgun and 16S sequencing

First of all, we visualize the overall distribution of organisms in each sample, after removing all the bacteria that MG-RAST did not assign to any category. Hence we consider as abundance the number of reads assigned by MG-RAST to a particular taxon (Section 2.2.1), and we visualize in Figure (3.1) the overall Relative Species Abundance distribution of a sample in the form of Preston plot (Section 2.2.2) by taking the logarithm to base 2 of genera abundances.



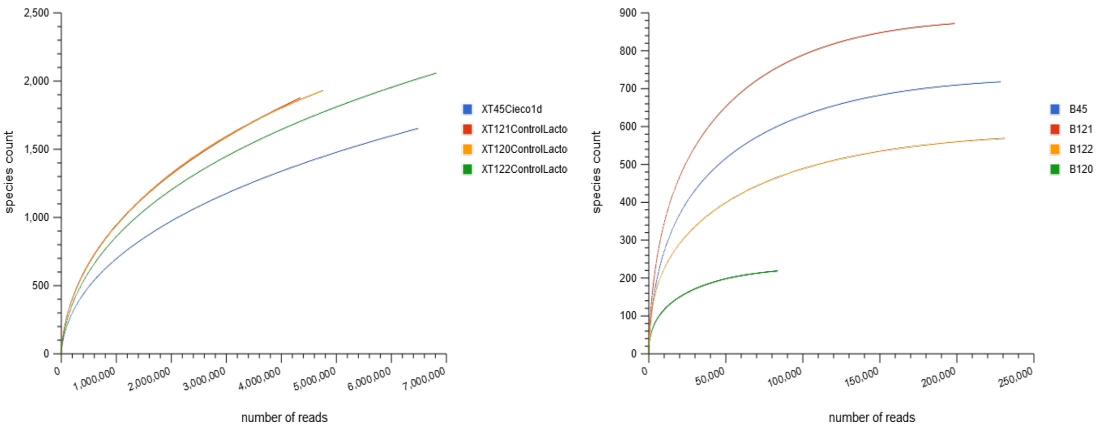
(a) (b)

Figure 3.1: Preston plot of genera abundances with default MG-RAST thresholds in shotgun sample XT69 (a) and its conjugate 16S B69 (b), from group caeca35dayCds. Median is shown as percent genus abundance.

We displayed a randomly selected sample, but the shapes exhibited are the same for most of the other samples, as in Supplementary Figure S1 and S2. We mean that shotgun samples are more bell-shaped, so that most of organisms are neither too rare nor too abundant; populations of 16S sets are positively more skewed. More precisely, we observe that the shape is strictly dependent on the number of sequences (as predicted by Preston[22]): in fact we found that a metagenome collected with shotgun metagenomics from both organs, behaves similarly to a 16S metataxonomic sample if it has less than 200 000 sequences, as we see again in S1 and S3.

Above this threshold of 200 000 reads, the distribution begins to develop a left tail. It seems that since no 16S sample has more than 100 000 sequences, none of these sets exhibits a double-tailed distribution, neither in caeca nor in crop. So we believe that the shape of the Preston plot of the RSA is strictly dependant on the coverage of the sequencing procedure, so it is determined by the sampling resolution of data collection. But for amplicon sequencing the number of sequences seems to provide already a high coverage, as we see in Figure 3.2, so it is likely that the rarest organisms are not accessible with 16S sequencing, even if the number of reads increased.

Anyway, now we may think that shotgun sequencing is able to detect rarer genera than 16S rRNA sequencing, which only recovers the most abundant genera. This would



(a) (b)

Figure 3.2: Rarefaction curves of samples from day1caeca, one of the groups with less shotgun sequences. We see that shotgun sequencing (a) is likely to increase its number of detected genera by adding more sequences, while amplicon sequencing (b) ha already reached a stable value.

explain why the Relative Species Abundance distribution in 16S is basically a Log-Series, while shotgun sequencing begins to portrait the shape of a bell (Log-Series[22]).

We can only infer a few conclusions looking directly at the sequences counts of genera, because the two methods rely on a very different amount of reads, so we choose to normalize each sample by the sum of reads and return a percent value.

A first interesting observation can be made about the overlapping of genera detected by the alignment of reads collected by both methods. So we consider those genera that are frequently (in median) found only by a single method and not by the other one, and compute their mean abundance among all samples.
Of course, we do not know a priori if the abundance of a bacterium is supposed to be constant among all sample; indeed we strongly believe the opposite, since samples are exposed to different treatments, so the operation of taking the mean abundance of a particular taxon is likely to give quite biologically inconsistent information by itself. However, for the sake of our actual aim of determining if species observed only by a method are rare or not, median information is quite significant anyway.

Cum.abund.(%)	in shotgun	in 16S
Gen. only in shotgun	26.39	0
Gen. only in 16S	0	12.43
Gen. in both	73.61	87.57

	# of genera
Only shotgun	440
Only 16S	52
Both	94
16S/shotgun	0.118

Table 3.1: CAECA samples: Cumulative percentage of the average abundance (left) and number (right) of genera detected only in shotgun samples, only in 16S samples and those detected by both methods on the same chicken. With 16S/shotgun we mean the ratio between genera detected only in metaxonomic sets and those found only in metagenomic sets.

From Tab 3.1 we acknowledge that the most consistent component (in terms of abundance) of the chickens metagenome is identified in the same way by the alignment of sequences taken from the two methods. In fact, the 94 genera they usually find in common in the same caeca of chicken, detain the biggest part of the abundance. We see that in shotgun samples there a lot more genera, and that the cumulative abundance of those detected only by this method is not negligible.

Cum.abund.(%)	in shotgun	in 16S
Gen. only in shotgun	11.44	0
Gen. only in 16S	0	2.63
Gen. in both	88.56	97.37

	# of genera
Only shotgun	331
Only 16S	35
Both	60
16S/shotgun	0.106

Table 3.2: CROP samples: Cumulative percent abundance and number (on average) of genera detected only in shotgun samples, only in 16S samples and those detected by both methods on the same chicken. With 16S/shotgun we mean the ratio between genera detected only in metataxonomic sets and those found only in metagenomic sets.

In Tab 3.2, we see that also in crop of chickens the majority of microbiome is recognized similarly by both sequencing methods, and the ratio between genera identified with metataxonomics and with metagenomics seems to be similar to caeca samples. Actually we suspect this similarity to be a coverage-based artefact. In fact, crop populations in this study sometimes suffer from low coverage (Supplementary FigureS3) and, for this reason, their RSA becomes similar to that of 16S samples. If we only consider sets with a number of sequences higher than the lowest threshold that consent a RSA typical of shotgun samples ($seqs_{min} = 2 \times 10^5$ from S1), caeca remains with 36 samples instead of 40 and crop with 16 samples instead of 38, so that while the ratio of 16S/shotgun detected genera remains almost the same for caeca, it falls to 0.04 for crop instead.
So, at similar minimum coverage for caeca and crop, the discrepancy between 16S and shotgun resolution is enhanced, in particular in crop of chickens, where very few genera (23) are individuated only by 16s while around 440 genera are detected in shotgun samples for both organs.

3.1.1 Correlation between conjugate sets and fitting 16S vs shotgun abundance profiles

To compute correlations between samples, we considered only common taxa in each pair of compared sets, normalizing before the removal in order not to inflate counts. In this case, we consider each genus which is not present in both sets as a not-measured variable because of under-sampling, and so we exclude it from the analysis. We mean that we do not believe that a null value measured in a shotgun sample is likely to be not null in its 16S coupled.
Furthermore, having only not-null values permits more operations to be computed, as quotient normalization or logarithmic scaling.

Now, we take the base 2 logarithm of not normalized data, so directly of the number of reads, and compute a scatter plot of the abundances of a set and its conjugate, that allows a visual guessing of the underlying correlations; for the sake of brevity, we show only an example in Figure 3.3.

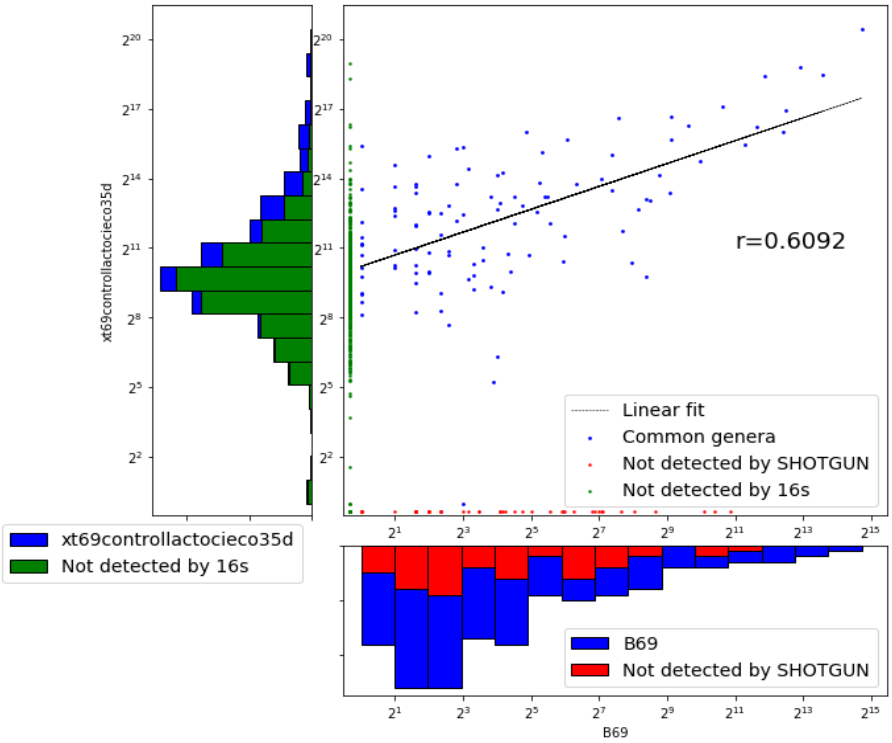


Figure 3.3: Scatter plot of genera abundances of sample XT69 and its conjugate B69, from group crop35dayCds. Correlation coefficients are computed by Pearson only on the common genera between conjugate sets, with default e-value threshold ($eval_T = -5$). Green and red observations do not count on Pearson coefficient's computation.

From Figure 3.3 we acknowledge that the trend between logarithms of genera abundances seems quite linear, even if Pearson coefficient is quite low, probably because of huge variability for rarest species (left-bottom corner). Anyway we exclude at the moment a non-linear dependence because even non-parametric statistics as Spearman correlation coefficient return low values of correlation.

Furthermore it's noticeable that getting to a less fine taxonomic level brings to higher correlations on average, as in Figure 3.4.

At higher taxonomic levels, as one can expect, the information is sort of averaged on more general taxa and phylogenetic distance between organisms belonging to different taxa becomes higher in terms of nucleotides sequences, so taxonomic misannotations and mistakes are more rare and taxa are less noisy.

Now, since we observed linear correlations between logarithm of genera abundances, another interesting observation is that we can try to fit this relation. We mean that with a linear fit we can get the parameters of:

$$y = mx + x_0 \tag{3.1}$$

where y is the logarithm of number of shotgun sequences and x of 16S ones. Then we can

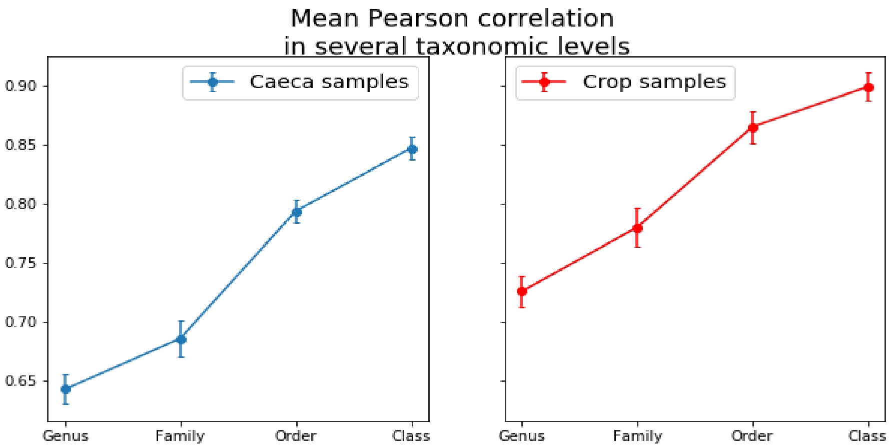


Figure 3.4: Mean Pearson's correlation between bacteria abundances at different taxonomic levels, computed on conjugate samples.

guess how many shotgun sequences correspond to the 16S counts with Formula (3.2).

$$Y = X^m 2^{x_0} \tag{3.2}$$

where in (3.2) we used Y for shotgun sequences and X for 16S ones.

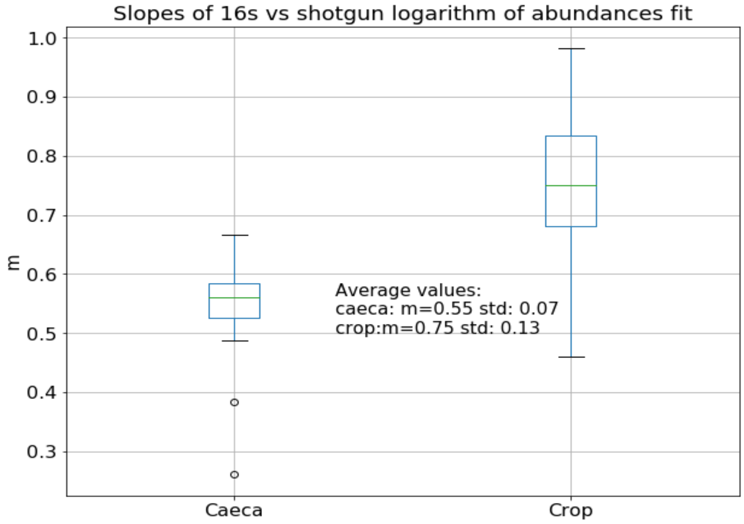


Figure 3.5: Boxplot of slopes of m according to Formula 3.1.

From Figure 3.5 we see that the slopes of the linear fit are not so variable for the same organ, even if it would be daring to asses a reliable equation to reconstruct a shotgun profile of a metagenome from its 16S abundances. Also, the non-linearity between real number of sequences produced by both methods has to be further investigated, before utilizing this fit. For now, instead of transforming the minimum abundance of 16S into its shotgun equivalent in order to set the same resolution of genera detection for both methods, we prefer to divide a shotgun sample in two parts: that composed of rarest genera that are not usually detected by 16S (that we call $\mathrm{X}_{\mathrm{left}}$, at the left of the first tertile) and that made of genera more abundant than the first tertile (called $\mathrm{X}_{\mathrm{right}}$).

So, for example, if we consider a full 16S set and its conjugate X_{right}, we obtain a scatter plot as in Figure 3.6.

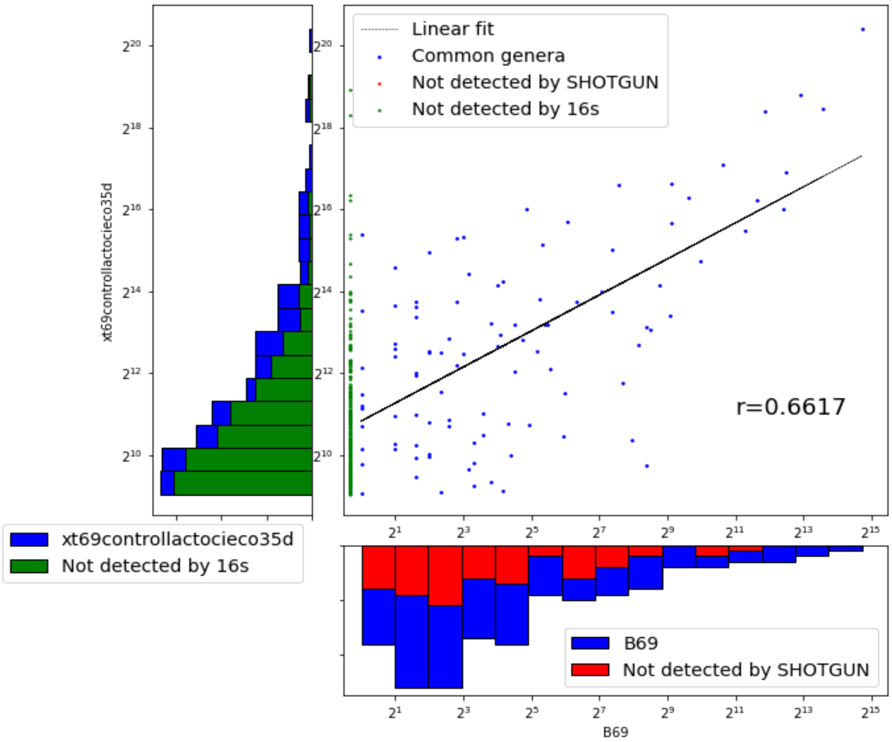


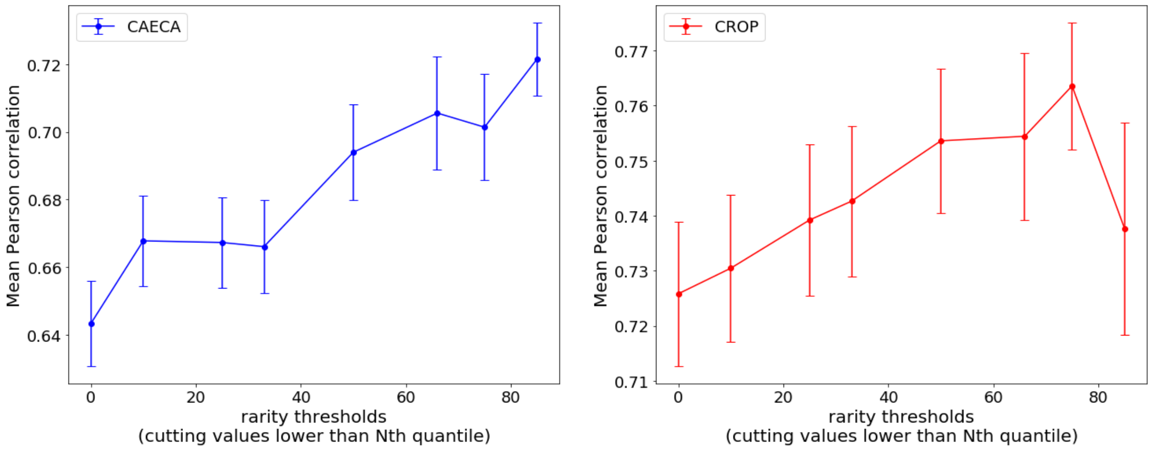
Figure 3.6: Scatter plot of genera abundances of sample XT69 and its conjugate B69, from group crop35dayCds. Correlation coefficients are computed by Pearson only on the common genera between conjugate sets, leaving in shotgun sample only those genera whose abundance is higher than the first tertile of the set. Green and red observations do not count on Pearson coefficient's computation.

As we can see, now the shotgun shape of X_{right} is like the right portion of the full shotgun population in 3.1(a), so more similar to a 16S sample as in 3.1(b), because we cut off rarest species.

It was not only a lucky accident that made the correlation between conjugate sets to increase passing from all genera (Figure 3.3) to only those rare at least as shotgun first tertile (Figure 3.6), because, on average, this behaviour is followed by other samples. In fact, it seems that the exclusion of rarest species increases the correlation as in 3.7, at least until too much information is lost.

Here, along x-axis we increased the quantile of abundance we cut off genera below. So if we do not consider rarest genera (moving to right), the accordance between shotgun and 16S is generally better, unless we excessively raise the threshold and Pearson correlation begins to be computed on too few genera. It means that mistakes on taxonomical annotations or wrong abundance estimation are more frequent on rarest genera, while they tend to decrease when we consider only abundant genera. In particular, setting the shotgun abundance threshold to the first tertile of the population seems a good choice (rar.tr=33 in Figure 3.7), that is the value that makes the RSA of the shotgun sample more similar to its 16S conjugate.

In terms of ratio between genera detected by metataxonomics in comparison to metage-



(a) (b)

Figure 3.7: Moving to right along x-axis we cut off genera in shotgun samples having lower abundance than increasing quantiles (from 0 to 80). On y-axis we have the average correlation between conjugate sets, with the standard deviations from mean as error bars.

nomics, raising the minimum abundance of shotgun samples leads to the graph in Figure 3.8.

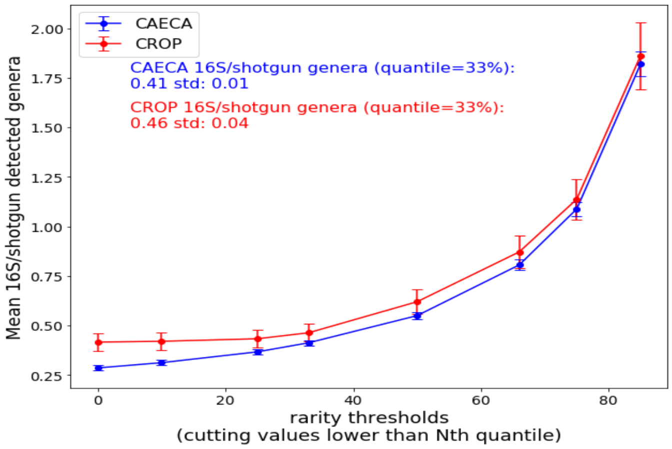
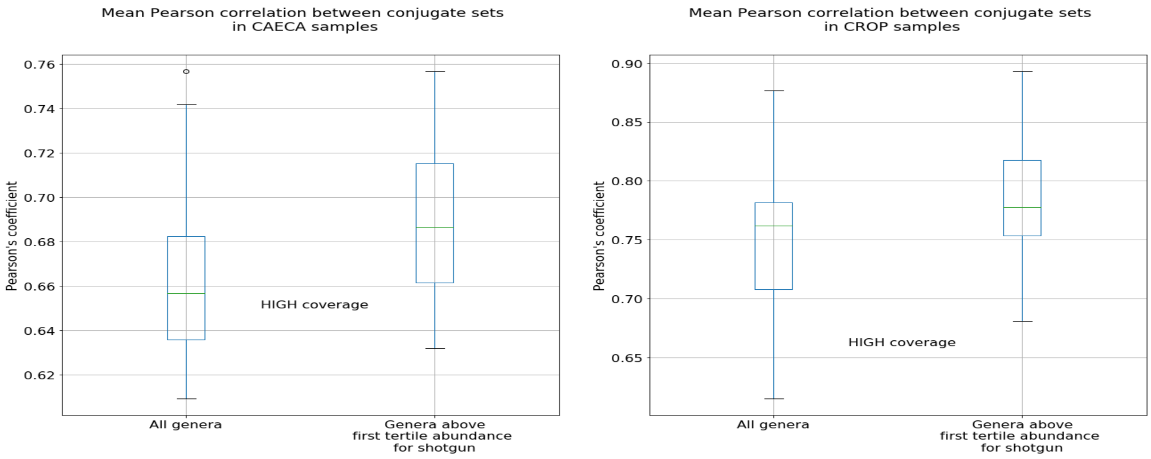


Figure 3.8: MMoving to right along x-axis we cut off genera in shotgun samples having lower abundance than increasing quantiles (from 0 to 80). On y-axis we have the average 16S/shotgun number of detected genera, explicitly annotated for genera more abundant than the first tertile.

We see that even if raising the thresholds on number of shotgun sequences increases linearly the 16S/shotgun number of detected genera similarly in both organs, setting this threshold to the first tertile ($rar.tr. = 33\%$) keeps this ratio lower than 1, as we see from the annotations on Figure 3.8. So, even if though we reduced the shotgun population to the right tail of the distribution, this sequencing technique still detects more genera than the other.

This means that abundance profiles produced by different sequencing methods do not correlate well, even if we consider only genera detected by both techniques; thus we have to determine the portion of metagenome were the noise or mistakes are minimized and we do that by raising the rarity threshold; in particular keeping only genera whose abundance is above the first tertile of the shotgun set seems a valid choice. Furthermore, even if there is a good correlation on abundant genera detected by both methods, a lot more genera are detected in shotgun samples exclusively.

The increase in correlation is particularly enhanced if we keep only shotgun samples with high coverage, so with a number of sequences higher than 200 000. In this case the gain goes from 0.663 ± 0.006 to 0.691 ± 0.005 for caeca and from 0.747 ± 0.015 to 0.784 ± 0.013 for crop, as in Figure 3.9.



(a) (b)

Figure 3.9: Mean correlation between conjugate sets before and after the exclusion in shotgun samples of genera with abundance lower than the median, for crop samples. Samples with number of sequences lower than $seqs_{min} = 2 \times 10^5$ are displayed on the left and those with higher values on right.

This means that, with low coverage, a shotgun sample consists almost totally in its $\mathrm{X}_{\mathrm{right}}$ (most abundant genera) so correlation is not really increased by the cut-off of rarest bacteria. So if our aim is to compare the detection sensitivity of the two sequencing techniques, it could be a right decision to exclude from the analysis the samples with low coverage, in particular in crop. However in this way we would loose full groups of samples (ciecο1day, crop1day, crop14day) for our further measurements and biologically driven analysis would loose consistency.

3.1.2 Not overlapping genera abundance

According to what said so far, we expected genera detected only by one of the two methods to be quite rare. Yet, this is not completely true. We intend to show the percent abundances of most populated genera detected by one of the two sequencing techniques, as in Tab 3.3 for caeca samples, referring to those genera already counted in Tab 3.1.

Found only in shotgun	abundance (%)	**Found only in 16S**	abundance (%)
Bacteroides	6.127	*Butyricicoccus*	4.201
Subdoligranulum	3.113	*Hespellia*	2.690
Anaerotruncus	1.336	*Robinsoniella*	2.427
Holdemania	1.281	*Sarcina*	0.345
Dorea	1.158	*Aneurinibacillus*	0.235
Coprococcus	0.952	*Lachnospira*	0.170
Providencia	0.454	*Tissierella*	0.166
Fusobacterium	0.373	*Desulfocaldus*	0.165
Thermoanaerobacter	0.347	*Gordonibacter*	0.131
Caldanaerobacter	0.318	*Pseudobutyrivibrio*	0.128

Table 3.3: Percent abundance of the ten most abundant genera that on median are detected only by shotgun (left) and only by amplicon sequencing (right), in caeca samples.

We chose to consider the genera that on median appeared only in one of two conjugate samples, but we could have chosen those that are never detected by one of the two techniques and we would get the same results, except for Bacteroides who would not appear in that case because it is detected once in an amplicon sample (in B121 with a relative abundance of 0.0063%).
Several displayed genera are not so rare, because they actually cover up a consistent portion of the overall sample population.
Integrating this information with that of Tab 3.1, we can say that MG-RAST pipeline seems to recognize, with both methods, only the most abundant genera (94 on average, with default thresholds), that usually represent the portion of metagenome where the correlation between conjugate sets is higher. As for the detection of rarer species instead, there is a significant difference between amplicon and shotgun sequencing because the latter seems to identify a number of rare genera that is four times larger than the number of rare genera detected by both methods in common. It is particularly interesting to see that this behaviour does not explain the failed detection of abundant genera as Bacteroides or Butyricicoccus, that can not be due to undersampling, since they are very abundant.
Nevertheless, even if it is quite clear, at genus level, that shotgun samples collect more species with less individuals than its rival technique, we can even take a glance at less fine taxonomic levels. In fact, the missed detection by MG-RAST of abundant genera, such as Bacteroides and Subdoligranulum in amplicon samples or Butyricicoccus, Hespellia and Robinsoniella in shotgun samples is quite interesting and triggers some doubts on the possibility of a trustworthy analysis at genus level for both methods. Repeating the measurements in Tab 3.1 and 3.3 at family level, we obtain Tab 3.4 and 3.5.

We see that, at this level, we find only few families in 16S samples that are not identified in shotgun samples too, while shotgun sequencing seems to find a lot of original bacteria. At this taxonomic resolution, we are quite confident that the misclassification

	Caeca: # of families
Only shotgun	147
Only 16S	7
Both	60
16S/shotgun	0.048

	Crop: # of families
Only shotgun	167
Only 16S	3
Both	33
16S/shotgun	0.018

Table 3.4: Number (on average) of families detected only in shotgun samples, only in 16S samples and those detected by both methods on the same chicken. With 16S/shotgun we mean the ratio between families detected only in metataxonomic sets and those found only in metagenomic sets. We are taking into account only high coverage samples from both organs.

rate is quite insignificant. However Tab 3.5 shows that the Bacteroidaceae family remains undetected by amplicon sequencing; for this reason, we think that the differences between two conjugate samples cannot only be explained by statistical paucity of sampling and that errors in the recognition of some important genera are made too.

Found only in shotgun	abundance (%)	**Found only in 16S**	abundance (%)
Bacteroidaceae	6.023	*Clostridiales Fam.XIV.Incertae Sedis*	0.095
Fusobacteriaceae	0.738	*Thermoactinomycetaceae*	0.079
Listeriaceae	0.271	*Planococcaceae*	0.062
Chlorobiaceae	0.191	*Anaeroplasmataceae*	0.052
Porphyromonodaceae	0.174	*Clostridiales Fam.XII.Incertae Sedis*	0.036
Geobacteraceae	0.172	*Dietziaceae*	0.017
Prevotellaceae	0.166	*Holosporaceae*	0.013
Burkholderiaceae	0.156	*Sporolactobacillaceae*	0.005
Xanthomonodaceae	0.151	*Spiroplasmataceae*	0.003
Rhodobacteraceae	0.139	*Desulfonatronumaceae*	0.0011

Table 3.5: Percent abundance of the ten most abundant families that on median are detected only by shotgun (left) and only by amplicon sequencing (right), in caeca samples.

Analogous measurements on crop samples are in Supplementary Tab S4, but in this case genera detected only by a method out of two are less abundant, in particular those found only in 16S samples and this is not caused by the lower coverage in crop samples. Bacteroidaceae family still remains an exclusive of shotgun sequencing in almost all samples.

3.2 Genera abundances in each group of samples

Since we dispose of a well structured database, with metadata that enable to distinguish the dosage of the probiotic supplemented to chickens' water and the day of treatment, it is interesting to make some comparisons between groups of samples based on the abundance of genera in each one. In particular, we have identified three interesting types of comparison, looking at the data descriptions: the first one is to compare the populations of the microbiome in the two organs we collected metagenomes from, the second is to study the differences in bacteria populations in chickens at day 1, 14 and 35, and the third is to compare at the same day of treatment, the microbiota in chickens subjected to high probiotic dose (H) against control samples (C).

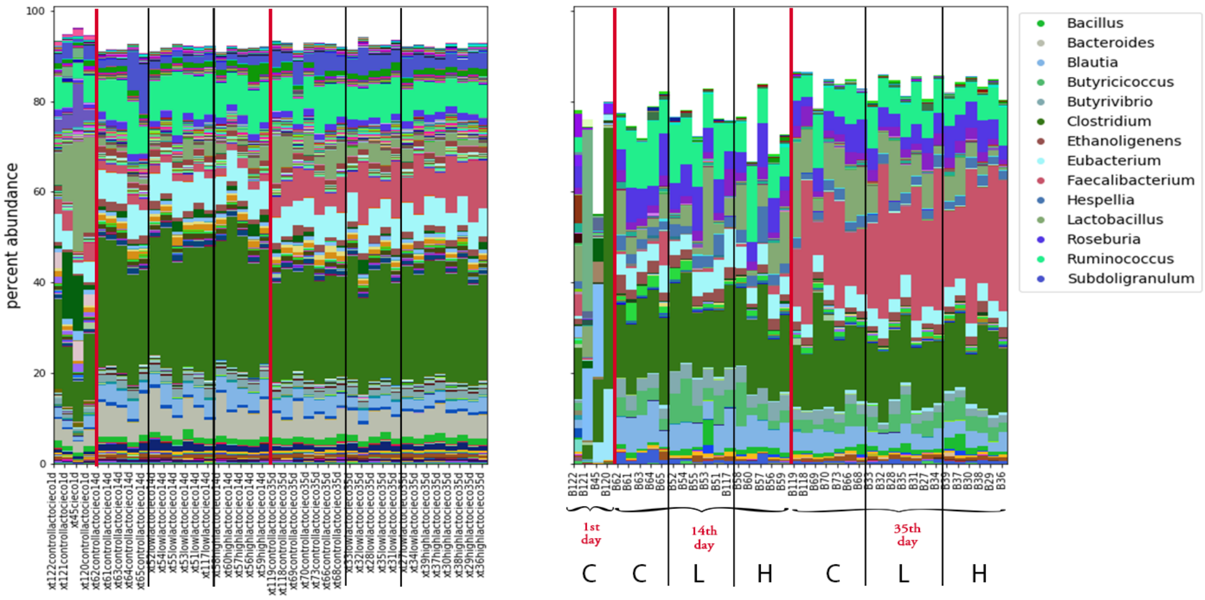


Figure 3.10: Genera abundance of caeca samples at day 35, 14 and 1, under Low, High and absent probiotc supplementation (Control), for shotgun and 16S sequencing. E-value thresholds were set to MGrast defaults [−5, −5]. The height of single coloured portion of a bar represents the percentage of genus abundance in that sample. Shotgun samples are on the left, 16S samples on the right. In the legend, we labelled only those genera whose abundance exceeded 1% on average.

In caeca samples, the most evident color-based separation between samples is on the day of life of chickens. In particular, 1st day samples (on the left of both images in Figure 3.10) are substantially different from other ones, and we can notice a difference between 35th and 14th day samples too, especially for 16S samples (right box). The biggest difference between days, by sight, is that *Faecalibacterium* increases with time, especially in the from day 14 to 35.
It's hard to notice, at the same day, differences between chickens treated differently, as if a different dose of the probiotic did not imply differences in the population of the gut, not even in comparison to control samples.

In general, as previously seen on average, the number of detected genera on each day is always higher for shotgun samples (Tab 3.7 and Tab 3.8), even on 1st day(Tab 3.6) although is very poor of sequences (as in Tab 2.1).

Shotgun genera	502
16S genera	193
Common genera	122
Total genera (shotgun+16S)	573

Table 3.6: Number of genera for all samples from caeca at 1st day.

Shotgun genera	560
16S genera	264
Common genera	165
Total genera (shotgun+16S)	659

Table 3.7: Number of genera for all samples from caeca at 14th day.

Shotgun genera	563
16S genera	278
Common genera	172
Total genera (shotgun+16S)	669

Table 3.8: Number of genera for all samples from caeca at 35th day.

Anyway, the number of total detected genera increases over time (especially with respect to 1st day, even considering only samples collected by amplicon sequencing), but we suspect this behaviour has not much to do with the treatment because even analysing only control samples, the same trend is achieved.

For sample taken from crop of chickens, the landscape seems radically different, with a single genus (*Lactobacillus*) that fills alone the biggest part of the population at 14th and 35th day.

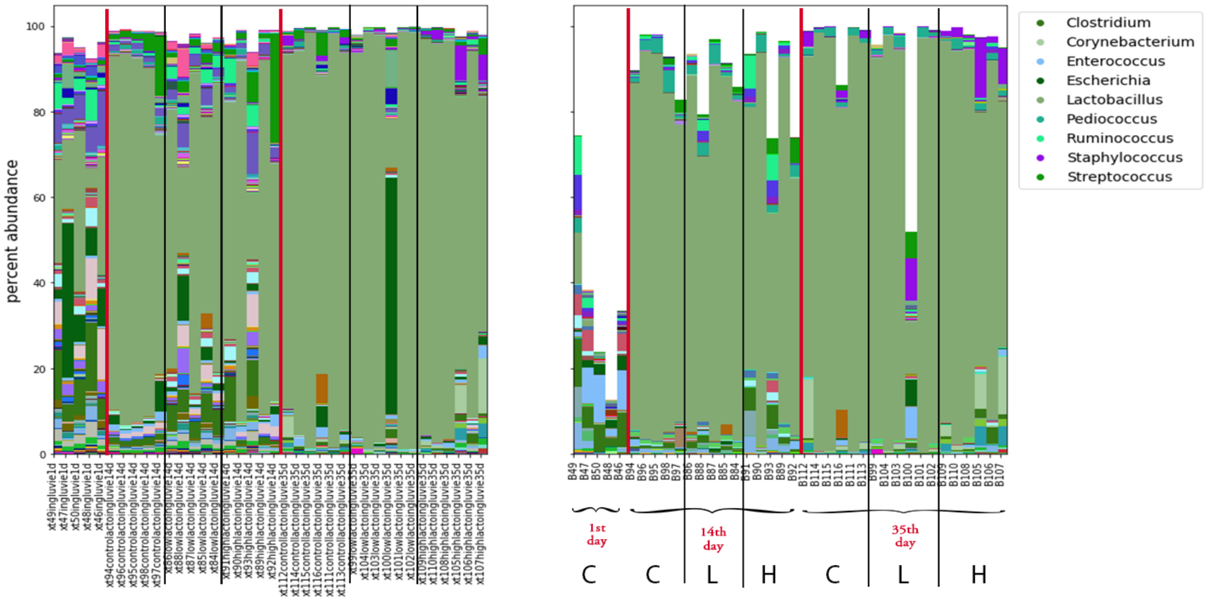


Figure 3.11: Genera abundance of crop samples at day 35, 14 and 1, under Low, High and absent probiotic supplementation (Control), for shotgun and 16S sequencing. E-value thresholds were set to MGrast defaults $[-5, -5]$. The height of single coloured portion of a bar represents the percentage of genus abundance in that sample. Shotgun samples are on the left, 16S samples on the right. In legend are labelled only those genera whose abundance exceeded 1% on average.

This particular bacterium proliferates a lot after the first day of sampling, while other evident patterns are not highlighted, so that we are not even able to distinguish between day 14 and 35 only by sight. Furthermore, we notice that first day samples taken from crop with amplicon sequencing are full of taxonomically unassigned bacteria, so that this set may not be really trustworthy.

3.3 PCoA and space segmentation

Since some driving patterns were evident in the stacked bar graph of bacteria abundances in Figure 3.10, we are willing to introduce some measurements of similarity between samples; this approach is more informative than visualizing heatmaps of Pearson's correlation coefficients between samples, that we show in Supplementary Section 3.5.
For example, if we want to overcome the exclusion of those genera that were not in common between pairs of correlated sets and so were ignored by Pearson coefficients, we can set up new measurements, based on β-diversity computation, introduced in chapter 2.2.2.

We can reduce the features space to two Principal Coordinates, as we do in Figure 3.12 and 3.13 for the entire dataset.

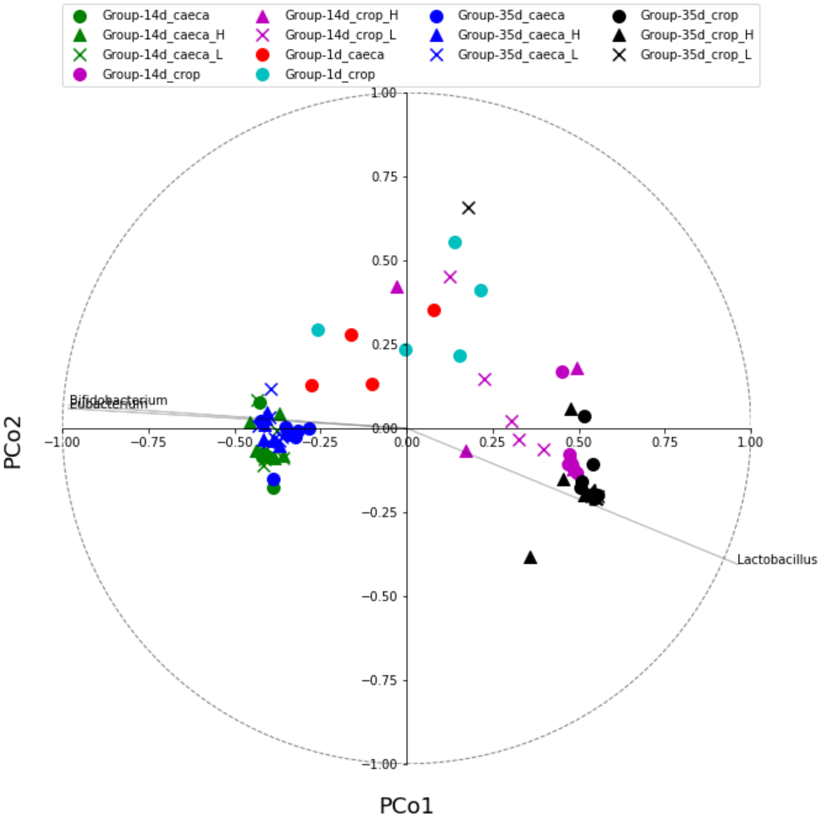


Figure 3.12: PCoA of genera abundances of all datasets normalized by total sum, sequenced by shotgun. Caeca samples follow ageing with shades R→G→B and crop samples with C→M→K.

Aside from sequencing method, the main separation in both figures is between samples collected in different locations (caeca (RGB) and crop (CMK)), especially at day 14 and day 35; for 1st day samples we see they are placed into a central column, parallel to y-axis, disregarding of the organ, so we can say that first day bacteria are similar in both caeca and crop of chickens. The split between samples collected from different location loads mostly on Lactobacillus (more abundant in crop) and Eubacterium, Bifidobacterium and Ruminococcus (more abundant in caeca).
We can see a clear split between 14th and 35th day in 16S samples taken from caeca (G vs B in Figure 3.13), while day 14 and 35 are quite crunched together in crop samples and

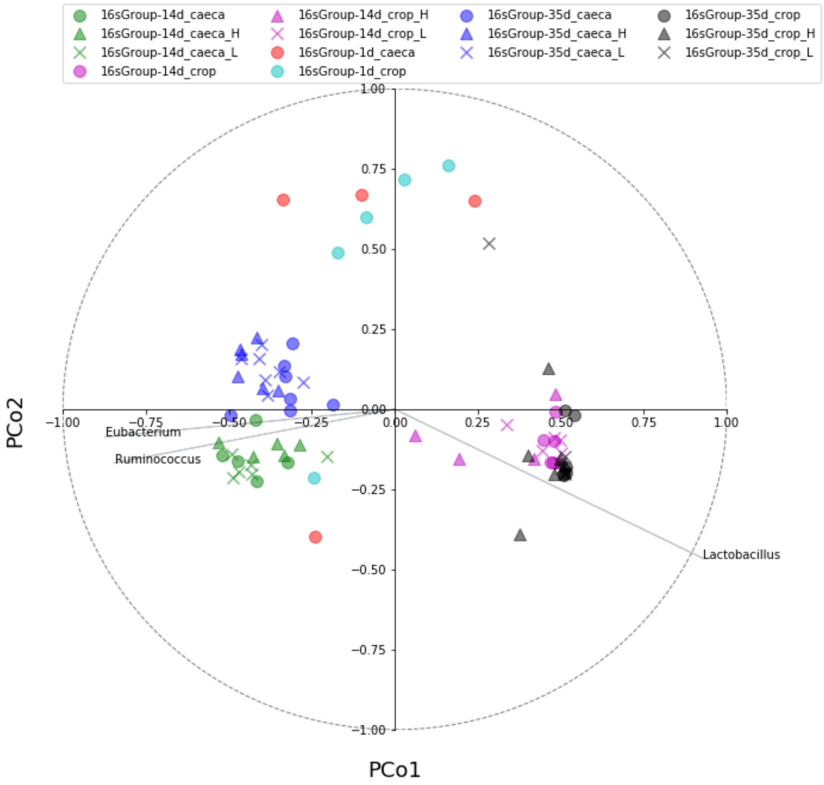


Figure 3.13: PCoA of genera abundances of all datasets normalized by total sum, sequenced by amplicon sequencing. Caeca samples follow ageing with shades R→G→B and crop with C→M→K.

in shotgun sequencing in general. It is quite difficult to detect any dosage-based pattern at this resolution, because so many other relevant differences are more highlighted. Even if we consider higher taxonomic levels as in Supplementary Figure S5, the situation does not get any better, so we continue to focus at genus level.

We can even add all samples in a unique plot, by operating a Procrustes rotation that tries to fit 16S samples to shotgun ones[30], trying to minimize M^2, reported in Figure 3.14. For brevity, we show always sample normalized by total sum because it is the scenario where lowest M^2 is achieved, but results about other normalizations are provided too.

In order to report some quantitative parameters of the spacial separation in the reduced features space, one can simply calculate the mean Silhouette Score (SS) of the samples, remaining at genus level. Silhouette scores are widely used to determine the goodness of a clustering procedure, but in our case we can simply use the true labels of the groups as cluster labels. Depending on how fine we want the analysis to be, groups labels can be those of organs ([caeca,crop]), of days ([14d,35d]) and of dosage ([Control,High]). We do not consider 1st day samples in the day-based segmentation because we believe they are different from other days mainly because of the paucity of sampling, so we prefer to compare sets with similar size as those of day 14 and 35.

The aim of this study is to assess which sequencing method is more trustworthy for the recognition of biological factors (such as organ of sampling, ageing or probiotic

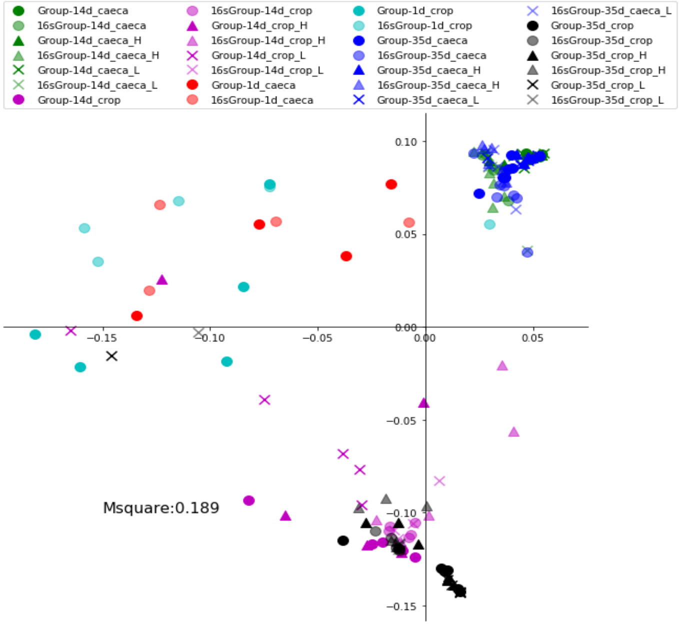


Figure 3.14: Procrustes rotation on PCoA of genera abundances of all datasets normalized by total sum. Caeca samples follow ageing with shades R→G→B and crop with C→M→K.

supplementation). We are going to compare the correct classification rates on 16S and shotgun samples separately. Furthermore, we would like to know which is the predictive component of a shotgun metagenome, since it contains a lot of more species than its counterpart.
We already saw, as in Figure 3.6, that with an opportune cut we can split each shotgun sample in two populations: one with rarer genera (left portion of the genera distribution) and one referring to the right portion of the distribution. In the first set (X_{left}), since we get to have the rarest species detected by shotgun, most of them are rarely individuated in 16 samples. In the second set instead (X_{right}), a good portion of genera is detected by both sequencing methods and since their abundance is strongly higher than that of X_{left} genera, X_{right} statistically covers X_{left} when they are joined together in X_{left}, so X_{right} accuracies are almost equal to those of the total shotgun set, yet not shown in the results.

Organ-based space segmentation

As we already pointed out, the samples are split in the PCoA space quite accordingly to the sequencing location, so silhouette scores for organ recognition are quite high in Tab 3.9.

At this resolution, it is easy to recognize the organ we collected the metagenome from, in both 16S and shotgun samples, so there is no real difference that can be highlighted between a full shotgun set and a 16S one, for both TSN and PQN. For not normalized data, SS is significantly lower in shotgun sample than 16S ($P < 0.05$), but this is probably due to great differences in sample size in not normalized shotgun data that make space segmentation not suitable. Therefore, we suppose that good scores for none normalization

SS	**TSN**	**PQN**	**NONE**
$\mathbf{X_{right}}$ **(Shotgun)**	0.816	0.735	0.696
B (16S)	0.811	0.738	0.767
$\mathbf{X_{left}}$	0.715	0.739	0.680
p_{value} $\mathbf{X_{right}vsB}$	0.86	0.95	0.05

Table 3.9: Euclidean Silhouette Score on Bray-Curtis PCoA of genera abundances for organ recognition task, on data from day 14 and 35 normalized by total sum (left) and PQN (right). Dataset were shotgun data (X), 16S (B), shotgun genera with abundance lower than the first tertile (X_{left}), shotgun genera with abundance higher than the first tertile (X_{right}). E-value thresholds were set to default $[-5, -5]$ for both organs.

are more due to the fact that, in our data, different groups have different sizes and could not be reproducible in other studies; in literature not normalizing data does not seem to be a good option[31], so we continue our studies focusing only in TSN (or PQN) that seems to perform well.
Furthermore, we learn that rare species are quite informative even alone (look at the SSs of Xleft).

Day-based space segmentation

To observe a day-based space segmentation, instead of taking into account all samples together as in Figure 3.12 and 3.13, it is better to study crop and caeca samples separately, in order to have only two classes (day14 and day35), if we overlook dosage for the moment, as in Figure 3.15.

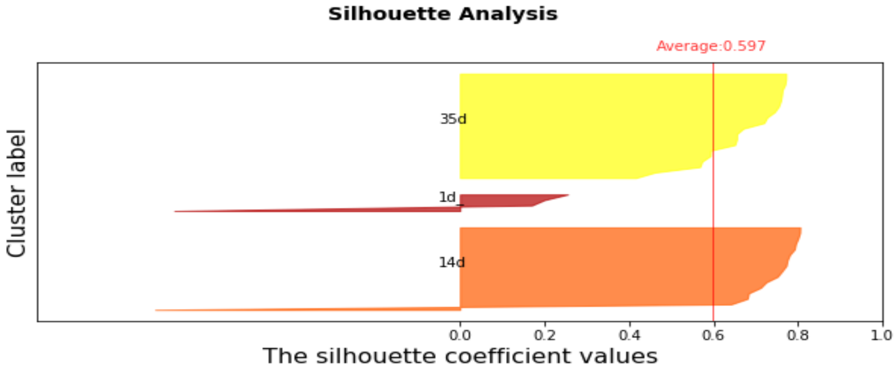


Figure 3.15: SS of genera abundance in shotgun samples taken from caeca and normalized by total sum, for day recognition task.

In Tab 3.10 we show SS scores for the day recognition task, computed on shotgun sets (X), 16S sets (B) and rare shotgun genera (Xleft).

From the low silhouette score on Xleft we learn that rarest genera in caeca are not so crucial for day recognition purpose. Instead, in crop samples, shotgun sequencing has slightly better space segmentations than 16S ($P < 0.22$) and rarest genera provide useful information, in fact the score of Xleft by itself is higher than both 16S and X_{right}. The evident space separation between day 14 and 35 in Xleft ($S_{score} = 0.520$) is not due to the fact that the two groups of samples are different in size, but it is because the rarest

SS (TSN)	**Caeca**	**Crop**
$\mathbf{X_{right}}$ **(Shotgun)**	0.524	0.274
B (16S)	0.538	0.166
$\mathbf{X_{left}}$	0.271	0.520
p_{value} $\mathbf{X_{right}vsB}$	0.73	0.22

Table 3.10: Euclidean Silhouette Score on Bray-Curtis PCoA of genera abundances for day recognition task (14d vs 35d), on data normalized by total sum. Dataset were shotgun data (X), 16S (B), shotgun genera with abundance lower than the first tertile (Xleft), shotgun genera with abundance higher than the first tertile (Xright). E-value thresholds were set to default $[-5, -5]$ for both organs.

genera play an important role in the detection of ageing in crop samples; in fact we find a $S_{score} = 0.865$ even for the segmentation of day 1 versus day 14, similar in size (not shown in tables).

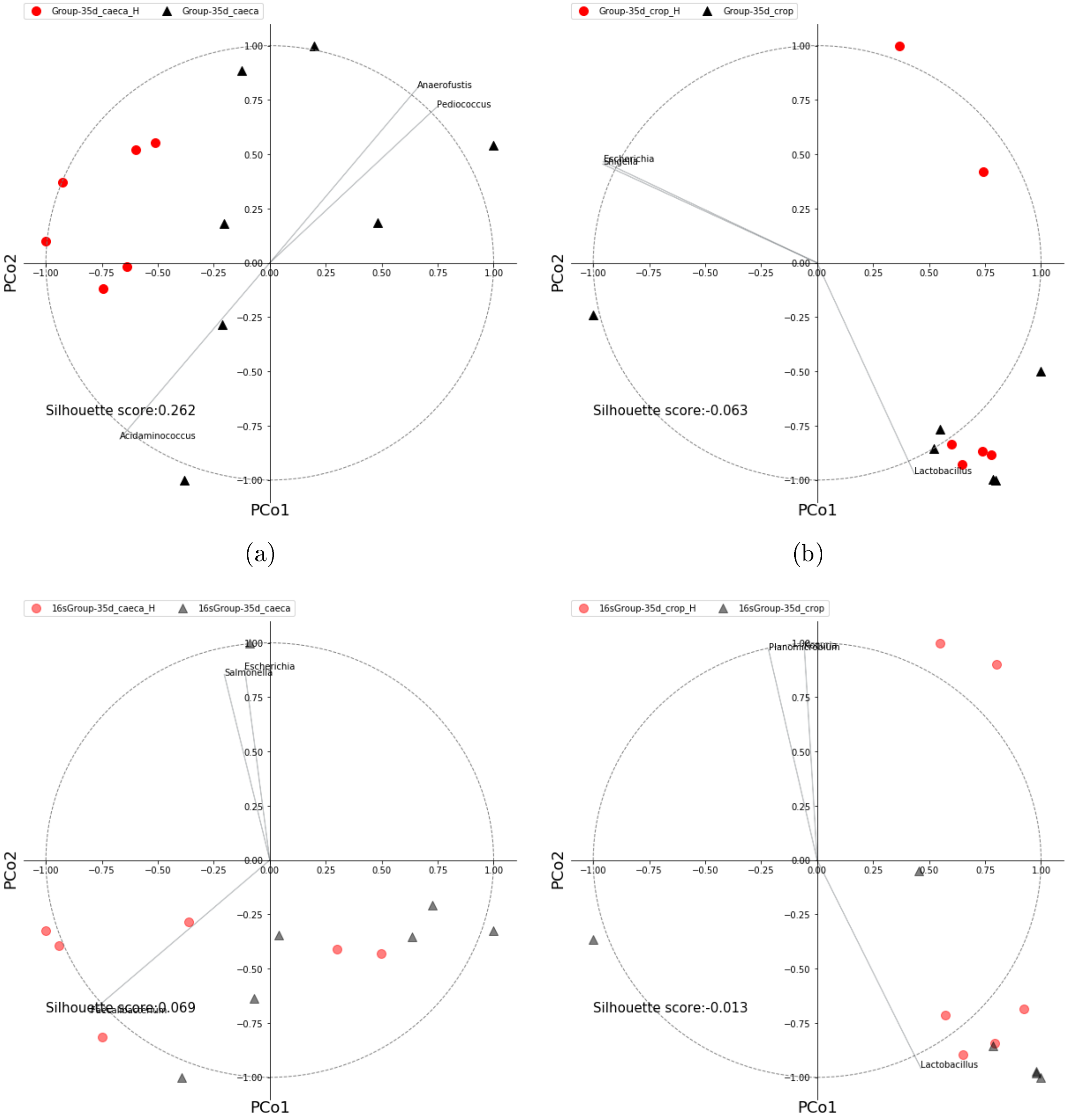
In the next paragraph we try to score the goodness of treatment-based space segmentation, that could be more difficult (looking at Figure 3.10 and 3.11).

Treatment-based space segmentation

As we said, while the passing of time seems to clearly discriminate between chickens, the dosage induces less relevant partitions instead (Figure 3.12 and 3.13). So it could be useful to look for a finer visualization, by considering only samples collected at the same day of treatment, in order to assess dosage-based distinctions. The procedure entails again to plot the samples on the reduced features plane, in order to compare their spacial separation with their groups of membership, that are now restricted to null dosage (Control) and High dosage, since it is quite hard to distinguish between Low and High dosage with simple clustering estimators as the ones we have been using so far.

As we see in Figure 3.16, observations at day 35 are split in the PCoA space according to probiotic dosage only for shotgun samples for caeca. As for crop samples, it does not seem too easy to discriminate between groups and, in some way, samples appear to be distributed randomly in the features space; in fact, looking at QDA scores, it seems that we are not able recognize a good space segmentation. It's not surprising that we can't distinguish the samples well by their dosage because we have already pinned that bacteria population does not seem to depend significantly on probiotic dose, according to Figure 3.10.
In Supplementary Table S5 we show that it is impossible to compare the silhouette scores of dosage recognition.

From the three types of space segmentation we studied, we can conclude that both Whole Genome and amplicon sequencing provide useful biomarkers; for specific tasks, however, as day recognition in crop, rarest genera in the microbiome could be significantly informative (high scores on $\mathrm{X_{left}}$); in this case metagenomics overcomes the other technique.
In addition, the fact that usually rarest shotgun genera are quite informative even alone, lets us know that mistakes on taxonomical annotations do not affect consistently the rarest genera and a lot of those genera recognized in shotgun samples are effectively inside the



(a)

(b)

(c)

(d)

Figure 3.16: Bray-Curtis PCoA of genera abundance at day 35 for High, Low and Control dose, in samples taken with shotgun sequencing from caeca (a) and crop (b). PCoA of genera abundance at day 35 for High, Low and Control dose, in samples taken with amplicon sequencing, from caeca (c) and crop (d). Abundances have default e-value thresholds and are normalized by total sum.

gut microbiome.

Conclusion

The analysis of abundance profiles of gut microbiota of 40 chickens gave interesting insights on the comparison between whole genome and amplicon sequencing. For samples collected from caeca, we found about 440 unique genera with shotgun against 52 found only by 16S sequencing, plus 94 genera recognized by both methods. In crop of chickens, with shotgun sequencing we detected around 331 unique genera, while with 16S only 35, plus a common group of 60 genera. Shotgun sequencing detects approximatively five times more genera than both techniques together, even if several shotgun sets have low coverage; thus, metagenomics offers an insight into the portion of the population that contains rarest genera, which in 16S samples is hidden (see section 3.1).

We also showed that the Pearson's correlation coefficients between the profiles of common genera in samples collected with the two methods are not much high (0.663 ± 0.006 for caeca and 0.747 ± 0.015 for crop); this is a sign of not-negligible differences in abundance estimation, in particular in the portion mostly composed of rare genera.

The last step of the analysis matched abundance profiles to biological metadata in order to estimate the ability of metagenomes to reveal important biomarkers (as ageing, probiotic dosage and organ). We found that silhouette scores of space segmentation with organ of collection as sample label are high even if we only consider genera whose abundance is less than the first tertile of the RSA distribution ($S_{score} > 0.68$ independently of normalization method). Furthermore using day of sampling as sample label (more fine segmentation), we see that rarest genera observed only in shotgun samples are particularly meaningful to tasks where 16S samples does not provide good silhouette scores (in crop we have $S_{score} = 0.52$ for rarest shotgun genera vs $S_{score} = 0.17$ for full 16S samples). In general, even if both methods provide similar scores on space segmentation according to metadata, we could verify that this information is recovered more reliably with metagenomics than with metataxonomics, thus demonstrating that one approach is more informative than the other.

Supplementary Figures

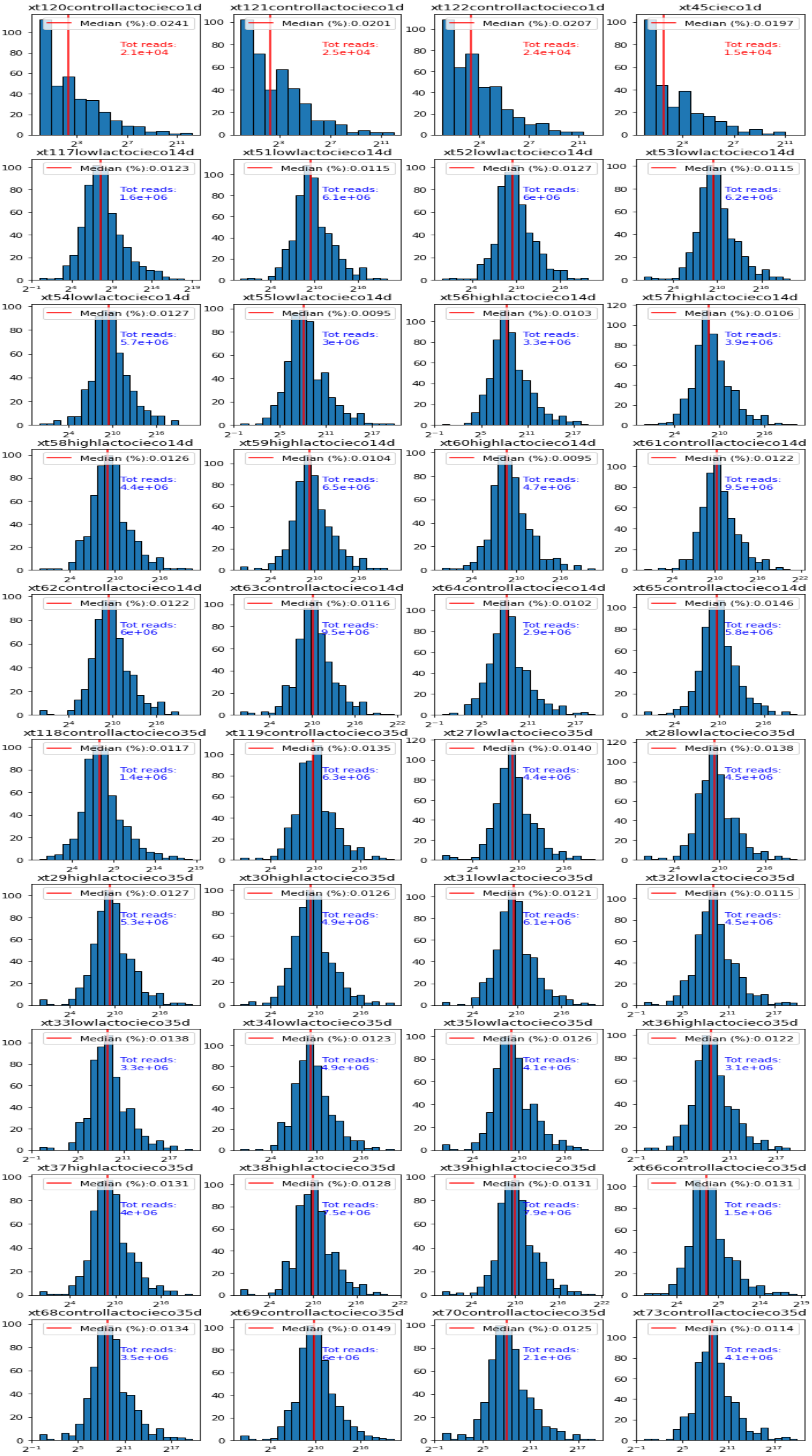


Figure S1: Logarithm in base 2 of genera abundances in all shotgun samples from caeca. Median is shown as percent genus abundance. Number of reads written in blue are those above the threshold that yield a log-normal distribution (200 000 for caeca), while those written in red are below it.

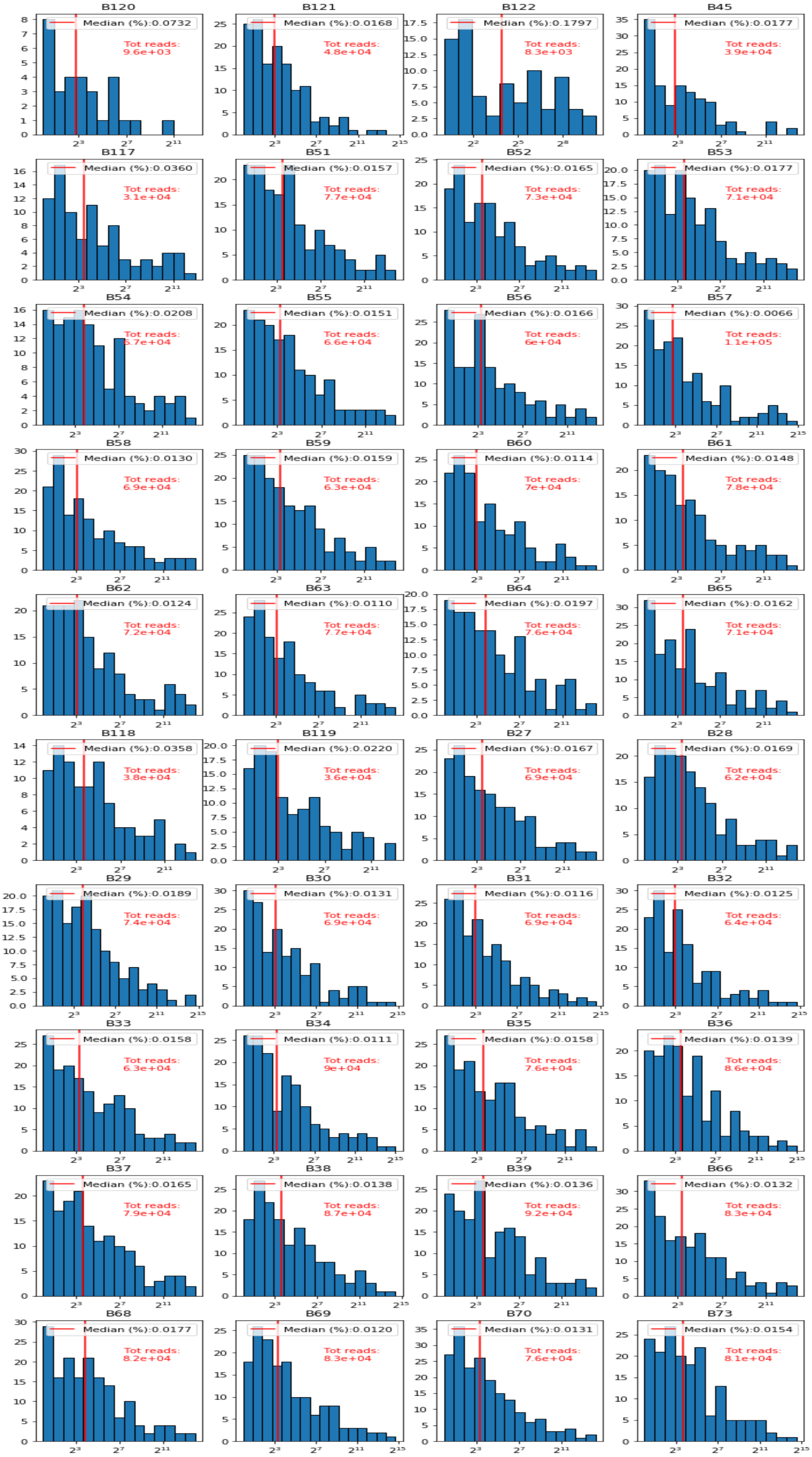


Figure S2: Logarithm in base 2 of genera abundances in all amplicon samples from caeca. Median is shown as percent genus abundance. Number of reads written in blue are those above the threshold that yield a log-normal distribution (200 000 for caeca), while those written in red are below it.

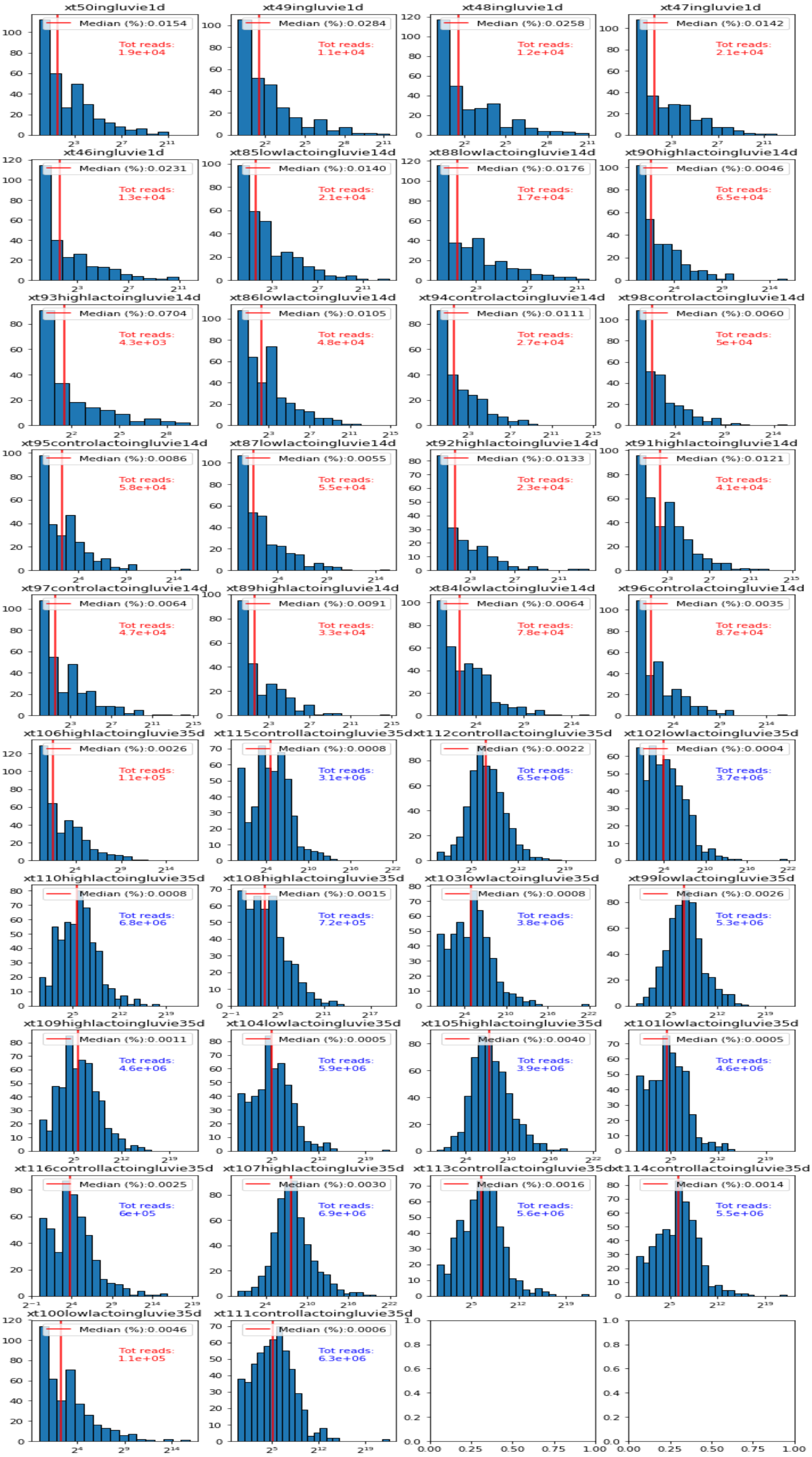


Figure S3: Logarithm in base 2 of genera abundances in all shotgun samples from crop. Median is shown as percent genus abundance. Number of reads written in blue are those above the threshold that yield a log-normal distribution (800 000 for crop), while those written in red are below it.

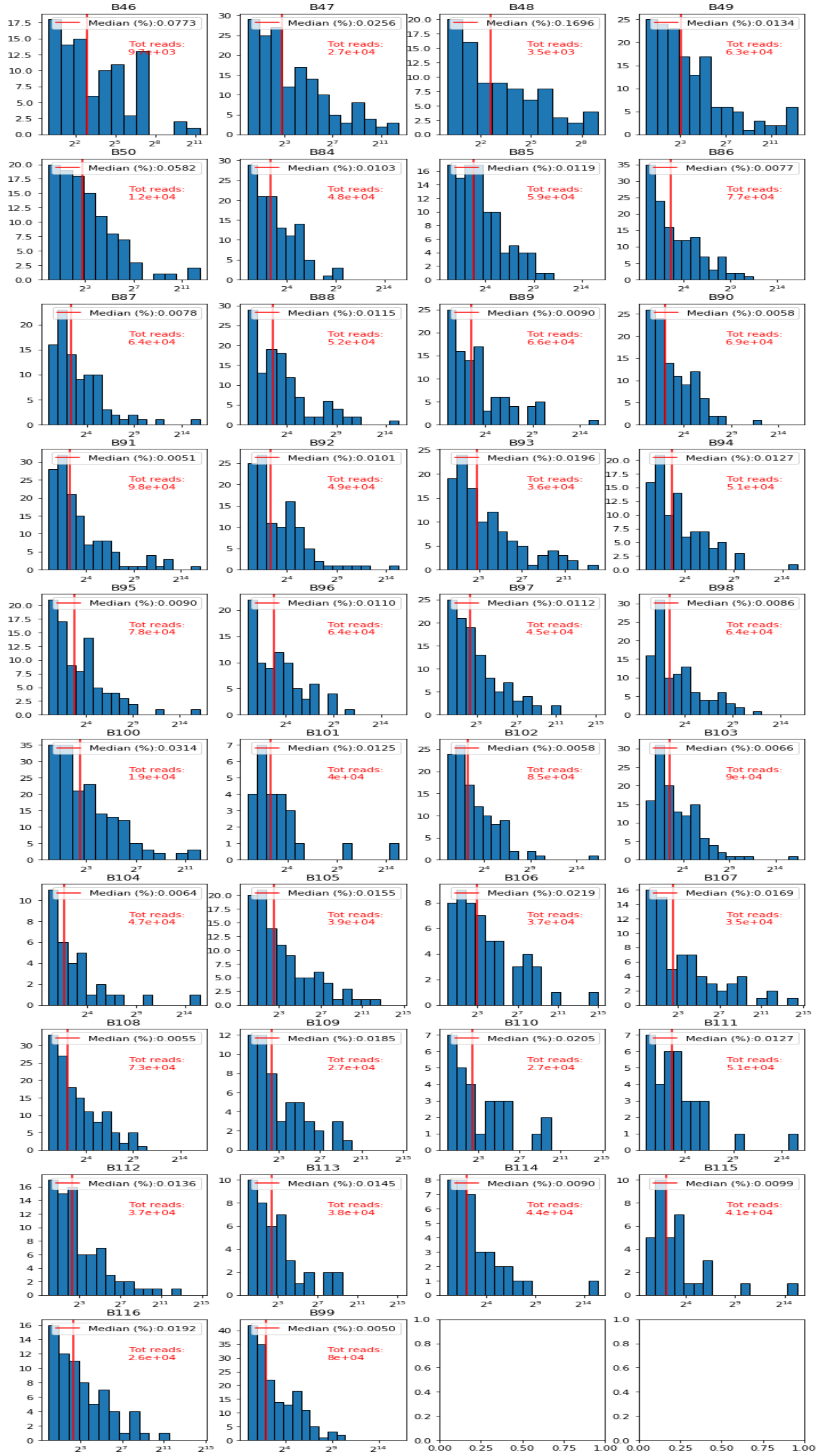


Figure S4: Logarithm in base 2 of genera abundances in all amplicon samples from crop. Median is shown as percent genus abundance. Number of reads written in blue are those above the threshold that yield a log-normal distribution (800 000 for crop), while those written in red are below it.

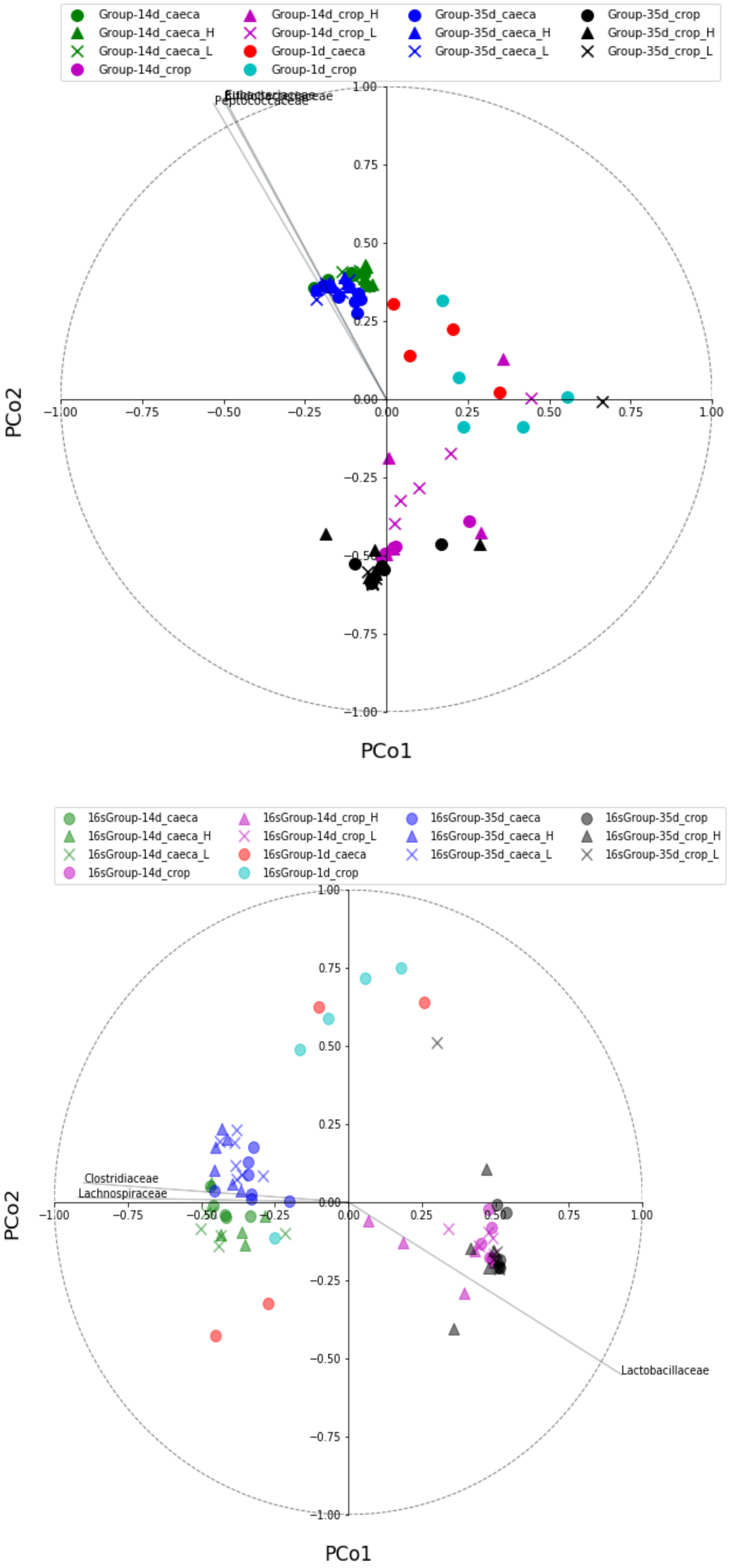
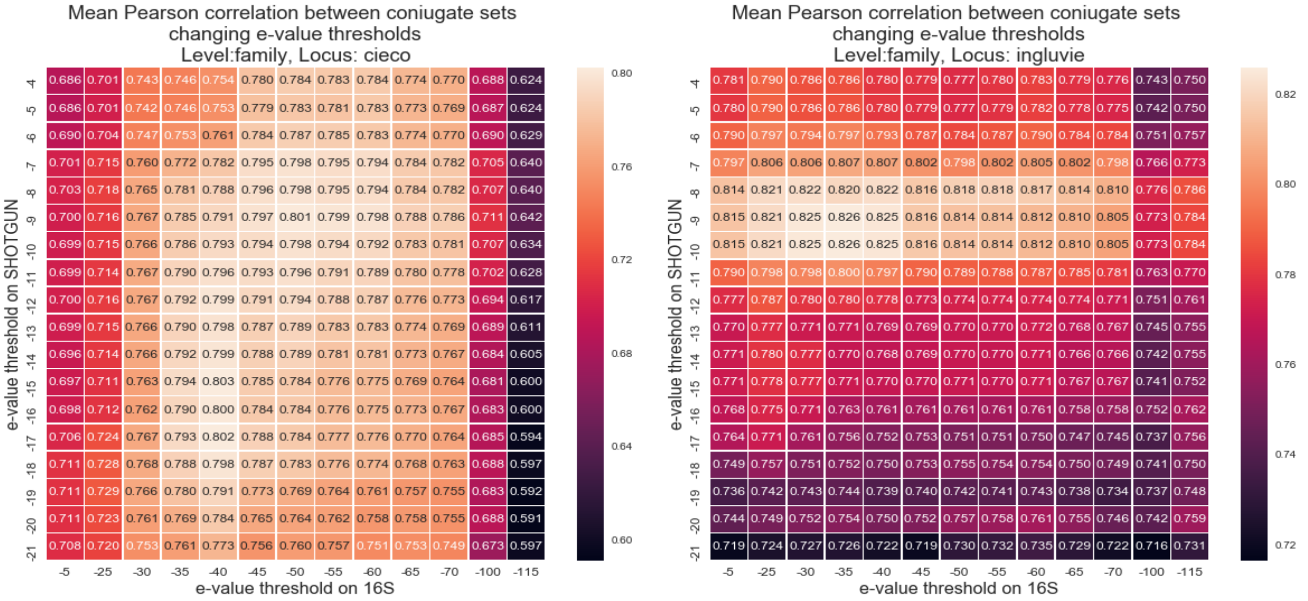
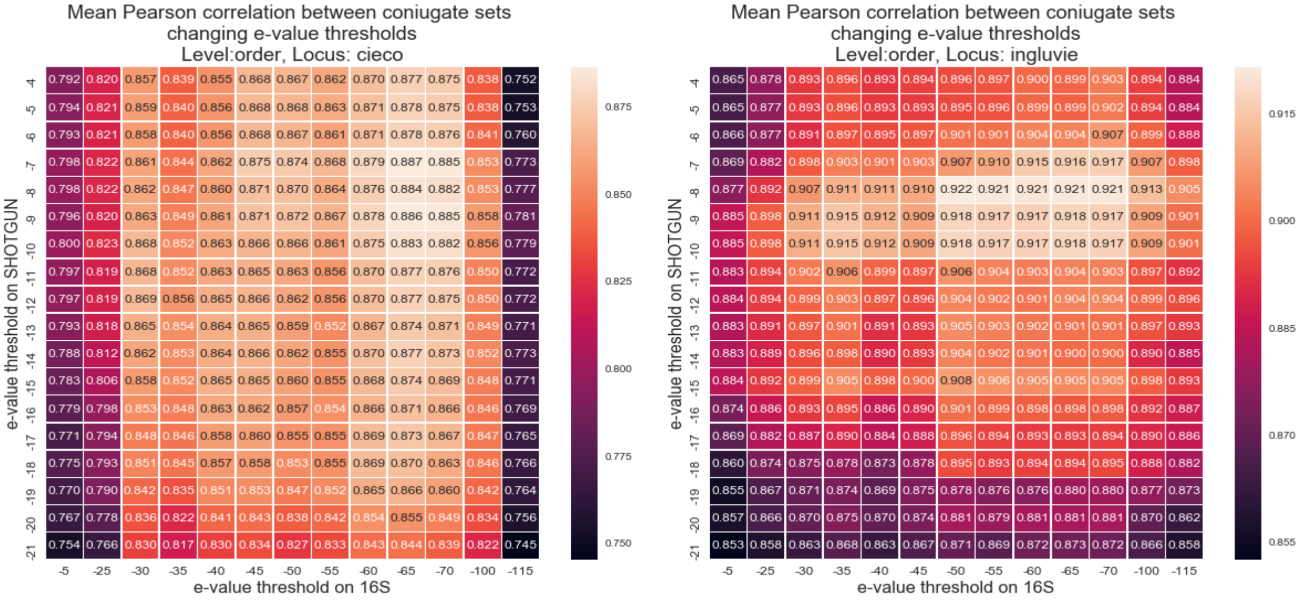


Figure S5: PCoA of family abundances of all datasets normalized by PQN. Full colors represent samples collected by shotgun sequencing, transparent color data collected by amplicon sequencing. Caeca samples follows the ageing with shades R→G→B and crop with C→M→K.



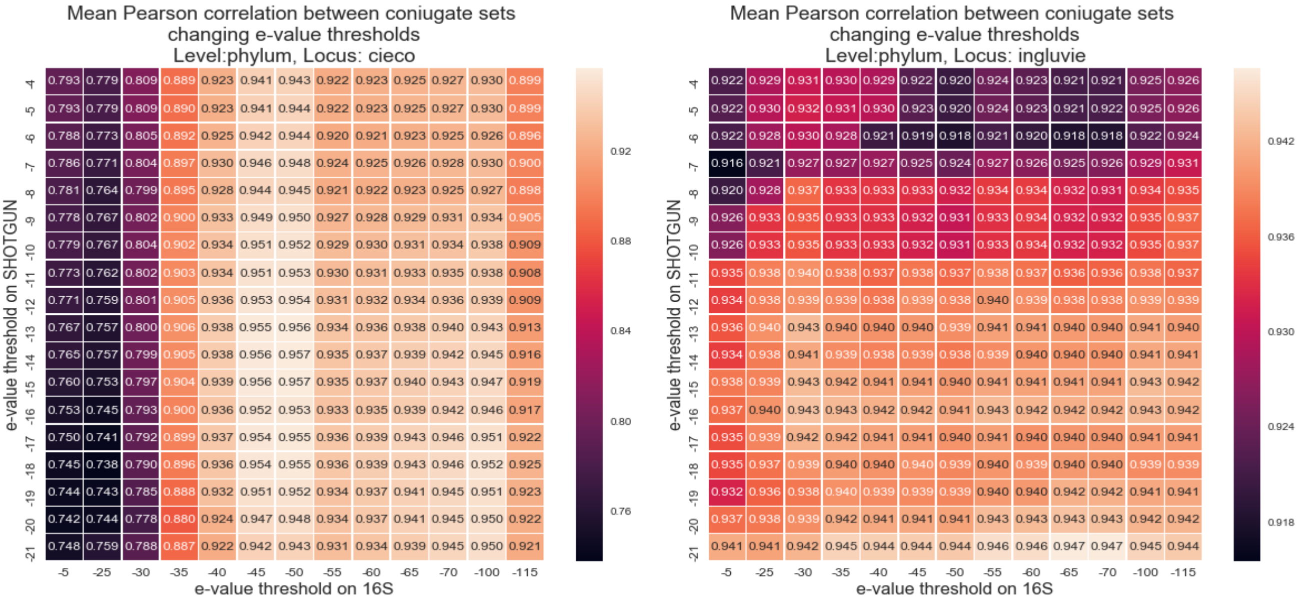
(a) (b)

Figure S6: Mean Pearson's correlation between bacteria abundances at family level computed on conjugate samples, with e-value threshold varying both on metagenomic and metataxonomic samples.



(a) (b)

Figure S7: Mean Pearson's correlation between bacteria abundances at order level computed on conjugate samples, with e-value threshold varying both on metagenomic and metataxonomic samples.



	-5	-25	-30	-35	-40	-45	-50	-55	-60	-65	-70	-100	-115
-4	0.793	0.779	0.809	0.889	0.923	0.941	0.943	0.922	0.923	0.925	0.927	0.930	0.899
-5	0.793	0.779	0.809	0.890	0.923	0.941	0.944	0.922	0.923	0.925	0.927	0.930	0.899
-6	0.788	0.773	0.805	0.892	0.925	0.942	0.944	0.920	0.921	0.923	0.925	0.926	0.896
-7	0.786	0.771	0.804	0.897	0.930	0.946	0.948	0.924	0.925	0.926	0.928	0.930	0.900
-8	0.781	0.764	0.799	0.895	0.928	0.944	0.945	0.921	0.922	0.923	0.925	0.927	0.898
-9	0.778	0.767	0.802	0.900	0.933	0.949	0.950	0.927	0.928	0.929	0.931	0.934	0.905
-10	0.779	0.767	0.804	0.902	0.934	0.951	0.952	0.929	0.930	0.931	0.934	0.938	0.909
-11	0.773	0.762	0.802	0.903	0.934	0.951	0.953	0.930	0.931	0.933	0.935	0.938	0.908
-12	0.771	0.759	0.801	0.905	0.936	0.953	0.954	0.931	0.932	0.934	0.936	0.939	0.909
-13	0.767	0.757	0.800	0.906	0.938	0.955	0.956	0.934	0.936	0.938	0.940	0.943	0.913
-14	0.765	0.757	0.799	0.905	0.938	0.956	0.957	0.935	0.937	0.939	0.942	0.945	0.916
-15	0.760	0.753	0.797	0.904	0.939	0.956	0.957	0.935	0.937	0.940	0.943	0.947	0.919
-16	0.753	0.745	0.793	0.900	0.936	0.952	0.953	0.933	0.935	0.939	0.942	0.946	0.917
-17	0.750	0.741	0.792	0.899	0.937	0.954	0.955	0.936	0.939	0.943	0.946	0.951	0.922
-18	0.745	0.738	0.790	0.896	0.936	0.954	0.955	0.936	0.939	0.943	0.946	0.952	0.925
-19	0.744	0.743	0.785	0.888	0.932	0.951	0.952	0.934	0.937	0.941	0.945	0.951	0.923
-20	0.742	0.744	0.778	0.880	0.924	0.947	0.948	0.934	0.937	0.941	0.945	0.950	0.922
-21	0.748	0.759	0.788	0.887	0.922	0.942	0.943	0.931	0.934	0.939	0.945	0.950	0.921

(a)

	-5	-25	-30	-35	-40	-45	-50	-55	-60	-65	-70	-100	-115
-4	0.922	0.929	0.931	0.930	0.929	0.922	0.920	0.924	0.923	0.921	0.921	0.925	0.926
-5	0.922	0.930	0.932	0.931	0.930	0.923	0.920	0.924	0.923	0.921	0.922	0.925	0.926
-6	0.922	0.928	0.930	0.928	0.921	0.919	0.918	0.921	0.920	0.918	0.918	0.922	0.924
-7	0.916	0.921	0.927	0.927	0.927	0.925	0.924	0.927	0.926	0.925	0.926	0.929	0.931
-8	0.920	0.928	0.937	0.933	0.933	0.933	0.932	0.934	0.934	0.932	0.931	0.934	0.935
-9	0.926	0.933	0.935	0.933	0.933	0.932	0.931	0.933	0.934	0.932	0.932	0.935	0.937
-10	0.926	0.933	0.935	0.933	0.933	0.932	0.931	0.933	0.934	0.932	0.932	0.935	0.937
-11	0.935	0.938	0.940	0.938	0.937	0.938	0.937	0.938	0.937	0.936	0.936	0.938	0.937
-12	0.934	0.938	0.939	0.939	0.938	0.939	0.938	0.940	0.939	0.938	0.938	0.939	0.939
-13	0.936	0.940	0.943	0.940	0.940	0.940	0.939	0.941	0.941	0.940	0.940	0.941	0.940
-14	0.934	0.938	0.941	0.939	0.938	0.939	0.938	0.939	0.940	0.940	0.940	0.941	0.941
-15	0.938	0.939	0.943	0.942	0.941	0.941	0.940	0.941	0.941	0.941	0.941	0.943	0.942
-16	0.937	0.940	0.943	0.943	0.942	0.942	0.941	0.943	0.942	0.942	0.942	0.943	0.942
-17	0.935	0.939	0.942	0.942	0.941	0.941	0.940	0.941	0.940	0.940	0.940	0.941	0.941
-18	0.935	0.937	0.939	0.940	0.940	0.940	0.939	0.940	0.939	0.940	0.940	0.939	0.939
-19	0.932	0.936	0.938	0.940	0.939	0.939	0.939	0.940	0.940	0.942	0.942	0.941	0.941
-20	0.937	0.938	0.939	0.942	0.941	0.941	0.941	0.943	0.943	0.943	0.943	0.942	0.942
-21	0.941	0.941	0.942	0.945	0.944	0.944	0.944	0.946	0.946	0.947	0.947	0.945	0.944

(b)

Figure S8: Mean Pearson's correlation between bacteria abundances at phylum level computed on conjugate samples, with e-value threshold varying both on metagenomic and metataxonomic samples.

Supplementary Tables

Day	Dose	ID	XSeqs	BSeqs
1st	**C**	xt120controllactociecold	21072.0	9592.0
		xt121controllactociecold	25288.0	47728.0
		xt122controllactociecold	24626.0	8432.0
		xt45ciecold	15512.0	39582.0
14th	**C**	xt61controllactociecol14d	9466554.0	78094.0
		xt62controllactociecol14d	6042612.0	72537.0
		xt63controllactociecol14d	9524410.0	77260.0
		xt64controllactociecol14d	2880760.0	76298.0
		xt65controllactociecol14d	5840218.0	71054.0
	L	xt117lowlactociecol14d	1615718.0	30660.0
		xt51lowlactociecol14d	6077376.0	76792.0
		xt52lowlactociecol14d	6040828.0	72980.0
		xt53lowlactociecol14d	6240844.0	70838.0
		xt54lowlactociecol14d	5664408.0	67590.0
		xt55lowlactociecol14d	3048038.0	66211.0
	H	xt56highlactociecol14d	3296138.0	60511.0
		xt57highlactociecol14d	3890376.0	106238.0
		xt58highlactociecol14d	4402968.0	69509.0
		xt59highlactociecol14d	6479056.0	63239.0
		xt60highlactociecol14d	4668384.0	70112.0
35th	**C**	xt118controllactocieco35d	1430492.0	37786.0
		xt119controllactocieco35d	6306610.0	36524.0
		xt66controllactocieco35d	1454336.0	83648.0
		xt68controllactocieco35d	3457116.0	81950.0
		xt69controllactocieco35d	5958374.0	83225.0
		xt70controllactocieco35d	2094574.0	76331.0
		xt73controllactocieco35d	4142770.0	81444.0
	L	xt27lowlactocieco35d	4399488.0	69038.0
		xt28lowlactocieco35d	4512624.0	62313.0
		xt31lowlactocieco35d	6098676.0	69074.0
		xt32lowlactocieco35d	4493552.0	64309.0
		xt33lowlactocieco35d	3337484.0	63423.0
		xt34lowlactocieco35d	4892594.0	90489.0
		xt35lowlactocieco35d	4100910.0	76331.0
	H	xt29highlactocieco35d	5276130.0	74245.0
		xt30highlactocieco35d	4930674.0	68850.0
		xt36highlactocieco35d	3052208.0	86411.0
		xt37highlactocieco35d	3992520.0	78983.0
		xt38highlactocieco35d	7455260.0	87075.0
		xt39highlactocieco35d	7875560.0	92025.0

Table S1: List of all CAECA samples. Format for amplicon samples is 'B+IDnumber'. Xseqs are shotgun sequences and Bseqs are 16S RNA sequences.

Day	Dose	ID	Xseqs	Bseqs
1st	**C**	xt46ingluvie1d	13266	9801
		xt47ingluvie1d	21478	27462
		xt48ingluvie1d	11912	3622
		xt49ingluvie1d	10836	63360
		xt50ingluvie1d	19752	12123
14th	**C**	xt94controlactoingluvie14d	27168	51405
		xt95controlactoingluvie14d	58308	78054
		xt96controlactoingluvie14d	87168	63739
		xt97controlactoingluvie14d	46880	44769
		xt98controlactoingluvie14d	50270	63896
	L	xt84lowlactoingluvie14d	78514	48538
		xt85lowlactoingluvie14d	21684	58879
		xt86lowlactoingluvie14d	47998	77628
		xt87lowlactoingluvie14d	54858	63866
		xt88lowlactoingluvie14d	17338	52148
	H	xt89highlactoingluvie14d	33288	66557
		xt90highlactoingluvie14d	65606	68792
		xt91highlactoingluvie14d	41524	98127
		xt92highlactoingluvie14d	22802	49472
		xt93highlactoingluvie14d	4450	35828
35th	**C**	xt111controllactoingluvie35d	6302668	51045
		xt112controllactoingluvie35d	6499020	36710
		xt113controllactoingluvie35d	5550608	37974
		xt114controllactoingluvie35d	5452562	44347
		xt115controllactoingluvie35d	3053780	40596
		xt116controllactoingluvie35d	602764	26072
	L	xt100lowlactoingluvie35d	109562	19262
		xt101lowlactoingluvie35d	4633232	40095
		xt102lowlactoingluvie35d	3684636	85589
		xt103lowlactoingluvie35d	3825402	90402
		xt104lowlactoingluvie35d	5912442	46785
		xt99lowlactoingluvie35d	5321692	79710
	H	xt105highlactoingluvie35d	3865150	38699
		xt106highlactoingluvie35d	114316	36558
		xt107highlactoingluvie35d	6868246	35527
		xt108highlactoingluvie35d	721320	73299
		xt109highlactoingluvie35d	4591154	27130
		xt110highlactoingluvie35d	6771762	26888

Table S2: List of all CROP samples. Format for amplicon samples is 'B+IDnumber'. Xseqs are shotgun sequences and Bseqs are 16S RNA sequences.

Cumulative abundance of genera:	in shotgun samples (%)	in 16S samples (%)	Rarity threshold (%)
Species rare in shotgun	1.519 on 264 genera	0.861 on 13 gen.	0.014
Species rare in 16S	4.430 on 45 genera	0.515 on 71 gen.	0.024

Table S3: Cumulative abundance of species below shotgun rarity threshold both in shotgun itself and in 16S, with default $[-5, -5]$ e-value thresholds. The analogous study is shown for species considered rare in 16S samples. Rarity thresholds shown in the table are the mean of each threshold used for the computation for each set, chosen as the median of all bacteria abundances in that sample. This analysis is carried only on caeca samples for brevity. Of the 223 genera identified as rare in samples collected by shotgun sequencing (last row in Tab) only 14 are found in their relative amplicon samples too, with similar abundance; on the contrary, the 71 species not abundant in 16S are quite populated in shotgun samples, reaching together around 3% of total organisms (on average). This behaviour may suggest that species rare for 16S are not so unpopulated in shotgun samples. On the other hand, the fact that species rare for shotgun seem to be detected also by 16S is false.

Found only on shotgun	abundance (%)	**Found only in 16S**	abundance (%)
Bacteroidaceae	0.704	*Planococcaceae*	0.061
Chlorobiaceae	0.507	*Sporolactobacillaceae*	0.032
Burkholderiaceae	0.238	*Dietziaceae*	0.025
Pasteurellaceae	0.219	*Thermoactinomycetaceae*	0.014
Rhodobacteraceae	0.207	*Rarobacteraceae*	0.010
Comamonadaceae	0.166	*Moritellaceae*	0.008
Bifidobacteriaceae	0.155	*Anaeroplasmataceae*	0.004
Listeriaceae	0.152	*Clostridiales Fam.XIV.Incertae Sedis*	0.003
Porphyromonodaceae	0.142	*Clostridiales Fam.XII.Incertae Sedis*	0.003
Helicobacteraceae	0.111	*Bacteriovoracaceae*	0.001

Table S4: Percent abundance of the ten most abundant families that on median are detected only by shotgun (left) and only by amplicon sequencing (right), in crop samples.

Day 14 TSN	**Caeca**	**Crop**
$\mathbf{X}_{\mathbf{right}}$	0.043	0.001
B	0.015	0.042
$\mathbf{X}_{\mathbf{left}}$	0.070	−0.027
p_{value} **XvsB**	0.87	0.83

Day 35 TSN	**Caeca**	**Crop**
$\mathbf{X}_{\mathbf{right}}$	0.262	−0.063
B	0.069	−0.013
$\mathbf{X}_{\mathbf{left}}$	0.124	−0.021
p_{value} **XvsB**	0.16	0.55

Table S5: Euclidean Silhouette Score on Bray-Curtis PCoA of genera abundances for dosage recognition task, on data normalized by total sum at day 14 (left) and 35 (right). Dataset were shotgun data (X), 16S (B), shotgun genera with abundance lower than the first tertile (X_{left}), shotgun genera with abundance higher than the first tertile (X_{right}). E-value thresholds were set to default $[-5, -5]$ for both organs.

Day 14 PQN	**Caeca**	**Crop**
$\mathbf{X_{right}}$	0.053	−0.009
B	−0.013	−0.053
$\mathbf{X_{left}}$	0.028	−0.078
p_{value} **XvsB**	0.66	0.73

Day 35 PQN	**Caeca**	**Crop**
$\mathbf{X_{right}}$	0.245	−0.058
B	0.100	−0.099
$\mathbf{X_{left}}$	0.180	0.050
p_{value} **XvsB**	0.35	0.63

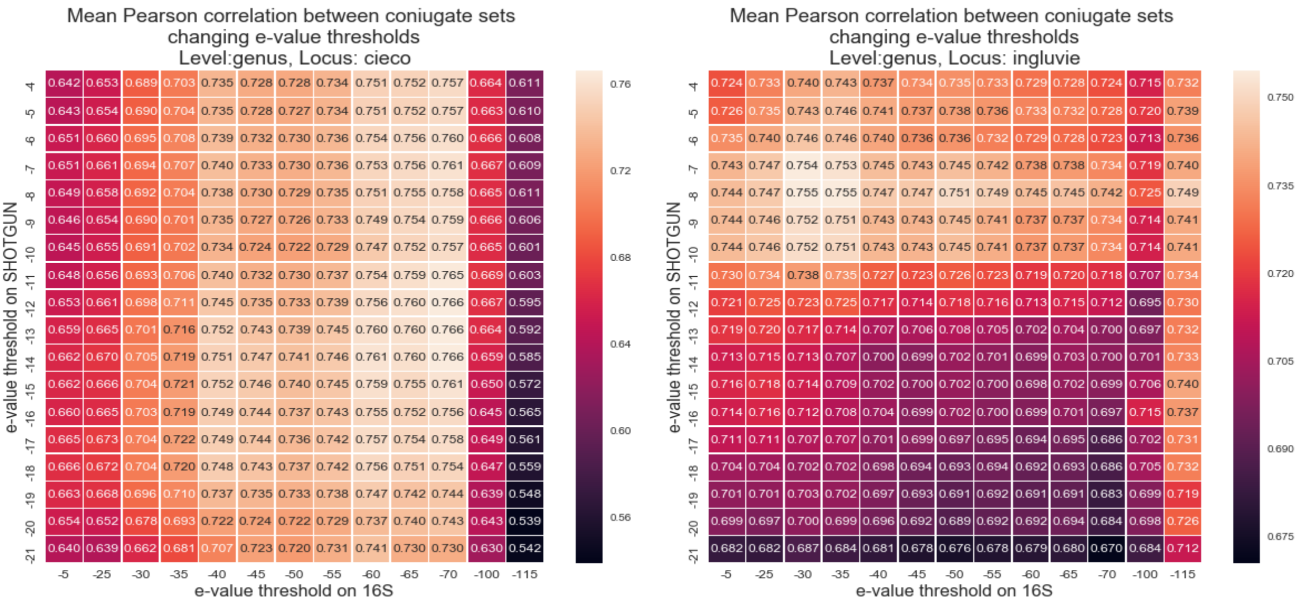
Table S6: Euclidean Silhouette Score on Bray-Curtis PCoA of genera abundances for dosage recognition task, on data normalized by PQN at day 14 (left) and 35 (right). Dataset were shotgun data (X), 16S (B), shotgun genera with abundance lower than the first tertile ($\mathrm{X_{left}}$), shotgun genera with abundance higher than the first tertile ($\mathrm{X_{right}}$). E-value thresholds were set to default $[-5, -5]$ for both organs.

Supplementary Sections

3.4 Tuning of alignment quality parameters

In order to choose optimal thresholds for the alignment quality parameters as e-value and percent identity, we decided to compute the mean value of correlation for each pair of conjugate datasets, where for *conjugate* we mean collected from the same chicken with two different sequencing techniques, and to determine how the thresholding affects the correlation. In particular we are looking for quality thresholds that maximize the correlation, believing that this operation will lead to a similar statistical significance for reads from both metagenomics and metataxonomics.

We let the e-value threshold vary both for metagenomic and metataxonomic samples, in order to see if a point of maximum correlation between conjugate sets is reached.



(a)

(b)

Figure S9: Mean Pearson's correlation between bacteria abundances at genus level computed on conjugate samples, with e-value thresholds varying both on metagenomic and metataxonomic samples.

In Figure S9 an optimal threshold for the e-value of reads alignment from caeca is individuated around $[-70, -12]$ where with the first number inside square brackets we refer to minimum e-value of 16S samples and with the latter to the minimum e-value of shotgun samples. For sequences from crop of chickens, the optimum seems to be around $[-35, -8]$

and increasing the quality only leads to deterioration (since many reads are not taken into account). To broaden the analysis, we present in Supplementary Figure S6, S7 and S8 the same graph computed on higher taxonomic levels, to see if thresholds are persistent.

For caeca sample, e-value thresholds that consent best correlations vary a bit respect to genus, while in crop samples, they are quite similar among all taxonomic levels except for phylum, that would suggest very high thresholds. In reality, the correlation at phylum level with very strict e-value thresholds is computed on very few phyla, so it is not surprising it can reach high values of Pearson coefficients; in caeca samples too we had a local minimum around the maximum values of thresholds, since it was less striking because we could find another good optimum anyway. In any case, we can even choose to keep lower thresholds, so $eval_T = [-60, -12]$ for caeca samples and $eval_T = [-35, -8]$ for crop ones, in order to not lose too much sequences being too strict, both for shotgun metagenomes, where some samples are very poor of DNA reads, and 16S, where the number of detected genera is already lower than shotgun.

Now the number of sequences becomes obviously lower than at the beginning, leading to the specifics in Tab S7.

SHOTGUN	**Seq.**	**st.dev. of seq.**	**e-value**	**Align. length**	**Percent id.**
caeca1day	1 523	630	-15.30	46	85.2
caeca14days	3 018 356	1 979 912	-15.39	51	76.7
caeca35days	2 031 072	1 293 162	-15.38	50	77.4
crop1day	7 584	3 226	-10.34	36	85.8
crop14days	34 705	19 711	-10.72	37	85.4
crop35days	3 950 779	2 184 641	-11.07	40	82.0

Table S7: Mean number of predicted sequences per chicken for shotgun samples at day 1, 14 and 35, with $eval_T = -12$ for caeca and $eval_T = -8$ for crop. We also show mean value of e-value, alignment length and percent identity, computed by MGrast.

16S	**Seq.**	**st. dev. of seq.**	**e-value**	**Align. length**	**Percent id.**
caeca1day	34 202	25 326	-124.30	240	96.5
caeca14days	56 910	12 758	-94.98	192	96.0
caeca35days	50 271	10 035	-94.50	191	96.0
crop1day	49 242	21 588	-78.90	156	97.9
crop14days	64 515	12 818	-87.71	171	98.0
crop35days	47 069	18 455	-101.47	198	97.2

Table S8: Mean number of predicted sequences per chicken for 16S samples at day 1, 14 and 35, with $eval_T = -60$ for caeca and $eval_T = -35$ for crop. We also show mean value of e-value, alignment length and percent identity, computed by MGrast.

At day 1 shotgun samples possess a very low number of metagenomic sequences (and in 14th day of crop samples too), but now quality parameters are quite high and also the number of detected genera, so one could think that reducing data like that is likely to offer important information anyway.
We can now inspect how the a scatter plot between conjugate sets is influenced by the

tuning of quality parameters, as in Figure S10.
As we see, tuning e-value in order to get a maximum of correlation has only led to the removal of the rarest species from 16S sets, so that they would not count in the computation of Pearson's coefficient. These observations are confirmed by Supplementary Tab S3 that investigates numerically the abundance of "rare" species both before and after tuning.

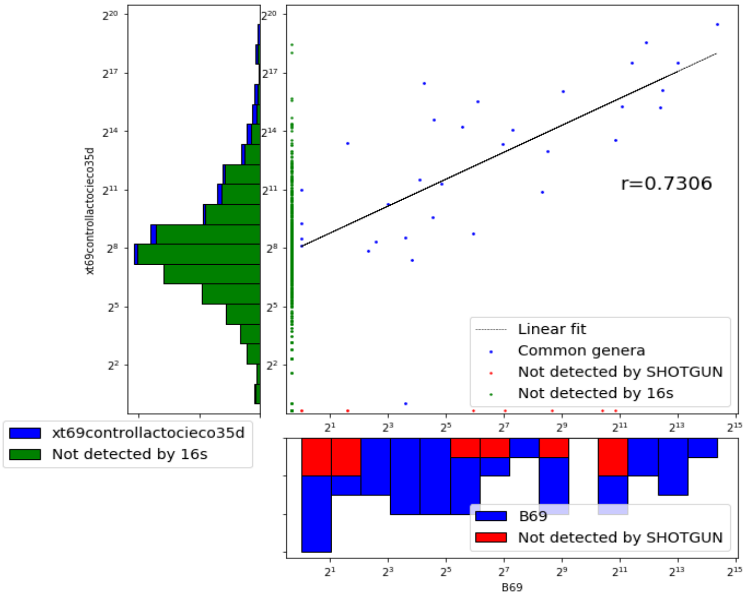


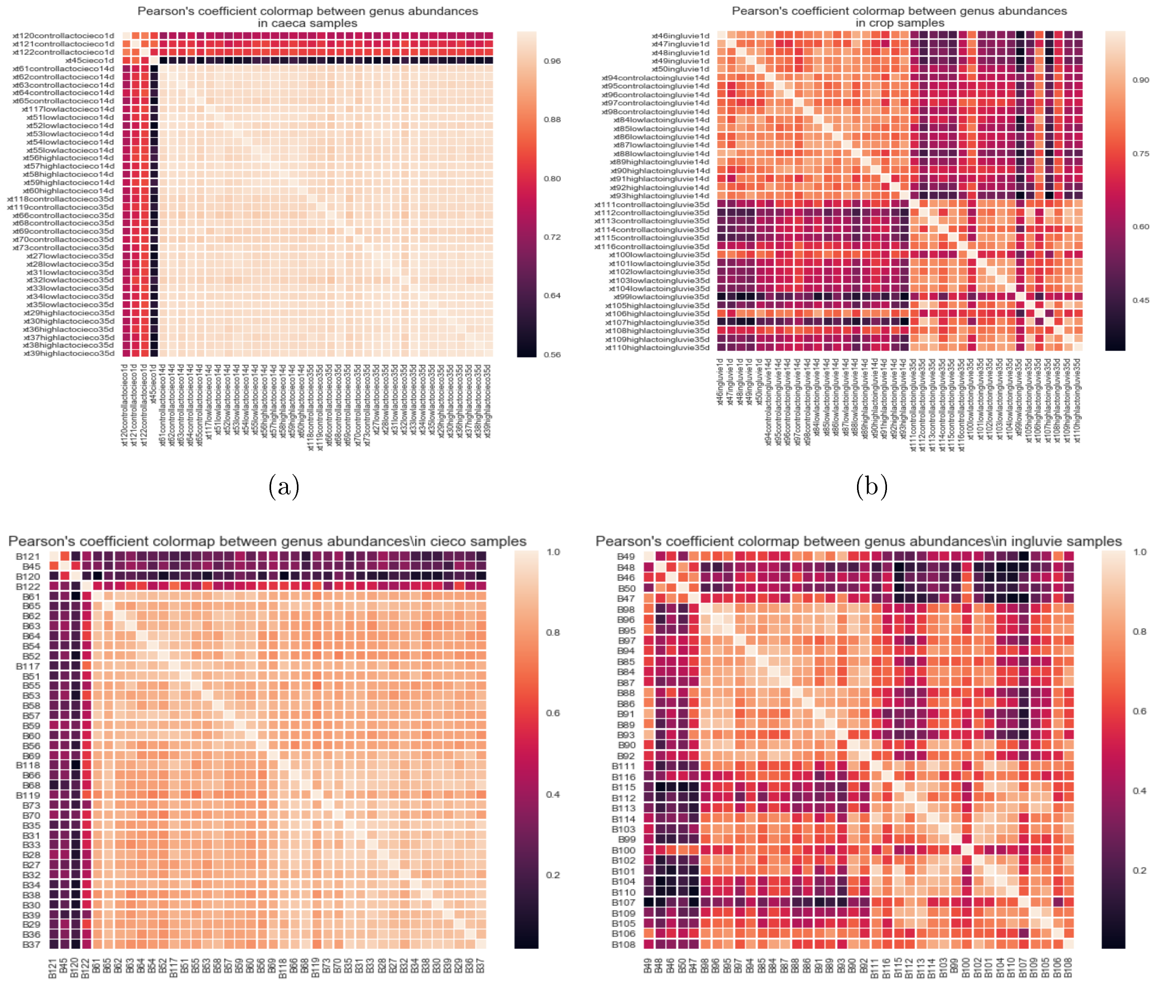
Figure S10: Scatter plot of genera abundances of sample XT69 and its conjugate B69, from group crop35dayCds. Correlation coefficients are computed by Pearson only on the common genera between conjugate sets, with tuned e-value threshold ($eval_T = [-60, -12]$). Green and red observations do not count on Pearson coefficient's computation.

For these reasons we prefer not to impose thresholds on parameter quality any higher than those already set by MGrast ($[-5; -5]$), in order to not lose sensitivity of detection in any of the two methods. Now our aim will be to assess the relationship between populations detected by both methods and assess their reliability, using biological metadata as reference (day of life and treatment dosage).

3.5 Correlation intra-group vs inter-group

Observing directly the genera abundances, as in Figure 3.10 and 3.11, we had an overview on biodiversity for each dataset, but it could be complicated to interpret, so we now try a more quantitative analysis to detect differences between group of samples. For example we can operate some measurements apt to detect similarity between samples based on the bacteria abundances, such as Pearson correlation or distance matrices.
In this case we use the same we build a heat map of correlations computed pairwise between all set sampled with the same sequencing technique, considering only those genera whose abundance is not null in both sets of the pair. Results are shown in Figure S11.

It's noticeable that 1st day samples are quite uncorrelated to others in every image, in particular in caeca of chickens, where for the remaining sample it's difficult to see other information. For crop samples, we can distinguish quite well the low correlation between



(a)

(b)

(c)

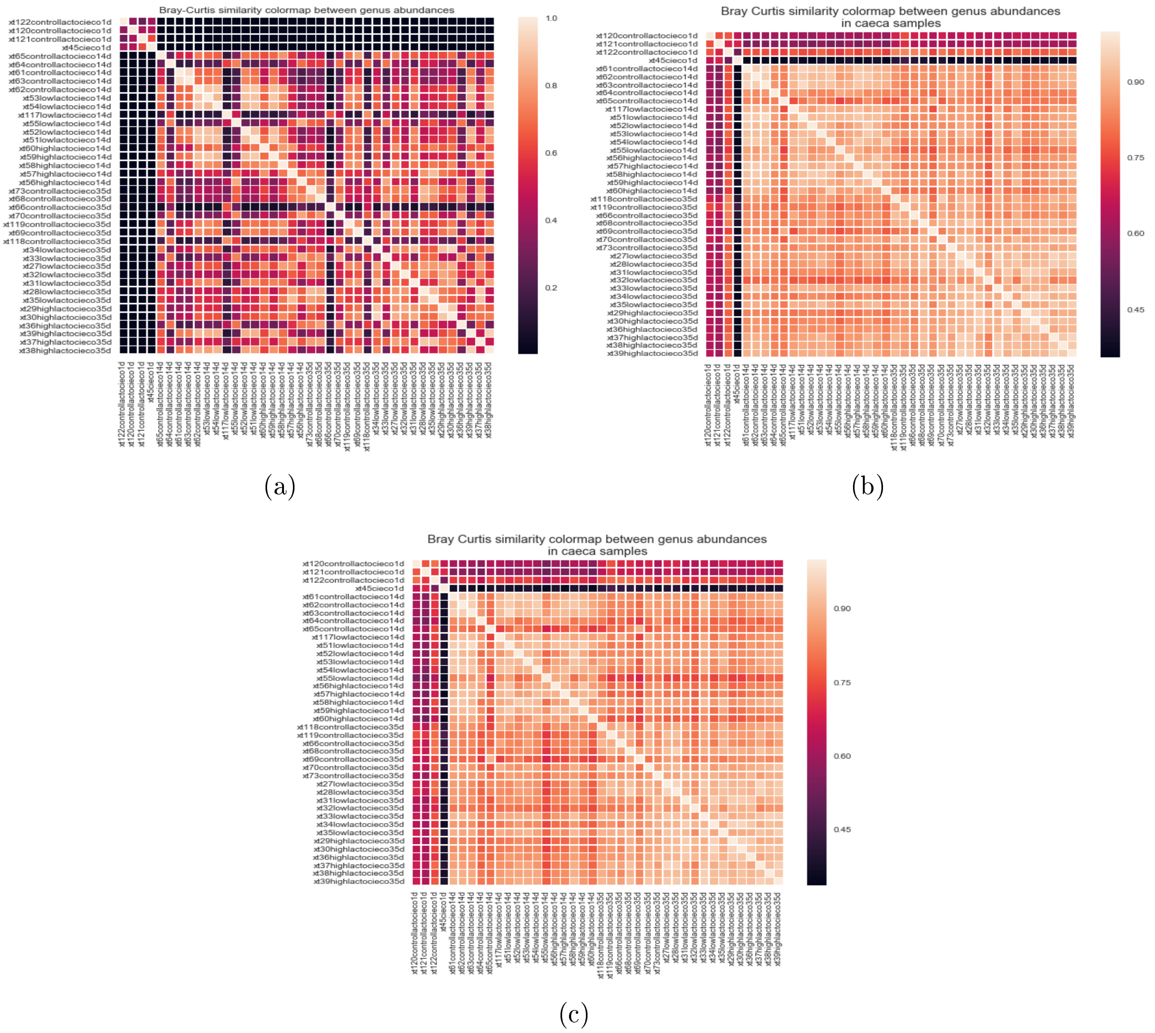
(d)

Figure S11: Pearson correlation between genera abundances for all shotgun samples from caeca (a) and crop (b). Pearson correlation between genera abundances for all 16S samples from caeca (c) and crop (d). E-value threshold are set to MGrast default of $[-5, -5]$.

14th and 35th datasets, that in caeca is strangely hidden.

Anyway, aside from the normalization we chose, we take the complement of the dissimilarity as $S_{B-C} = 1 - BC$, so that the comparison with Pearson correlation gets easier.

In Figure S12, not normalized abundances (a) seem to highlight only first day samples as a group distinct from others, as Pearson correlation did in Figure S11, but we cannot exclude that this behaviour is due to the significant difference in terms of number of reads between this group and others, that now becomes relevant, differently from Pearson correlation.
Normalizing by total sum of reads (b) leads to the detection of slightly paler square corresponding to intra-group correlations at 14, but the visualization gets cleaner in (c), where with PQN the intra-group similarity at day 35 is visualized better and it seems that a compromise is reached between taking into account the volume of samples and a genuine comparison in terms of normalized abundances. But we need more quantitative criteria in order to assess with normalization is more suitable for our purposes, as we do



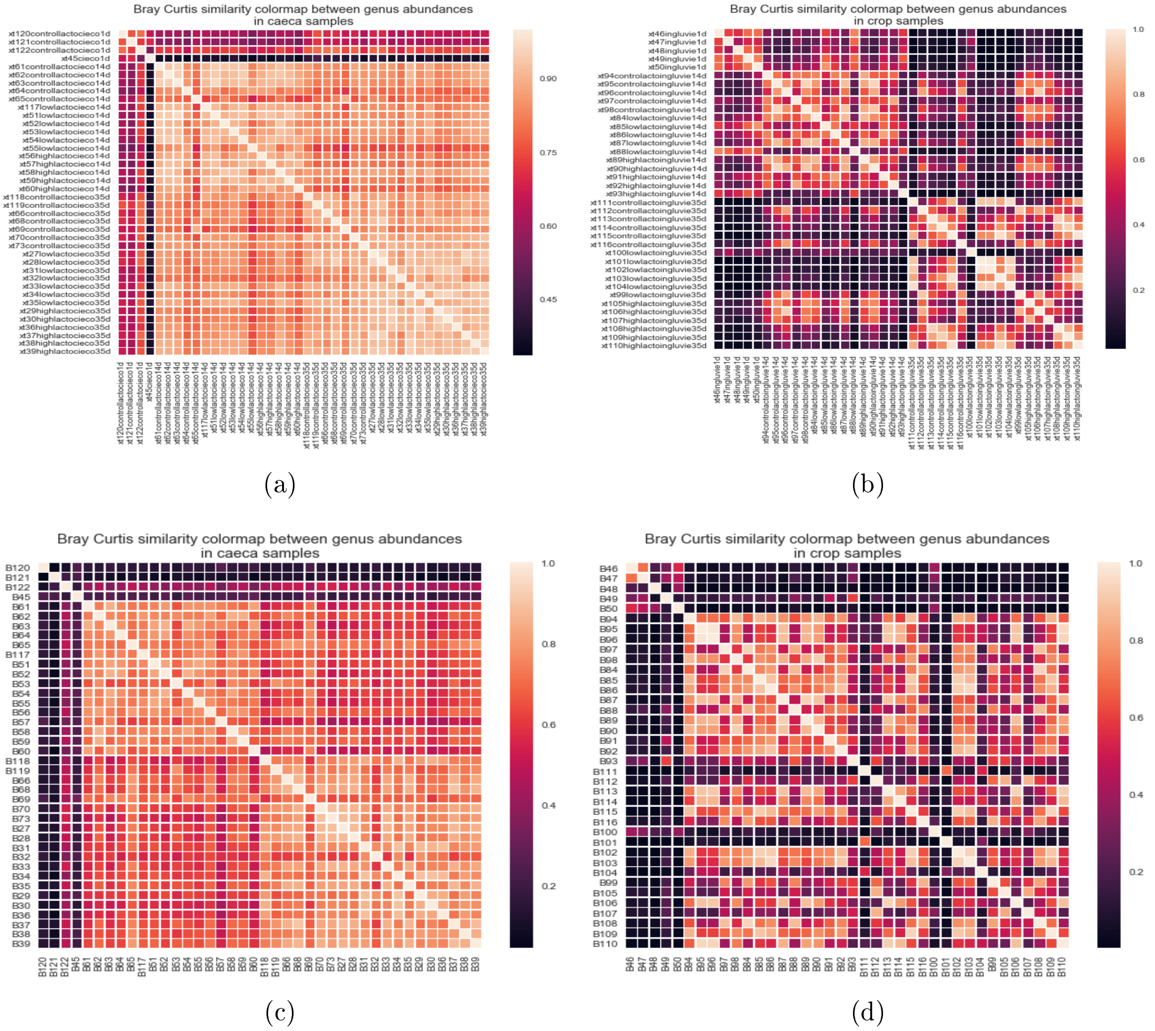
(a) (b)

(c)

Figure S12: Bray-Curtis similarity between genera abundances for all shotgun samples from caeca. Abundances are considered as number of reads in (a), normalized by total sum in (b) and by PQN in (c).

in the next section.
In the meantime, if we want to build a comparison with Pearson correlations in Figure S11, we plot the Bray-Curtis similarities for the same sets and display it in Figure S13, with PQN normalization.

Respect to Pearson correlation, now the dissimilarity between day 14 and 35 is highlighted even in caeca samples in (a) and (c). About the metagenome collected from crop, the similarity intra-group is still visible, in particular for shotgun sequences (b), but the heatmap appears to be more noisy, so less clean than those of Pearson coefficients (Figure S11).
Normalizing with total sum produced no remarkable differences, as we can see in Supplementary Figure S14, except for shotgun samples taken for crop that are a lot worse visualized with total sum normalization.



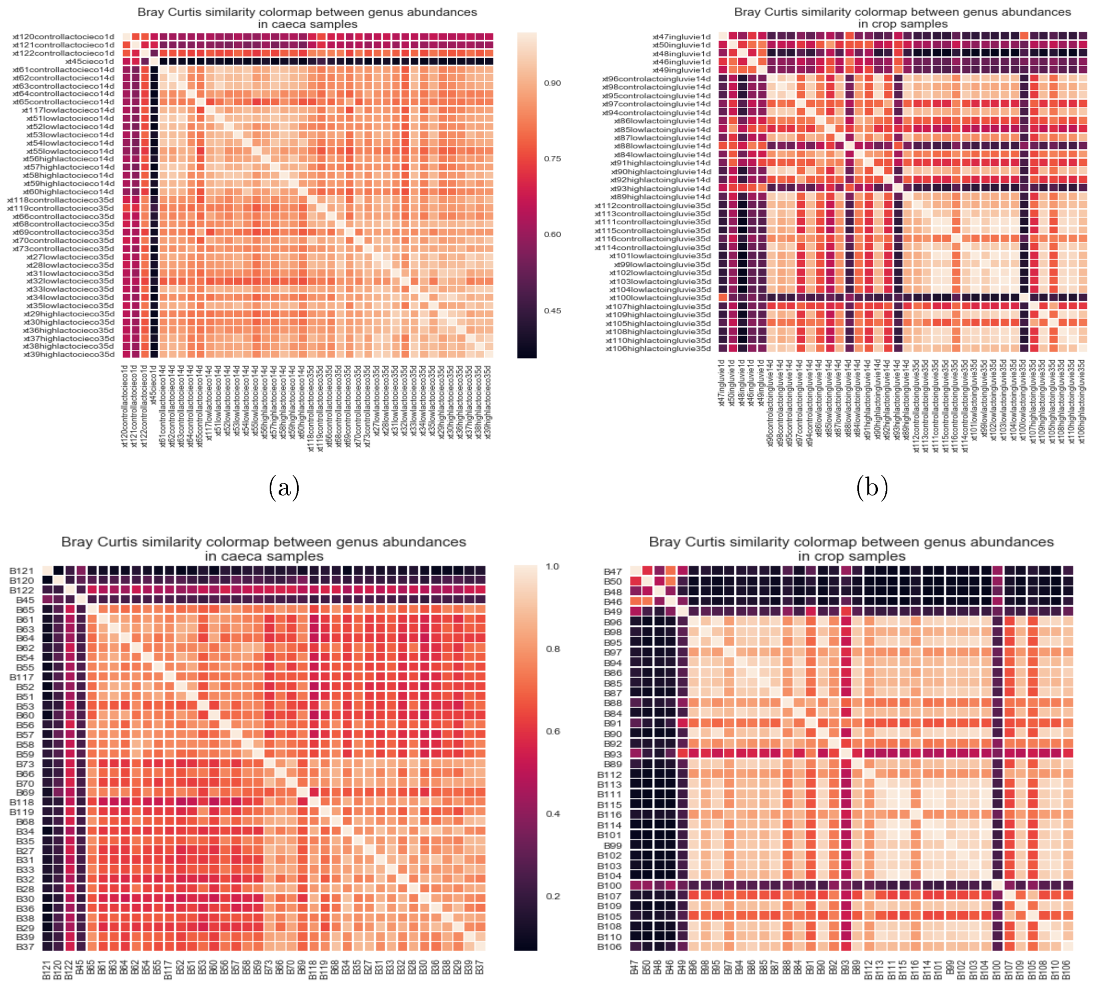
(a)

(b)

(c)

(d)

Figure S13: Bray-Curtis similarity between genera abundances for all shotgun samples from caeca (a) and crop (b) normalized with PQN. Bray-Curtis similarity between genera abundances for all 16S samples from caeca (c) and crop (d) normalized with PQN.



(a)

(b)

(c)

(d)

Figure S14: Bray-Curtis similarity between genera abundances for all shotgun samples from caeca (a) and crop (b) normalized with total sum normalization. Bray-Curtis similarity between genera abundances for all 16S samples from caeca (c) and crop (d) normalized with total sum normalization. E-value threshold are set to MGrast default ([−5, −5].

Bibliography

[1] J. M. Kinross, A. W. Darzi, J. K. Nicholson, *Gut microbiome-host interactions in health and disease*, Genome Medicine, Vol. 3, No. 14, 2011.

[2] R. E. Ley, R. Knight, J. I. Gordon, *The human microbiome: eliminating the biomedical/environmental dichotomy in microbial ecology*, Environ. Microbiol., Vol. 9, 2007.

[3] A. H. Nishida, H. Ochman, *Rates of gut microbiome divergence in mammals*, Molecular Ecology, Vol. 27, 2018.

[4] A. Spor, O. Koren, R. Ley, *Unravelling the effects of the environment and host genotype on the gut microbiome*, Nature Reviews Microbiology, Vol. 9, 2011.

[5] S. Maccaferri, E. Biagi, P. Brigidi, *Metagenomics: Key to Human Gut Microbiota*, Digestive Diseases, Vol. 29, 2011.

[6] D. Stanley, R. J. Hughes, R. J. Moore, *Microbiota of the chicken gastrointestinal tract: influence on health, productivity and disease*, Applied Microbiology and Biotechnology, Is. 10, 2014.

[7] Jun L., Haihong H., Guyue C., Chunbei L., Saeed A., Muhammad A. B. S., Hafiz I. H., Menghong D., Zonghui Y., *Microbial Shifts in the Intestinal Microbiota of Salmonella Infected Chickens in Response to Enrofloxacin*, Front Microbiol., Vol.8, 2017.

[8] Amit-Romach E., Sklan D., Uni Z., *Microflora ecology of the chicken intestine using 16S ribosomal DNA primers*, Poultry Science, Vol. 83, 2004.

[9] B. B. Oakley, H. S. Lillehoj, M. H. Kogut, W. K. Kim, J. J. Maurer, A. Pedroso, M. D. Lee, S. R. Collett, T. J. Johnson, N. A. Cox, *The chicken gastrointestinal microbiome*, FEMS Microbiology letters, Vol. 360, 2014.

[10] Sekelja M., Rud I., Knutsen S. H., Denstadli V., Westereng B., Naes T., Rudi K., *Abrupt temporal fluctuations in the chicken fecal microbiota are explained by its gastrointestinal origin*, Appl. Environ. Microbiol., Vol. 78, 2012.

[11] Nakphaichit M., Thanomwongwattana S., Phraephaisarn C., Sakamoto N., Keawsompong S., Nakayama J., Nitisinprasert S., *The effect of including Lactobacillus reuteri KUB-AC5 during post-hatch feeding on the growth and ileum microbiota of broiler chickens*, Poult. Sci., Vol. 90, 2011.

[12] A. De Cesare, F. Sirri, G. Manfreda, P. Moniaci, A. Giardini, M. Zampiga, A. Meluzzi, *Effect of dietary supplementation with Lactobacillus acidophilus D2/CSL (CECT 4529) on caecum microbioma and productive performance in broiler chickens*, PLOS ONE, 2017.

[13] A. M. Lesk, *Introduction to genomics*, Oxford University Press, Second edition, 2012.

[14] F. P. Breitwieser, J. Lu, S. L. Salzberg, *A review of methods and databases for metagenomic classification and assembly*, Briefings in Bioinformatics, 2017.

[15] J. Jovel, J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. L. Mason, K. L. Madsen and G. K. S. Wong, *Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics*, Frontiers in Microbiology, Vol. 7, 2016.

[16] S. Campanaro, L. Treu, P. G. Kougias, X. Zhu, I. Angelidaki, *Taxonomy of anaerobic digestion microbiome reveals biases associated with the applied high throughput sequencing strategies*, Scientific reports, Vol. 8, 2018.

[17] M. Tessler, J. S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L. F. M. Velho, B. T. Segovia, F. A. Lansac-Toha, M. Lemke, R. DeSalle, C. E. Mason, M. R. Brugler, *Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing*, Scientific Reports, Vol. 7, 2017.

[18] A. Wilke1, W. Gerlach, T. Harrison, T. Paczian, W. L. Trimble,F. Meyer, *MG-RAST Manual for version 4, revision 3*, 2017.

[19] F. Meyer, D. Paarmann, M. D'Souza, R. Olson , E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards, *The Metagenomics RAST server — A public resource for the automatic phylogenetic and functional analysis of metagenomes*, BMC Bioinformatics 2008, http://www.biomedcentral.com/1471-2105/9/386 .

[20] NCBI, *BLAST FAQ*,
https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ

[21] NCBI, *The Statistics of Sequence Similarity Scores*, https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

[22] F. W. Preston, *The Commonness, And Rarity, of Species*, Ecology, Vol. 29, 1948.

[23] C. Sala, *Ecological modelling for next generation sequencing data*, Master degree dissertation, University of Bologna, 2013.

[24] K. R. Clarke, P. J. Somerfield, M. G. Chapman, *On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages*, Journal of Experimental Marine Biology and Ecology, Vol.330, Issue 1, 2017.

[25] M. Templ, K. Hron, P. Filzmoser, A. Gardlo *Imputation of rounded zeros for high-dimensional compositional data*, Chemometrics and Intelligent Laboratory Systems, Vol. 155, 2016.

[26] P. Filzmoser, B. Walczak, *What can go wrong at the data normalization step for identification of biomarkers?*, Journal of Cromatography A, 1362, 2014.

[27] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, *Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics*, Analytical Chemistry, Vol. 78, No. 13, 2006.

[28] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.

[29] F. Wickelmaier, *An Introduction to MDS*, Sound Quality Research Unit - Aalborg University, Denmark, 2003.

[30] SciPy.org, *scipy.spatial.procrustes*, https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.procrustes.html

[31] P. J. McMurdie, S. Holmes, *Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible*, PLOS Computational Biology, Vol.10, Is.4, 2014.

[32] O. H. Tuovinen, J. C. Hsu, *Aerobic and Anaerobic Microorganisms in Tubercles of the Columbus, Ohio, Water Distribution System*, Applied and Environmental Microbiology, Vol.44, No.3, 1982.